

D

Data and Dimensionality Reduction in Data Analysis and System Modeling

WITOLD PEDRYCZ^{1,2}

¹ Department of Electrical and Computer Engineering,
University of Alberta, Edmonton, Canada

² Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Information Granularity and Granular Computing](#)

[Data Reduction](#)

[Dimensionality Reduction](#)

[Co-joint Data and Dimensionality Reduction](#)

[Conclusions](#)

[Future Directions](#)

[Acknowledgments](#)

[Appendix: Particle Swarm Optimization \(PSO\)](#)

[Bibliography](#)

Glossary

Reduction process A suite of activities leading to the reduction of available data and/or reduction of features.

Feature selection An algorithmic process in which a large set of features (attributes) is reduced by choosing a certain relatively small subset of them. The reduction is a combinatorial optimization task which is NP-complete. Given this, quite often it is realized in a suboptimal way.

Feature transformation A process of transforming a highly dimensional feature space into a low-dimensional counterpart. These transformations are linear or nonlinear and can be guided by some optimization criterion. The commonly encountered methods uti-

lize Principal Component Analysis (PCA) which is an example of a linear feature transformation.

Dimensionality reduction A way of converting a large data set into a representative subset. Typically, the data are grouped into clusters whose prototypes are representatives of the overall data set.

Curse of dimensionality A phenomenon of a rapid exponential increase of computing related with the dimensionality of the problem (say, the number of data or the number of features) which prevents us from achieving an optimal solution. The curse of dimensionality leads to the construction of sub-optimal solutions.

Data mining A host of activities aimed at discovery of easily interpretable and experimentally sound findings in huge data sets.

Biologically-inspired optimization An array of optimization techniques realizing searches in highly-dimensional spaces where the search itself is guided by a collection of mechanisms (operators) inspired by biological search processes. Genetic algorithms, evolutionary methods, particle swarm optimization, and ant colonies are examples of biologically-inspired search techniques.

Definition of the Subject

Data and dimensionality reduction are fundamental pursuits of data analysis and system modeling. With the rapid growth of sizes of data sets and the diversity of data themselves, the use of some reduction mechanisms becomes a necessity. Data reduction is concerned with a reduction of sizes of data sets in terms of the number of data points. This helps reveal an underlying structure in data by presenting a collection of groups present in data. Given a number of groups which is very limited, the clustering mechanisms become effective in terms of data reduction. Dimensionality reduction is aimed at the reduction of the number of attributes (features) of the data which leads to a typically small subset of features or brings the data from a highly dimensional feature space to a new one of a far

lower dimensionality. A joint reduction process involves data and feature reduction.

Introduction

In the information age, we are continuously flooded by enormous amounts of data that need to be stored, transmitted, processed and understood. We expect to make sense of data and find general relationships within them. On the basis of data, we anticipate to construct meaningful models (classifiers or predictors). Being faced with this ongoing quest, there is an intensive research along this line with a clear cut objective to design effective and computationally feasible algorithms to combat a curse of dimensionality which comes associated with the data. Data mining has become one of the dominant developments in data analysis. The term intelligent data analysis is another notion which provides us with a general framework supporting thorough and user-oriented processes of data analysis in which reduction processes play a pivotal role.

Interestingly enough, the problem of dimensionality reduction and complexity management is by no means a new endeavor. It has been around for a number of decades almost from the very inception of computer science. One may allude here to pattern recognition, data visualization as the two areas in which we were faced with inherent data dimensionality. This has led to a number of techniques which as of now are regarded classic and are used quite intensively. There have been a number of approaches deeply rooted in classic statistical analysis. The ideas of principal component analysis, Fisher analysis and alike are the techniques of paramount relevance. What has changed quite profoundly over the decades is the magnitude of the problem itself which has forced us to the exploration of new ideas and optimization techniques involving advanced techniques of global search including tabu search and biologically-inspired optimization mechanisms.

In a nutshell, we can distinguish between reduction processes involving (a) data and (b) features (attributes). Data reduction is concerned with grouping data and revealing their structure in the form of clusters (groups). Clustering is regarded as one of the fundamental techniques within the domain of data reduction. Typically, we start with thousands of data points and arrive at 10–15 clusters. Feature or attribute reduction deals with a (a) transformation of the feature space into another feature space of a far lower dimensionality or (b) selection of a subset of features that are regarded to be the most essential with respect to a certain predefined objective function. Considering the underlying techniques of feature transfor-

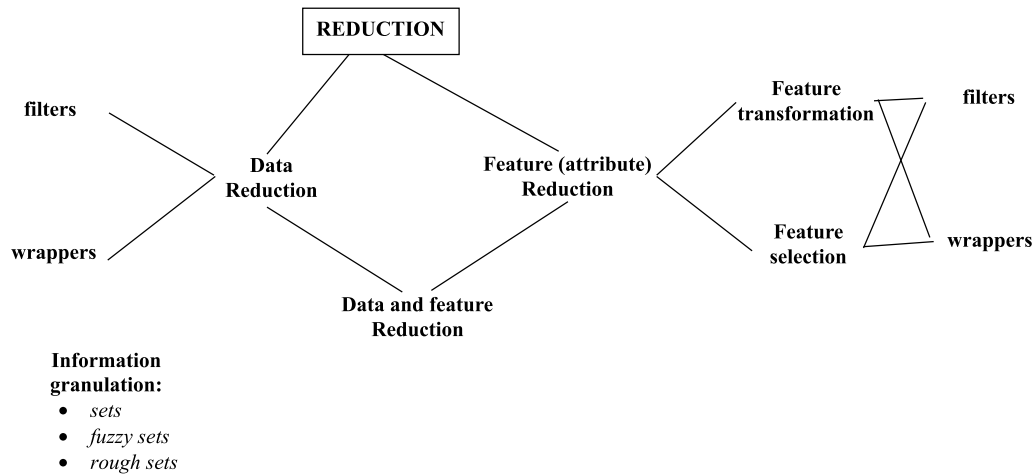
mation, we encounter a number of classic linear statistical techniques such as e. g., principal component analysis or more advanced nonlinear mapping mechanisms realized by e. g., neural networks.

The criteria used to assess the quality of the resulted (reduced) feature space give rise to the two general categories, namely *filters* and *wrappers*. Using filters we consider some criterion that pertains to the statistical characteristics of the selected attributes and evaluate them with this respect. In contrast, when dealing with wrappers, we are concerned with the effectiveness of the features as a vehicle to carry out classification so in essence there is a mechanism (e. g., a certain classifier) which effectively evaluates the performance of the selected features with respect to their discriminating capabilities.

In addition to feature and data reduction being regarded as two separate processes, we may consider their combinations. A general roadmap of the dimensionality reduction is outlined in Fig. 1. While the most essential components have been already described, we note that all reduction processes are guided by various criteria. The reduction activities are established in some formal frameworks of information granules.

In the study, we will start with an idea of information granularity and information granules, Sect. “[Introduction](#)”, in which we demonstrate their fundamental role of the concept that leads to the sound foundations and helps establish a machinery of data and feature reduction. We start with data reduction (Sect. “[Information Granularity and Granular Computing](#)”) where we highlight the role of clustering and fuzzy clustering as a processing vehicle leading to the discovery of structural relationships within the data. We also cover several key measures used to the evaluation of reduction process offering some thoughts on fuzzy quantization which brings a quantitative characterization of the reconstruction process. In this setting, we show that a tandem of encoding and decoding is guided by well-grounded optimization criteria. Dimension reduction is covered in Sect. “[Data Reduction](#)”. Here we start with linear and nonlinear transformations of the feature space including standard methods such as a well-known Principal Component Analysis (PCA). In the sequel, we stress a way in which biologically inspired optimization leads to the formation of the optimal subset of features. Some mechanisms of co-joint data and dimensionality reduction are outlined in Sect. “[Dimensionality Reduction](#)” in which we discuss a role of biclustering in these reduction problems.

In this paper, we adhere to the standard notation. Individual data are treated as n -dimensional vectors of real numbers, say $\mathbf{x}_1, \mathbf{x}_2, \dots$ etc. We consider a collection of



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 1

A general roadmap of reduction processes; note various ways of dealing with data and dimensionality along with a way of providing some evaluation mechanisms (performance indexes)

N data points which could be arranged in a matrix form where data occupy consecutive rows of the matrix. Furthermore we will be using the terms attribute and feature interchangeably.

Information Granularity and Granular Computing

Information granules permeate numerous human endeavors [1,19,20,22,23,24,25,26]. No matter what problem is taken into consideration, we usually express it in a certain conceptual framework of basic entities, which we regard to be of relevance to the problem formulation and problem solving. This becomes a framework in which we formulate generic concepts adhering to some level of abstraction, carry out processing, and communicate the results to the external environment. Consider, for instance, image processing. In spite of the continuous progress in the area, a human being assumes a dominant and very much uncontested position when it comes to understanding and interpreting images. Surely, we do not focus our attention on individual pixels and process them as such but group them together into semantically meaningful constructs – familiar objects we deal with in everyday life. Such objects involve regions that consist of pixels or categories of pixels drawn together because of their proximity in the image, similar texture, color, etc. This remarkable and unchallenged ability of humans dwells on our effortless ability to construct information granules, manipulate them and arrive at sound conclusions. As another example, consider a collection of time series. From our perspective we can describe them in a semi-qualitative manner by pointing at

specific regions of such signals. Specialists can effortlessly interpret ECG signals. They distinguish some segments of such signals and interpret their combinations. Experts can interpret temporal readings of sensors and assess the status of the monitored system. Again, in all these situations, the individual samples of the signals are not the focal point of the analysis and the ensuing signal interpretation. We always granulate all phenomena (no matter if they are originally discrete or analog in their nature). Time is another important variable that is subjected to granulation. We use seconds, minutes, days, months, and years. Depending which specific problem we have in mind and who the user is, the size of information granules (time intervals) could vary quite dramatically. To the high level management time intervals of quarters of year or a few years could be meaningful temporal information granules on basis of which one develops any predictive model. For those in charge of everyday operation of a dispatching plant, minutes and hours could form a viable scale of time granulation. For the designer of high-speed integrated circuits and digital systems, the temporal information granules concern nanoseconds, microseconds, and perhaps microseconds. Even such commonly encountered and simple examples are convincing enough to lead us to ascertain that (a) information granules are the key components of knowledge representation and processing, (b) the level of granularity of information granules (their size, to be more descriptive) becomes crucial to the problem description and an overall strategy of problem solving, (c) there is no universal level of granularity of information; the size of granules is problem-oriented and user dependent.

What has been said so far touched a qualitative aspect of the problem. The challenge is to develop a computing framework within which all these representation and processing endeavors could be formally realized. The common platform emerging within this context comes under the name of Granular Computing. In essence, it is an emerging paradigm of information processing. While we have already noticed a number of important conceptual and computational constructs built in the domain of system modeling, machine learning, image processing, pattern recognition, and data compression in which various abstractions (and ensuing information granules) came into existence, Granular Computing becomes innovative and intellectually proactive in several fundamental ways

- It identifies the essential commonalities between the surprisingly diversified problems and technologies used there which could be cast into a unified framework we usually refer to as a granular world. This is a fully operational processing entity that interacts with the external world (that could be another granular or numeric world) by collecting necessary granular information and returning the outcomes of the granular computing.
- With the emergence of the unified framework of granular processing, we get a better grasp as to the role of interaction between various formalisms and visualize a way in which they communicate.
- It brings together the existing formalisms of set theory (interval analysis) [8,10,15,21], fuzzy sets [6,26], rough sets [16,17,18], etc. under the same roof by clearly visualizing that in spite of their visibly distinct underpinnings (and ensuing processing), they exhibit some fundamental commonalities. In this sense, Granular Computing establishes a stimulating environment of synergy between the individual approaches.
- By building upon the commonalities of the existing formal approaches, Granular Computing helps build heterogeneous and multifaceted models of processing of information granules by clearly recognizing the orthogonal nature of some of the existing and well established frameworks (say, probability theory coming with its probability density functions and fuzzy sets with their membership functions).
- Granular Computing fully acknowledges a notion of variable granularity whose range could cover detailed numeric entities and very abstract and general information granules. It looks at the aspects of compatibility of such information granules and ensuing communication mechanisms of the granular worlds.
- Interestingly, the inception of information granules is highly motivated. We do not form information granules without reason. Information granules arise as an evident realization of the fundamental paradigm of abstraction.

Granular Computing forms a unified conceptual and computing platform. Yet, it directly benefits from the already existing and well-established concepts of information granules formed in the setting of set theory, fuzzy sets, rough sets and others. A selection of a certain formalism from this list depends upon the problem at hand. While in dimensionality reduction set theory becomes more visible, in data reduction we might encounter both set theory and fuzzy sets.

Data Reduction

In data reduction, we transform large collections of data into a limited, quite small family of their representatives which capture the underlying structure (topology). Clustering has become one of the fundamental tools being used in this setting. Depending upon the distribution of data, we may anticipate here the use of set theory or fuzzy sets. As an illustration, consider a two-dimensional data portrayed in Fig. 2.

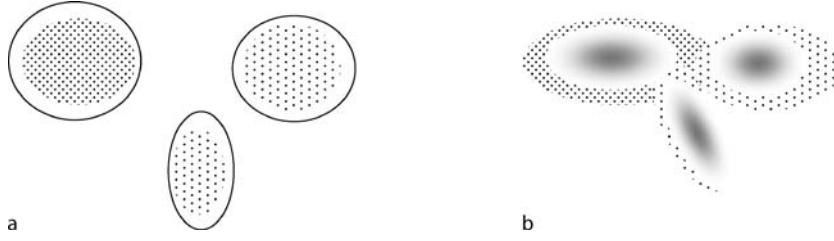
The data in Fig. 2a exhibit a clearly delineated structure with three well separated clusters. Each cluster can be formalized as a set of elements. A situation illustrated in Fig. 2b is very different: while there are some clusters visible, they overlap to some extent and a number of points are “shared” by several clusters. In other words, some data may belong to more than a single cluster and hence the emergence of fuzzy sets as a formal setting in which we can formalize the resulting information granules.

In what follows, we briefly review the essence of fuzzy clustering and then move on with the ideas of the evaluation of quality of the results.

Fuzzy C-means as an Algorithmic Vehicle of Data Reduction Through Fuzzy Clusters

Fuzzy sets can be formed on a basis of numeric data through their clustering (groupings). The groups of data give rise to membership functions that convey a global more abstract view at the available data. With this regard Fuzzy C-Means (FCM) is one of the commonly used mechanisms of fuzzy clustering [2,3,20].

Let us review its formulation, develop the algorithm and highlight the main properties of the fuzzy clusters. Given a collection of n -dimensional data set $\{\mathbf{x}_k\}$, $k =$



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 2

Data clustering and underlying formalism of set theory suitable for handling well-delineated structures of data (a) and the use of fuzzy sets in capturing the essence of data with significant overlap (b)

$1, 2, \dots, N$, the task of determining its structure – a collection of “ c ” clusters, is expressed as a minimization of the following objective function (performance index) Q being regarded as a sum of the squared distances between data and their representatives (prototypes)

$$Q = \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m \|x_k - v_i\|^2. \quad (1)$$

Here v_i s are n -dimensional prototypes of the clusters, $i = 1, 2, \dots, c$ and $U = [u_{ik}]$ stands for a partition matrix expressing a way of allocation of the data to the corresponding clusters; u_{ik} is the membership degree of data x_k in the i th cluster. The distance between the data z_k and prototype v_i is denoted by $\| \cdot \|$. The fuzzification coefficient m (> 1.0) expresses the impact of the membership grades on the individual clusters.

A partition matrix satisfies two important and intuitively appealing properties

$$\begin{aligned} \text{(a)} \quad & 0 < \sum_{k=1}^N u_{ik} < N, \quad i = 1, 2, \dots, c \\ \text{(b)} \quad & \sum_{i=1}^c u_{ik} = 1, \quad k = 1, 2, \dots, N. \end{aligned} \quad (2)$$

Let us denote by U a family of matrices satisfying (a)–(b). The first requirement states that each cluster has to be nonempty and different from the entire set. The second requirement states that the sum of the membership grades should be confined to 1.

The minimization of Q completed with respect to $U \in \mathcal{U}$ and the prototypes v_i of $V = \{v_1, v_2, \dots, v_c\}$ of the clusters. More explicitly, we write it down as follows

$$\min Q \quad \text{with respect to} \quad U \in \mathcal{U}, v_1, v_2, \dots, v_c \in \mathbf{R}^n. \quad (3)$$

From the optimization standpoint, there are two individual optimization tasks to be carried out separately for the

partition matrix and the prototypes. The first one concerns the minimization with respect to the constraints given the requirement of the form (2b) which holds for each data point x_k . The use of Lagrange multipliers converts the problem into its constraint-free version. The augmented objective function formulated for each data point, $k = 1, 2, \dots, N$, reads as

$$V = \sum_{i=1}^c u_{ik}^m d_{ik}^2 + \lambda \left(\sum_{i=1}^c u_{ik} - 1 \right) \quad (4)$$

where $d_{ik}^2 = \|x_k - v_i\|^2$.

It is now instructive to go over the optimization process. Proceeding with the necessary conditions for the minimum of V for $k = 1, 2, \dots, N$, one has

$$\frac{\partial V}{\partial u_{st}} = 0 \quad \frac{\partial V}{\partial \lambda} = 0 \quad (5)$$

$s = 1, 2, \dots, c, t = 1, 2, \dots, N$. Now we calculate the derivative of V with respect to the elements of the partition matrix in the following way

$$\frac{\partial V}{\partial u_{st}} = m u_{st}^{m-1} d_{st}^2 + \lambda. \quad (6)$$

Given this relationship, using (5) we calculate u_{st} to be equal to

$$u_{st} = - \left(\frac{\lambda}{m} \right)^{\frac{1}{m-1}} d_{st}^{\frac{2}{m-1}}. \quad (7)$$

Given the normalization condition $\sum_{j=1}^c u_{jt} = 1$ and plugging it into (2) one has

$$- \left(\frac{\lambda}{m} \right)^{\frac{1}{m-1}} \sum_{j=1}^c d_{jt}^{\frac{2}{m-1}} = 1 \quad (8)$$

we compute

$$- \left(\frac{\lambda}{m} \right)^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^c d_{jt}^{\frac{2}{m-1}}}. \quad (9)$$

Inserting this expression into (7), we obtain the successive entries of the partition matrix

$$u_{st} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{st}^2}{d_{jt}^2} \right)^{\frac{1}{m-1}}} . \quad (10)$$

The optimization of the prototypes \mathbf{v}_i is carried out assuming the Euclidean distance between the data and the prototypes that is $\|\mathbf{x}_k - \mathbf{v}_i\|^2 = \sum_{j=1}^n (x_{kj} - v_{ij})^2$. The objective function reads now as follows $Q = \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m \sum_{j=1}^n (x_{kj} - v_{ij})^2$ and its gradient with respect to \mathbf{v}_i , $\nabla_{\mathbf{v}_i} Q$ made equal to zero yields the system of linear equations

$$\sum_{k=1}^N u_{ik}^m (x_{kt} - v_{it}) = 0 \quad (11)$$

$s = 1, 2, \dots, c; t = 1, 2, \dots, n$.

Thus

$$v_{it} = \frac{\sum_{k=1}^N u_{ik}^m x_{kt}}{\sum_{k=1}^N u_{ik}^m} . \quad (12)$$

Overall, the FCM clustering is completed through a sequence of iterations where we start from some random allocation of data (a certain randomly initialized partition matrix) and carry out the following updates by adjusting the values of the partition matrix and the prototypes. The iterative process is continued until a certain termination criterion has been satisfied. Typically, the termination condition is quantified by looking at the changes in the membership values of the successive partition matrices.

Denote by $U(t)$ and $U(t+1)$ the two partition matrices produced in two consecutive iterations of the algorithm. If the distance $\|U(t+1) - U(t)\|$ is less than a small predefined threshold ε (say, $\varepsilon = 10^{-5}$ or 10^{-6}), then we terminate the algorithm. Typically, one considers the Tchebyshev distance between the partition matrices meaning that the termination criterion reads as follows

$$\max_{i,k} |u_{ik}(t+1) - u_{ik}(t)| \leq \varepsilon . \quad (13)$$

The key components of the FCM and a quantification of their impact on the form of the produced results are summarized in Table 1.

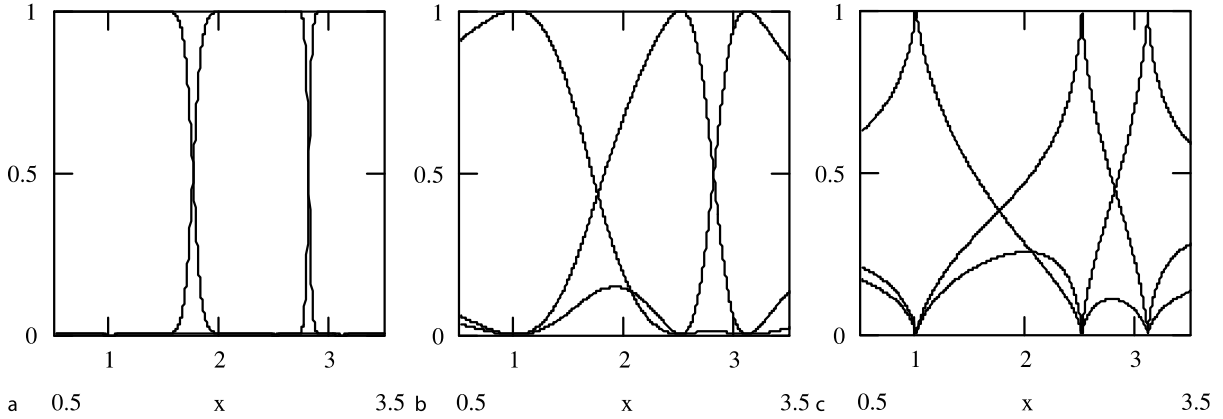
The fuzzification coefficient exhibits a direct impact on the geometry of fuzzy sets generated by the algorithm. Typically, the value of “ m ” is assumed to be equal to 2.0. Lower values of m (that are closer to 1) yield membership functions that start resembling characteristic functions of sets; most of the membership values become localized around 1 or 0. The increase of the fuzzification coefficient ($m = 3, 4$, etc.) produces “spiky” membership functions with the membership grades equal to 1 at the prototypes and a fast decline of the values when moving away from the prototypes. Several illustrative examples of the membership functions are included in Fig. 3. In addition to the varying shape of the membership functions, observe that the requirement put on the sum of membership grades imposed on the fuzzy sets yields some rippling effect: the membership functions are not unimodal but may exhibit some ripples whose intensity depends upon the distribution of the prototypes and the values of the fuzzification coefficient.

The membership functions offer an interesting feature of evaluating an extent to which a certain data point is shared between different clusters and in this sense become

Data and Dimensionality Reduction in Data Analysis and System Modeling, Table 1

The main features of the Fuzzy C-Means (FCM) clustering algorithm

Feature of the FCM algorithm	Representation and optimization aspects
Number of clusters (c)	Structure in the data set and the number of fuzzy sets estimated by the method; the increase in the number of clusters produces lower values of the objective function however given the semantics of fuzzy sets one should maintain this number quite low (5–9 information granules)
Objective function Q	Develops the structure aimed at the minimization of Q ; iterative process supports the determination of the local minimum of Q
Distance function $\ \cdot\ $	Reflects (or imposes) a geometry of the clusters one is looking for; essential design parameter affecting the shape of the membership functions
Fuzzification coefficient (m)	Implies a certain shape of membership functions present in the partition matrix; essential design parameter. Low values of “ m ” (being close to 1.0) induce characteristic function. The values higher than 2.0 yield spiky membership functions
Termination criterion	Distance between partition matrices in two successive iterations; the algorithm terminated once the distance below some assumed positive threshold (ε) that is $\ U(\text{iter} + 1) - U(\text{iter})\ < \varepsilon$



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 3

Examples of membership functions of fuzzy sets; the prototypes are equal to 1, 3.5 and 5 while the fuzzification coefficient assumes values of 1.2 (a), 2.0 (b) and 3.5 (c). The intensity of the rippling effect is affected by the values of “ m ” and increases with the higher values of “ m ”

difficult to allocate to a single cluster (fuzzy set). Let us introduce the following index which serves as a certain separation measure between the clusters

$$\varphi(u_1, u_2, \dots, u_c) = 1 - c^c \prod_{i=1}^c u_i \quad (14)$$

where u_1, u_2, \dots, u_c are the membership degrees for some data point. If only one of membership degrees, say $u_i = 1$, and the remaining are equal to zero, then the separation index attains its maximum equal to 1. On the other extreme, when the data point is shared by all clusters to the same degree equal to $1/c$, then the value of the index drops down to zero. This means that there is no separation between the clusters as reported for this specific point.

While the number of clusters is typically limited to a few information granules, we can easily proceed with successive refinements of fuzzy sets. This can be done by splitting fuzzy clusters of the highest heterogeneity. Let us assume that we have already constructed “ c ” fuzzy clusters. Each of them can be characterized by the performance index

$$V_i = \sum_{k=1}^N u_{ik}^m \|x_k - v_i\|^2 \quad (15)$$

$i = 1, 2, \dots, c$. The higher the value of V_i , the more heterogeneous the i th cluster. The one with the highest value of V_i , that is the one for which we have $i_0 = \operatorname{argmax}_i V_i$ is refined by being split into two clusters. Denote the set of data associated with the i_0 th cluster by $X(i_0)$,

$$X(i_0) = \{x_k \in X | u_{i_0 k} = \max_i u_{ik}\} \quad (16)$$

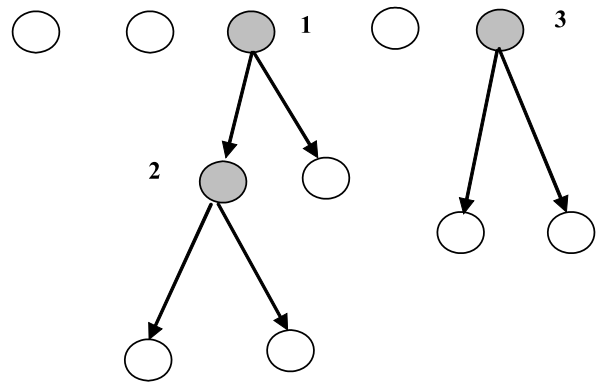
We cluster the elements in $X(i_0)$ by forming two clusters which leads to two more specific (detailed) fuzzy sets. This

gives rise to a hierarchical structure of the family of fuzzy sets as illustrated in Fig. 4. The relevance of this construct in the setting of fuzzy sets is that it emphasizes the essence of forming a hierarchy of fuzzy sets rather than working with a single level structure of a large number of components whose semantics could not be retained.

The process of further refinements is realized in the same by picking up the cluster of the highest heterogeneity and its split into two consecutive clusters.

Evaluation of Performance of Data Reduction

There are two fundamental ways of evaluating the quality of data reduction achieved through clustering or fuzzy clustering. In the first category of methods, we assess the



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 4

Successive refinements of fuzzy sets through fuzzy clustering applied to the clusters of the highest heterogeneity. The numbers indicate the order of the splits

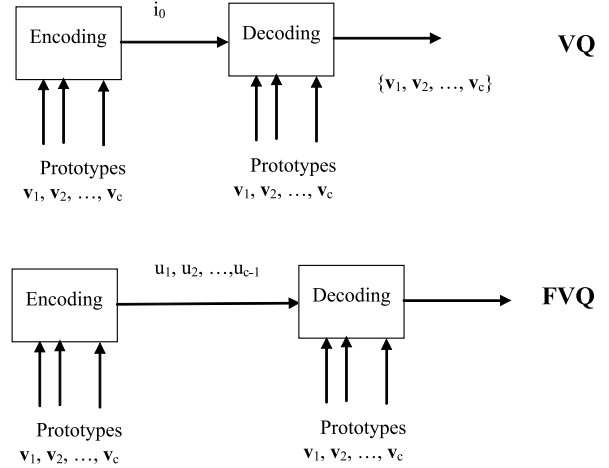
quality of the obtained clusters in terms of their compactness, low fuzziness (entropy measure of fuzziness, in particular), structural separability, and alike. We also look for a “plausible” number of clusters which can be descriptive of the structure of the overall data. The second category of methods stresses the aspects of data representation by its reduced counterpart. In this case, a reconstruction criterion plays a pivotal role. We elaborate on this method by looking at it as a useful vehicle of fuzzy quantization. Furthermore we contrast fuzzy quantization with its set-based counterpart.

Exploring Fuzzy Quantization: Moving Beyond a Winner-Takes-All Scheme

The crux of the considerations here dwells upon the following observation. While the set-based codebook developed through clustering [5,7] leads to a decoding scheme which decodes the result using a single element of the codebook (which in essence becomes a manifestation of the well-known concept of the *winner-takes-all* strategy), here we are interested in the exploitation of the nonzero degrees of membership of several elements (fuzzy sets) of the codebook while representing the input datum. In other words, rather than using a single prototype as a sole representative of a collection of neighboring data, our conjecture is that involving several prototypes at different degrees of activation (weights) could be beneficial to the ensuing decoding. Along the line of abandoning the winner-takes-all principle, let us start with a collection of weights $u_i(\mathbf{x})$, $i = 1, 2, \dots, c$. Adhering to the vector notation, we have $\mathbf{u}(\mathbf{x}) = [u_1(\mathbf{x}) \ u_2(\mathbf{x}), \dots, u_c(\mathbf{x})]^T$. The weights (membership degrees) express an extent to which the corresponding datum \mathbf{x} is encoded in the language of the given prototypes (elements of the codebook) should be involved in the decoding (decompression) scheme. We require that these membership degrees are in-between 0 and 1 and sum up to 1. Thus at the encoding end of the overall scheme, we represent each vector \mathbf{x} by $c - 1$ values of $u_i(\mathbf{x})$. The decoding is then based upon a suitable aggregation of the degrees of membership and the prototypes. Denote this operation by $\hat{\mathbf{x}} = D(\mathbf{u}(\mathbf{x}), \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c)$ where $\hat{\mathbf{x}}$ denotes the result of the decoding. On the other hand, the formation of the membership degrees (encoding) can be succinctly described in the form of the mapping $E(\mathbf{x}, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c)$.

The overall development process is split into two fundamental phases, namely

- (a) local optimization activities confined to the individual datum \mathbf{x}_k that involve (i) encoding each \mathbf{x}_k leading to the vector of the membership degrees (u_i), (ii) decod-



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 5

VQ and FVQ – a view contrasting the essence of the process and showing the key elements of the ensuing encoding and decoding

- ing being realized in terms of the membership degrees, and
- (b) global optimization activities concerning the formation of the codebook in which case we take all data into consideration thus bringing the design to the global level.

The overall process described above will be referred to as fuzzy vector quantization (FVQ) or a fuzzy granulation-degranulation. The latter terminology emphasizes a position of fuzzy sets played in this construct. In contrast, the winner-takes-all strategy is a cornerstone of the vector quantization (VQ). To contrast the underlying computing in the VQ and FVQ schemes, we portray the essential computational facets in Fig. 5.

We focus on the detailed formulas expressing the coding and decoding schemes. Let us emphasize that the coding and decoding emerge as solutions to the well-defined optimization problems. At this point let us assume that the codebook – denoted here as $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c\}$ has been already formed and become available.

Encoding Mechanism A way of encoding (representing) the original datum \mathbf{x} is done through the collection of the degrees of activation of the elements of the codebook. We require that the membership degrees are confined to the unit interval and sum up to 1. We determine their individual values by minimizing the following performance index

$$Q_1(\mathbf{x}) = \sum_{i=1}^c u_i^m \|\mathbf{x} - \mathbf{v}_i\|^2 \quad (17)$$

subject to the following constraints already stated above, that is

$$u_i(\mathbf{x}) \in [0, 1], \quad \sum_{i=1}^c u_i(\mathbf{x}) = 1. \quad (18)$$

The distance function is denoted by $\|\cdot\|^2$. The fuzzification coefficient (m , $m > 1$), standing in the above expression is used to adjust the level of contribution of the impact of the prototypes on the result of the encoding. The collection of “ c ” weights $\{u_i(\mathbf{x})\}$ is then used to encode the input datum \mathbf{x} . These membership degrees along with the corresponding prototypes are afterwards used in the decoding scheme.

The minimization of (17) is straightforward and follows a standard way of transforming the problem to unconstrained optimization using Lagrange multipliers. Once solved, the resulting weights (membership degrees) read as

$$u_i(\mathbf{x}) = \frac{1}{\sum_{j=1}^c \left(\frac{\|\mathbf{x} - \mathbf{v}_i\|}{\|\mathbf{x} - \mathbf{v}_j\|} \right)^{\frac{2}{m-1}}}. \quad (19)$$

The fuzzification coefficient used in the formulation of the encoding problem plays an important role. To demonstrate this, let us show how the values of the coefficient impact the distribution of the generated membership degrees. The plots of these degrees in the two-dimensional case ($n = 2$) and $c = 3$ elements of the codebook are included in Fig. 6. The prototypes have been set up to $\mathbf{v}_1 = [3.0 \ 1.5]^T$, $\mathbf{v}_2 = [2.5 \ 0.8]^T$, $\mathbf{v}_3 = [1.0 \ 3.7]^T$. The values of “ m ” close to 1 result in a visible set-like character of the distribution of the membership degrees.

The Decoding Mechanism The decoding is concerned with the “reconstruction” of \mathbf{x} , denoted here by $\hat{\mathbf{x}}$. It is based on some aggregation of the elements of the codebook and the associated membership grades $u(\mathbf{x})$. The

proposed way of forming $\hat{\mathbf{x}}$ is accomplished through the minimization of the following expression

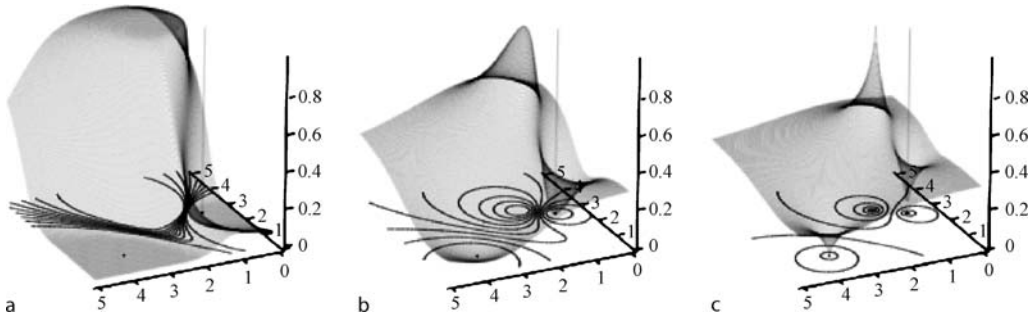
$$Q_2(\hat{\mathbf{x}}) = \sum_{i=1}^c u_i^m \|\hat{\mathbf{x}} - \mathbf{v}_i\|^2. \quad (20)$$

Given the Euclidean distance, the problem of unconstrained optimization leads to a straightforward solution expressed as a combination of the prototypes weighted by the membership degrees, that is

$$\hat{\mathbf{x}} = \frac{\sum_{i=1}^c u_i^m \mathbf{v}_i}{\sum_{i=1}^c u_i^m}. \quad (21)$$

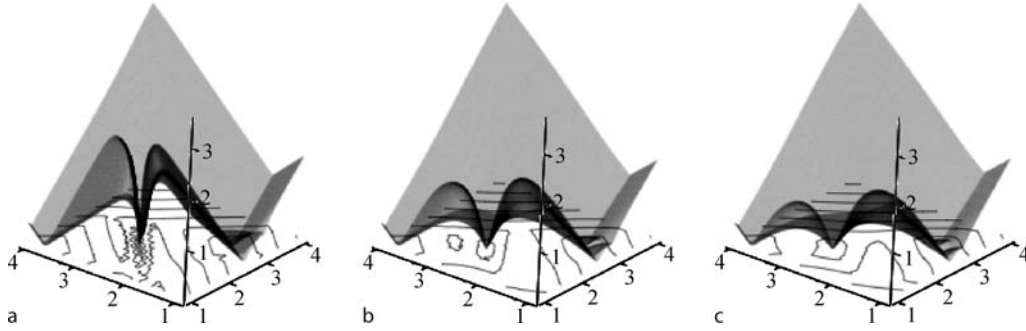
Interestingly, all prototypes contribute to the decoding process and this way of computing stands in a sharp contrast with the winner-takes-all decoding scheme encountered in the VQ where $\hat{\mathbf{x}} = \mathbf{v}_l$ where l denotes the index of the winning prototype that was identified during the decoding phase. Some other VQ decoding schemes are available in which several dominant elements of the codebook are involved. Nevertheless because of their binary involvement in the decoding phase, the possible ways of aggregation to be realized by the decoder are quite limited.

The quality of the reconstruction depends on a number of essential parameters including the size of the codebook as well as the value of the fuzzification coefficient. The impact of these particular parameters is illustrated in Fig. 7 in which we quantify the distribution of the decoding error by showing values of the Hamming distance between \mathbf{x} and $\hat{\mathbf{x}}$. Evidently, what could have been anticipated, all \mathbf{x} ’s positioned in a close vicinity of the prototypes exhibit low values of the decoding error. Low values of “ m ” lead to significant jumps of the error which result from the abrupt switching in-between the prototypes used in the decoding process. For the higher value of the fuzzification coefficient, say $m = 2.0$, the jumps are significantly



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 6

Plots of $u_1(\mathbf{x})$ (3D and contour plots) for selected values of the fuzzification coefficient: a $m = 1.2$, b $m = 2.0$, c $m = 3.5$



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 7

Plots of the decoding error expressed in the form of the Hamming distance between x and \hat{x} (3D and contour plots) for selected values of the fuzzification coefficient: **a** $m = 1.2$, **b** $m = 2.0$, **c** $m = 3.5$

reduced. Further reduction is achieved with the larger values of “ m ”, see Fig. 7c.

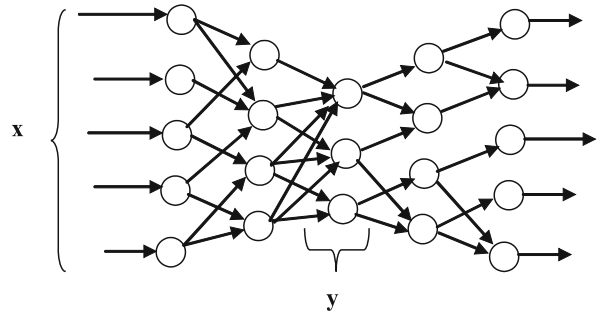
Obviously in the regions which are quite remote from the location of the prototypes, the error increases quite quickly. For instance, this effect is quite visible for the values of x close to $[4.0 \ 4.0]^T$.

Dimensionality Reduction

As indicated earlier, dimensionality reduction is concerned with a reduction of the set of features. One of the two fundamental directions that are taken here involves transformations of the original, highly dimensional feature vectors. The other group of methods focuses on a selection of an “optimal” subset of original attributes.

Linear and Nonlinear Transformations of the Feature Space

Making use of such transformations, we position the original vectors in some other spaces of far lower dimensionality than the original one. There are a number of statistical methods devoted to the formation of optimal linear transformations. For instance, the commonly used Principal Component Analysis (PCA) [11] deals with a linear mapping of original data in such a way that the reduced data vectors y are expressed as $y = Tx$ where T denotes a transformation matrix and y is positioned in the reduced space of dimensionality “ m ” where $m \ll n$. The transformation matrix is constructed using eigenvectors associated with the largest eigenvalues of the covariance matrix of the data, that is $\Sigma = \sum_{k=1}^N (x_k - m)(x_k - m)^T$ where m is a mean vector of all data. The PCA method does not consider the labels of data in case they become available. To accommodate class information, one may consider Fisher discriminant analysis [3].



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 8

Neural network in nonlinear dimensionality reduction; the inner shortest layer produces a reduced version of the original data (y)

Neural networks can serve as nonlinear transformations of the original feature space. More specifically, the dimensionality reduction is accomplished through a sand glass topology of the network as illustrated in Fig. 8. Notice that the shortest hidden layer of the network produces the reduced version of data. The optimization criterion deals with the minimal reconstruction error. For each input x , we strive for the output of the network (NN) to be made as close as possible to the input, that is $NN(x) \approx x$. The learning (adjustments) of the weights helps minimize the error (objective function) of the form $\sum_{k=1}^N ||NN(x_k) - x_k||^2$.

Given the evident nonlinear nature of neurocomputing, we often allude to these transformations as a nonlinear component analysis (NLCA).

Neural networks are also used as a visualization vehicle in which we display original data in some low-dimensional space. The typical architecture comes in the form of Kohonen self-organizing maps [13]. Data that are distant in an original space remain distant in the low-dimensional

space. Likewise, if the two data points are originally close to each other, this closeness is also present in the transformed feature space. While Kohonen maps realize a certain form of clustering, their functioning is quite distinct from the FCM algorithm described before. In essence, self-organizing maps bring a high involvement of humans in the final interpretation of results. For instance, the number of clusters is not predefined in advance, as encountered in the FCM, but, its choice is left to the end user to determine on a basis of the structure displayed on the map.

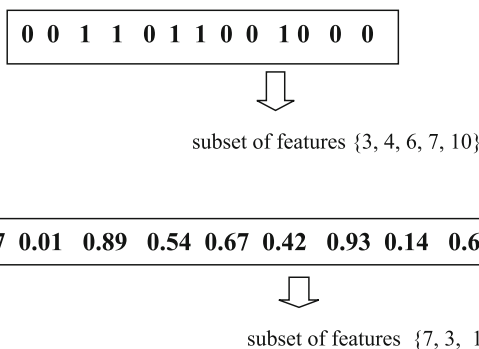
The Development of Optimal Subsets of Features

An objective is to select a subset of attributes so that a certain predefined performance index (selection criterion) becomes optimized (either minimized or maximized). The choice of a subset of attributes is a combinatorial problem. As a matter of fact, it is NP complete. The only optimal solution comes through exhaustive search. Evidently, such an enumeration of possibilities is feasible for problems of a very low dimensionality. In essence, the practicality of such enumeration is out of question. This has led to numerous techniques which rely on some sort of heuristics which help reduce the size of the search space. While the heuristics could be effective, its usage could easily build some bias into the search process. Subsequently, the results produced by such methods are suboptimal. In addition to a number of heuristics (which contribute to the effectiveness of the search process), there are more advanced optimization techniques such as those available in the realm of Evolutionary Computing [14] such as genetic algorithms, evolutionary strategies, particle swarm optimization are examples of biologically inspired techniques

that are aimed at structural optimization using which we construct an optimal (or sub-optimal) collection of attributes. Any potential solution is represented in a form of a chromosome that identifies attributes contributing to the formation of the optimal space. We may encounter either binary or real-number chromosomes whose size is equal to the number of features. In case of the binary chromosome, we choose attributes which are identified as 1, see Fig. 9a. For the required “ m ” features, we choose the first ones encountered in the chromosome. For the real-number coding of the chromosome, we typically have far more alternatives to take advantage of. For instance, we could consider the first “ m ” largest entries of the chromosome. Then the coordinates of these entries form a reduced subset of the features; refer to Fig. 9b.

Optimization Environment in Dimensionality Reduction: Wrappers and Filters

Attribute selection is guided by some criterion. There are numerous ways of expressing a way in which a collection of features is deemed “optimal”. There is a generally accepted taxonomy of the methods in which we distinguish between so-called wrappers and filters. Filters embrace criteria that directly express the quality of the subset of features in terms of their homogeneity, compactness and alike. Wrappers, as the name implies, evaluate the quality of the subset of features after being used by some mechanism of classification or prediction. In other words, the assessment of the attributes is done once they have been “wrapped” in some construct. This means that the quality of the feature space is expressed indirectly. The essence of filters and wrappers is illustrated schematically in Fig. 10.



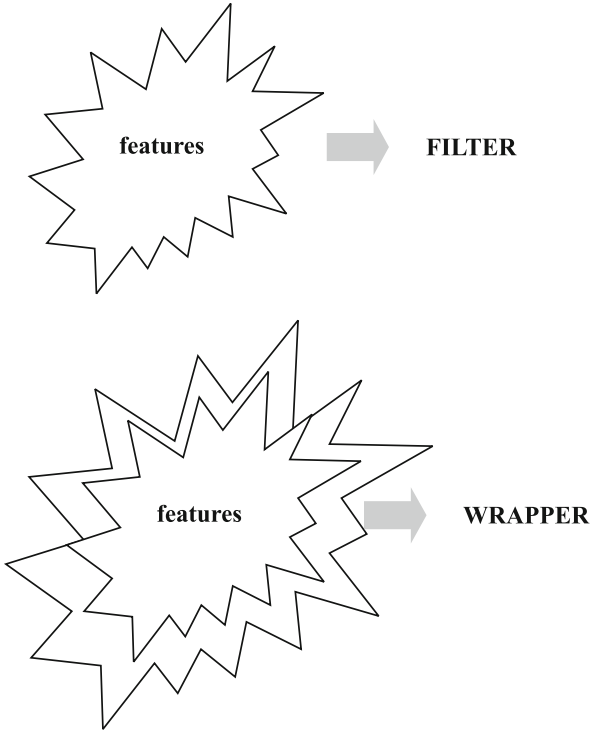
Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 9

Examples of forming a subset of features for a given chromosome: a binary coding, and b real-number coding (here we are concerned with a 3 dimensional feature space by choosing three dominant features)

Co-joint Data and Dimensionality Reduction

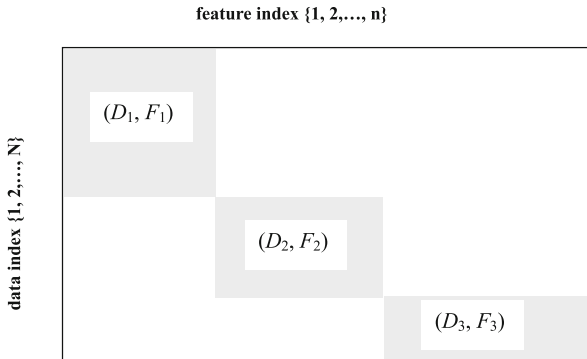
So far our investigations were concerned with data or dimensionality reduction carried out individually. An interesting question arises as to the possibility of carrying out concurrent reduction in both ways. One among possible alternatives can be envisioned in the form of so-called bi-clustering (or co-clustering). The crux of this process is to carry out simultaneous clustering for data and features. The result is a collection of groups of data and groups of features.

We start with a basic notation. Let us denote by D_1, D_2, \dots, D_c subsets of data. Similarly, by F_1, F_2, \dots, F_c we denote subsets of features. We may regard D_i and F_i as two subsets of indexes of data and features $\{1, 2, \dots, N\}$ and $\{1, 2, \dots, n\}$. $\{D_i\}$ and $\{F_i\}$ form partitions so that the fundamental requirements of disjointness and total space



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 10

Filters and wrappers in the assessment of the process of dimensionality reduction



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 11

An example heatmap of biclusters ($c = 3$)

coverage are satisfied. By biclusters of data set we mean a collection of pairs (D_i, F_i) $i = 1, 2, \dots, c$.

Graphically, biclusters are visualized in the form of so-called heatmaps where each bicluster is displayed as a certain region in the data and feature space, Fig. 11. Here the data are represented as a two-dimensional array X .

There have been a number of biclustering algorithms. The method proposed by Hartigan [9] and Cheng and Church [4] are just two examples of biclustering algorithms. The objective function introduced by Hartigan is concerned with the overall variance of the “ c ” biclusters

$$Q = \sum_{k=1}^c \sum_{i \in D_k} \sum_{j \in F_k} (x_{ij} - m_k)^2 \quad (22)$$

where m_k is the mean of the k th bicluster

$$m_k = \frac{1}{|D_k| |F_k|} \sum_{i \in D_k} \sum_{j \in F_k} x_{ij} \quad (23)$$

where $|\cdot|$ denotes a cardinality of clusters in data space and feature space. The minimization of Q leads to the development of the biclusters.

The idea of Cheng and Church is to form bi-clusters in such a way so that we minimize the mean squared residue of each bi-cluster. More specifically, denote an average taken over data in D_i and features in F_i as follows

$$\begin{aligned} \mu_{ij} &= \frac{1}{|D_i|} \sum_{k \in D_i} x_{kj} \\ v_{ik} &= \frac{1}{|F_i|} \sum_{j \in F_i} x_{kj} \end{aligned} \quad (24)$$

The residue of x_{ij} taken with respect to the k th bicluster is defined as

$$r_{ij} = x_{ij} - \mu_{ij} - v_{ij} + m_k$$

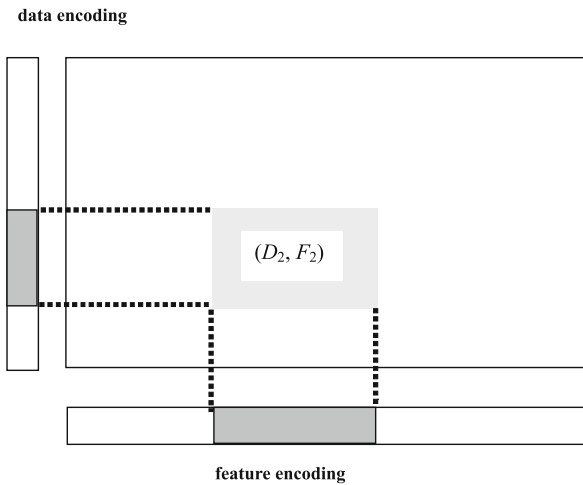
and then

$$V_k = \sum_{i \in D_k} \sum_{j \in F_k} r_{ij} \quad (25)$$

The overall minimization of $V_1 + V_2 + \dots + V_c$ gives rise to the formation of biclusters.

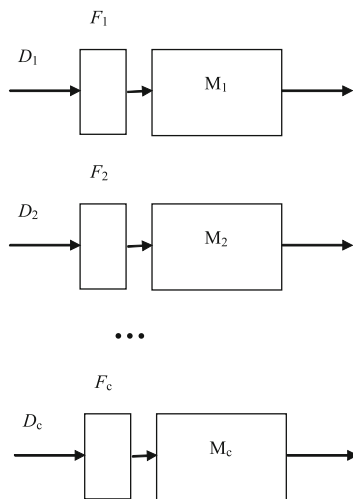
Biologically-inspired optimization can be used here to construct biclusters. Again, as before, a formation of a suitable structure (content) of a chromosome becomes crucial to the effectiveness of the obtained structure. One possible encoding could be envisioned as follows, refer to Fig. 12. Note that a chromosome consists of two parts where the first portion encodes the data while the other one deals with the features.

Assuming that a chromosome has entries in-between 0 and 1 and “ c ” biclusters are to be developed, we convert entries in the interval $[0, 1/c)$ to identify elements belonging to the first biclusters. Those entries that assume values in $[1/c, 2/c)$ indicate elements (either data or features) belonging to the second bicluster, etc.



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 12

Encoding of biclusters in terms of the chromosome with entries confined to the unit interval



Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 13

System modeling realized through a collection of local models developed for individual biclusters

Biclustering brings an interesting issue of simultaneous reduction of data and dimensionality. Each cluster of data is expressed in terms of its own (reduced) feature space. This interesting phenomenon gives rise to a collection of models which are locally formed by using subset of data and a subset of features. Each model is developed on a basis of the given bicluster, see Fig. 13.

In this way, the complexity of modeling has been substantially reduced as instead of global models (which are more difficult to construct), we develop a family of com-

pact models involving far fewer attributes and dealing with less data.

Conclusions

Reduction problems in data are essential to a number of fundamental developments of data analysis and system modeling. These activities are central to data understanding and constitute a core of generic activities of data mining and human-centric data interpretation. They are also indispensable in system modeling which otherwise being carried out for high-dimensional data make the development of the models highly inefficient and lead to the models that are of lower quality in particular when it comes to their generalization capabilities. The two facets of the reduction process involve data and features (attributes) and as such they tend to be highly complementary. When discussing individual techniques aimed at either data or dimensionality reduction, we indicated that there are interesting alternatives of dealing with these two aspects making use of biclustering.

Future Directions

There are a number of directions in which the area can expand either in terms of the reduction concept itself or the techniques being sought. Several of them might offer some tangible benefits to the future pursuits:

- The use of stratified approaches to data and dimensionality reduction in which the reduction processes are carried out on a hierarchical basis meaning that further reductions are contemplated bases on results developed at the lower level of generality.
- More active involvement and integration of techniques of visualization of results. While those are in place now, we may envision their more vital role in the future.
- Distributed mode of reduction activities where we simultaneously encounter several sources of data and intend to process them at the same time.
- Exploitation of joint data and dimensionality reduction approaches in which one tackles both the reduction of the number of data and the number of attributes. These two are quite intertwined so their mutual handling could be advantageous. In particular, one could envision their role in the formation of a suite of simpler models.
- Given the NP complete nature of many reduction problems, the use of some heuristics is inevitable. In this way, the role of techniques of biologically inspired optimization will be anticipated and it is very likely that their role will become more visible.

Acknowledgments

Support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chair (CRC) is gratefully acknowledged.

Appendix: Particle Swarm Optimization (PSO)

In the studies on dimensionality and data reduction, we consider the use of Particle Swarm Optimization (PSO). PSO is an example of a biologically-inspired and population-driven optimization. Originally, the algorithm was introduced by Kennedy and Eberhart [12] where the authors strongly emphasized an inspiration coming from the swarming behavior of animals as well as some inspiration coming from the as well as human social behavior, cf. also [19]. In essence, a particle swarm is a population of particles – possible solutions in the multidimensional search space. Each particle explores the search space and during this search adheres to some quite intuitively appealing guidelines navigating the search process: (a) it tries to follow its previous direction, and (b) it looks back at the best performance both at the level of the individual particle and the entire population. In this sense there is some collective search of the problem space along with some component of memory incorporated as an integral part of the search mechanism.

The performance of each particle during its movement is assessed by means of some performance index. A position of a swarm in the search space, is described by some vector $\mathbf{z}(t)$ where “ t ” denotes consecutive discrete time moments. The next position of the particle is governed by the following update expressions concerning the particle, $\mathbf{z}(t + 1)$ and its speed, $\mathbf{v}(t + 1)$

$$\begin{aligned}\mathbf{z}(t + 1) &= \mathbf{z}(t) + \mathbf{v}(t + 1) \\ &\quad // \text{ update of position of the particle} \\ \mathbf{v}(t + 1) &= \xi \mathbf{v}(t) + \phi_1(\mathbf{p} - \mathbf{x}(t)) + \phi_2(\mathbf{p}_{\text{total}} - \mathbf{x}(t)) \\ &\quad // \text{ update of speed of the particle}\end{aligned}$$

where \mathbf{p} denotes the best position (the lowest performance index) reported so far for this particle, $\mathbf{p}_{\text{total}}$ is the best position overall developed so far across the whole population. ϕ_1 and ϕ_2 are random number drawn from the uniform distribution $U[0, 2]$ that help build a proper mix of the components of the speed; different random numbers affect the individual coordinates of the speed. The second expression governing the change in the velocity of the particle is particularly interesting as it nicely captures the relationships between the particle and its history as well as the history of the overall population in terms of their performance reported so far.

There are three components determining the updated speed of the particle. First, the current speed $\mathbf{v}(t)$ is scaled by the inertial weight (ξ) smaller than 1 whose role is to articulate some tendency to a drastic change of the current speed. Second, we relate to the memory of this particle by recalling the best position of the particle achieved so far. Thirdly, there is some reliance on the best performance reported across the whole population (which is captured by the last component of the expression governing the speed adjustment).

Bibliography

Primary Literature

1. Bargiela A, Pedrycz W (2003) Granular Computing: An Introduction. Kluwer, Dordrecht
2. Bezdek JC (1992) On the relationship between neural networks, pattern recognition and intelligence. *Int J Approx Reason* 6(2):85–107
3. Duda RO, Hart PE, Stork DF (2001) Pattern Classification, 2nd edn. Wiley, New York
4. Cheng Y, Church GM (2000) Biclustering of expression data. *Proc 8th Int Conf on Intelligent Systems for Molecular Biology*, pp 93–103
5. Gersho A, Gray RM (1992) Vector Quantization and Signal Compression. Kluwer, Boston
6. Gottwald S (2005) Mathematical fuzzy logic as a tool for the treatment of vague information. *Inf Sci* 172(1–2):41–71
7. Gray RM (1984) Vector quantization. *IEEE Acoust Speech Signal Process* 1:4–29
8. Hansen E (1975) A generalized interval arithmetic. *Lecture Notes in Computer Science*, vol 29. Springer, Berlin, pp 7–18
9. Hartigan JA (1972) Direct clustering of a data matrix. *J Am Stat Assoc* 67:123–129
10. Jaulin L, Kieffer M, Didrit O, Walter E (2001) Applied Interval Analysis. Springer, London
11. Jolliffe IT (1986) Principal Component Analysis. Springer, New York
12. Kennedy J, Eberhart RC (1995) Particle swarm optimization, vol 4. *Proc IEEE Int Conf on Neural Networks*. IEEE Press, Piscataway, pp 1942–1948
13. Kohonen T (1989) Self Organization and Associative Memory, 3rd edn. Springer, Berlin
14. Mitchell M (1996) An Introduction to Genetic Algorithms. MIT Press, Cambridge
15. Moore R (1966) Interval Analysis. Prentice Hall, Englewood Cliffs
16. Pawlak Z (1982) Rough sets. *Int J Comput Inform Sci* 11:341–356
17. Pawlak Z (1991) Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht
18. Pawlak Z, Skowron A (2007) Rough sets: some extensions. *Inf Sci* 177(1):28–40
19. Pedrycz W (ed) (2001) Granular Computing: An Emerging Paradigm. Physica, Heidelberg
20. Pedrycz W (2005) Knowledge-based Clustering. Wiley, Hoboken
21. Warmus M (1956) Calculus of approximations. *Bull Acad Pol Sci* 4(5):253–259

22. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta MM, Ragade RK, Yager RR (eds) *Advances in Fuzzy Set Theory and Applications*. North Holland, Amsterdam, pp 3–18
23. Zadeh LA (1996) Fuzzy logic = Computing with words. *IEEE Trans Fuzzy Syst* 4:103–111
24. Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst* 90:111–117
25. Zadeh LA (1999) From computing with numbers to computing with words—from manipulation of measurements to manipulation of perceptions. *IEEE Trans Circ Syst* 45:105–119
26. Zadeh LA (2005) Toward a generalized theory of uncertainty (GTU) – an outline. *Inf Sci* 172:1–40
27. Zimmermann HJ (1996) *Fuzzy Set Theory and Its Applications*, 3rd edn. Kluwer, Norwell

Books and Reviews

- Baldi P, Hornik K (1989) Neural networks and principal component analysis: learning from examples without local minima. *Neural Netw* 2:53–58
- Bortolan G, Pedrycz W (2002) Fuzzy descriptive models: an interactive framework of information granulation. *IEEE Trans Fuzzy Syst* 10(6):743–755
- Busygin S, Prokopyev O, Pardalos PM (2008) Biclustering in data mining. *Comput Oper Res* 35(9):2964–2987
- Fukunaga K (1990) *Introduction to Statistical Pattern Recognition*. Academic Press, Boston
- Gifi A (1990) *Nonlinear Multivariate Analysis*. Wiley, Chichester
- Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. *ACM Comput Surv* 31(3):264–323
- Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artif Intell* 97(1–2):273–324
- Lauro C, Palumbo F (2000) Principal component analysis of interval data: a symbolic data analysis approach. *Comput Stat* 15:73–87
- Manly BF, Bryan FJ (1986) *Multivariate Statistical Methods: A Primer*. Chapman and Hall, London
- Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity. *IEEE Trans Pattern Anal Mach Intell* 24(3):301–312
- Monahan AH (2000) Nonlinear principal component analysis by neural networks: Theory and applications to the Lorenz system. *J Clim* 13:821–835
- Muni DP, Das Pal NR (2006) Genetic programming for simultaneous feature selection and classifier design. *IEEE Trans Syst Man Cybern Part B* 36(1):106–117
- Pawlak Z (1991) *Rough Sets – Theoretical Aspects of Reasoning about Data*. Kluwer, Dordrecht
- Pedrycz W, Vukovich G (2002) Feature analysis through information granulation and fuzzy sets. *Pattern Recognit* 35:825–834
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(1):2323–2326
- Setiono R, Liu H (1977) Neural-network feature selector. *IEEE Trans. Neural Netw* 8(3):654–662
- Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(1):2319–2323
- Watada J, Yabuuchi Y (1997) Fuzzy principal component analysis and its application. *Biomed Fuzzy Human Sci* 3:83–92

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets

LECH POLKOWSKI^{1,2}

¹ Polish–Japanese Institute of Information Technology, Warsaw, Poland

² Department of Mathematics and Computer Science, University of Warmia and Mazury, Olsztyn, Poland

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Rough Set Theory: Extensions](#)

[Rough Set Theory: Applications](#)

[Nearest Neighbor Method](#)

[Case-Based Reasoning](#)

[Complexity Issues](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Knowledge This is a many-faceted and difficult notion to define and it is used very frequently without any attempt at definition as a notion that explains per se. One can follow J. M. Bocheński in claiming that the world is a system of states of things, related to themselves by means of the network of relations; things, their features, relations among them and states are reflected in knowledge: things in objects or notions, features and relations in notions (or, concepts), states of things in sentences. Sentences constitute knowledge. Knowledge allows its possessor to classify new objects, model processes, make predictions etc.

Reasoning Processes of reasoning include an effort by means of which sentences are created; various forms of reasoning depend on the chosen system of notions, symbolic representation of notions, forms of manipulating symbols etc.

Knowledge representation This is a chosen symbolic system (language) by means of which notions are encoded and reasoning is formalized.

Boolean functions An n -ary Boolean function is a mapping $f: \{0, 1\}^n \rightarrow \{0, 1\}$ from the space of binary sequences of length n into the doubleton $\{0, 1\}$. An equivalent representation of the function f is as a formula ϕ_f of propositional calculus; a Boolean function

can be thus represented either in DNF form or in CNF form. The former representation: $\bigvee_{i \in I} \bigwedge_{j \in J_i} l_j^i$, where l_j^i is a literal, i. e., either a propositional variable or its negation, is instrumental in Boolean Reasoning: when the problem is formulated as a Boolean function, its solutions are searched for as prime implicants $\bigwedge_{j \in J_i} l_j^i$. Applications are to be found in various reduct induction algorithms.

Information systems One of the languages for knowledge representation is the attribute-value language in which notions representing things are described by means of attributes (features) and their values; information systems are pairs of the form (U, A) where U is a set of objects – representing things – and A is a set of attributes; each attribute a is modeled as a mapping $a: U \rightarrow V_a$ from the set of objects into the value set V_a . For an attribute a and its value v , the descriptor $(a = v)$ is a formula interpreted in the set of objects U as $[(a = v)] = \{u \in U: a(u) = v\}$. Descriptor formulas are the smallest set containing all descriptors and closed under sentential connectives $\vee, \wedge, \neg, \Rightarrow$. Meanings of complex formulas are defined recursively: $[\alpha \vee \beta] = [\alpha] \cup [\beta]$, $[\alpha \wedge \beta] = [\alpha] \cap [\beta]$, $[\neg \alpha] = U \setminus [\alpha]$, $[\alpha \Rightarrow \beta] = [\neg \alpha \vee \beta]$. In descriptor language each object $u \in U$ can be encoded over a set B of attributes as its information vector $\text{Inf}_B(u) = \{(a = a(u)): a \in B\}$.

Indiscernibility The Leibnizian Principle of Identity of Indiscernibles affirms that two things are identical in case they are indiscernible, i. e., no available operator acting on both of them yields distinct values; in the context of information systems, indiscernibility relations are induced from sets of attributes: given a set $B \subseteq A$, the indiscernibility relation relative to B is defined as $\text{Ind}(B) = \{(u, u'): a(u) = a(u') \text{ for each } a \in B\}$. Objects u, u' in relation $\text{Ind}(B)$ are said to be B -indiscernible and are regarded as identical with respect to knowledge represented by the information system (U, B) . The class $[u]_B = \{u': (u, u') \in \text{Ind}(B)\}$ collects all objects identical to u with respect to B .

Exact, inexact notion An exact notion is a set of objects in the considered universe which can be represented as the union of a collection of indiscernibility classes; otherwise, the set is inexact. In this case, there exist a boundary about the notion consisting of objects which can be with certainty classified neither into the notion nor into its complement (Pawlak, Frege).

Decision systems A particular form of an information system, this is a triple (U, A, d) in which d is the decision, the attribute not in A , that does express the evaluation of objects by an external oracle, an expert. At-

tributes in A are called conditional in order to discern them from the decision d .

Classification task The problem of assigning to each element in a set of objects (test sample) of a class (of a decision) to which the given element should belong; it is effected on the basis of knowledge induced from the given collection of examples (the training sample). To perform this task, objects are mapped usually onto vectors in a multi-dimensional real vector space (feature space).

Decision rule A formula in descriptor language that does express a particular relation among conditional attributes in the attribute set A and the decision d , of the form: $\bigwedge_{a \in A} (a = v_a) \Rightarrow (d = v)$ with the semantics defined in (Glossary: “Indiscernibility”). The formula is true in case $[\bigwedge_{a \in A} (a = v_a)] = \bigcap_{a \in A} [(a = v_a)] \subseteq [(d = v)]$. Otherwise, the formula is partially true. An object o which matches the rule, i. e., $a(o) = v_a$ for $a \in A$ can be classified to the class $[(d = v)]$; often a partial match based on a chosen distance measure has to be performed.

Distance functions (metrics) A metric on a set X is a non-negative valued function $\rho: X \times X \rightarrow R$ where R is the set of reals, which satisfies conditions: 1. $\rho(x, y) = 0$ if and only if $x = y$. 2. $\rho(x, y) = \rho(y, x)$. 3. $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ for each z in X (the triangle inequality); when in 3. $\rho(x, y)$ is bound by $\max\{\rho(x, z), \rho(z, y)\}$ instead of by the sum of the two, one says of non-archimedean metric.

Object closest to a set For a metric ρ on a set X , and a subset Y of X , the distance from an object x in X and the set Y is defined as $\text{dist}(x, Y) = \inf\{\rho(x, y): y \in Y\}$; when Y is a finite set, then infimum \inf is replaced with minimum \min .

A nearest neighbor For an object $x_0 \in X$, this is an object $n(x_0)$ such that $\rho(x_0, n(x_0)) = \text{dist}(x_0, X \setminus \{x_0\})$; $n(x_0)$ may not be unique. In plain words, $n(x_0)$ is the object closest to x_0 and distinct from it.

K-nearest neighbors For an object $x_0 \in X$, and a natural number $K \geq 1$, this is a set $n(x_0, K) \subseteq X \setminus \{x_0\}$ of cardinality K such that for each object $y \in X \setminus [n(x_0, K) \cup \{x_0\}]$ one has $\rho(y, x_0) \geq \rho(z, x_0)$ for each $z \in n(x_0, K)$. In plain words, objects in $n(x_0, K)$ for a set of K objects that are closest to x_0 among objects distinct from it.

Rough inclusion A ternary relation μ on a set $U \times U \times [0, 1]$ which satisfies conditions: 1. $\mu(x, x, 1)$. 2. $\mu(x, y, 1)$ is a binary partial order relation on the set X . 3. $\mu(x, y, 1)$ implies that for each object z in U : if $\mu(z, x, r)$ then $\mu(z, y, r)$. 4. $\mu(x, y, r)$ and $s < r$ imply that $\mu(x, y, s)$. The formula $\mu(x, y, r)$ is read as “the object x is a part in object y to a degree at least r ”.

The partial containment idea does encompass the idea of an exact part, i. e., mereological theory of concepts (Leśniewski, Leonard–Goodman).

Similarity An extension and relaxing of an equivalence relation of indiscernibility. Among similarity relations, we single out the class of tolerance relations τ (Poincaré, Zeeman) which are reflexive, i. e., $\tau(x, x)$, and symmetric, i. e., $\tau(x, y)$ implies necessarily $\tau(y, x)$. The basic distinction between equivalences and tolerances is that the former induce partitions of their universes into disjoint classes whereas the latter induce covering of their universes by their classes; for this reason they are more difficult in analysis. The symmetry condition may be over-restrictive, for instance, rough inclusions are usually not symmetric.

Case A case is informally a part of knowledge that describes a certain state of the world (the context), along with a query (problem) and its solution, and a description of the outcome, i. e., the state of the world after the solution is applied.

Retrieve A process by means of which a case similar in a sense to the currently considered is recovered from the case base.

Reuse A process by means of which the retrieved case's solution is re-used in the current new problem.

Revise A process by means of which the retrieved solution is adapted in order to satisfactorily solve the current problem.

Retain A process by means of which the adapted solution is stored in the case base as the solution to the current problem.

Definition of the Subject

Basic ideas of Rough Set Theory were proposed by Zdzisław Pawlak [60,62], as a formal set of notions aimed at carrying out tasks of reasoning, in particular about classification of objects, in conditions of uncertainty. Conditions of uncertainty are imposed by incompleteness, imprecision and ambiguity of knowledge. Originally, the basic notion proposed was that of a knowledge base, understood as a collection \mathcal{R} of equivalence relations on a universe of objects; each relation $r \in \mathcal{R}$ induces on the set U a partition ind_r into equivalence classes. Knowledge so encoded is meant to represent the classification ability. As objects for analysis and classification come most often in the form of data, a useful notion of an information system is commonly used in knowledge representation; knowledge base in that case is defined as the collection of indiscernibility relations. Exact concepts are defined as unions of indiscernibility classes whereas inex-

act concepts are only approximated from below (lower approximations) and from above (upper approximations) by exact ones. Each inexact concept is thus perceived as a pair of exact concepts between which it is sandwiched. In terms of indiscernibility it is possible to define dependencies among attributes: the best case, of functional dependence, comprises instances in which a set say B of attributes depends on a set C of attributes, meaning that values of attributes from B are function-dependent on values of attributes in C ; a weaker form of dependence is partial dependence in which there is a non-functional relation between values of attributes in B and those of C . The most important case of dependency is the one between conditional attributes in a decision system and the decision; these dependencies are expressed by means of decision rules which altogether constitute a classifier. From the application point of view, one of most important tasks is to induce from data a satisfactory classifier. A plethora of ideas were put forth in this area. First, the idea of knowledge reduction appeared; here, the notion of a reduct in an information system comes foremost: a reduct $B \subseteq A$ is a minimal with respect to inclusion subset of A such that indiscernibility induced by B is the indiscernibility induced by A . Finding a reduct means that one can dispense with attributes in $A \setminus B$ and induce decision rules from conditional attributes restricted to B . Further developments in reduction area encompass notions of a relative reduct, a local reduct, a dynamic reduct, an entropy reduct and so on. Subsequent developments include discretization algorithms, similarity measures, granulation of knowledge.

Research on classifiers has led to the emergence of a number of systems for rule induction that have been applied in many tasks in areas of, e.g., data classification, medical diagnosis, economic analysis, signal classification; in some cases rough set classifiers have been coupled with fuzzy, wavelet or neural reasoning systems to produce hybrid systems.

Matching rules to new objects is done on the basis of closeness of the object and the rule measured often in terms of common descriptor fraction in the object description and in the rule and in many classifiers it is the closest rule that sets the value of decision class to the object.

The idea of value assignment to a new object by the closest already classified object is at the heart of the Nearest Neighbor Method. The idea goes back to Fix and Hodges [21,22], Skellam [88] and Clark and Evans [12]. Nearest neighbor technique in its simplest form consists of imposing a metric ρ on the object set in question and assigning the class value $c(x_0)$ to a currently examined test object x_0 by selecting the closest to x_0 with re-

spect to the metric ρ training object $n(x_0)$ and letting $c(x_0) = c(n(x_0))$, the class already assigned to $n(x_0)$. This simplest variant is also named 1-nearest neighbor method. A generalization consists of a choice of a natural number k and selecting from among training objects the set of k closest to x_0 objects, $n(x_0) = \{x_1, \dots, x_k\}$, eventually breaking ties randomly. Then the class value $c(x_0)$ is defined on the basis of class values $c(x_1), \dots, c(x_k)$ by majority voting; eventual ties are broken randomly. Nearest neighbor methods can be perceived as heuristics implementing the idea of the Parzen window: a region about x_0 (the “window”), e.g., rectangular, in which the count of objects leads to an estimate of density of probability distribution of objects in space. The problem of choosing the “right” window is in case of nearest neighbor techniques delegated to the training set itself and the class value is assigned without any probability assessment. Nearest neighbor idea is a basis as well for prototype methods, in particular k-means and clustering methods and is also applied in other approaches, e.g. rough set classifiers. Applications of this technique can be found in many areas like satellite images analysis, plant ecology, forestry, cluster identification in medical diagnosis etc.

Difficulties in extracting from data an explicit model on basis of which one could make inferences, already observed with the nearest neighbor method, led to emergence of the ideology of Case-Based Reasoning, traced back to Schank and Abelson’s [84] discussion of methods for describing the knowledge about situations. Closely related to reasoning by analogy, experiential learning, and philosophical and psychological theories of concepts claiming that real-world concepts are best classified by sets of representative cases (Wittgenstein [117]), Case-Based Reasoning emerged as a methodology for problem solving based on cases, i.e., capsules of knowledge containing the description of the current state of the world at the moment the problem is posed (the situation), the solution to the problem, and the state of the world after the solution is implemented. Solved cases stored in case base form a repository to which new cases are addressed. A new case is matched against the case base and cases in the case base most similar to the given case are retrieved; retrieval is usually based on a form of analogy reasoning, nearest neighbor selection, inductive procedures, or template retrieval. Once the similar case is retrieved, its solution is reused, or adapted, to provide a solution for the new case; various strategies are applied in the reuse process, from the simplest null strategy, i.e., using directly the retrieved solution in the new problem to derivational strategies that take the procedure that generated the retrieved solution and modify it to yield a new solution to

the current problem. Finally, the new completed case is retained (or, stored) in the case base. Case-based reasoning has been applied in many systems in the area of expert systems, to mention CYRUS (Kolodner [37]), MEDATOR (Simpson [87]), PERSUADER (Sycara [106]) and many others (Watson and Marir [114]). CBR systems have found usage in many areas of expertise like complex diagnosis, legal procedures, civil engineering, manufacture planning.

Introduction

Three paradigms covered in this article have been arranged in the previous section in the order corresponding to the specificity of their scope: rough sets are a paradigm intended to reason under uncertainty from data and they exploit the idea of the nearest neighbor in common with many other paradigms; rough set techniques make use of metrics and similarity relations in their methods and consequently we discuss nearest neighbor method as based on metrics and finally case-based reasoning as based on similarity ideas.

Introduced in Pawlak [60] rough set theory is based on ideas that go back to Gottlob Frege, Gottfried Wilhelm Leibniz, Jan Łukasiewicz, Stanisław Leśniewski, to mention a few names of importance. Its characteristics are that they divide notions (concepts) into two classes: exact as well as inexact. Because of the Fregean idea (Frege [23]), an inexact concept should possess a boundary into which objects that can be classified with certainty neither to the concept nor to its complement fall. This boundary to a concept is constructed from indiscernibility relations induced by attributes (features) of objects (see Glossary: “Indiscernibility”). Given a set of objects U and attributes in a set A , each set B of attributes does induce the B -indiscernibility relation $\text{ind}(B)$ (see “Glossary: Indiscernibility”), according to the Leibnizian Principle of Identity of Indiscernibles, see [98]. Each class $[u]_B = \{v \in U : (u, v) \in \text{ind}(B)\}$ of $\text{ind}(B)$ is B -definable, i.e., the decision problem whether $v \in [u]_B$ is decidable: $[u]_B$ is exact (see “Glossary”). Unions of classes $[u]_B$ for some u, B are exact as well. Concepts $\subseteq U$ that are not of this form are inexact, they possess a non-empty boundary. To express the B -boundary of a concept X induced by the set B of attributes, approximations over B are introduced, i.e.,

$$\underline{B}X = \bigcup \{[u]_B : [u]_B \subseteq X\} \quad (\text{the } B\text{-lower approximation})$$

$$\overline{B}X = \bigcup \{[u]_B : [u]_B \cap X \neq \emptyset\} \quad (\text{the } B\text{-upper approximation}).$$

The difference $Bd_B X = \overline{B}X \setminus \underline{B}X$ is the B -boundary of X ; when non-empty it does witness that X is inexact over B .

The information system (U, A) (see Glossary: “Information systems”) does represent knowledge about a certain aspect of the world. This knowledge can be reduced: a reduct B of the set A of attributes is a minimal subset of A with the property that $\text{ind}(B) = \text{ind}(A)$. An algorithm for finding reducts based on Boolean reasoning was proposed in Skowron and Rauszer [90]; given input (U, A) with $U = \{u_1, \dots, u_n\}$ it starts with the discernibility matrix,

$$M_{U,A} = [c_{i,j} = \{a \in A : a(u_i) \neq a(u_j)\}]_{1 \leq i, j \leq n}$$

and builds the Boolean function,

$$f_{U,A} = \bigwedge_{c_{i,j} \neq \emptyset, i < j} \bigvee_{a \in c_{i,j}} \bar{a},$$

where \bar{a} is the Boolean variable assigned to the attribute $a \in A$.

The function $f_{U,A}$ is converted to its DNF form:

$$f_{U,A}^* : \bigvee_{j \in J} \bigwedge_{k \in K_j} \bar{a}_{j,k}.$$

Then: sets of the form $R_j = \{a_{j,k} : k \in K_j\}$ for $j \in J$, corresponding to prime implicants (see Glossary: “Boolean functions”), are all reducts of A . Choosing a reduct R , and forming the reduced information system (U, R) one is assured that no information encoded in (U, A) has been lost.

Moreover, as $\text{ind}(R) \subseteq \text{ind}(B)$ for any subset B of the attribute set A , one can establish a functional dependence of B on R : as for each object $u \in U$, $[u]_R \subseteq [u]_B$, the assignment $f_{R,B} : \text{Inf}_R(u) \rightarrow \text{Inf}_B(u)$ (see Glossary: “Information systems”) is functional, i. e., values of attributes in R determine functionally values of attributes on U , object-wise. Thus, any reduct determines functionally the whole data system.

A decision system (U, A, d) (see Glossary: “Decision systems”) encodes information about the external classification d (by an oracle, expert etc.). Methods based on rough sets aim as well at finding a description of the concept d in terms of conditional attributes in A in the language of descriptors (see Glossary: “Information systems”). The simpler case is when the decision system is deterministic, i. e., $\text{ind}(A) \subseteq \text{ind}(d)$. In this case the relation between A and d is functional, given by the unique assignment $f_{A,d}$ or in the decision rule form (see Glossary: “Decision rule”), as the set of rules: $\bigwedge_{a \in A} (a = a(u)) \Rightarrow (d = d(u))$. In place of A any reduct R of A can be substituted leading to shorter rules. In the contrary case, some

classes $[u]_A$ are split into more than one decision class $[v]_d$ leading to ambiguity in classification. In that case, decision rules are divided into certain (or, exact) and possible; to induce the certain rule set, the notion of a δ -reduct was proposed in Skowron and Rauszer [90]; it is called a relative reduct in Bazan et al. [7]. To define δ -reducts, first the generalized decision δ_B is defined: for $u \in U$, $\delta_B(u) = \{v \in V_d : d(u') = v \wedge (u, u') \in \text{ind}(B) \text{ for some } u' \in U\}$. A subset B of A is a δ -reduct to d when it is a minimal subset of A with respect to the property that $\delta_B = \delta_A$.

δ -reducts can be obtained from the modified Skowron and Rauszer algorithm [90]: it suffices to modify the entries $c_{i,j}$ to the discernibility matrix, by letting $c_{i,j}^d = \{a \in A \cup \{d\} : a(u_i) \neq a(u_j)\}$ and then setting $c_{i,j}' = c_{i,j}^d \setminus \{d\}$ in case $d(u_i) \neq d(u_j)$ and $c_{i,j}' = \emptyset$ in case $d(u_i) = d(u_j)$. The algorithm described above input with entries $c_{i,j}'$ forming the matrix $M_{U,A}^\delta$ outputs all δ -reducts to d encoded as prime implicants of the associated Boolean function $f_{U,A}^\delta$.

An Example of Reduct Finding and Decision Rule Induction

We conclude the first step into rough sets with a simple example of a decision system, its reducts and decision rules.

Table 1 shows a simple decision system.

Reducts of the information system $(U, A = \{a_1, a_2, a_3, a_4\})$ can be found from the discernibility matrix $M_{U,A}$ in Table 2; by symmetry, cells $c_{i,j} = c_{j,i}$ with $i > j$ are not filled. Each attribute a_i is encoded by the Boolean variable i .

After reduction by means of absorption rules of sentential calculus: $(p \vee q) \wedge p \Leftrightarrow p$, $(p \wedge q) \vee p \Leftrightarrow p$, the DNF form $f_{U,A}^*$ is $1 \wedge 2 \wedge 3 \vee 1 \wedge 2 \wedge 4 \vee 1 \wedge 3 \wedge 4$. Reducts of A in the information system (U, A) are:

$\{a_1, a_2, a_3\}, \{a_1, a_2, a_4\}, \{a_1, a_3, a_4\}$.

δ -reducts of the decision d in the decision system Simple, can be found from the modified discernibility matrix $M_{U,A}^\delta$ in Table 3.

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 1

Decision system Simple

Obj.	a_1	a_2	a_3	a_4	d
u_1	1	0	0	1	0
u_2	0	1	0	0	1
u_3	1	1	0	0	1
u_4	1	0	0	1	1
u_5	0	0	0	1	1
u_6	1	1	1	1	0

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 2

Discernibility matrix $M_{U,A}$ for reducts in (U,A)

Obj.	u_1	u_2	u_3	u_4	u_5	u_6
u_1	\emptyset	$\{1, 2, 4\}$	$\{2, 4\}$	\emptyset	$\{1\}$	$\{2, 3\}$
u_2	—	\emptyset	$\{1\}$	$\{1, 2, 3\}$	$\{2, 4\}$	$\{1, 3, 4\}$
u_3	—	—	\emptyset	$\{2, 4\}$	$\{2, 4\}$	$\{3, 4\}$
u_4	—	—	—	\emptyset	$\{1\}$	$\{2, 3\}$
u_5	—	—	—	—	\emptyset	$\{1, 2, 3\}$
u_6	—	—	—	—	—	\emptyset

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 3

Discernibility matrix $M_{U,A}^\delta$ for δ -reducts in (U, A, d)

Obj.	u_1	u_2	u_3	u_4	u_5	u_6
u_1	\emptyset	$\{1, 2, 4\}$	$\{2, 4\}$	\emptyset	$\{1\}$	\emptyset
u_2	—	\emptyset	\emptyset	\emptyset	\emptyset	$\{1, 3, 4\}$
u_3	—	—	\emptyset	\emptyset	\emptyset	$\{3, 4\}$
u_4	—	—	—	\emptyset	\emptyset	$\{2, 3\}$
u_5	—	—	—	—	\emptyset	$\{1, 2, 3\}$
u_6	—	—	—	—	—	\emptyset

From the Boolean function $f_{U,A}^\delta$ we read off δ -reducts $R_1 = \{a_1, a_2, a_3\}$, $R_2 = \{a_1, a_2, a_4\}$, $R_3 = \{a_1, a_3, a_4\}$.

Taking R_1 as the reduct for inducing decision rules, we read the following certain rules:

$$\begin{aligned} r_1: (a_1 = 0) \wedge (a_2 = 1) \wedge (a_3 = 0) &\Rightarrow (d = 1); \\ r_2: (a_1 = 1) \wedge (a_2 = 1) \wedge (a_3 = 0) &\Rightarrow (d = 1); \\ r_3: (a_1 = 0) \wedge (a_2 = 0) \wedge (a_3 = 0) &\Rightarrow (d = 1); \\ r_4: (a_1 = 1) \wedge (a_2 = 1) \wedge (a_3 = 1) &\Rightarrow (d = 0); \end{aligned}$$

and two possible rules

$$\begin{aligned} r_5: (a_1 = 1) \wedge (a_2 = 0) \wedge (a_3 = 0) &\Rightarrow (d = 0); \\ r_6: (a_1 = 1) \wedge (a_2 = 0) \wedge (a_3 = 0) &\Rightarrow (d = 1), \end{aligned}$$

each with certainty factor = .5 as there are two objects with $d = 0$.

Consider a new object v : $\text{Inf}_A(v) = \{(a_1 = 0), (a_2 = 1), (a_3 = 1), (a_4 = 0)\}$.

The Nearest Neighbor Approach

We define the metric ρ on objects encoded by means of their information vectors (see Glossary: “Information systems”) as the reduced Hamming metrics on information vectors over the reduct R_1 ; for an object x we let $\text{Desc}(x)$ to denote the set of descriptors that occur in the information vector $\text{Inf}_{R_1}(x)$ of x . The metric is

then $\rho_{R_1}(x, y) = 1 - (|\text{Desc}(x) \cap \text{Desc}(y)|)/3$; the nearest neighbors of v with respect to ρ_{R_1} are u_2 and u_6 described by, respectively, rules r_1, r_4 ; however, these rules point to distinct decision classes: 1, 0. The majority voting must be implemented by random choice between class values 0, 1 with probability of error of .5. We can still resort to the idea of the most similar case.

The Case-Based Reasoning Approach

This ambiguity may be resolved if we treat objects as cases defined over the set A with solutions as the values of the decision d ; thus formally a case is of the form $(x, d(x))$. Similarity among cases will be defined by means of the Hamming distance reduced over the whole set A ; thus, $\rho_A(x, y) = 1 - (|\text{Desc}(x) \cap \text{Desc}(y)|)/4$. The most similar case to v is u_6 ($\rho_A(v, u_6) = 1/4$, $\rho(v, u_2) = 1/2$) hence we adapt the solution $d = 0$ to the case u_6 as the solution to v and we classify v as $d = 0$; the new case in our case base is then $(v, 0)$.

Rough Set Theory: Extensions

Basic principles of the rough set approach: knowledge reduction, indiscernibility, functional and partial dependence, decision rules are exposed in Sect. “Introduction”. In this section, we augment our discussion with more detailed and specific aspects, dividing the exposition into parts concerned with particular aspects.

Reducts

The notion of a reduct as a minimal with respect to set inclusion subset of the set of attributes A preserving a certain property of the whole attribute set, has undergone an evolution from the basic notions of a reduct and δ -reduct to more specialized variants. The problem of finding a reduct of minimal length is NP-hard [90], therefore one may foresee that no polynomial algorithm is available for computing reducts. Thus, the algorithm based on the discernibility matrix has been proposed with stop rules that permits one to stop the algorithm and obtain a partial set of reducts [7]. In order to find stable reducts, robust to perturbations in data, the idea of a dynamic reduct was proposed in Bazan et al. [7]; for a decision system $D = (U, A, d)$, a family F of subsystems of the form (U_1, A, d) with $U_1 \subseteq U$ is derived (e.g., by random choice) and given $\varepsilon \in (0, 1)$, a generalized dynamic ε -reduct B of D is a set B which is also found as a reduct in at least $(1 - \varepsilon)$ -fraction of subsystems in F .

In order to precisely discriminate between certain and possible rules, the notion of a positive region along with

the notion of a relative reduct was proposed and studied in Skowron and Rauszer [90].

Positive region $\text{pos}_B(d)$ is the set $\{u \in U: [u]_B \subseteq [u]_d\} = \bigcup_{v \in V_d} B[(d = v)]$; $\text{pos}_B(d)$ is the greatest subset of X of U such that (X, B, d) is deterministic; it generates certain rules. Objects in $U \setminus \text{pos}_B(d)$ are subjected to ambiguity: given such u , and the collection v_1, \dots, v_k of decision d values on the class $[u]_B$, the decision rule describing u can be formulated as, $\bigwedge_{a \in B} (a = a(u)) \Rightarrow \bigvee_{i=1, \dots, k} (d = v_i)$; each of the rules $\bigwedge_{a \in B} (a = a(u)) \Rightarrow (d = v_i)$ is possible but not certain as only for a fraction of objects in the class $[u]_B$ the decision take the value v_i on.

Relative reducts are minimal sets B of attributes with the property that $\text{pos}_B(d) = \text{pos}_A(d)$; they can also be found by means of discernibility matrix $M_{U,A}^*$ [90]: $c_{i,j}^* = c_{i,j}^d \setminus \{d\}$ in case either $d(u_i) \neq d(u_j)$ and $u_i, u_j \in \text{pos}_A(d)$ or $\text{pos}(u_i) \neq \text{pos}(u_j)$ where pos is the characteristic function of $\text{pos}_A(d)$; otherwise, $c_{i,j}^* = \emptyset$.

For a relative reduct B , certain rules are induced from the deterministic system $(\text{pos}_B(d), A, d)$, possible rules are induced from the non-deterministic system $(U \setminus \text{pos}_B(d), A, d)$. In the last case, one can find δ -reducts to d in this system and turn the system into a deterministic one $(U \setminus \text{pos}_B(d), A, \delta)$ inducing certain rules of the form $\bigwedge_{a \in B} (a = a(u)) \Rightarrow \bigvee_{v \in \delta(u)} (d = v)$.

Localization of reducts was taken care of by means of the notion of a local reduct in [7]: given an object u_i , a local reduct to decision d is a minimal set $B(u_i)$ such that for each object u_j , if $d(u_i) \neq d(u_j)$ then $a(u_i) \neq a(u_j)$ for some $a \in B(u_i)$. An algorithm for finding a covering of the set of objects by local reducts for each of the objects is given in [7].

A more general notion of a reduct as a minimal set of attributes preserving a certain characteristic of the distribution of objects into decision classes, e. g., frequencies of objects, entropy of classification distribution, etc., is discussed in [94].

Decision Rules

Already defined as descriptor formulas of the form $\bigwedge_{a \in B} (a = v_a) \Rightarrow (d = v)$ in descriptor language of the decision system (U, A, d) , decision rules provide a description of the decision d in terms of conditional attributes in the set A . Forming a decision rule is searching in the pool of available semantically non-vacuous descriptors for their combination that describes as well as possible a chosen decision class. The very basic idea of inducing rules consists of considering a set B of attributes: the lower approximation $\text{pos}_B(d)$ allows for rules which are certain, the

upper approximation $\bigcup_{v \in V_d} \overline{B}[(d = v)]$ adds rules which are possible.

We write down a decision rule in the form $\phi/B, u \Rightarrow (d = v)$ where ϕ/B is a descriptor formula $\bigwedge_{a \in B} (a = a(u))$ over B . A method for inducing decision rules in the systematic way of Pawlak and Skowron [63] and Skowron [89] consists of finding the set of all δ -reducts $R = \{R_1, \dots, R_m\}$, and defining for each reduct R_j and each object $u \in U$, the rule $\phi/R_j, u \Rightarrow (d = d(u))$. Rules obtained by this method are not minimal usually in the sense of the number of descriptors in the premise ϕ .

A method for obtaining decision rules with a minimal number of descriptors [63,89], consists of reducing a given rule $r: \phi/B, u \Rightarrow (d = v)$ by finding a set $R_r \subseteq B$ consisting of irreducible attributes in B only, in the sense that removing any $a \in R_r$ causes the inequality $[\phi/R_r, u \Rightarrow (d = v)] \neq [\phi/R_r \setminus \{a\}, u \Rightarrow (d = v)]$ to hold. In case $B = A$, reduced rules $\phi/R_r, u \Rightarrow (d = v)$ are called optimal basic rules (with a minimal number of descriptors). The method for finding all irreducible subsets of the set A (Skowron [89]) consists of considering another modification of the discernibility matrix: for each object $u_k \in U$, the entry $c'_{i,j}$ into the matrix $M_{U,A}^\delta$ for δ -reducts is modified into $c'_{i,j} = c'_{i,j}$ in case $d(u_i) \neq d(u_j)$ and $i = k \vee j = k$, otherwise $c'_{i,j} = \emptyset$. Matrices $M_{U,A}^k$ and associated Boolean functions $f_{U,A}^k$ for all $u_k \in U$ allow for finding all irreducible subsets of the set A and in consequence all basic optimal rules (with a minimal number of descriptors).

Decision rules are judged by their quality on the basis of the training set and by quality in classifying new as yet unseen objects, i. e., by their performance on the test set. Quality evaluation is done on the basis of some measures: for a rule $r: \phi \Rightarrow (d = v)$, and an object $u \in U$, one says that u matches r in case $u \in [\phi]$. $\text{match}(r)$ is the number of objects matching r . Support $\text{supp}(r)$ of r is the number of objects in $[\phi] \cap [(d = v)]$; the fraction $\text{cons}(r) = (\text{supp}(r))/(\text{match}(r))$ is the consistency degree of r $\text{cons}(r) = 1$ and means that the rule is certain.

Strength, $\text{strength}(r)$, of the rule r is defined (Michalski et al. [49], Bazan [6], Grzymala-Busse and Ming Hu [29]), as the number of objects correctly classified by the rule in the training phase; relative strength is defined as the fraction $\text{rel-strength}(r) = (\text{supp}(r))/(|[(d = v)]|)$. Specificity of the rule r , $\text{spec}(r)$, is the number of descriptors in the premise ϕ of the rule r [29].

In the testing phase, rules vie among themselves for object classification when they point to distinct decision classes; in such a case, negotiations among rules or their sets are necessary. In these negotiations rules with better characteristics are privileged.

For a given decision class $c: d = v$, and an object u in the test set, the set $\text{Rule}(c, u)$ of all rules matched by u and pointing to the decision v , is characterized globally by $\text{Support}(\text{Rule}(c, u)) = \sum_{r \in \text{Rule}(c, u)} \text{strength}(r) \cdot \text{spec}(r)$. The class c for which $\text{Support}(\text{Rule}(c, u))$ is the largest wins the competition and the object u is classified into the class $c: d = v$ [29].

It may happen that no rule in the available set of rules is matched by the test object u and partial matching is necessary, i.e., for a rule r , the matching factor $\text{match} - \text{fact}(r, u)$ is defined as the fraction of descriptors in the premise ϕ of r matched by u to the number $\text{spec}(r)$ of descriptors in ϕ . The rule for which the partial support $\text{Part} - \text{Support}(\text{Rule}(c, u)) = \sum_{r \in \text{Rule}(c, u)} \text{match} - \text{fact}(r, u) \cdot \text{strength}(r) \cdot \text{spec}(r)$ is the largest wins the competition and it does assign the value of decision to u [29].

In a similar way, notions based on relative strength can be defined for sets of rules and applied in negotiations among them [7].

Dependency

Decision rules are particular cases of dependencies among attributes or their sets; certain rules of the form $\phi/B \Rightarrow (d = v)$ establish functional dependency of decision d on the set B of conditional attributes. Functional dependence of the set B of attributes on the set C , $C \mapsto B$, in an information system (U, A) means that $\text{Ind}(C) \subseteq \text{Ind}(B)$. Minimal sets $D \subseteq C$ of attributes such that $D \mapsto B$ can be found from a modified discernibility matrix $M_{U,A}$: letting $\langle B \rangle$ to denote the global attribute representing B : $\langle B \rangle(u) = \langle b_1(u), \dots, b_m(u) \rangle$ where $B = \{b_1, \dots, b_m\}$, for objects u_i, u_j , one sets $c_{i,j} = \{a \in C \cup \{\langle B \rangle\} : a(u_i) \neq a(u_j)\}$ and then $c_{i,j}^B = c_{i,j} \setminus \{\langle B \rangle\}$ in case $\langle B \rangle$ is in $c_{i,j}$; otherwise $c_{i,j}^B$ is empty. The associated Boolean function $f_{U,A}^B$ gives all minimal subsets of C on which B depends functionally; in particular, when $B = \{b\}$, one obtains in this way all subsets of the attribute set A on which b depends functionally (Skowron and Rauszer [90]). A number of contributions by Pawlak, Novotny and Pawlak and Novotny are devoted to this topic in an abstract setting of semi-lattices: Novotny and Pawlak [56], Pawlak [61].

Partial dependence of the set B on the set C of attributes takes place where there is no functional dependence $C \mapsto B$; in that case, some measures of a degree to which B depends on C were proposed in Novotny and Pawlak [55]: the degree can be defined, e.g., as the fraction $\gamma_{B,C} = (|\text{pos}_C B|)/|U|$ where the C -positive region of B is defined in analogy to already discussed positive region for decision, i.e., $\text{pos}_C(B) = \{u \in U : [u]_C \subseteq [u]_B\}$; then,

B depends on C partially to the degree $\gamma_{B,C} : C \mapsto_{\gamma_{B,C}} B$. The relation $C \mapsto_r B$ of partial dependency is transitive in the sense: if $C \mapsto_r B$ and $D \mapsto_s C$, then $D \mapsto_{\max\{0, r+s-1\}} B$, where $t(r, s) = \max\{0, r + s - 1\}$ is the Łukasiewicz t-norm (or, tensor product) [55].

A logical characterization of functional dependence between attribute sets was given as a fragment of intuitionistic logic in Rauszer [80].

Similarity

Analysis based on indiscernibility has allowed for extracting the most important notions of rough set theory; further progress has been obtained inter alia by departing from indiscernibility to more general similarity relations (see Glossary: “Similarity”). There have been various methods for introducing similarity relations.

An attempt at introducing some degrees of comparison among objects with respect to particular concepts consisted of defining rough membership functions in Pawlak and Skowron [64]: for an object u , an attribute set B and a concept $X \subseteq U$, the value $\mu_{B,X}(u)$ of the rough membership function $\mu_{B,X}$ was defined as $\mu_{B,X}(u) = (|[u]_B \cap X|)/(|[u]_B|)$. Informally, one can say that objects u, v are ε, B -similar with respect to the concept X in case $|\mu_{B,X}(u) - \mu_{B,X}(v)| < \varepsilon$. This relation is reflexive and symmetric, i.e., it is a tolerance relation, see Poincaré [66] and Zeeman [125], for each ε, B, X .

Tolerance relations were introduced into rough sets in Polkowski, Skowron, and Żytkow [77] and also studied in Krawiec et al. [43], Skowron and Stepaniuk [91]. It is possible to build a parallel theory of rough sets based on tolerance or similarity relations in analogy to indiscernibility relations. $\mu_{B,X}$ does characterize partial containment of objects in U into concepts X ; a further step consists of considering general relations of partial containment in the form of predicates “to be a part to a degree”.

A general form of partial containment was proposed as an extension of mereological theory of concepts by Leśniewski [44]; mereology takes the predicate “to be a part of” as its primitive notion, requiring of it to be irreflexive and transitive on the universe of objects U . The primitive notion of a (proper) part is relaxed to the notion of an ingredient (an improper part) $\text{ing} = \text{part} \cup =$.

The extension consists of considering the predicate “to be a part to a degree of”, formally introduced as the generic notion of a rough inclusion μ in Polkowski and Skowron [74,75], a ternary predicate with the semantic domain of $U \times U \times [0, 1]$ and satisfying the requirements that (1) $\mu(u, v, 1)$ is a binary relation of ingredient on objects in U , i.e., $\mu(u, v, 1)$ sets an exact scheme of decom-

position of objects in U into (possibly improper) parts. (2) $\mu(u, v, 1)$ implies informally that the object u is “inside” object v hence for any other object w : $\mu(w, u, r)$ implies $\mu(w, v, r)$. (3) $\mu(u, v, r)$ implies $\mu(u, v, s)$ for any $s < r$.

Similarity of objects with respect to μ can be judged by the degree $\text{sim}(u, v) = \max\{\arg \max \mu(u, v, r), \arg \max \mu(v, u, s)\}$; $\text{sim}(u, v) = 1$ means that u, v are identical with respect to the parts; the smaller $\text{sim}(u, v)$ the less similar u, v are.

Rough inclusions in an information system (U, A) can be induced in some distinct ways as in Polkowski [67,68,69,71]; Archimedean t-norms, i.e., t-norms $t(x, y)$ that are continuous and have no idempotents, i.e., values x with $t(x, x) = x$ except 0, 1 offer one way; it is well-known from Mostert and Shields [50] and Faucett [18] that up to isomorphism, there are two Archimedean t-norms: the Łukasiewicz t-norm $L(x, y) = \max\{0, x + y - 1\}$ and the product (Menger) t-norm $P(x, y) = x \cdot y$. Archimedean t-norms admit a functional characterization shown in Ling [48]: $t(x, y) = g(f(x) + f(y))$, where the function $f: [0, 1] \rightarrow R$ is continuous decreasing with $f(1) = 0$, and $g: R \rightarrow [0, 1]$ is the pseudo-inverse to f , i.e., $f \circ g = \text{id}$. The t-induced rough inclusion μ_t is defined [68] as $\mu_t(u, v, r) \Leftrightarrow g((|\text{DIS}(u, v)|)/(|A|)) \geq r$ where $\text{DIS}(u, v) = \{a \in A: a(u) \neq a(v)\}$. With the Łukasiewicz t-norm, $f(x) = 1 - x = g(x)$ and $\text{IND}(u, v) = U \times U \setminus \text{DIS}(u, v)$, the formula becomes: $\mu_L(u, v, r) \Leftrightarrow (|\text{IND}(u, v)|)/(|A|) \geq r$; thus in the case of Łukasiewicz logic, μ_L becomes the similarity measure based on the Hamming distance between information vectors of objects reduced modulo $|A|$; from a probabilistic point of view it is based on the relative frequency of descriptors in information sets of u, v . This formula permeates data mining algorithms and methods, see Klösgen and Żytkow [36].

Another method exploits residua of t-norms [69]. For a continuous t-norm $t(x, y)$, the *residuum*, $x \Rightarrow_t y$ is defined as $\max\{z: t(x, z) \leq y\}$. Rough inclusions are defined in this case with respect to objects selected as standards, e.g., objects u representing desired (in the given domain) combinations of attribute values. Given a standard s , the rough inclusion $v_t^{\text{IND},s}$ is defined:

$$v_t^{\text{IND},s}(u, v, r) \text{ iff } \frac{|\text{IND}(u, s)|}{|A|} \Rightarrow \frac{|\text{IND}(v, s)|}{|A|} \geq r.$$

In case of the t-norm L , $v_L^{\text{IND},s}(x, y, r) \text{ iff } |\text{IND}(y, s)| - |\text{IND}(x, s)| \geq (1 - r)|A|$.

In case of the t-norm P , $v_P^{\text{IND},s}(x, y, 1) \text{ iff } |\text{IND}(x, s)| \leq |\text{IND}(y, s)|$, and $v_P^{\text{IND},s}(x, y, r) \text{ iff } |\text{IND}(x, s)| \geq |\text{IND}(y, s)| \geq r \cdot |A|$.

Finally, in the case of $t = \min$, we have $v_{\min}^{\text{IND},s}(x, y, r) \text{ iff } |\text{IND}(y, s)|/|\text{IND}(x, s)| \geq r$.

All these rough inclusions satisfy the transitivity property [68,71]: From $\mu(u, v, r)$ and $\mu(v, w, s)$ it follows that $\mu(u, w, t(r, s))$, where μ is induced by the t-norm t .

Each metric d on the set of objects U [see Glossary: “Distance functions (metrics)”] induces the rough inclusion: $\mu_d(u, v, r)$ if and only if $d(u, v) \leq 1 - r$; conversely, any symmetric rough inclusion μ does induce the metric: $d_\mu(u, v) \leq r \Leftrightarrow \mu(u, v, 1 - r)$. For instance in the case of the t-norm L , the metric $d_L(u, v) = (|\text{DIS}(u, v)|)/(|A|)$, i.e., it is the Hamming distance on information vectors of objects reduced modulo $|A|$.

Let $\phi: [0, 1] \rightarrow [0, 1]$ be an increasing mapping with $\phi(1) = 1$; for a rough inclusion μ , we define a new predicate μ_ϕ by letting, $\mu_\phi(u, v, r) \Leftrightarrow \mu(u, v, \phi(r))$. μ_ϕ is a rough inclusion. It is transitive with respect to the t-norm $T_\phi(r, s) = \phi^{-1}[t(\phi(r), \phi(s))]$ whenever μ is transitive with respect to the t-norm t . A particular case takes place when one replaces $(|\text{DIS}|)/(|A|)$ with a sum $\sum_{a \in \text{DIS}} w_a$ where $\{w_a: a \in A\}$ is a convex family of positive weights. Further, one can replace weights w_a with values of rough inclusions μ_a defined on sets of values V_a . The corresponding formulas will be

$$\begin{aligned} \mu_L(u, v, r) &\Leftrightarrow \sum_{a \in \text{DIS}(u, v)} \mu_a^*(a(u), a(v)) \geq r, \\ \mu_P(u, v, r) &\Leftrightarrow \exp \left[- \sum_{a \in \text{DIS}(u, v)} \mu_a^*(a(u), a(v)) \right] \geq r, \end{aligned}$$

where

$$\mu_a^*(u, v) = \frac{\arg \max \mu_a(a(u), a(v), r)}{\sum_{b \in \text{DIS}} \arg \max \mu_b(b(u), b(v), r)}.$$

This discussion shows that rough inclusions encompass basically all forms of distance functions applied in Data Mining.

Other methods of introducing similarity relations into the realm of information systems, besides rough inclusions, include methods based on templates and on quasi-distance functions in Nguyen S. H. [53]; a template is any conjunct of the form $T: \bigwedge_{a \in B} (a \in W_a)$ where B is a set of attributes, and $W_a \subseteq V_a$ is a subset of the value set V_a of the attribute a . Semantics of templates is defined as with descriptor formulas (see Glossary: “Information systems”). Templates are judged by some parameters: length, i.e., the number of generalized descriptors ($a \in W_a$); support, i.e., number of matching objects; approximated length, *applength*, i.e., $\sum_{a \in B} (1/|W_a \cap a(U)|)$. Quality of the template is given by a combination of some of the parameters, e.g., *quality*(T) = *support*(T) + *applength*(T), etc. Templates are used

in classification problems in an analogous way to decision rules.

A quasi-metric (a similarity measure) [53] is a family $\Delta: \{\Delta_a: a \in A\}$ of functions where $\Delta_a(u, u) = 0$ and $\Delta_a(u, v) = \Delta_a(v, u)$ for $u, v \in U$. By means of these functions tolerance relations are built with the help of standard metric-forming operators like $\max, \sum: \tau_1(u, v) \Leftrightarrow \max_a \{\Delta_a(u, v) \leq \varepsilon\}, \tau_2(u, v) \Leftrightarrow \sum_a \Delta_a(u, v) \leq \varepsilon\}$ for a given threshold ε are examples of such similarity relations. These similarity relations are applied in finding classifiers in [53].

Discretization

The important problem of treating continuous values of attributes has been resolved in rough sets with the help of the discretization of attributes technique, common to many paradigms like decision trees etc.; for a decision system (U, A, d) , a cut is a pair (a, c) , where $a \in A, c$ in reals. The cut (a, c) induces the binary attribute $b_{a,c}(u) = 1$ if $a(u) \geq c$ and it is 0, otherwise. Given a finite sequence $p_a = c_0^a < c_1^a < \dots < c_m^a$ of reals, the set V_a of values of a is split into disjoint intervals: $(\leftarrow, c_0^a), [c_0^a, c_1^a), \dots, [c_m^a, \rightarrow)$; the new attribute $D_a(u) = i$ when $b_{c_{i-1}^a}^a = 0, b_{c_i^a}^a = 1$, is a discrete counterpart to the continuous attribute a . Given a collection $P = \{p_a: a \in A\}$ (a cut system), the set $D = \{D_a: a \in A\}$ of attributes transforms the system (U, A, d) into the discrete system (U, D, d) called the P -segmentation of the original system. The set P is consistent in case the generalized decision in both systems is identical, i. e., $\delta_A = \delta_{D_P}$; a consistent P is irreducible if P' is not consistent for any proper subset $P' \subset P$; P is optimal if its cardinality is minimal among all consistent cut systems, see Nguyen Hung Son [51], Nguyen Hung Son and Skowron [52].

MD-heuristics based on maximal discernibility of objects and decision trees for optimal sets of cuts were proposed in [51,52]; see also [7].

Granulation of Knowledge

The issue of granulation of knowledge as a problem on its own, has been posed by L.A. Zadeh [124]. Granulation can be regarded as a form of clustering, i. e., grouping objects into aggregates characterized by closeness of certain parameter values among objects in the aggregate and greater differences in those values from aggregate to aggregate. The issue of granulation has been a subject of intensive studies within the rough set community in, e. g., T.Y. Lin [47], Polkowski and Skowron [75,76], Skowron and Stepaniuk [92], Y. Y. Yao [123].

The rough set context offers a natural venue for granulation, and indiscernibility classes were recognized as *elementary granules* whereas their unions serve as *granules of knowledge*; these granules and their direct generalizations to various similarity classes induced by binary relations were subject to research by T.Y. Lin [46], Y.Y. Yao [121]. Granulation of knowledge by means of rough inclusions was studied as well in Polkowski and Skowron [74,75], Skowron and Stepaniuk [92], Polkowski [67,68,69,71].

For an information system (U, A) , and a rough inclusion μ on U , granulation with respect to similarity induced by μ is formally performed (Polkowski, op.cit.; op.cit.) by exploiting the class operator Cls of mereology [44]. The class operator is applied to any non-vacuous property F of objects (i. e. a distributive entity) in the universe U and produces the object $ClsF$ (i. e., the collective entity) representing wholeness of F . The formal definition of Cls is: assuming a part relation in U and the associated ingredient relation in $ClsF$ does satisfy conditions (1) if $u \in F$ then u is an ingredient of $ClsF$; (2) if v is an ingredient of $ClsF$ then some ingredient w of v is an ingredient as well of a T that is in F ; in plain words, each ingredient of $ClsF$ has an ingredient in common with an object in F . An example of part relation is the proper subset \subset relation on a family of sets; then the subset relation \subseteq is the ingredient relation, and the class of a family F of sets is its union $\bigcup F$. The merit of class operator is in the fact that it always projects hierarchies onto the collective entity plane containing objects.

For an object u and a real number $r \in [0, 1]$, we define the granule $g_\mu(u, r)$ about u of the radius r , relative to μ , as the class $ClsF(u, r)$, where the property $F(u, r)$ is satisfied with an object v if and only if $\mu(v, u, r)$ holds.

It was shown [68] that in the case of a transitive μ , v is an ingredient of the granule $g_\mu(u, r)$ if and only if $\mu(v, u, r)$. This fact allows for writing down the granule $g_\mu(u, r)$ as a distributive entity (a set, a list) of objects v satisfying $\mu(v, u, r)$.

Granules of the form $g_\mu(u, r)$ have regular properties of a neighborhood system [69]. Granules generated from a rough inclusion μ can be used in defining a compressed form of the decision system: a granular decision system [69]; for a granulation radius r , and a rough inclusion μ , we form the collection $U_{r,\mu}^G = \{g_\mu(u, r)\}$. We apply a strategy G to choose a covering $Cov_{r,\mu}^G$ of the universe U by granules from $U_{r,\mu}^G$. We apply a strategy S in order to assign the value $a^*(g)$ of each attribute $a \in A$ to each granule $g \in Cov_{r,\mu}^G$: $a^*(g) = S(\{a(u): u \in g\})$. The granular counterpart to the decision system (U, A, d) is a tuple $(U_{r,\mu}^G, G, S, \{a^*: a \in A\}, d^*)$. The heuristic principle that H : objects, similar with respect to conditional at-

tributes in the set A , should also reveal similar (i. e., close) decision values, and therefore, granular counterparts to decision systems should lead to classifiers satisfactorily close in quality to those induced from original decision systems that is at the heart of all classification paradigms, can be also formulated in this context [69]. Experimental results bear out the hypothesis [73].

Rough Set Theory: Applications

Applications of rough set theory are mainly in classification patterns recognition under uncertainty; classification is performed on test sets by means of judiciously chosen sets of decision rules or generalized decision rules (see Subsects. “Decision Rules”, “Similarity”, and “Granulation of Knowledge” of Sect. “Rough Set Theory. Extensions”) induced from training samples.

Classification

Classification methods can be divided according to the adopted methodology, into classifiers based on reducts and decision rules, classifiers based on templates and similarity, classifiers based on descriptor search, classifiers based on granular descriptors, hybrid classifiers.

For a decision system (U, A, d) , classifiers are sets of decision rules. Induction of rules was a subject of research in rough set theory since its beginning. In most general terms, building a classifier consists of searching in the pool of descriptors for their conjuncts that describe sufficiently well decision classes. As distinguished in Stefanowski [99], there are three main kinds of classifiers searched for: *minimal*, i. e., consisting of the minimum possible number of rules describing decision classes in the universe, *exhaustive*, i. e., consisting of all possible rules, *satisfactory*, i. e., containing rules tailored to a specific use. Classifiers are evaluated globally with respect to their ability to properly classify objects, usually by *error* which is the ratio of the number of correctly classified objects to the number of test objects, *total accuracy* being the ratio of the number of correctly classified cases to the number of recognized cases, and *total coverage*, i.e., the ratio of the number of recognized test cases to the number of test cases.

Minimum size algorithms include the LEM2 algorithm due to Grzymala-Busse [28] and covering algorithm in the RSES package [82]; exhaustive algorithms include, e.g., the LERS system due to Grzymala-Busse [27], systems based on discernibility matrices and Boolean reasoning by Skowron [89], Bazan [6], Bazan et al. [7], implemented in the RSES package [82].

Minimal consistent sets of rules were introduced in Skowron and Rauszer [90]; they were shown to coin-

cide with rules induced on the basis of local reducts in Wróblewski [120]. Further developments include dynamic rules, approximate rules, and relevant rules as described in [6,7] as well as local rules [6,7], effective in implementations of algorithms based on minimal consistent sets of rules. Rough set-based classification algorithms, especially those implemented in the RSES system [82], were discussed extensively in [7]; [101] contains a discussion of rough set classifiers along with some attempt at analysis of granulation in the process of knowledge discovery.

In [6], a number of techniques were verified in experiments with real data, based on various strategies:

discretization of attributes (codes: N-no discretization, S-standard discretization, D-cut selection by dynamic reducts, G-cut selection by generalized dynamic reducts);

dynamic selection of attributes (codes: N-no selection, D-selection by dynamic reducts, G-selection based on generalized dynamic reducts);

decision rule choice (codes: A-optimal decision rules, G-decision rules on the basis of approximate reducts computed by Johnson’s algorithm, simulated annealing and Boltzmann machines etc., N-without computing of decision rules);

approximation of decision rules (codes: N-consistent decision rules, P-approximate rules obtained by descriptor dropping);

negotiations among rules (codes: S-based on strength, M-based on maximal strength, R-based on global strength, D-based on stability).

Any choice of a strategy in particular areas yields a compound strategy denoted with the alias being concatenation of symbols of strategies chosen in consecutive areas, e.g., NNAND etc.

We record here in Table 4 an excerpt from the comparison (Tables 8, 9, 10 in [6]) of the best of these strategies with results based on other paradigms in classification for two sets of data: Diabetes and Australian credit from UCI Repository [109].

An adaptive method of classifier construction was proposed in [121]; reducts are determined by means of a genetic algorithm [7] and in turn reducts induce subtables of data regarded as classifying agents; the choice of optimal ensembles of agents is done by a genetic algorithm.

Classifiers constructed by means of similarity relations are based on templates matching a given object or closest to it with respect to a certain distance function, or on coverings of the universe of objects by tolerance classes and assigning the decision value on basis of some of them [53];

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 4

A comparison of errors in classification by rough set and other paradigms

Paradigm	System/method	Diabetes	Austr. credit
Stat. Methods	Logdisc	0.223	0.141
Stat. Methods	SMART	0.232	0.158
Neural Nets	Backpropagation2	0.248	0.154
Neural Networks	RBF	0.243	0.145
Decision Trees	CART	0.255	0.145
Decision Trees	C4.5	0.270	0.155
Decision Trees	ITrule	0.245	0.137
Decision Rules	CN2	0.289	0.204
Rough Sets	NNANR	0.335	0.140
Rough Sets	DNANR	0.280	0.165
Rough Sets	Best result	0.255 (DNAPM)	0.130 (SNAPM)

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 5

Accuracy of classification by template and similarity methods

Paradigm	System/method	Diabetes	Austr. credit
Rough Sets	Simple.templ./Hamming	0.6156	0.8217
Rough Sets	Gen.templ./Hamming	0.742	0.855
Rough Sets	Simple.templ./Euclidean	0.6312	0.8753
Rough Sets	Gen.templ./Euclidean	0.7006	0.8753
Rough Sets	Match.tolerance	0.757	0.8747
Rough Sets	Clos.tolerance	0.743	0.8246

we include in Table 5 excerpts from classification results in [53].

A combination of rough set methods with the k-nearest neighbor idea is a further refinement of the classification based on similarity or analogy in Wojna [119]. In this approach, training set objects are endowed with a metric, and the test objects are classified by voting by k nearest training objects for some k that is subject to optimization.

A far-reaching step in classification based on similarity or analogy is granular classification in Polkowski and Artiemjew [73]; in this approach, a granular counterpart (see Subsect. “Granulation of Knowledge” of Sect. “Rough Set Theory. Extensions”) $(U_{r,\mu}^G, G, S, \{a*: a \in A\}, d^*)$ to a decision system (U, A, d) is constructed; strategy S can be majority voting, and G either a random choice of a covering or a sequential process of selection. To the granular system standard classifiers like LEM2 or RSES classifiers have been applied. This approach does involve a compression of both universes of objects and rule sets. Table 6 shows a comparison of results with other rough set methods on Australian credit data.

Descriptor search is applied in the LEM2 algorithm [28] aimed at issuing a set of rules with a minimal number of descriptors. This system is applied in medical diagnosis, particularly in the problem of missing values, i.e., data in which some attribute values are lacking.

A number of software systems for inducing classifiers were proposed based on rough set methodology, among them LERS by Grzymala-Busse [27]; TRANCE due to Kowalczyk [42]; RoughFamily by Słowiński and Stefanowski [95,96]; TAS by Suraj [102]; PRIMEROSE due to Tsumoto [108]; KDD-R by Ziarko [126]; RSES by Skowron et al. [82]; ROSETTA due to Komorowski, Skowron et al. [39]; RSDM by Fernandez-Baizan et al. [19]; GROBIAN due to Dumentsch and Gediga [17]; Rough-FuzzyLab by Swiniarski [104] and MODLEM by Stefanowski [100].

Rough set techniques were applied in many areas of data exploration, among them in exemplary works:

Processing of audio signals:

Czyżewski et al. [14], Kostek [40].

Handwritten digit recognition:

Nguyen Tuan Trung [54].

Pattern recognition:

Skowron and Swiniarski [93].

Signal classification:

Wojdyłło [118].

Image processing:

Swiniarski and Skowron [105].

Synthesis and analysis of concurrent processes; Petri nets:

Suraj [103].

Conflict analysis, game theory:

Deja [15], Polkowski and Araszkiewicz [72].

Rough neural computation modeling:

Polkowski [70].

Self organizing maps:

Pal, Dasgupta and Mitra [57].

Software engineering:

Peters and Ramanna [65].

Multicriteria decision theory;

operational research

Greco et al. [26].

Programming in logic:

Vitoria [112].

Missing value problem:

Grzymala-Busse [28].

Learning cognitive concepts:

M. Semeniuk-Polkowska [85].

Nearest Neighbor Method

The nearest neighbor method goes back to Fix and Hodges [21,22], Skellam [88] and Clark and Evans [12], who used the idea in statistical tests of significance in nearest neighbor statistics in order to discern between random patterns and clusters. The idea is to assign a class value to a tested object on the basis of the class value of the near-

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 6

Best results for Australian credit by some rough set-based algorithms; in case *, reduction in object size is 49.9 percent, reduction in rule number is 54.6 percent; in case **, resp., 19.7, 18.2; in case ***, resp., 3.6, 1.9

Source	Method	Accuracy	Coverage
(Bazan,1998)	SNAPM(0.9)	error=0.130	–
(Nguyen S.H.,2000)	simple.templates	0.929	0.623
(Nguyen S.H., 2000)	general.templates	0.886	0.905
(Nguyen S.H.,2000)	closest.simple.templates	0.821	1.0
(Nguyen S.H., 2000)	closest.gen.templates	0.855	1.0
(Nguyen S.H., 2000)	tolerance.simple.templ.	0.842	1.0
(Nguyen S.H., 2000)	tolerance.gen.templ.	0.875	1.0
(Wroblewski, 2004)	adaptive.classifier	0.863	–
(Polkowski Artiemjew, 2007)	granular*.r=0.642	0.8990	1.0
(Polkowski Artiemjew, 2007)	granular**.r=0.714	0.964	1.0
(Polkowski Artiemjew, 2007)	granular***.concept.r=0.785	0.9970	0.9995

est with respect to a chosen metric already classified object. The method was generalized to k-nearest neighbors in Patrick and Fischer [59].

The idea of a closest classified object permeates as already mentioned, many rough set algorithms and heuristics; it is also employed by many other paradigms as a natural choice of strategy. The underlying assumption is that data are organized according to some form of continuity and finding a proper metric or similarity measure secures the correctness of the principle that close in that metric or satisfactorily similar objects should have close class values.

To reach for a formal justification of these heuristics, we recall the rudiments of statistical decision theory in its core fragment, the Bayesian decision theory, see, e.g., Duda, Hart and Stork [16]. Assuming that objects in question are assigned to k decision classes c_1, \dots, c_k , with probabilities $p(c_1), \dots, p(c_k)$ (priors) and the probability that an object with feature vector x will be observed in the class c_i is $p(x|c_i)$ (the likelihood), the probability of the class c_i when feature vector x has been observed (posterior) is given by the Bayes formula of Bayes [5]: $p(c_i|x) = (p(x|c_i) \cdot p(c_i))/p(x)$ where the evidence $p(x) = \sum_{j=1}^k p(x|c_j) \cdot p(c_j)$. Bayesian decision theory assigns to an object with feature vector x the class c_i with $p(c_i|x) > p(c_j|x)$ for all $j \neq i$. The Bayesian error in this case is $p^B = 1 - p(c_i|x)$. This decision rule may be conveniently expressed by means of discriminants assigned to classes: to the class c_i the discriminant $g_i(x)$ is assigned which by the Bayes formula can be expressed as an increasing function of the product $p(x|c_i) \cdot p(c_i)$, e.g., $g_i(x) = \log p(x|c_i) + \log p(c_i)$; an object with features x is classified to c_i for which $g_i(x) = \max\{g_j(x): j = 1, 2, \dots, k\}$. The most important case happens when conditional densities $p(x|c_i)$ are distributed normally un-

der $N(\mu_i, \sum_i)$; under simplifying assumptions that prior probabilities are equal and features are independent, the Bayes decision rule becomes the following recipe: for an object with feature vector x , the class c_i is assigned for which the distance $\|x - \mu_i\|$ from x to the mean μ_i is the smallest among all distances $\|x - \mu_j\|$ [16]. In this way the nearest neighbor method can be formally introduced and justified.

In many real cases, estimation of densities $p(x|c_i)$ is for obvious reasons difficult; nonparametric techniques were proposed to bypass this difficulty. The idea of a Parzen window in Parzen [58], see [16], consists of considering a sample S of objects in a d -space potentially increasing to infinity their number, and a sequence $R_1, R_2, \dots, R_n, \dots$ of d -hypercubes of edge lengths h_n hence of volumes $V_n = h_n^d$, R_n containing k_n objects from the sample; the estimate $p_n(x)$ for density induced from R_n is clearly equal to $p_n(x) = (k_n \cdot |S|^{-1})/(V_n)$. Under additional conditions that $\lim_{n \rightarrow \infty} V_n = 0$ and $\lim_{n \rightarrow \infty} n \cdot V_n = \infty$, the sequence $p_n(x)$ does converge to density $p(x)$ in case the latter is continuous [16].

The difficulty with this approach lies in the question of how to choose regions R_i and their parameters and the idea of nearest neighbors returns: the modification rests on the idea that the sampling window should be dependent on the training sample itself and adapt to its structure; the way of implementing this idea can be as follows: one could center the region R about the object with features x and resize it until it absorbs k training objects (k nearest neighbors); if k_i of them falls into the class c_i then the estimate for probability density is $p(c_i|x) = (k_i)/(k)$ [16]. Letting k variable with $\lim_{n \rightarrow \infty} \frac{k}{n} = 0$ and $\lim_{n \rightarrow \infty} k = \infty$ secures that the estimate sequence $p_n(c|x)$ would converge in probability to $p(c_i|x)$ at continuity points of the lat-

ter [16]. The k-nearest neighbor method finds its justification through considerations like the one recalled above.

The 1-nearest neighbor method is the simplest variant of k-nearest neighbors; in the sample space, with the help of a selected metric, it does build neighborhoods of training objects, splitting the space into cells composing together the Voronoi tessellation (Voronoi diagram) of the space, see [79]. The basic theorem by Cover and Hart [13] asserts that the error P in classification by the 1-nearest neighbor method is related to the error P^B in classification by the Bayesian decision rule by the inequality: $P^B \leq P \leq P^B \cdot (2 - (k - 1)/(k) \cdot P^B)$, where k is the number of decision classes. Although the result is theoretically valid under the assumption of infinite sample, yet it can be regarded as an estimate of limits of error by the 1-nearest neighbor method as at most twice the error by Bayes classifier also in case of finite reasonably large samples.

A discussion of analogous results for the k-nearest neighbor method can be found in Ripley [81].

Metrics [see Glossary: "Distance functions (metrics)"] used in nearest neighbor-based techniques can be of varied form; the basic distance function is the Euclidean metric in a d-space: $\rho_E(x, y) = [\sum_{i=1}^d (x_i - y_i)^2]^{1/2}$ and its generalization to the class of Minkowski metrics $L_p(x, y) = [\sum_{i=1}^d |x_i - y_i|^p]^{1/p}$ for $p \geq 1$ with limiting cases of $L_1 = \sum_{i=1}^d |x_i - y_i|$ (the Manhattan metric) and $L_\infty(x, y) = \max\{|x_i - y_i| : i = 1, 2, \dots, d\}$. These metrics can be modified by scaling factors (weights) applied to coordinates, e. g., $L_1^w(x, y) = \sum_{i=1}^d w_i \cdot |x_i - y_i|$ is the Manhattan metric modified by the non-negative weight vector w [119] and subject to adaptive training.

Metrics like the above can be detrimental to the nearest neighbor method in the sense that the nearest neighbors are not invariant with respect to transformations like translations, shifts, rotations. A remedy for this difficulty was proposed as the notion of the tangent distance by Simard, Le Cun and Denker [86]. The idea consists of replacing each training as well as each test object, represented as a vector x in the feature space R^k , with its invariance manifold, see Hastie, Tibshirani and Friedman [32], consisting of x along with all its images by allowable transformations: translation, scaling of axes, rotation, shear, line thickening; instead of measuring distances among object representing vectors x, y , one can find shortest distances among invariance manifolds induced by x and y ; this task can be further simplified by finding for each x the tangent hyperplane at x to its invariance manifold and measuring the shortest distance between these tangents. For a vector x and the matrix T of basic tangent vectors at x to the invariance manifold, the equation of the tangent

hyperplane is $H(x): y = x + Ta$, where a in R^k . The simpler version of the tangent metric method, so-called "one-sided", assumes that tangent hyperplanes are produced for training objects x whereas for test objects x' no invariance manifold is defined but the nearest to x' tangent hyperplane is found and x' is classified as x that defined the nearest tangent hyperplane. In this case the distance to the tangent hyperplane is given by $\arg \min_a \rho_E(x', x + Ta)$ in case the Euclidean metric is chosen as the underlying distance function. In case of a "two-sided" variant, the tangent hyperplane at x' is found as well and x' is classified as the training object x for which the distance between tangent hyperplanes $H(x), H(x')$ is the smallest among distances from $H(x')$ to tangent hyperplanes at training objects.

A simplified variant of the tangent distance method was proposed by Abu-Mostafa, see [32], consisting of producing all images of training and test objects representations in the feature space and regarding them as respectively training and test objects and using standard nearest neighbor techniques.

Among problems related to metrics is also the dimensionality problem; in high-dimensional feature spaces, nearest neighbors can be located at large distances from the test point in question, thus violating the basic principle justifying the method as window choice; as pointed out in [32], the median of the radius R of the sphere about the origin containing the nearest neighbor of n objects uniformly distributed in the cube $[-1/2, 1/2]^p$ is equal to $(1)/(V_p^p) \cdot (1 - 1/2^{1/N})^{1/p}$ where $V_p \cdot r^p$ is the volume of the sphere of radius r in p -space, and it approaches the value of 0.5 with an increase of p for each n , i.e., a nearest neighbor is asymptotically located on the surface of the cube. A related result in Aggarwal et al. [3] asserts that the expected value of the quotient $k^{-1/2} \cdot (\max\{|x|_p, |y|_p\} - \min\{|x|_p, |y|_p\})/(\min\{|x|_p, |y|_p\})$ for vectors x, y uniformly distributed in the cube $(0, 1)^k$ is asymptotically ($k \rightarrow \infty$) equal to $C \cdot (2 \cdot p + 1)^{-1/2}$ (this phenomena are collectively known as the dimensionality curse).

To cope with this problem, a method called discriminant adaptive nearest neighbor was proposed by Hastie and Tibshirani [31]; the method consists of adaptive modification of the metric in neighborhoods of test vectors in order to assure that probabilities of decision classes do not vary much; the direction in which the neighborhood is stretched is the direction of the least change in class probabilities.

This idea that the metric used in nearest neighbor finding should depend locally on the training set was also used in constructions of several metrics in the realm of deci-

sion systems, i. e., objects described in the attribute-value language; for nominal values, the metric *VDM* (Value Difference Metric) in Stanfill and Waltz [97] takes into account conditional probabilities $P(d = v|a_i = v_i)$ of decision value given the attribute value, estimated over the training set *Trn*, and on this basis constructs in the value set V_i of the attribute a_i a metric $\rho_i(v_i, v'_i) = \sum_{v \in V_d} |P(d = v|a_i = v_i) - P(d = v|a_i = v'_i)|$. The global metric is obtained by combining metrics ρ_i for all attributes $a_i \in A$ according to one of many-dimensional metrics, e. g., Minkowski metrics.

This idea was also applied to numerical attributes in Wilson and Martinez [116] in metrics *IVDM* (Interpolated VDM) and *WVDM* (Windowed VDM), see also [119]. A modification of the *WVDM* metric based again on the idea of using probability densities in determining the window size was proposed as the *DBVDM* metric [119].

Implementation of the k-nearest neighbor method requires satisfactorily fast methods for finding k nearest neighbors; the need for accelerating search in training space led to some indexing methods. The method due to Fukunaga and Narendra [24] splits the training sample into a hierarchy of clusters with the idea that in a search for the nearest neighbor the clusters are examined in a top-down manner and clusters evaluated as not fit in the search are pruned from the hierarchy. The alternative is the bottom-up scheme of clustering due to Ward [113]. Distances between test objects and clusters are evaluated as distances between test objects and clusters centers that depend on the chosen strategy.

For features that allow one to map objects onto vectors in a vector space, and clusters in the form of hypercubes a number of tree-based indexing methods exist: k-d trees due to Bentley [8] and quad trees in Finkel and Bentley [20] among them; for high-dimensional feature spaces where the dimensionality curse plays a role, more specialized tree-based indexing methods were proposed: X-trees by Berchtold, Keim and Kriegel [9], SR-trees in Katayama and Satoh [34], TV-trees by Lin, Jagadish and Faloutsos [45]. For general, also non-numerical features, more general cluster regions are necessary and a number of indexing methods like BST due to Kalantari and McDonald [33], GHT by Uhlmann [110], GNAT in Brin [10], SS-tree due to White and Jain [115] and M-trees in Ciacia, Patella and Zezula [11] were proposed. A selection of cluster centers for numerical features may be performed as mean values of feature values of vectors in the cluster; in a general case it may be based on minimization of variance of distance within the cluster [119]. Search for k nearest neighbors of a query object is done in the tree in depth-first order from the root down.

Applications of the k nearest neighbor method include satellite images recognition, protein sequence matching, spatial databases, information retrieval, stock market and weather forecasts, see Aha [4] and Veloso [111].

Case-Based Reasoning

Nearest neighbor techniques can be seen abstractly as tasks of classification of an entity e by means of recovering from the base of already classified entities the nearest one E to e and adopting the class of E as the class of e . Regarded in this light, classification and more generally, problem solving, tasks may be carried out as tasks of recovering from the base of cases already solved the case most similar to the case being currently solved and adopting its solution, possibly with some modifications, as the solution of the current problem. The distinction between this general idea and nearest neighbor techniques is that in the former, new cases are expected to be used in classification of the next cases.

This is the underlying idea of Case-Based Reasoning. The Case-Based Reasoning methodology emerged in response to some difficulties that were encountered by Knowledge-Based Reasoning systems that dominated in the first period of growth of Artificial Intelligence like DENDRAL, MYCIN, or PROSPECTOR. These systems relied on knowledge of the subject in the form of decision rules or models of entities; the problems were of varied nature: complexity of knowledge extraction; difficulty of managing large amounts of information, difficulties with maintenance and refreshing knowledge, see [114]. In spite of search for improvements and a plethora of ideas and concepts aimed at solving the mentioned difficulties, new ideas based on analogy reasoning came forth. Resorting to analogous cases eliminates the need for modeling knowledge as knowledge is cases along with their solutions and methods; implementing consists in identifying features determining cases; large volumes of information can be managed; learning is acquiring new cases [114].

Case-Based Reasoning (CBR) can be traced back to few sources of origin; on a philosophical level, L. Wittgenstein [117] is quoted in Aamodt and Plaza [2] due to his idea of natural concepts and the claim that they be represented as collections of cases rather than features. An interest in cognitive aspects of learning due partially to the contemporary growth of interest in theories of language inspired in large part by work of N. Chomsky and in psychological aspects of learning and exploiting the concept of situation, was reflected in the machine-learning area by works of Schank and Abelson and Schank [84], [83], regarded as pioneering in the CBR area in [114]. Schank and

Abelson's idea was to use scripts as the tool to describe memory of situation patterns. The role of memory in reasoning by cases was analyzed by Schank [83] leading to memory organization packets (MOP) and by Porter as the case memory model. These studies led to the first CBR-based systems: CYRUS by Kolodner [37], MEDIATOR by Simpson [87], PROTOS by Porter and Bareiss [78], among many others; see [114] for a more complete listing.

The methodology of CBR systems was described in [2] as consisting of four distinct processes repeated in cycles: (1) RETRIEVE: the process consisting of matching the currently being solved case against the case base and fetching cases most similar according to adopted similarity measure to the current case; (2) REUSE: the process of making solutions of most similar cases fit to solve the current case; (3) REVISE: the process of revision of fetched solutions to adapt it to the current case taking place after the REUSED solution was proposed, evaluated and turned to be REVISED; (4) RETAIN: the process of RETAINING the REVISED solution in the case base along with the case as the new solved case that in turn can be used in solving future cases.

Whereas the process of reasoning by cases can be adapted to model reasoning by nearest neighbors yet the difference between the two is clear from Aamodt and Plaza's description of the CBR process: in CBR, case base is incrementally enlarged and retained cases become its valid members used in the process of solving new cases; in nearest neighbors the sharp distinction between training and test objects is kept throughout the classification process.

From an implementation point of view, the problem of representing cases is the first to comment upon. According to Kolodner [38], case representation should secure functionality and ease of information acquisition from the case. The structure of a case is a triple (problem, solution, outcome): the problem is a query along with the state of the world (situation) at the moment the query is posed; the solution is the proposed change in the state of the world and the outcome is the new state after the query is answered. The solution may come along with the method in which it was obtained which is important when revision is necessary. In representing cases many formalisms were used like frames/objects, rules, semantic nets, predicate calculi/first order logics etc.

Retrieval of cases from the case base requires a form of case indexing; Kolodner [38] recommended hand-chosen indices as more human-oriented than automated methods but many automated strategies for indexing were introduced and verified, among them (listed in [114]):

Checklist-based indexing in Kolodner [38] that indexes cases by features and dimensions like MEDIATOR indexing which takes into account types and functions of objects being disputed and relationships among disputants;

Difference-based indexing that selects features discerning best among cases like in CYRUS by Kolodner [37];

Methods based on inductive learning and rule induction which extract features used in indexing;

Similarity methods which produce abstractions of cases sharing common sets of features and use remaining features to discern among cases;

Explanation-based methods which select features by inspecting each case and deciding on the set of relevant features.

The retrieval mechanism relies on indexing and memory organization to retrieve cases; the process of selection of the most similar case uses some measures of similarity and a strategy to find next matching cases in the case base. Among these strategies the nearest neighbor strategy can be found, based on chosen similarity measures for particular features; an exemplary similarity measure is the one adopted in the system ReMind in Kolodner [38]:

$$\text{sim}(I, R) = \frac{\sum_{\text{features } f} w_f \cdot \text{sim}_f(f(I), f(R))}{\sum_{\text{features } f} w_f}$$

where w_f is a weight assigned to the feature f , sim_f is the chosen similarity measure on values of feature f , I is the new case and R is the retrieved case. Other methods are based on inductive learning, template retrieval etc., see [114].

Memory organization should provide a trade-off between wealth of information stored in the case base and efficiency of retrieval; two basic case memory models emerged: the dynamic memory model by Schank [83], Kolodner [37] and the category-exemplar model due to Porter and Bareiss [78]. On the basis of Schank's MOP (Memory Organization Packets), Episodic Memory Organization Packets in Kolodner [37] and Generalized Episodes (GE) in Koton [41] were proposed; a generalized episode consists of cases, norms and indices where norms are features shared by all cases in the episode and indices are features discerning cases in the episode. The memory is organized as a network of generalized episodes, cases, index names and index values. New cases are dynamically incorporated into new episodes. The category-exemplar model divides cases (exemplars) into categories, according to case features, and indexes cases by case links from

categories to cases in them; other indices are feature links from features to categories or cases and difference links from categories to similar cases that differ to a small degree in features. Categories are related within a semantic network.

Reusing and adaptation follows two main lines: structural adaptation, see [38]; in this methodology, the solution of the retrieved case is directly adapted by the adaptation rules. Derivational adaptation in Simpson [87] in which the methods that produced the retrieved solution undergo adaptation in order to yield the solution for the new case is the other. Various techniques were proposed and applied in adaptation process, see, e. g., [114].

Applications based on CBR up to the early 1990s are listed in [38]; to mention a few: JUDGE (Bain): the system for sentencing in murder, assault and man-slaughter; KICS (Yang and Robertson): the system for deciding about building regulations; CASEY (Koton): the system for heart failure diagnosis; CADET (Sycara K): the system for assisting in mechanical design; TOTLEC (Costas and Kashyap): the system for planning in the process of manufacturing design; PLEXUS (Alterman): the system for plan adaptation; CLAVIER (Hennessy and Hinkle): the system implemented at Lockheed to control and modify the autoclave processes in part manufacturing; CaseLine (Magaldi): the system implemented at British Airways for maintenance and repair of the Boeing fleet.

Complexity Issues

Complexity of Computations in Information Systems

The basic computational problems: (DM) Computing the discernibility matrix $M_{U,A}$ from an information system (U, A) ; (MLA) The membership in the lower approximation; (RD) Rough definability of sets are of polynomial complexity: (DM) and (RD) in time $O(n^2)$; (MLA) in time $O(n)$ [90].

The core, $CORE(U, A)$, of an information system (U, A) , is the set of all indispensable attributes, i. e., $CORE(U, A) = \{a \in A : \text{ind}(A) \neq \text{ind}(A \setminus \{a\})\}$. As proved in [90], $CORE(U, A) = \{a \in A : c_{i,j} = \{a\} \text{ for some entry } c_{i,j} \text{ into the discernibility matrix. Thus, finding the core requires } O(n^2) \text{ time [90].}$

The reduct membership Problem, i. e., checking whether a given set B of attributes is a reduct, requires $O(n^2)$ time [90].

The reduct set Problem, i. e., finding the set of reducts, is polynomially equivalent to the problem of converting a conjunctive form of a monotone Boolean function into the reduced disjunctive form [90]. The number of reducts

in an information system of n attributes can be exponential with the upper bound of $\binom{n}{\lfloor n/2 \rfloor}$ reducts [89].

The problem of finding a reduct of minimal cardinality is NP-hard [90]; thus, heuristics for reduct finding are in the general case necessary; they are based on the Johnson algorithm, simulated annealing etc. Genetic algorithms-based hybrid algorithms give short reducts in relatively short time [120].

Complexity of Template-Based Computations

Templates (see Subsect. “Similarity” of Sect. “Rough Set Theory. Extensions”) were used in rule induction based on similarity relations [53]. Decision problems related to templates are [53]:

TEMPLATE SUPPORT (TS)

Instance: information system (U, A) ;
natural numbers s, l

Q.: does there exist a template of
length l and support s ?

OPTIMAL TEMPLATE SUPPORT (OTS)

Input: information system (U, A) ; natural
number l

Output: a template T of length l and maximal
support.

The problem TS is NP-complete and the problem OTS is NP-hard [53].

TEMPLATE QUALITY PROBLEM (TQP)

Instance: an information system (U, A) ;
a natural number k

Q.: does there exist a template of
quality greater than k ?

In the case $\text{quality}(T) = \text{support}(T) + \text{length}(T)$, the problem TQP can be solved in polynomial time; in the case $\text{quality}(T) = \text{support}(T) \cdot \text{length}(T)$, the problem TQP is conjectured to be NP-hard.

In [53] some methods for template extraction of satisfactory quality are discussed.

Complexity of Discretization

Discretization (see Subsect. “Discretization” of Sect. “Rough Set Theory. Extensions”) offers a number of decision problems. In the case of numeric values of attributes, the decision problem is to check for an irreducible set of cuts.

IRREDUCIBLE SET OF CUTS (ISC)

Instance: a decision system (U, A, d)

with $|A| \geq 2$; a natural number k

Q.: does there exist an irreducible set
of cuts of cardinality less than k ?

ISC is NP-complete and its optimization version, i.e., if there exists an optimal set of cuts, is NP-hard [53]; in the case $|A| = 1$, ISC is of complexity $O(|U| \cdot \log|U|)$ [53].

For non-numerical values of attributes, the counterpart of a cut system is a partition of the value set; for a partition P_a of an attribute value set V_a into k sets, the rank $r(P_a)$ is k ; for a family $P: \{P_a: a \in A\}$ of partitions for all value sets of attributes in a decision system (U, A, d) , the rank of P is $r(P) = \sum_{a \in A} r(P_a)$; the notion of a consistent partition P mimics that of a consistent cut system. The corresponding problem is

Symbolic Value Partition Problem (SVPP)
 Input: a decision system (U, A, d)
 a set B of attributes; a natural
 number k
 Output: a B -consistent partition of rank
 less or equal k .

The problem SVPP is NP-complete [53].

Problems of generating satisfactory cut systems require some heuristics; Maximal Discernibility Heuristics (MD) [51] is also discussed in [7,53].

Complexity of Problems Related to K-Nearest Neighbors

In the case of n training objects in d -space, the search for the nearest neighbor ($k = 1$) does require $O(dn^2)$ time and $O(n)$ space. A parallel implementation [16] is $O(1)$ in time and $O(n)$ in space. This implementation checks for each of Voronoi regions induced from the training sample whether the test object falls inside it and thus would receive its class label by checking for each of the d "faces" of the region whether the object is on the region side of the face.

The large complexity associated with storage of n training objects, called for eliminating some "redundant" training objects; a technique of editing in Hart [30] consists of editing away, i.e., eliminating from the training set of objects that are surrounded in the Voronoi diagram by objects in the same decision class; the complexity of the editing algorithm is $O(d^3 \cdot n^{\lfloor d/2 \rfloor} \cdot \log n)$.

The complexity of problems related to Voronoi diagrams has been researched in many aspects; the Voronoi diagram for n points in 2-space can be constructed in $O(n \cdot \log n)$ [79]; in d -space the complexity is $\Theta(n^{\lfloor d/2 \rfloor})$ (Klee [35]). Such also is the complexity of finding the nearest neighbor in the corresponding space. There were proposed other editing techniques based on graphs: Gabriel graphs in Gabriel and Sokal [25] and relative neighborhood graphs in Toussaint [107].

Future Directions

It seems reasonable to include the following among problems that could occupy the attention of researchers notwithstanding the paradigm applied:

Compression of knowledge encoded in the training set in order to treat large volumes of data.

Granulation of knowledge as a compression means. This should call for efficient formal granulation models based on similarity relations rather than on metrics that would preserve large parts of information encoded in original data.

Search for the most critical decision/classification rules among the bulk induced from data. This search should lead to a deeper insight into relations in data.

Missing values problem, important especially for medical and biological data; its solution should also lead to new deeper relations and dependencies among attributes.

Working out the methods for analyzing complex data like molecular, genetic and medical data that could include signals, images.

Bibliography

Primary Literature

1. Aamodt A (1991) A knowledge intensive approach to problem solving and sustained learning. Dissertation, University Trondheim, Norway. University Microfilms PUB 92-08460
2. Aamodt A, Plaza E (1994) Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications* 7:39-59
3. Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. In: *Proceedings of the eighth international conference on database theory*, London, pp 420-434
4. Aha DW (1998) The omnipresence of case-based reasoning in science and applications. *Knowl-Based Syst* 11:261-273
5. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Philos Trans R Soc (London)* 53:370-418

6. Bazan JG (1998) A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery*, vol 1. Physica, Heidelberg, pp 321–365
7. Bazan JG et al (2000) Rough set algorithms in classification problems. In: Polkowski L, Tsumoto S, Lin TY (eds) *Rough set methods and applications. New developments in knowledge discovery in information systems*. Physica, Heidelberg, pp 49–88
8. Bentley JL (1975) Multidimensional binary search trees used for associative searching. *Commun ACM* 18:509–517
9. Berchtold S, Keim D, Kriegel HP (1996) The X-tree: an index structure for high dimensional data. In: *Proceedings of the 22nd International Conference on Very Large Databases VLDB'96 1996* Mumbai, Morgan Kaufmann, San Francisco, pp 29–36
10. Brin S (1995) Near neighbor search in large metric spaces. In: *Proceedings of the 21st International Conference on Very Large Databases VLDB'95* Zurich, Morgan Kaufmann, San Francisco, pp 574–584
11. Ciaccia P, Patella M, Zezula P (1997) M-tree: an efficient access method for similarity search in metric spaces. In: *Proceedings of the 23rd International Conference on Very Large Databases VLDB'97*, Athens, Morgan Kaufmann, San Francisco, pp 426–435
12. Clark P, Evans F (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 35:445–453
13. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. *IEEE Trans Inf Theory* IT-13(1):21–27
14. Czyżewski A et al (2004) Musical phrase representation and recognition by means of neural networks and rough sets. In: *Transactions on rough sets*, vol 1. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 254–278
15. Deja R (2000) Conflict analysis. In: Polkowski L, Tsumoto S, Lin TY (eds) *Rough set methods and applications. New developments in knowledge discovery in information systems*. Physica, Heidelberg, pp 491–520
16. Duda RO, Hart PE, Stork DG (2001) *Pattern classification*. Wiley, New York
17. Düntsch I, Gediga G (1998) GROBIAN. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 555–557
18. Faucett WM (1955) Compact semigroups irreducibly connected between two idempotents. *Proc Am Math Soc* 6:741–747
19. Fernandez-Baizan MC et al (1998) RSDM: Rough sets data miner. A system to add data mining capabilities to RDBMS. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 558–561
20. Finkel R, Bentley J (1974) Quad trees: a data structure for retrieval and composite keys. *Acta Inf* 4:1–9
21. Fix E, Hodges JL Jr (1951) Discriminatory analysis: Nonparametric discrimination: Consistency properties. *USAF Sch Aviat Med* 4:261–279
22. Fix E, Hodges JL Jr (1952) Discriminatory analysis: Nonparametric discrimination: Small sample performance. *USAF Sch Aviat Med* 11:280–322
23. Frege G (1903) *Grundlagen der Arithmetik II*. Jena, Hermann Pohle
24. Fukunaga K, Narendra PM (1975) A branch and bound algorithm for computing k-nearest neighbors. *IEEE Trans Comput* 24:750–753
25. Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. *Syst Zool* 18:259–278
26. Greco S, Matarazzo B, Słowiński R (1999) On joint use of indiscernibility, similarity and dominance in rough approximation of decision classes. In: *Proceedings of the 5th international conference of the decision sciences institute*, Athens, Greece, pp 1380–1382
27. Grzymala-Busse JW (1992) LERS – a system for learning from examples based on rough sets. In: Słowiński R (ed) *Intelligent decision support. Handbook of Advances and Applications of the Rough Sets Theory*. Kluwer, Dordrecht, pp 3–18
28. Grzymala-Busse JW (2004) Data with missing attribute values: Generalization of indiscernibility relation and rule induction. In: *Transactions on rough sets*, vol 1. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 78–95
29. Grzymala-Busse JW, Ming H (2000) A comparison of several approaches to missing attribute values in data mining. In: *Lecture notes in AI*, vol 2005. Springer, Berlin, pp 378–385
30. Hart PE (1968) The condensed nearest neighbor rule. *IEEE Trans Inf Theory* IT-14(3):515–516
31. Hastie T, Tibshirani R (1996) Discriminant adaptive nearest-neighbor classification. *IEEE Pattern Recognit Mach Intell* 18:607–616
32. Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning*. Springer, New York
33. Kalantari I, McDonald G (1983) A data structure and an algorithm for the nearest point problem. *IEEE Trans Softw Eng* 9:631–634
34. Katayama N, Satoh S (1997) The SR-tree: an index structure for high dimensional nearest neighbor queries. In: *Proceedings of the 1997 ACM SIGMOD international conference on management of data*, Tucson, AZ, pp 369–380
35. Klee V (1980) On the complexity of d-dimensional Voronoi diagrams. *Arch Math* 34:75–80
36. Klösgen W, Żytkow J (eds) (2002) *Handbook of data mining and knowledge discovery*. Oxford University Press, Oxford
37. Kolodner JL (1983) Maintaining organization in a dynamic long-term memory. *Cogn Sci* 7:243–80
38. Kolodner JL (1993) *Case-based reasoning*. Morgan Kaufmann, San Mateo
39. Komorowski J, Skowron A et al (1998) The ROSETTA software system. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 572–575
40. Kostek B (2007) The domain of acoustics seen from the rough set perspective. In: *Transactions on rough sets*, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 133–151
41. Koton P (1989) Using experience in learning and problem solving. Ph D Dissertation MIT/LCS/TR-441, MIT, Laboratory of Computer Science, Cambridge
42. Kowalczyk W (1998) TRANCE: A tool for rough data analysis, classification and clustering. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 566–568
43. Krawiec K et al (1998) Learning decision rules from similarity based rough approximations. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery*, vol 2. Physica, Heidelberg, pp 37–54
44. Leśniewski S (1916) *Podstawy Ogólnej Teorii Mnogosci* (On

- the Foundations of Set Theory), in Polish. The Polish Scientific Circle, Moscow; see also a later digest: (1982) *Topoi* 2:7–52
45. Lin KI, Jagadish HV, Faloutsos C (1994) The TV-tree: an index structure for high dimensional data. *Vldb J* 3:517–542
 46. Lin TY (1997) From rough sets and neighborhood systems to information granulation and computing with words. In: 5th European Congress on Intelligent Techniques and Soft Computing, 1997 Aachen, Verlagshaus Mainz, Aachen, pp 1602–1606
 47. Lin TY (2005) Granular computing: Examples, intuitions, and modeling. In: Proceedings of IEEE 2005 conference on granular computing GrC05, Beijing, China. IEEE Press, pp 40–44, IEEE Press, New York
 48. Ling C-H (1965) Representation of associative functions. *Publ Math Debrecen* 12:189–212
 49. Michalski RS et al (1986) The multi-purpose incremental learning system AQ15 and its testing to three medical domains. In: Proceedings of AAAI-86. Morgan Kaufmann, San Mateo, pp 1041–1045
 50. Mostert PS, Shields AL (1957) On the structure of semigroups on a compact manifold with a boundary. *Ann Math* 65:117–143
 51. Nguyen SH (1997) Discretization of real valued attributes: Boolean reasoning approach. Ph D Dissertation, Warsaw University, Department of Mathematics, Computer Science and Mechanics
 52. Nguyen SH, Skowron A (1995) Quantization of real valued attributes: Rough set and Boolean reasoning approach. In: Proceedings 2nd annual joint conference on information sciences, Wrightsville Beach, NC, pp 34–37
 53. Nguyen SH (2000) Regularity analysis and its applications in Data Mining. In: Polkowski L, Tsumoto S, Lin TY (eds) *Rough set methods and applications. New developments in knowledge discovery in information systems*. Physica, Heidelberg, pp 289–378
 54. Nguyen TT (2004) Handwritten digit recognition using adaptive classifier construction techniques. In: Pal SK, Polkowski L, Skowron A (eds) *Rough – neural computing. Techniques for computing with words*. Springer, Berlin, pp 573–586
 55. Novotny M, Pawlak Z (1988) Partial dependency of attributes. *Bull Pol Acad Ser Sci Math* 36:453–458
 56. Novotny M, Pawlak Z (1992) On a problem concerning dependence spaces. *Fundam Inform* 16:275–287
 57. Pal SK, Dasgupta B, Mitra P (2004) Rough-SOM with fuzzy discretization. In: Pal SK, Polkowski L, Skowron A (eds) *Rough – neural computing. Techniques for computing with words*. Springer, Berlin, pp 351–372
 58. Parzen E (1962) On estimation of a probability density function and mode. *Ann Math Stat* 33(3):128–152
 59. Patrick EA, Fisher FP (1970) A generalized k-nearest neighbor rule. *Inf Control* 16(2):128–152
 60. Pawlak Z (1982) Rough sets. *Int J Comput Inf Sci* 11:341–356
 61. Pawlak Z (1985) On rough dependency of attributes in information systems. *Bull Pol Acad Ser Sci Tech* 33:551–559
 62. Pawlak Z (1991) *Rough sets: theoretical aspects of reasoning about data*. Kluwer, Dordrecht
 63. Pawlak Z, Skowron A (1993) A rough set approach for decision rules generation. In: Proceedings of IJCAI'93 workshop W12. The management of uncertainty in AI. also: ICS Research Report 23/93 Warsaw University of Technology
 64. Pawlak Z, Skowron A (1994) Rough membership functions. In: Yaeger RR, Fedrizzi M, Kasprzyk J (eds) *Advances in the Dempster–Shafer theory of evidence*. Wiley, New York, pp 251–271
 65. Peters J, Ramanna S (2004) Approximation space for software models. In: *Transactions on rough sets, vol I. Lecture Notes in Computer Science*, vol 3100. Springer, Berlin, pp 338–355
 66. Poincaré H (1902) *Science et hypothese and l'Hypothese*. Flammarion, Paris
 67. Polkowski L (2003) A rough set paradigm for unifying rough set theory and fuzzy set theory. In: Proceedings RSFDGrC03, Chongqing, China, 2003. Lecture Notes in AI, vol 2639. Springer, Berlin, pp 70–78; also: *Fundam Inf* 54:67–88
 68. Polkowski L (2004) Toward rough set foundations. Mereological approach. In: Proceedings RSCTC04, Uppsala, Sweden. Lecture Notes in AI, vol 3066. Springer, Berlin, pp 8–25
 69. Polkowski L (2005) Formal granular calculi based on rough inclusions. In: Proceedings of IEEE 2005 conference on granular computing GrC05, Beijing, China. IEEE Press, New York, pp 57–62
 70. Polkowski L (2005) Rough-fuzzy-neurocomputing based on rough mereological calculus of granules. *Int J Hybrid Intell Syst* 2:91–108
 71. Polkowski L (2006) A model of granular computing with applications. In: Proceedings of IEEE 2006 conference on granular computing GrC06, Atlanta, USA May 10–12. IEEE Press, New York, pp 9–16
 72. Polkowski L, Araszkiewicz B (2002) A rough set approach to estimating the game value and the Shapley value from data. *Fundam Inf* 53(3/4):335–343
 73. Polkowski L, Artiemjew P (2007) On granular rough computing: Factoring classifiers through granular structures. In: Proceedings RSEISP'07, Warsaw. Lecture Notes in AI, vol 4585, pp 280–290
 74. Polkowski L, Skowron A (1994) Rough mereology. In: Proceedings of ISMIS'94. Lecture notes in AI, vol 869. Springer, Berlin, pp 85–94
 75. Polkowski L, Skowron A (1997) Rough mereology: a new paradigm for approximate reasoning. *Int J Approx Reason* 15(4):333–365
 76. Polkowski L, Skowron A (1999) Towards an adaptive calculus of granules. In: Zadeh LA, Kasprzyk J (eds) *Computing with words in information/intelligent systems, vol 1*. Physica, Heidelberg, pp 201–228
 77. Polkowski L, Skowron A, Żytrowski J (1994) Tolerance based rough sets. In: Lin TY, Wildberger M (eds) *Soft Computing: Rough sets, fuzzy logic, neural networks, uncertainty management, knowledge discovery*. Simulation Councils Inc., San Diego, pp 55–58
 78. Porter BW, Bareiss ER (1986) PROTOS: An experiment in knowledge acquisition for heuristic classification tasks. In: Proceedings of the first international meeting on advances in learning (IMAL), Les Arcs, France, pp 159–174
 79. Preparata F, Shamos MI (1985) *Computational geometry: an introduction*. Springer, New York
 80. Rauszer C (1985) An equivalence between indiscernibility relations in information systems and a fragment of intuitionistic logic. *Bull Pol Acad Ser Sci Math* 33:571–579
 81. Ripley BD (1997) *Pattern recognition and neural networks*. Cambridge University Press, Cambridge
 82. Skowron A et al (1994) A system for data analysis. <http://logic.mimuw.edu.pl/~rses/>
 83. Schank RC (1982) *Dynamic memory: A theory of reminding*

- and learning in computers and people. Cambridge University Press, Cambridge
84. Schank RC, Abelson RP (1977) Scripts, plans, goals and understanding. Lawrence Erlbaum, Hillsdale
 85. Semeniuk-Polkowska M (2007) On conjugate information systems: A proposition on how to learn concepts in humane sciences by means of rough set theory. In: Transactions on rough sets, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 298–307
 86. Simard P, Le Cun Y, Denker J (1993) Efficient pattern recognition using a new transformation distance. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems, vol 5. Morgan Kaufmann, San Mateo, pp 50–58
 87. Simpson RL (1985) A computer model of case-based reasoning in problem solving: An investigation in the domain of dispute mediation. Georgia Institute of Technology, Atlanta
 88. Skellam JG (1952) Studies in statistical ecology, I, Spatial pattern. *Biometrika* 39:346–362
 89. Skowron A (1993) Boolean reasoning for decision rules generation. In: Komorowski J, Ras Z (eds) Proceedings of ISMIS'93. Lecture Notes in AI, vol 689. Springer, Berlin, pp 295–305
 90. Skowron A, Rauszer C (1992) The discernibility matrices and functions in decision systems. In: Słowiński R (ed) Intelligent decision support. Handbook of applications and advances of the rough sets theory. Kluwer, Dordrecht, pp 311–362
 91. Skowron A, Stepaniuk J (1996) Tolerance approximation spaces. *Fundam Inf* 27:245–253
 92. Skowron A, Stepaniuk J (2001) Information granules: towards foundations of granular computing. *Int J Intell Syst* 16:57–85
 93. Skowron A, Swiniarski RW (2004) Information granulation and pattern recognition. In: Pal SK, Polkowski L, Skowron A (eds), *Rough – Neural Computing. Techniques for computing with words*. Springer, Berlin, pp 599–636
 94. Slezak D (2000) Various approaches to reasoning with frequency based decision reducts: a survey. In: Polkowski L, Tsumoto S, Lin TY (eds) *Rough set methods and applications. New developments in knowledge discovery in information systems*. Physica, Heidelberg, pp 235–288
 95. Słowiński R, Stefanowski J (1992) “RoughDAS” and “Rough-Class” software implementations of the rough set approach. In: Słowiński R (ed) *Intelligent decision support: Handbook of advances and applications of the rough sets theory*. Kluwer, Dordrecht, pp 445–456
 96. Słowiński R, Stefanowski J (1998) Rough family – software implementation of the rough set theory. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 580–586
 97. Stanfill C, Waltz D (1986) Toward memory-based reasoning. *Commun ACM* 29:1213–1228
 98. Mackie M (2006) Stanford encyclopedia of philosophy: Transworld identity <http://plato.stanford.edu/entries/identity-transworld> Accessed 6 Sept 2008
 99. Stefanowski J (1998) On rough set based approaches to induction of decision rules. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 1*. Physica, Heidelberg, pp 500–529
 100. Stefanowski J (2007) On combined classifiers, rule induction and rough sets. In: Transactions on rough sets, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 329–350
 101. Stepaniuk J (2000) Knowledge discovery by application of rough set models. In: Polkowski L, Tsumoto S, Lin TY (eds) *Rough set methods and applications. New developments in knowledge discovery in information systems*. Physica, Heidelberg, pp 138–233
 102. Suraj Z (1998) TAS: Tools for analysis and synthesis of concurrent processes using rough set methods. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery, vol 2*. Physica, Heidelberg, pp 587–590
 103. Suraj Z (2000) Rough set methods for the synthesis and analysis of concurrent processes. In: Polkowski L, Tsumoto S, Lin TY (eds) *Rough set methods and applications. New developments in knowledge discovery in information systems*. Physica, Heidelberg, pp 379–490
 104. Swiniarski RW (1998) RoughFuzzyLab: A system for data mining and rough and fuzzy sets based classification. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 591–593
 105. Swiniarski RW, Skowron A (2004) Independent component analysis, principal component analysis and rough sets in face recognition. In: Transactions on rough sets, vol I. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 392–404
 106. Sycara EP (1987) Resolving adversarial conflicts: An approach to integrating case-based and analytic methods. Georgia Institute of Technology, Atlanta
 107. Toussaint GT (1980) The relative neighborhood graph of a finite planar set. *Pattern Recognit* 12(4):261–268
 108. Tsumoto S (1998) PRIMEROSE. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 594–597
 109. UCI Repository <http://www.ics.uci.edu/mllearn/databases/> University of California, Irvine, Accessed 6 Sept 2008
 110. Uhlmann J (1991) Satisfying general proximity/similarity queries with metric trees. *Inf Process Lett* 40:175–179
 111. Veloso M (1994) Planning and learning by analogical reasoning. Springer, Berlin
 112. Vitoria A (2005) A framework for reasoning with rough sets. In: Transactions on rough sets, vol IV. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 178–276
 113. Ward J (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58:236–244
 114. Watson I, Marir F (1994) Case-based reasoning: A review <http://www.ai-cbr.org/classroom/cbr-review.html> Accessed 6 Sept 2008; see also: Watson I (1994). *Knowl Eng Rev* 9(4):327–354
 115. White DA, Jain R (1996) Similarity indexing with the SS-tree. In: Proceedings of the twelve international conference on data engineering, New Orleans LA, pp 516–523
 116. Wilson DR, Martinez TR (1997) Improved heterogeneous distance functions. *J Artif Intell Res* 6:1–34
 117. Wittgenstein L (1953) Philosophical investigations. Blackwell, London
 118. Wojdyłło P (2004) WaRS: A method for signal classification. In: Pal SK, Polkowski L, Skowron A (eds) *Rough – neural computing. Techniques for computing with words*. Springer, Berlin, pp 649–688
 119. Wojna A (2005) Analogy-based reasoning in classifier construction. In: Transactions on rough sets, vol IV. Lecture Notes in Computer Science, vol 3700. Springer, Berlin, pp 277–374
 120. Wróblewski J (1998) Covering with reducts – a fast algorithm

- for rule generation. In: *Lecture notes in artificial intelligence*, vol 1424. Springer, Berlin, pp 402–407
121. Wróblewski J (2004) Adaptive aspects of combining approximation spaces. In: Pal SK, Polkowski L, Skowron A (eds) *Rough – neural computing. Techniques for computing with words*. Springer, Berlin, pp 139–156
 122. Yao YY (2000) Granular computing: Basic issues and possible solutions. In: *Proceedings of the 5th Joint Conference on Information Sciences I. Assoc Intell Machinery*, Atlantic NJ, pp 186–189
 123. Yao YY (2005) Perspectives of granular computing. In: *Proceedings of IEEE 2005 Conference on Granular Computing GrC05*, Beijing, China. IEEE Press, New York, pp 85–90
 124. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yaeger RR (eds) *Advances in fuzzy set theory and applications*. North-Holland, Amsterdam, pp 3–18
 125. Zeeman EC (1965) The topology of the brain and the visual perception. In: Fort MK (ed) *Topology of 3-manifolds and selected topics*. Prentice Hall, Englewood Cliffs, pp 240–256
 126. Ziarko W (1998) KDD-R: Rough set-based data mining system. In: Polkowski L, Skowron A (eds) *Rough sets in knowledge discovery 2*. Physica, Heidelberg, pp 598–601

Books and Reviews

- Aviz D, Bhattacharya BK (1983) Algorithms for computing d-dimensional Voronoi diagrams and their duals. In: Preparata FP (ed) *Advances in computing research: Computational geometry*. JAI Press, Greenwich, pp 159–180
- Bocheński JM (1954) *Die Zeitgenössischen Denkmethode*. A. Francke, Bern
- Dasarathy BV (ed) (1991) *Nearest neighbor (NN) norms: NN Pattern classification techniques*. IEEE Computer Society, Washington
- Friedman J (1994) *Flexible metric nearest-neighbor classification*. Technical Report, Stanford University
- Polkowski L (2002) *Rough sets. Mathematical foundations*. Physica, Heidelberg
- Russell SJ, Norvig P (2003) *Artificial intelligence. A modern approach*, 2nd edn. Prentice Hall Pearson Education, Upper Saddle River
- Toussaint GT, Bhattacharya BV, Poulsen RS (1984) Application of voronoi diagrams to nonparametric decision rules. In: *Proceedings of Computer Science and Statistics: The Sixteenth Symposium on the Interface*. North Holland, Amsterdam, pp 97–108
- Watson I (1997) *Applying case-based reasoning. Techniques for enterprise systems*. Morgan Kaufmann, Elsevier, Amsterdam

vant information leading to the knowledge discovery process for extracting meaningful patterns, rules and models from raw data making discovered patterns understandable. Applications include medicine, politics, games, business, marketing, bioinformatics and many other areas of science and engineering. It is an area of research activity that stands at the intellectual intersection of statistics, computer science, machine learning and database management. It deals with very large datasets, tries to make fewer theoretical assumptions than has traditionally been done in statistics, and typically focuses on problems of classification, prediction, description and profiling, clustering, and regression. In such domains, data mining often uses decision trees or neural networks as models and frequently fits them using some combination of techniques such as bagging, boosting/arcing, and racing. Data mining techniques include data visualization, neural network analysis, support vector machines, genetic and evolutionary algorithms, case based reasoning, etc. Other activities in data mining focus on issues such as causation in large-scale systems, and this effort often involves elaborate statistical models and, quite frequently, Bayesian methodology and related computational techniques. Also data mining covers evaluation of the top down approach of model building, starting with an assumed mathematical model solving dynamical equations, with the bottom up approach of time series analysis, which takes measured data as input and provides as output a mathematical model of the system. Their difference is termed measurement error. Data mining allows characterization of chaotic dynamics, involves Lyapunov exponents, fractal dimension and Kolmogorov–Sinai entropy. Data mining comes in two main directions: directed and undirected. Directed data mining tries to categorize or explain some particular target field, while undirected data mining attempts to find patterns or similarities among groups of records without the specific goal.

Pedrycz (see ► [Data and Dimensionality Reduction in Data Analysis and System Modeling](#)) points out that data and dimensionality reduction are fundamental pursuits of data analysis and system modeling. With the rapid growth of the size of data sets and the diversity of data themselves, the use of some reduction mechanisms becomes a necessity. Data reduction is concerned with a reduction of sizes of data sets in terms of the number of data points. This helps reveal an underlying structure in data by presenting a collection of groups present in data. Given the number of groups which are very limited, the clustering mechanisms become effective in terms of data reduction. Dimensionality reduction is aimed at the reduction of the number of attributes (features) of the data which leads to a typically

Data-Mining and Knowledge Discovery, Introduction to

PETER KOKOL

Department of Computer Science, University of Maribor,
Maribor, Slovenia

Data mining and knowledge discovery is the principle of analyzing large amounts of data and picking out rele-

small subset of features or brings the data from a highly dimensional feature space to a new one of a far lower dimensionality. A joint reduction process involves data and feature reduction.

Brameier (see ► [Data-Mining and Knowledge Discovery, Neural Networks](#) in) describes neural networks or, more precisely, artificial neural networks as mathematical and computational models that are inspired by the way biological nervous systems process information. A neural network model consists of a larger number of highly interconnected, simple processing nodes or units which operate in parallel and perform functions collectively, roughly similar as in biological neural networks. Artificial neural networks like their biological counterpart, are adaptive systems which learn by example. Learning works by adapting free model parameters, i. e., the signal strength of the connections and the signal flow, to external information that is presented to the network. In other terms, the information is stored in the weighted connections. If not stated explicitly, let the term “neural network” mean “artificial neural network” in the following. In more technical terms, neural networks are non-linear statistical data modeling tools. Neural networks are generally well suited for solving problem tasks that involve classification of (necessarily numeric) data vectors, pattern recognition and decision-making. The power and usefulness of neural networks have been demonstrated in numerous application areas, like image processing, signal processing, biometric identification – including handwritten character, fingerprint, face, and speech recognition – robotic control, industrial engineering, and biomedicine. In many of these tasks neural networks outperform more traditional statistical or artificial intelligence techniques or may even achieve human-like performance. The most valuable characteristics of neural networks are adaptability and tolerance to noisy or incomplete data. Another important advantage is in solving problems that do not have an algorithmic solution or for which an algorithmic solution is too complex or time-consuming to be found. Brameier describes his chapter as providing a concise introduction to the two most popular neural network types used in application, back propagation neural networks and self-organizing maps. The former learn high-dimensional non-linear functions from given input-output associations for solving classification and approximation (regression) problems. The latter are primarily used for data clustering and visualization and for revealing relationships between clusters. These models are discussed in context with alternative learning algorithms and neural network architectures.

Polkowski (see ► [Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and](#)

[Rough Sets](#)) describes rough set theory as a formal set of notions aimed at carrying out tasks of reasoning, in particular about classification of objects, in conditions of uncertainty. Conditions of uncertainty are imposed by incompleteness, imprecision and ambiguity of knowledge. Applications of this technique can be found in many areas like satellite images analysis, plant ecology, forestry, conflict analysis, game theory, and cluster identification in medical diagnosis. Originally, the basic notion proposed was that of a knowledge base, understood as a collection of equivalence relations on a universe of objects; each relation induces on the set a partition into equivalence classes. Knowledge so encoded is meant to represent the classification ability. As objects for analysis and classification come most often in the form of data, a useful notion of an information system is commonly used in knowledge representation; knowledge base in that case is defined as the collection of indiscernibility relations. Exact concepts are defined as unions of indiscernibility classes whereas inexact concepts are only approximated from below (lower approximations) and from above (upper approximations) by exact ones. Each inexact concept is, thus, perceived as a pair of exact concepts between which it is sandwiched.

Povalej, Verlic and Stiglig (see ► [Discovery Systems](#)) point out that results from simple query for “discovery system” on the World Wide Web returns different types of discovery systems: from knowledge discovery systems in databases, internet-based knowledge discovery, service discovery systems and resource discovery systems to more specific, like for example drug discovery systems, gene discovery systems, discovery system for personality profiling, and developmental discovery systems among others. As illustrated variety of discovery systems can be found in many different research areas, but they focus on knowledge discovery and knowledge discovery systems from the computer science perspective.

A decision tree in data mining or machine learning is a predictive model or a mapping process from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. Podgorelec and Zorman (see ► [Decision Trees](#)) point out that the term “decision trees” has been used for two different purposes: in decision analysis as a decision support tool for modeling decisions and their possible consequences to select the best course of action in situations where one faces uncertainty, and in machine learning or data mining as a predictive model;

that is, a mapping from observations about an item to conclusions about its target value. Their article concentrates on the machine learning view.

Orlov, Sipper and Hauptman (see ► [Genetic and Evolutionary Algorithms and Programming: General Introduction and Application to Game Playing](#)) cover genetic and evolutionary algorithms, which are a family of search algorithms inspired by the process of (Darwinian) evolution in Nature. Common to all the different family members is the notion of solving problems by evolving an initially random population of candidate solutions, through the application of operators like crossover and mutation inspired by natural genetics and natural selection, such that in time “fitter” (i. e., better) solutions emerge. The field, whose origins can be traced back to the 1950s and 1960s, has come into its own over the past two decades, proving successful in solving multitudinous problems from highly diverse domains including (to mention but a few): optimization, automatic programming, electronic-circuit design, telecommunications, networks, finance, economics, image analysis, signal processing, music, and art.

Berkhin and Dhillon (see ► [Knowledge Discovery: Clustering](#)) discuss the condition where data found in scientific and business applications usually do not fit a particular parametrized probability distribution. In other words, the data are complex. Knowledge discovery starts with exploration of this complexity in order to find inconsistencies, artifacts, errors, etc. in the data. After data are cleaned, it is usually still extremely complex. Descriptive data mining deals with comprehending and reducing this complexity. Clustering is a premier methodology in descriptive unsupervised data mining. A cluster could represent an important subset of the data such as a galaxy in astronomical data or a segment of customers in marketing applications. Clustering is important as a fundamental technology to reduce data complexity and to find data patterns in an unsupervised fashion. It is universally used as a first technology of choice in data exploration.

Džeroski, Panov and Zenko (see ► [Machine Learning, Ensemble Methods in](#)) cover ensemble methods, which are machine learning methods that construct a set of predictive models and combine their outputs into a single prediction. The purpose of combining several models together is to achieve better predictive performance, and it has been shown in a number of cases that ensembles can be more accurate than single models. While some work on ensemble methods has already been done in the 1970s, it was not until the 1990s, and the introduction of methods such as bagging and boosting that ensemble methods started to be more widely used. Today, they represent a standard ma-

chine learning method which has to be considered whenever good predictive accuracy is demanded.

Liu and Zhao (see ► [Manipulating Data and Dimension Reduction Methods: Feature Selection](#)) cover feature selection, which is the study of algorithms for reducing dimensionality of data for various purposes. One of the most common purposes is to improve machine learning performance. The other purposes include simplifying data description, streamlining data collection, improving comprehensibility of the learned models, and helping gain insight through learning. The objective of feature selection is to remove irrelevant and/or redundant features and retain only relevant features. Irrelevant features can be removed without affecting learning performance. Redundant features are a type of irrelevant features. The distinction is that a redundant feature implies the co-presence of another feature; individually, each feature is relevant, but the removal of either one will not affect learning performance. As a plethora of data are generated in every possible means with the exponential decreasing costs of data storage and computer processing power, data dimensionality increases on a scale beyond imagination in cases ranging from transactional data to high-throughput data. In many fields such as medicine, health care, Web search, and bioinformatics, it is imperative to reduce high dimensionality such that efficient data processing and meaningful data analysis can be conducted in order to mine nuggets from high-dimensional, massive data.

Data-Mining and Knowledge Discovery, Neural Networks in

MARKUS BRAMEIER

Bioinformatics Research Center, University of Aarhus, Århus, Denmark

Article Outline

- [Glossary](#)
- [Definition of the Subject](#)
- [Introduction](#)
- [Neural Network Learning](#)
- [Feedforward Neural Networks](#)
- [Backpropagation](#)
- [Other Learning Rules](#)
- [Other Neural Network Architectures](#)
- [Self-organizing Maps](#)
- [Future Directions](#)
- [Bibliography](#)

Glossary

Artificial neural network An artificial neural network is a system composed of many simple, but highly interconnected processing nodes or neurons which operate in parallel and collectively. It resembles biological nervous systems in two basic functions: (1) Experiential knowledge is acquired through a learning process and can be retrieved again later. (2) The knowledge is stored in the strength (weights) of the connections between the neurons.

Artificial neuron An artificial neuron receives a number of inputs, which may be either external inputs to the neural network or outputs of other neurons. Each input connection is assigned a weight, similar to the synaptic efficacy of a biological neuron. The weighted sum of inputs is compared against an activation level (threshold) to determine the activation value of the neuron.

Activation function The activation or transfer function transforms the weighted inputs of a neuron into an output signal. Activation functions often have a “squashing” effect. Common activation functions used in neural networks are: threshold, linear, sigmoid, hyperbolic, and Gaussian.

Learning rule The learning rule describes the way a neural network is trained, i. e., how its free parameters undergo changes to fit the network to the training data.

Feedforward network Feedforward neural networks are organized in one or more layers of processing units (neurons). In a feedforward neural network the signal is allowed to flow one-way only, i. e., from inputs to outputs. There are no feedback loops, i. e., the outputs of a layer do not affect its inputs.

Feedback networks In feedback or recurrent networks signals may flow in both directions. Feedback networks are dynamic such that they have a state that is changing continuously until it reaches an equilibrium point.

Definition of the Subject

Neural networks (NNs) or, more precisely, *artificial neural networks* (ANNs) are mathematical and computational models that are inspired by the way biological nervous systems process information. A neural network model consists of a larger number of highly interconnected, simple processing nodes or units which operate in parallel and perform functions collectively, roughly similar to biological neural networks. ANNs, like their biological counterpart, are adaptive systems that learn by example. Learning works by adapting free model parameters, i. e., the signal

strength of the connections and the signal flow, to external information that is presented to the network. In other terms, the information is stored in the weighted connections. If not stated explicitly, let the term “neural network” mean “artificial neural network” in the following.

In more technical terms, neural networks are non-linear statistical data modeling tools. Neural networks are generally well suited for solving problem tasks that involve classification, pattern recognition and decision making. The power, and usefulness of neural networks have been demonstrated in numerous application areas, like image processing, signal processing, biometric identification – including handwritten character, fingerprint, face, and speech recognition – robotic control, industrial engineering, and biomedicine. In many of these tasks neural networks outperform more traditional statistical or artificial intelligence techniques or may even achieve human-like performance. The most valuable characteristics of neural networks are adaptability and tolerance to noisy or incomplete data. Another important advantage is in solving problems that do not have an algorithmic solution or for which an algorithmic solution is too complex or time-consuming to be found.

The first neural network model was developed by McCulloch and Pitts in the 1940s [32]. In 1958 Rosenblatt [41] described the first learning algorithm for a single neuron, the perceptron model. After Rumelhart et al. [46] invented the popular backpropagation learning algorithm for multi-layer networks in 1986, the field of neural networks gained incredible popularity in the 1990s.

Neural networks is regarded as a method of *machine learning*, the largest subfield of *artificial intelligence* (AI). Conventional AI mainly focuses on the development of expert systems and the design of intelligent agents. Today neural networks also belong to the more recent field of *computational intelligence* (CI), which also includes *evolutionary algorithms* (EAs) and *fuzzy logic*.

Introduction

Herein I provide a concise introduction to the two most popular neural network types used in application, backpropagation neural networks (BPNNs) and self-organizing maps (SOMs). The former learn high-dimensional non-linear functions from given input-output associations for solving classification and approximation (regression) problems. The latter are primarily used for data clustering and visualization and for revealing relationships between clusters. These models are discussed in context with alternative learning algorithms and neural network architectures.

Biological Motivation

Artificial neural networks are inspired by biological nervous systems, which are highly distributed and interconnected networks. The human brain is principally composed of a very large number of relatively simple neurons (approx. 100 billion), each of which is connected to several thousand other *neurons*, on average. A neuron is a specialized cell that consists of the cell body (the *soma*), multiple spine-like extensions (the *dendrites*) and a single nerve fiber (the *axon*). The axon connects to the dendrites of another neuron via a synapse. When a neuron is activated, it transmits an electrical impulse (activation potential) along its axon. At the synapse the electric signal is transformed into a chemical signal such that a certain number of neurotransmitters cross the synaptic gap to the post synaptic neuron, where the chemical signal is converted back to an electrical signal to be transported along the dendrites. The dendrites receive signals from the axons of other neurons. One very important feature of neurons is that they react delayed. A neuron combines the strengths (energies) of all received input signals and sends out its own signal (“fires”) only if the total signal strength exceeds a certain critical activation level. A synapse can either be excitatory or inhibitory. Input signals from an excitatory synapse increase the activation level of the neuron while inputs from an inhibitory synapse reduce it. The strength of the input signals critically depends on modulations at the synapses. The brain learns basically by adjusting number and strength of the synaptic connections.

Artificial neural networks copy only a small amount of the biological complexity by using a much smaller number of simpler neurons and connections. Nevertheless, artificial neural networks can perform remarkably complex tasks by applying a similar principle, i. e., the combination of simple and local processing units, each calculating a weighted sum of its inputs and sending out a signal if the sum exceeds a certain threshold.

History and Overview

The history of artificial neural networks begins with a discrete mathematical model of a biological neural network developed by pioneers McCulloch and Pitts in 1943 [32]. This model describes neurons as threshold logic units (TLUs) or binary decision units (BDNs) with multiple binary inputs and a single binary output. A neuron outputs 1 (is activated) if the sum of its unweighted inputs exceeds a certain specified threshold, otherwise it outputs 0. Each neuron can only represent simple logic functions like OR or AND, but any boolean function can be realized by combinations of such neurons.

In 1958 Rosenblatt [41] extended the McCulloch–Pitts model to the perceptron model. This network was based on a unit called the perceptron, which produces an output depending on the weighted linear combination of its inputs. The weights are adapted by the perceptron learning rule. Another single-layer neural network that is based on the McCulloch–Pitts neuron is the ADALINE (ADaptive Linear Element) which was invented in 1960 by Widrow and Hoff [55] and employs a Least-Mean-Squares (LMS) learning rule.

In 1969 Minsky and Papert [33] provided mathematical proofs that single-layer neural networks like the perceptron are incapable of representing functions which are linearly inseparable, including in particular the exclusive-or (XOR) function. This fundamental limitation led the research on neural networks to stagnate for many years, until it was found that a perceptron with more than one layer has far greater processing power.

The backpropagation learning method was first described by Werbos in 1974 [53,54], and further developed for multi-layer neural networks by Rumelhart et al. in 1986 [46]. Backpropagation networks are by far the most well known and most commonly used neural networks today.

Recurrent auto-associative networks were first described independently by Anderson [2] and Kohonen [21] in 1977. Invented in 1982, the Hopfield network [17] is a recurrent neural network in which all connections are symmetric. All neurons are both input and output neurons and update their activation values asynchronously and independently from each other. For each new input the network converges dynamically to a new stable state. A Hopfield network may serve as an associative, i. e., content-addressable, memory.

The Boltzmann machine by Ackley et al. [1] can be seen as an extension of the Hopfield network. It uses a stochastic instead of a deterministic update rule that simulates the physical principle of annealing. The Boltzmann machine is one of the first neural networks to demonstrate learning of an internal representation (hidden units).

It was also in 1982 when Kohonen first published his self-organizing maps [22], a neural networks model based on unsupervised learning and competitive learning. SOMs produce a low-dimensional representation (feature map) of high-dimensional input data while preserving their most important topological features.

Significant progress was made in the 1990s in the field of neural networks which attracted a great deal of attention both in research and in many application domains (see Sect. “[Definition of the Subject](#)”). Faster computers allowed a more efficient solving of more complex problems.

Hardware implementations of larger neural networks were realized on parallel computers or in neural network chips with multiple units working simultaneously.

Characteristics of Neural Networks

Interesting general properties of neural networks are that they (1) mimic the way the brain works, (2) are able to learn by experience, (3) make predictions without having to know the precise underlying model, (4) have a high fault tolerance, i. e., can still give the correct output to missing, noisy or partially correct inputs, and (5) can work with data they have never seen before, provided that the underlying distribution fits the training data.

Computation in neural networks is local and highly distributed throughout the network such that each node operates by itself, but tries to minimize the overall network (output) error in cooperation with other nodes. Working like a cluster of interconnected processing nodes, neural networks automatically distribute both the problem and the workload among the nodes in order to find and converge to a common solution. Actually, parallel processing was one of the original motivations behind the development of artificial neural networks.

Neural networks have the topology of a directed graph. There are only one-way connections between nodes, just like in biological nervous systems. A two-way relationship requires two one-way connections. Many different network architectures are used, often with hundreds or thousands of adjustable parameters.

Neural networks are typically organized in layers which each consist of a number of interconnected nodes (see Fig. 2 below). Data patterns are presented to the system via the *input layer*. This non-computing layer is connected to one or more *hidden layers* where the actual processing is done. The hidden layers then link to an *output layer* which combines the results of multiple processing units to produce the final response of the network.

The network acts as a high-dimensional vector function, taking one vector as input and returning another vector as output. The modeled functions are general and complex enough to solve a large class of non-linear classification and estimation problems. No matter which problem domain a neural network is operating in, the input data always have to be encoded into numbers which may be continuous or discrete.

The high number of free parameters in a neural network and the high degree of collinearity between the neuron outputs let individual parameter settings (weight coefficients) become meaningless and make the network for the most part uninterpretable. High-order interactions be-

tween neurons also do not permit simpler substructures in the model to be identified. Therefore, an extraction of the acquired knowledge from such *black box* predictors to understand the underlying model is almost impossible.

Neural Network Learning

The human brain learns by practice and experience. The learned knowledge can change if more information is received. Another important element of learning is the ability to infer knowledge, i. e., to make assumptions based on what we know and to apply what we have learned in the past to similar problems and situations. One theory of the physiology of human learning by repetition is that repeated sequences of impulses strengthen connections between neurons and form memory paths. To retrieve the learned information, nerve impulses follow these paths to the correct information. If we get out of practice these paths may diminish over time and we forget what we have learned.

The information stored in a neural network is contained in its free parameters. In general, the network architecture and connections are held constant and only the connection weights are variable during training. Once the numbers of (hidden) layers and units have been selected, the free parameters (weights) are set to fit the model or function represented by the network to the training data, following a certain *training algorithm* or *learning rule*. A neural network learns, i. e., acquires knowledge, by adjusting the weights of connections between its neurons. This is also referred to as *connectionist learning*. Training occurs iteratively in multiple cycles during which the training examples are repeatedly presented to the network. During one *epoch* all data patterns pass through the network once.

The three major *learning paradigms* include supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning, in general, means learning by an external teacher using global information. The problem the network is supposed to solve is defined through a set of training examples given. The learning algorithm searches the solution space F , the class of possible functions, for a function $f^* \in F$ that matches this set of input-output associations (\vec{x}, \vec{y}) best. In other words, the mapping implied by the sample data has to be inferred.

Training a neural network means to determine a set of weights which minimizes its prediction error on the training set. The *cost* or *error function* $E: F \rightarrow \mathbb{R}$ measures the error between the desired output values \vec{y} and the predicted network outputs $f(\vec{x})$ over all input vec-

tors \vec{x} . That means, it calculates how far away the current state f of the network is from the optimal solution f^* with $E(f^*) \leq E(f) \forall f \in F$.

A neural network cannot perfectly learn a mapping if the input data does not contain enough information to derive the desired outputs. It may also not converge if there is not enough data available (see also below).

Unsupervised learning uses no external teacher and only local information. It is distinguished from supervised learning by the fact that there is no a priori output. In unsupervised learning we are given some input data \vec{x} , and the cost function to be minimized can be any function of \vec{x} and the network output $f(\vec{x})$. Unsupervised learning incorporates self-organization, i. e., organizes the input data by using only their inherent properties to reveal their emergent collective properties.

A neural network learns *offline* if learning phase and operation (application) phase are separated. A neural network learns *online* if both happens at the same time. Usually, supervised learning is performed offline, whereas unsupervised learning is performed online.

In *reinforcement learning*, neither inputs \vec{x} nor outputs \vec{y} are given explicitly, but are generated by the interactions of an agent within an environment. The agent performs an action \vec{y} with costs c according to an observation \vec{x} made in the environment. The aim is to discover a policy or plan for selecting actions that minimizes some measure of the expected total costs.

Overtraining and Generalization

The overall motivation and most desirable property of neural networks is their ability to generalize to new unknown data, i. e., to classify patterns correctly on which they have not been trained. Minimizing the network error on the training examples only, does not automatically minimize the real error of the unknown underlying function. This important problem is called *overfitting* or *overtraining*.

A regular distribution of training examples over the input data space is important. Generalization is reasonable only as long as the data inputs remain inside the range for which the network was trained. If the training set only included vectors from a certain part of the data space, predictions on other parts are random and likely wrong.

Overtraining may occur also when the iterative training algorithm is run for too long and if the network is too complex for the problem to solve or the available quantity of data. A larger neural network with more weights models a more complex function and invariably achieves a lower error, but is prone to overfitting. A network with

less weights, on the other hand, may not be sufficiently powerful to model the underlying function.

A simple heuristic, called *early stopping*, helps to ensure that the network will generalize well to examples not in the training set. One solution is to check progress during training against an independent data set, the *validation set*. As training progresses, the training error naturally decreases monotonically and, providing training is minimizing the true error function, also the validation error decreases. However, if the validation error stops dropping or even starts to increase again, this is an indication that the network is starting to overfit the data. Then the optimization process has become stuck in a local minima and training should be stopped. The weights that produced the minimum validation error are then used for the final model.

In this case of overtraining, the size of the network, i. e., the number of hidden units and/or hidden layers, may be decreased. Neural networks typically involve experimenting with a large number of different configurations, training each one a number of times while observing the validation error. A problem with repeated experimentation is that the validation set is actually part of the training process. One may just find a network by chance that happens to perform well on the validation set. It is therefore normal practice to reserve a third set of examples for testing the final model on this *test set*.

In many cases a sufficient amount of data is not available, however. Then we have to get around this problem by resampling techniques, like *cross validation*. In principle, multiple experiments are conducted, each using a different division of the available data into training and validation set. This should remove any sampling bias. For small data sets, where splitting the data would leave too few observations for training, *leave-one-out validation* may be used to determine when to stop training or the optimal network size.

Machine-learning techniques, like neural networks, require both positive and negative training examples for solving classification problems. Because they minimize an overall error, the proportion of positive and negative examples in the training set is critical. Ideally, the relation should be close to the (usually unknown) real distribution in the data space. Otherwise, it may bias the network's decision to be more often wrong on unknown data.

Feedforward Neural Networks

In *feedforward neural networks* the information is passed in only one direction (forward) from the inputs, through the hidden nodes (if any) to the output nodes. There are

no connections backwards to neurons of upper layers i. e., there are no feedback loops or cycles in the network.

In feedforward networks with a single-layer of weights, the inputs are directly connected to the output units (*single-layer neural networks*). *Multi-layer feedforward networks* use additional intermediate layers of hidden units. Neural networks with two or more processing layers may have far greater processing power than networks with only one layer. Single-layer neural networks are only capable of learning linearly separable patterns and functions.

The Perceptron

The most simple kind of feedforward neural network is a *perceptron*, which consists of a single pseudo input layer and one or more processing nodes in the output layer. All inputs are weighted and fed directly to the output neuron(s) (see Fig. 1). Each node calculates the sum of the products of weights and inputs. If this value is above some threshold (typically 0) the neuron takes the activated value 1, otherwise it outputs 0. Neurons with this kind of activation function are called *threshold units*. In the literature the term *perceptron* often refers to networks consisting of only one (output) neuron.

More formally, the perceptron is a linear binary classifier that maps a binary n -dimensional input vector $\vec{x} \in \{0, 1\}^n$ to a binary output value $f(\vec{w} \cdot \vec{x}) \in \{0, 1\}$ calculated as

$$f(s) = \begin{cases} 1 & \text{if } s > T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $s = \vec{w} \cdot \vec{x} = \sum_{i=1}^n w_i x_i$ is the input sum and \vec{w} is a vector of real-valued weights. The constant threshold T does not depend on any input value.

In addition to the network topology, the learning rule is an important component of neural networks. Perceptrons can be trained by a simple learning algorithm, called the *delta rule* or *perceptron learning rule*. This realizes

a simple stochastic *gradient descent* where the weights of the network are adjusted depending on the error between the predicted outputs of the network and the example outputs. The delta rule changes the weight vector such that the output error is minimized. McClelland and Rumelhart [30] proved that a neural network using the delta rule can learn associations whenever the inputs are linearly independent.

All neurons of a perceptron share the same structure and learning algorithm. Each weight w_{ij} , representing the influence of input x_i on neuron j , is updated at time t according to the rule:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij} \quad (2)$$

$$\Delta w_{ij} = \alpha(o_i - y_i)x_{ij} \quad (3)$$

The network learns by updating the weight vector after each iteration (training example) by an amount proportional to the difference between given output o_i and calculated output $y_i = f(s_i)$. The *learning rate* α is a constant with $0 < \alpha < 1$ and regulates the learning speed.

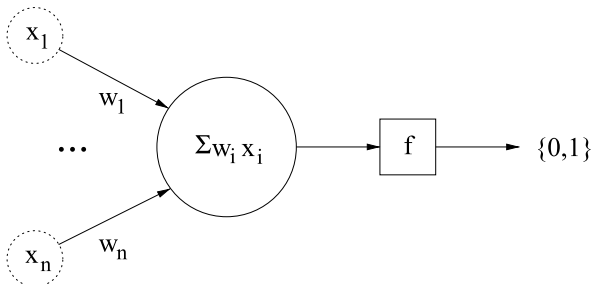
The training data set is *linearly separable* in n -dimensional data space if its two-classes of vectors \vec{x} can be separated by an $(n-1)$ -dimensional hyperplane. If the training examples are not linearly separable, the perceptron learning algorithm is not guaranteed to converge. Linear classifiers, like single-unit perceptrons, are only able to learn, i. e., perfectly classify, linearly separable patterns because they can only implement a simple decision surface (single hyperplane) [33]. The same is true for single-layer neural networks with more than one output unit. This makes these *linear neural networks* unable to learn, for example, the XOR function [33]. Nevertheless a problem that is thought to be highly complex may still be solved as well by a linear network as by a more powerful (non-linear) neural network.

Single-Layer Neural Networks

In general, the state of a neuron is represented by its activation value. An *activation* or *transfer function* f calculates the activation value of a unit from the weighted sum s of its inputs. In case of the perceptron f is called *step* or *threshold function*, with the activation value being 1 if the network sum is greater than a constant T , and 0 otherwise (see Eq. (1)). Another common form of non-linear activation function is the *logistic* or *sigmoid function*:

$$f(s) = \frac{1}{1 + e^{-s}} \quad (4)$$

This enables a neural network to compute a continuous output between 0 and 1 instead of a step function. With



Data-Mining and Knowledge Discovery, Neural Networks in, Figure 1

Principle structure of single-unit perceptron network

this choice, a single-layer network is identical to a logistic regression model. If the activation functions is linear, i. e., the *identity*, then this is just a multiple linear regression and the output is proportional to the total weighted sum s .

Multi-Layer Neural Networks

The limitation that non-linearly separable functions cannot be represented by a single-layer network with fixed weights can be overcome by adding more layers. A multi-layer network is a feedforward network with two or more layers of computational units, interconnected such that the neurons' outputs of one layer serve as inputs only to neurons of the directly subsequent layer (see Fig. 2). The input layer is not considered a real layer with processing neurons.

The number of units in the input layer is determined by the problem, i. e., the dimension of the input data space. The number of output units also depends on the output encoding (see Subsect. "Application Issues").

By using hidden layers, the partitioning of the data space can be more effective. In principle, each hidden unit adds one hyperplane to divide the space and discriminate the solution. Only if the outputs of at least two neurons are combined in a third neuron, the XOR problem is solvable. Important issues in multi-layer NN design are, thus, the specification of the number of hidden layers and the number of units in these layers (see also Subsect. "Application Issues"). Both numbers determine the complexity of functions that can be modeled. There is no theoretical limitation on the number of hidden layers, but usually one or two are used.

The *universal approximation theorem* for neural networks states that any continuous function that maps intervals of real numbers to an output interval of real numbers

can be approximated arbitrarily closely by a multi-layer neural network with only one hidden layer and certain types of non-linear activation functions. This gives, however, no indication about how fast or likely a solution is found. Networks with two hidden layers may work better for some problems. However, more than two hidden layers usually provide only marginal benefit compared to the significant increase in training time.

Any multi-layer network with fixed weights and linear activation function is equivalent to a single-layer (linear) network: In the case of a two-layer linear system, for instance, let all input vectors to the first layer form matrix X and W_1 and W_2 be the weight matrices of the two processing layers. Then the output $Y_1 = W_1 \cdot X$ of the first layer is input to the second layer, which produces output $Y_2 = W_2 \cdot (W_1 \cdot X) = (W_2 \cdot W_1) \cdot X$. This is equivalent to a single-layer network with weight matrix $W = W_2 \cdot W_1$.

Only a multi-layer network that is non-linear can provide more computational power. In many applications these networks use a sigmoid function as non-linear activation function. This is the case at least for the hidden units. For the output layer, the sigmoid activation function is usually applied with classification problems, while a linear transfer function is applied with regression problems.

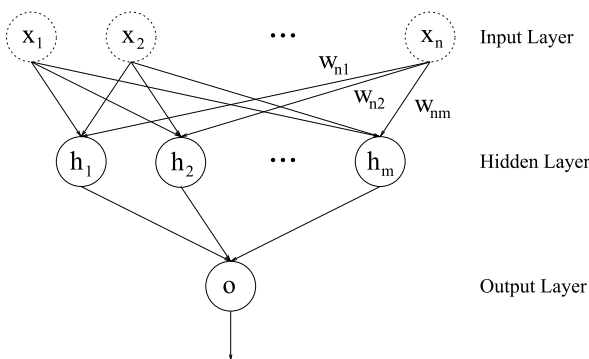
Error Function and Error Surface

The *error function* derives the overall network error from the difference of the network's output y_{ij} and target output o_{ij} over all examples i and output units j . The most-common error function is the *sum squared error*:

$$E = \frac{1}{2} \sum_i \sum_j (o_{ij} - y_{ij})^2. \quad (5)$$

Neural network training performs a search within the space of solution, i. e., all possible network configurations, towards a global minimum of the error surface. The *global minimum* is the best overall solution with the lowest possible error.

A helpful concept for understanding NN training is the *error surface*: The n weight parameters of the network model form the n dimensions of the search space. For any possible state of the network or configuration of weights the error is plotted in the $(n + 1)$ th dimension. The objective of training is to find the lowest point on this n -dimensional surface. The error surface is seldom smooth. Indeed, for most problems, the surface is quite rugged with numerous hills and valleys which may cause the network search to run into a *local minimum*, i. e., a Suboptimum solution.



Data-Mining and Knowledge Discovery, Neural Networks in, Figure 2

Principle structure of multi-layer feedforward neural network

The speed of learning is the rate of convergence between the current solution and the global minimum.

In a linear network with a sum-squared error function, the error surface is an multi-dimensional parabola, i. e., has only one minimum. In general, it is not possible to analytically determine where the global minimum of the error surface is. Training is essentially an exploration of the error surface. Because of the probabilistic and often highly non-linear modeling by neural networks, we cannot be sure that the error could not be lower still, i. e., that the minimum we found is the absolute one. Since the shape of the error space cannot be known a priori, neural network analysis requires a number of independent runs to determine the best solution. When different initial values for the weights are selected, different network models will be derived.

From an initially random configuration of the network, i. e., a random point on the error surface, the training algorithm starts seeking for the global minimum. Small random values are typically used to initialize the network weights. Although neural networks resulting from different initial weights may have very different parameter settings, their prediction errors usually do not vary dramatically. Training is stopped when a maximum number of epochs has expired or when the network error does not improve any further.

Backpropagation

The best known and most popular training algorithm for multi-layer networks is *backpropagation*, short for backwards error propagation and also referred to as the *generalized delta rule* [46]. The algorithm involves two phases:

Forward pass. During the first phase, the free parameters (weights) of the network are fixed. An example pattern is presented to the network and the input signals are propagated through the network layers to calculate the network output at the output unit(s).

Backward pass. During the second phase, the model parameters are adjusted. The error signals at the output units, i. e., the differences between calculated and expected outputs, are propagated back through the network layers. In doing so, the error at each processing unit is calculated and used to make adjustments to its connecting weights such that the overall error of the network is reduced by some small amount.

After iteratively repeating both phases for a sufficiently large number of training cycles (epochs) the network will converge to a state where its output error is small enough. The backpropagation rule involves the repeated use of the *chain rule*, saying that the output error of a neuron can be

ascribed partly to errors in the weights of its direct inputs and partly to errors in the outputs of higher-level (hidden) nodes [46].

Moreover, backpropagation learning may happen in two different modes. In *sequential mode* or *online mode* weight adjustments are made example by example, i. e., each time an example pattern has been presented to the network. The *batch mode* or *offline mode* adjustments are made epoch by epoch, i. e., only after all example patterns have been presented. Theoretically, the backpropagation algorithm performs gradient descent on the total error only if the weights are updated epoch-wise. There are empirical indications, however, that a pattern-wise update results in faster convergence. The training examples should be presented in random order. Then the precision of predictions will be more similar over all inputs.

Backpropagation learning requires a differentiable activation function. Besides adding non-linearity to multi-layer networks, the sigmoid activation function (see Eq. (4)) is often used in backpropagation networks because it has a continuous derivative that can be calculated easily:

$$f'(s) = f(s)(1 - f(s)). \quad (6)$$

We further assume that there is only one hidden layer in order to keep notations and equations clear. A generalization to networks with more than one hidden layer is straightforward.

The backpropagation rule is a generalization of the delta learning rule (see Eq. (3)) to multi-layer networks with non-linear activation function. For an input vector \vec{x} the output $y = f(s)$ is calculated at each output neuron of the network and compared with the desired target output o , resulting in an error δ . Each weight is adjusted proportionally to its effect on the error. The weight of a connection between a unit i and a unit j is updated depending on the output of i (as input to j) and the error signal at j :

$$\Delta w_{ij} = \alpha \delta_j y_i. \quad (7)$$

For an output node j the error signal (error surface gradient) is given by:

$$\delta_j = (o_j - y_j) f'(s_j) = (o_j - y_j) y_j (1 - y_j). \quad (8)$$

If the error is zero, no changes are made to the connection weight. The larger the absolute error, the more the responsible weight is changed, while the sign of the error determines the direction of change.

For a hidden neuron j the error signal is calculated recursively using the signals of all directly connected output neurons k .

$$\delta_j = f'(s_j) \sum_k \delta_k w_{jk} = y_j (1 - y_j) \sum_k \delta_k w_{jk}. \quad (9)$$

The partial derivative of the error function with respect to the network weights can be calculated purely locally, such that each neuron needs information only from neurons directly connected to it. A theoretical foundation of the backpropagation algorithm can be found in [31].

The backpropagation algorithm performs a *gradient descent* by calculating the gradient vector of the error surface at the current search point. This vector points into the direction of the steepest descent. Moving in this direction will decrease the error and will eventually find a new (local) minimum, provided that the *step size* is adapted appropriately. Small steps slow down learning speed, i. e., require a larger number of iterations. Large steps may converge faster, but may also overstep the solution or make the algorithm oscillate around a minimum without convergence of the weights. Therefore, the step size is made proportional to the slope δ , i. e., is reduced when the search point approaches a minimum, and to the learning rate α . The constant α allows one to control the size of the gradient descent step and is usually set to be between 0.1 and 0.5. For practical purposes, it is recommended to choose the learning rate as large as possible without leading to oscillation.

Momentum

One possibility to avoid oscillation and to achieve faster convergence is in the addition of a *momentum term* that is proportional to the previous weight change:

$$\Delta w_{ij}(t+1) = \alpha \delta_j y_i + \beta \Delta w_{ij}(t). \quad (10)$$

The algorithm increases learning speed step size if it has taken several steps in the same direction. This gives it the ability to overcome obstacles in the error surface, e. g., to avoid and escape from local minima, and to move faster over larger plateaus.

Finding the optimum learning rate α and momentum scale parameter β , i. e., the best trade-off between longer training time and instability, can be difficult and might require many experiments. Global or local adaptation techniques use, for instance, the partial derivative to automatically adapt the learning rate. Examples here are the Delta-Bar-Delta rule [18] and the SuperSAB algorithm [51].

Other Learning Rules

The backpropagation learning algorithm is computationally efficient in that its time complexity is linear in the number of weight parameters. Its learning speed is comparatively low, however, on the basis of epochs. This may result in long training times, especially for difficult and

complex problems requiring larger networks or larger amounts of training data.

Another major limitation is that backpropagation does not always converge. Still, it is a widely used algorithm and has its advantages: It is relatively easy to apply and to configure and provides a quick, though not absolutely perfect solution. Its usually pattern-wise error adjustment is hardly affected by data that contains a larger number of redundant examples. Standard backpropagation also generalizes equally well on small data sets as more advanced algorithms, e. g., if there is insufficient information available to find a more precise solution.

There are many variations of the backpropagation algorithm, like resilient propagation (Rprop) [42], quick propagation (Quickprop) [13], conjugate gradient descent [6], Levenberg–Marquardt [16], Delta-Bar-Delta [18], to mention the most popular. All these second-order algorithms are designed to deal with some of the limitations on the standard approach. Some work substantially faster in many problem domains, but require more control parameters than backpropagation, which makes them more difficult to use.

Resilient Propagation

Resilient propagation (Rprop) as proposed in [42,43] is a variant of standard backpropagation with very robust control parameters that are easy to adjust. The algorithm converges faster than the standard algorithm without being less accurate.

The size of the weight step Δw_{ij} taken by standard backpropagation not only depends on the learning rate α , but also on the size of the partial derivative (see Eq. (7)). This may have an unpredictable influence during training that is difficult to control. Therefore, Rprop uses only the sign of derivative to adjust the weights. It necessarily requires learning by epoch, i. e., all adjustments take place after each epoch only.

One iteration of the Rprop algorithm involves two steps, the adjustment of the step size and the update of the weights. The amount of weight change is found by the following update rule:

$$\Delta_{ij}(t) = \begin{cases} \eta^+ \Delta_{ij}(t-1) & \text{if } d_{ij}(t-1) \cdot d_{ij}(t) > 0 \\ \eta^- \Delta_{ij}(t-1) & \text{if } d_{ij}(t-1) \cdot d_{ij}(t) < 0 \\ \Delta_{ij}(t-1) & \text{otherwise} \end{cases} \quad (11)$$

with $0 < \eta^- < 1 < \eta^+$ and derivative $d_{ij} = -\delta_j y_i$. Every time t the derivative term changes its sign, indicating that the last update (at time $t-1$) was too large and the algo-

rithm has jumped over a local minimum, the update value (step size) $\Delta_{ij}(t-1)$ is decreased by a constant factor η^- .

The rule for updating the weights is straightforward:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \quad (12)$$

$$\Delta w_{ij}(t) = \begin{cases} -\Delta_{ij}(t) & \text{if } d_{ij}(t) > 0 \\ +\Delta_{ij}(t) & \text{if } d_{ij}(t) < 0 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

One advantage of the Rprop algorithm, compared to, for example, Quickprop [13], is its small set of parameters that hardly requires adaptation. Standard values for decrease factor η^- and increase factor η^+ are 0.5 and 1.2, respectively. To avoid too large or too small weight values, this is bounded above by Δ_{\max} and bounded below by Δ_{\min} , set by default to 50 and 10^{-6} . The same initial value $\Delta_0 = 0.1$ is recommended for all Δ_{ij} . While the choice of the parameter settings is not critical, for most problems no other choice is needed to obtain the optimal or at least a nearly optimal solution.

Application Issues

The architecture of a neural network, i.e., the number of (hidden) neurons and layers, is an important decision. If a neural network is highly redundant and overparameterized, it might adapt too much to the data. Thus, there is a trade-off between reducing bias (fitting the training data) and reducing variance (fitting unknown data). The most-common procedure is to select a network structure that has more than enough parameters and neurons and then to avoid overfitting only over the training algorithm (see Subject. “[Overtraining and Generalization](#)”).

There is no general best network structure for a particular type of application. There are only general rules for selecting the network architecture: (1) The more complex the relationships between input and output data are, the higher the number of hidden units should be selected. (2) If the modeled process is separable into multiple stages, more than one hidden layer may be beneficial. (3) An upper bound for the total number of hidden units may be set by the number of data examples divided by the number of input and output units and multiplied by a scaling factor. A simpler rule is to start with one hidden layer and half as many hidden units as there are input and output units.

One would expect that for a given data set there would be an optimal network size, lying between a minimum of one hidden neuron (high bias, low variance) and a very large number of neurons (low bias, high variance). While this is true for some data sets, in many cases increasing the number of hidden nodes continues to improve prediction

accuracy, as long as cross validation is used to stop training in time.

For classification problems, the neural network assigns to each input case a class label or, more generally, estimates the probability of the case to fall into each class. The various output classes of a problem are normally represented in neural networks using one of two techniques, including *binary encoding* and *one-out-of-n encoding*.

A binary encoding is only possible for two-class problems. A single unit calculates class 1 if its output is above the acceptance threshold. If the output is below the rejection threshold, class 0 is predicted. Otherwise, the output class is undecided. Should the network output be always defined, both threshold values must be equal (e.g. 0.5).

In one-out-of-n encoding one unit is allocated for each class. A class is selected if the corresponding output is above the acceptance threshold and all the other outputs are below the rejection threshold. If this condition is not met, the class is undecided. Alternatively, instead of using a threshold, a *winner-takes-all* decision may be made such that the unit with the highest output gives the class.

For regression problems, the objective is to estimate the value of a continuous output variable, given the input variables. Particularly important issues in regression are output scaling and interpolation. The most-common NN architectures produce outputs in a limited range. Scaling algorithms may be applied to the training data to ensure that the target outputs are in the same range. Constraining the network's outputs limits its generalization performance. To overcome this, a linear activation function may be used for the output units. Then there is often no need for output scaling at all, since the units can in principle calculate any value.

Other Neural Network Architectures

This section summarizes some alternative NN architectures that are variants or extensions of multi-layer feed-forward networks.

Cascade Correlation Networks

Cascade-correlation is a neural network architecture with variable size and topology [14]. The initial network has no hidden layer and grows during training by adding new hidden units one at a time. In doing so, a near minimal network topology is built. In the cascade architecture the outputs from all existing hidden neurons in the network are fed into a new neuron. In addition, all neurons – including the output neurons – receive all input values.

For each new hidden unit, the learning algorithm tries to maximize the correlation between this unit's output and

the overall network error using an ordinary learning algorithm, like, e. g., backpropagation. After that the input-side weights of the new neuron are frozen. Thus, it does not change anymore and becomes a permanent feature detector.

Cascade correlation networks have several advantages over multi-layer perceptrons: (1) Training time is much shorter already because the network size is relatively small. (2) They require only little or no adjustment of parameters, especially not in terms of the number of hidden neurons to use. (3) They are more robust and training is less likely to become stuck in local minima.

Recurrent Networks

A network architecture with cycles is adopted by *recurrent* or *feedback neural networks* such that outputs of some neurons are fed back as extra inputs. Because past outputs are used to calculate future outputs, the network is said to “remember” its previous state. Recurrent networks are designed to process sequential information, like time series data. Processing depends on the state of the network at the last time step. Consequently, the response to the current input depends on previous inputs.

Two similar types of recurrent network are extensions of the multi-layer perceptron: *Jordan networks* [19] feed back all network outputs into the input layer; *Elman networks* [12] feed back from the hidden units. State or context units are added to the input layer for the feedback connections which all have constant weight one. At each time step t , an input vector is propagated in a standard feedforward fashion, and then a learning rule (usually backpropagation) is applied. The extra units always maintain a copy of the previous outputs at time step $t - 1$.

Radial Basis Function Networks

Radial basic function (RBF) networks [8,34,37,38] are another popular variant of two-layer feedforward neural networks which uses radial basis functions as activation functions. The idea behind radial basis functions is to approximate the unknown function $f(\vec{x})$ by a weighted sum of non-linear basis functions ϕ , which are often Gaussian functions with a certain standard deviation σ .

$$f(\vec{x}) = \sum_i w_i \phi(\|\vec{x} - \vec{c}_i\|) \quad (14)$$

The basis functions operate on the Euclidean distance between n -dimensional input vector \vec{x} and center vector \vec{c}_i . Once the center vectors \vec{c}_i are fixed, the weight coefficients w_i are found by simple linear regression.

The architecture of RBF networks is fixed to two processing layers. Each unit in the hidden layer represents a center vector and a basis function which realizes a non-linear transformation of the inputs. Each output unit calculates a weighted sum (linear combination) of the non-linear outputs from the hidden layer. Only the connections between hidden layer and output layer are weighted.

The use of a linear output layer in RBF networks is motivated by *Cover's theorem* on the separability of patterns. The theorem states that if the transformation from the data (input) space to the feature (hidden) space is non-linear and the dimensionality of the feature space is relatively high compared to that of the data space, then there is a high likelihood that a non-separable pattern classification task in the input space is transformed into a linearly separable one in the feature space.

The center vectors are selected from the training data, either randomly or uniformly distributed in the input space. In principle, as many centers (and hidden units) may be used as there are data examples. Another method is to group the data in space using, for example, k -means clustering, and select the center vectors close to the cluster centers.

RBF learning is considered a curve-fitting problem in high-dimensional space, i. e., approximates a surface with the basis functions that fits and interpolates the training data points best. The basis functions are well-suited to online learning applications, like adaptive process control. Adapting the network to new data and changing data statistics only requires a retraining by linear regression which is fast. RBF networks are more local approximators than multi-layer perceptrons. New training data from one region of the input space have less effect on the learned model and its predictions in other regions.

Self-organizing Maps

A *self-organizing map* (SOM) or *Kohonen map* [22,23] applies an *unsupervised* and *competitive* learning scheme. That means that the class labels of the data vectors are unknown or not used for training and that each neuron improves through competition with other neurons. It is a non-deterministic machine-learning approach to data clustering that implements a mapping of the high-dimensional input data into a low-dimensional feature space. In doing so, SOMs filter and compress information while preserving the most relevant features of a data set. Complex, non-linear relationships and dependencies are revealed between data vectors and between clusters and are transformed into simple geometric distances. Such an ab-

straction facilitates both visualization and interpretation of the clustering result.

Typically, the units of a SOM network are arranged in a two-dimensional regular grid, the *topological feature map*, which defines a two-dimensional Euclidean distance between units. Each unit is assigned a center vector from the n -dimensional data space and represents a certain cluster.

Algorithm 1 describes the basic principle behind SOM training. Starting with an initially random set of center vectors, the algorithm iteratively adjusts them to reflect the clustering of the training data. In doing so, the two-dimensional order of the SOM units is imposed on the input vectors such that more similar clusters (center vectors) in the input space are closer to each other on the two-dimensional grid structure than more different clusters. One can think of the topological map to be folded and distorted into the n -dimensional input space, so as to preserve as much as possible the original structure of the data.

Algorithm 1 (Self-organizing Map)

1. Initialize the n -dimensional center vector $\vec{c}_i \in \mathbb{R}^n$ of each cluster randomly.
2. For each data point $\vec{p} \in \mathbb{R}^n$ find the nearest center vector \vec{c}_w (called the “winner”) in n -dimensional space according to a distance metric d .
3. Move \vec{c}_w and all centers \vec{c}_i within its local neighborhood closer to \vec{p} :

$$\vec{c}_i(t+1) = \vec{c}_i(t) + \alpha(t) \cdot h_{r(t)} \cdot (\vec{p} - \vec{c}_i(t))$$

with learning rate $0 < \alpha < 1$ and neighborhood function h depending on a neighborhood radius r .

4. $\alpha(t+1) = \alpha(t) - \Delta\alpha$ where $\Delta\alpha = \alpha_0/t_{\max}$
 $r(t+1) = r(t) - \Delta r$ where $\Delta r = r_0/t_{\max}$
5. Repeat steps 2.–4. for each epoch $t = 1, \dots, t_{\max}$.
6. Assign each data point to the cluster with the nearest center vector.

Each iteration involves randomly selecting a data point \vec{p} and moving the closest center vector a bit in the direction of \vec{p} . Only distance metric d (usually Euclidean) defined on the data space influences the selection of the closest cluster. The adjustment of centers is applied not just to the winning neuron, but to all the neurons of its neighborhood. The *neighborhood function* is often Gaussian. A simple definition calculates $h = 1$ if the Euclidean distance between the grid coordinates of the winning cluster w and a cluster i is below a radius r , i.e., $|(x_w, y_w) - (x_i, y_i)| < r$, and $h = 0$ otherwise.

Both learning rate α and *neighborhood radius* r monotonically decrease over time. Initially quite large areas of the network are affected by the neighborhood update, leading to a rather rough topological order. As epochs pass, less neurons are altered with lower intensity and finer distinctions are drawn within areas of the map.

Unlike hierarchical clustering [52] and k -means clustering [29], which are both deterministic – apart from the randomized initialization in k -means clustering – and operate only locally, SOMs are less likely to become stuck in local minima and have a higher robustness and accuracy.

Once the network has been trained to recognize structures in the data, it can be used as a visualization tool and for exploratory data analysis. The example map in Fig. 3 shows a clustering of time series of gene expression values [48]. Clearly higher similarities between neighboring clusters are revealed when comparing the mean vectors.

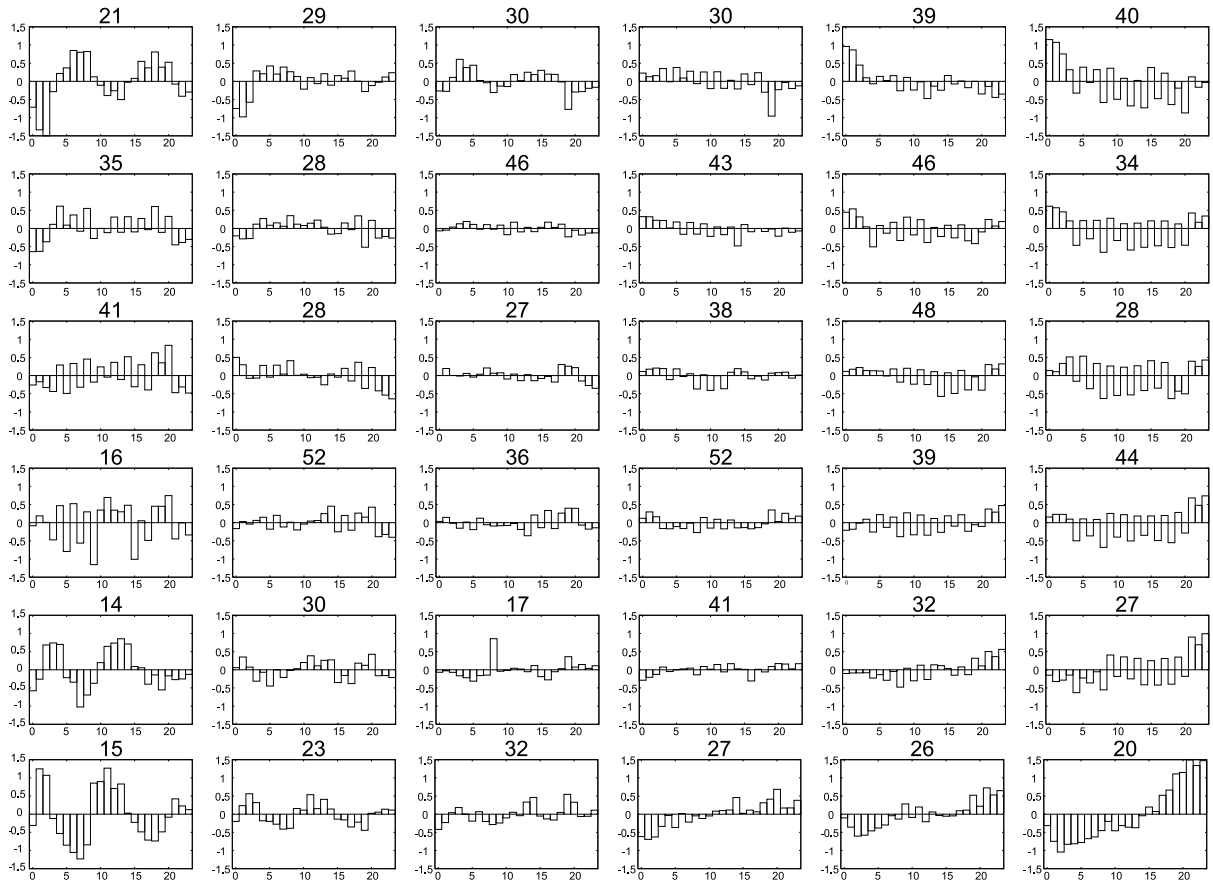
If neurons in the feature map can be labeled, i.e., if common meaning can be derived from the vectors in a clusters, the network becomes capable of classification. If the winning neuron of an unknown input case has not been assigned a class label, labels of clusters in close or direct neighborhood may be considered. Ideally, higher similarities between neighboring data clusters are reflected in similar class labels. Alternatively, the network output is undefined in this case. SOM classifiers also make use of the distance of the winning neuron from the input case. If this distance exceeds a certain maximum threshold, the SOM is regarded as undecided. In this way, a SOM can be used for detecting novel data classes.

Future Directions

To date, neural networks are widely accepted as an alternative to classical statistical methods and are frequently used in medicine [4,5,11,25,27,44,45] with many applications related to cancer research [10,26,35,36,49]. In the first place, these comprise diagnostics and prognosis (i.e. classification) tasks, but also image analysis and drug design. Cancer prediction is often based on clustering of gene expression data [15,50] or microRNA expression profiles [28] which may involve both self-organizing maps and multi-layer feedforward neural networks.

Another broad application field of neural networks today is bioinformatics and, in particular, the analysis and classification of gene and protein sequences [3,20,56]. A well-known successful example is protein (secondary) structure prediction from sequence [39,40].

Even though the NN technology is clearly established today, the current period is rather characterized by stagnation. This is partly because of a redirection of research



Data-Mining and Knowledge Discovery, Neural Networks in, Figure 3

6 x 6 SOM example clustering of gene expression data (time series over 24 time points) [48]. Mean expression vector plotted for each cluster. Cluster sizes indicate the number of vectors (genes) in each cluster

to newer and often – but not generally – more powerful paradigms, like the popular *support vector machines* (SVMs) [9,47], or to more open and flexible methods, like *genetic programming* (GP) [7,24]. In many applications, for example, in bioinformatics, SVMs have already replaced conventional neural networks as the state-of-the-art black-box classifier.

Bibliography

Primary Literature

1. Ackley DH, Hinton GF, Sejnowski TJ (1985) A learning algorithm for Boltzman machines. *Cogn Sci* 9:147–169
2. Anderson JA et al (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychol Rev* 84:413–451
3. Baldi P, Brunak S (2001) *Bioinformatics: The machine learning approach*. MIT Press, Cambridge
4. Baxt WG (1995) Applications of artificial neural networks to clinical medicine. *Lancet* 346:1135–1138
5. Begg R, Kamruzzaman J, Sarkar R (2006) *Neural networks in healthcare: Potential and challenges*. Idea Group Publishing, Hershey
6. Bishop CM (1995) *Neural networks for pattern recognition*. Oxford University Press, London
7. Brameier M, Banzhaf W (2007) *Linear genetic programming*. Springer, New York
8. Broomhead DS, Lowe D (1988) Multivariable functional interpolation and adaptive networks. *Complex Syst* 2:321–355
9. Cristianini N (2000) *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, London
10. Dybowski R (2000) Neural computation in medicine: Perspectives and prospects. In: Malmgren H, Borga M, Niklasson L (eds) *Proceedings of the Conference on Artificial Neural Networks in Medicine and Biology (ANNIMAB)*. Springer, Berlin, pp 26–36
11. Dybowski R, Gant V (2001) *Clinical Applications of Artificial Neural Networks*. Cambridge University Press, London
12. Elman JL (1990) Finding structure in time. *Cogn Sci* 14:179–211
13. Fahlman SE (1989) Faster learning variations on backpropagation: An empirical study. In: Touretzky DS, Hinton GE, Sejnowski TJ (eds) *Proceedings of the (1988) Connectionist Models Summer School*. Morgan Kaufmann, San Mateo, pp 38–51

14. Fahlman SE, Lebiere C (1990) The cascade-correlation learning architecture. In: Touretzky DS (ed) *Advances in Neural Information Processing Systems 2*. Morgan Kaufmann, Los Altos
15. Golub TR et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439):531–537
16. Hagan MT, Menhaj M (1994) Training feedforward networks with the Marquardt algorithm. *IEEE Trans Neural Netw* 5(6):989–993
17. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 79(8):2554–2558
18. Jacobs RA (1988) Increased rates of convergence through learning rate adaptation. *Neural Netw* 1:295–307
19. Jordan MI (1986) Attractor dynamics and parallelism in a connectionist sequential machine. In: *Proceedings of the Eighth Annual Conf of the Cogn Sci Society*. Lawrence Erlbaum, Hillsdale, pp 531–546
20. Keedwell E, Narayanan A (2005) *Intelligent bioinformatics: The application of artificial intelligence techniques to bioinformatics problems*. Wiley, New York
21. Kohonen T (1977) *Associative Memory: A System-Theoretical Approach*. Springer, Berlin
22. Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern* 43:59–69
23. Kohonen T (1995) *Self-organizing maps*. Springer, Berlin
24. Koza JR (1992) *Genetic programming: On the programming of computer programs by natural selection*. MIT Press, Cambridge
25. Lisboa PJG (2002) A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw* 15:11–39
26. Lisboa PJG, Taktak AFG (2006) The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw* 19(4):408–415
27. Lisboa PJG, Ifeachor EC, Szczepaniak PS (2001) *Artificial neural networks in biomedicine*. Springer, Berlin
28. Lu J et al (2005) MicroRNA expression profiles classify human cancers. *Nature* 435:834–838
29. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol 1. University of California Press, Berkeley, pp 281–297
30. McClelland JL, Rumelhart DE (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press, Cambridge
31. McClelland J, Rumelhart D (1988) *Explorations in parallel distributed processing*. MIT Press, Cambridge
32. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5:115–133
33. Minsky ML, Papert SA (1969/1988) *Perceptrons*. MIT Press, Cambridge
34. Moody J, Darken CJ (1989) Fast learning in networks of locally-tuned processing units. *Neural Comput* 1:281–294
35. Naguib RN, Sherbet GV (1997) Artificial neural networks in cancer research. *Pathobiology* 65(3):129–139
36. Naguib RNG, Sherbet GV (2001) *Artificial neural networks in cancer diagnosis, prognosis, and patient management*. CRC Press, Boca Raton
37. Poggio T, Girosi F (1990) Regularization algorithms for learning that are equivalent to multi-layer networks. *Science* 247:978–982
38. Poggio T, Girosi F (1990) Networks for approximation and learning. *Proc IEEE* 78:1481–1497
39. Quian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 202:865–884
40. Rost B (2001) Review: Protein secondary structure prediction continues to rise. *J Struct Biol* 134:204–218
41. Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* 65(6):386–408
42. Riedmiller M, Braun H (1992) Rprop – a fast adaptive learning algorithm. In: *Proceedings of the International Symposium on Computer and Information Science VII*
43. Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: The Rprop algorithm. In: *Proceedings of the IEEE International Conference on Neural Networks*. IEEE Press, Piscataway, pp 586–591
44. Ripley BD, Ripley RM (2001) Neural Networks as statistical methods in survival analysis. In: Dybowski R, Gant V (eds) *Clinical Applications of Artificial Neural Networks*. Cambridge Univ Press, London
45. Robert C et al (2004) Bibliometric overview of the utilization of artificial neural networks in medicine and biology. *Scientometrics* 59:117–130
46. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by backpropagating errors. *Nature* 323:533–536
47. Schölkopf B, Smola AJ (2001) *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge
48. Spellman PT et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9(12):3273–97
49. Taktak AFG, Fisher AC (2007) *Outcome prediction in cancer*. Elsevier Science, London
50. Tamayo P et al (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS* 96(6):2907–2912
51. Tollenaere T (1990) SuperSAB: Fast adaptive backpropagation with good scaling properties. *Neural Netw* 3:561–573
52. Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244
53. Werbos PJ (1974) *Beyond regression: New tools for prediction and analysis in the behavioral science*. PhD Thesis, Harvard University
54. Werbos PJ (1994) *The roots of backpropagation*. Wiley, New York
55. Widrow B, Hoff ME (1960) Adaptive switching circuits. In: *IRE WESCON Convention Record*, Institute of Radio Engineers (now IEEE), vol 4, pp 96–104
56. Wu CH, McLarty JW (2000) *Neural networks and genome informatics*. Elsevier Science, Amsterdam

Books and Reviews

- Abdi H (1994) *A neural network primer*. J Biol Syst 2:247–281
- Bishop CM (2008) *Pattern recognition and machine learning*. Springer, Berlin
- Fausett L (1994) *Fundamentals of neural networks: Architectures, algorithms, and applications*. Prentice Hall, New York

- Freeman JA, Skapura DM (1991) *Neural networks: Algorithms, applications, and programming techniques*. Addison, Reading
- Gurney K (1997) *An Introduction to neural networks*. Routledge, London
- Hastie T, Tibshirani R, Friedman JH (2003) *The elements of statistical learning*. Springer, Berlin
- Haykin S (1999) *Neural networks: A comprehensive foundation*. Prentice Hall, New York
- Hertz J, Krogh A, Palmer R (1991) *Introduction to the theory of neural computation*. Addison, Redwood City
- Kröse B, van der Smagt P (1996) *An introduction to neural networks*. University of Amsterdam, Amsterdam
- Masters T (1993) *Practical neural network recipes in C++*. Academic Press, San Diego
- Masters T (1995) *Advanced algorithms for neural networks: A C++ sourcebook*. Wiley, New York
- Parks R, Levine D, Long D (1998) *Fundamentals of neural network modeling*. MIT Press, Cambridge
- Patterson D (1996) *Artif neural networks*. Prentice Hall, New York
- Peretto P (1992) *An introduction to the modeling of neural networks*. Cambridge University Press, London
- Ripley BD (1996/2007) *Pattern recognition and neural networks*. Cambridge University Press, London
- Smith M (1993) *Neural networks for statistical modeling*. Van Nostrand Reinhold, New York
- Wasserman PD (1989) *Neural Computing: Theory and Practice*. Van Nostrand Reinhold, New York
- Wasserman PD (1993) *Advanced methods in neural computing*. Van Nostrand Reinhold, New York
- De Veaux RD, Ungar LH (1997) *A brief introduction to neural networks*. Technical Report, Williams College, University of Pennsylvania
- Hinton GE (1992) How neural networks learn from experience. *Sci Am* 267:144–151
- Lippman RP (1987) An introduction to computing neural networks. *IEEE ASSP Mag* 4(2):4–22
- Reilly D, Cooper L (1990) An overview of neural networks: Early models to real world systems. *Neural Electron Netw* 2:229–250

Glossary

Accuracy The most important quality measure of an induced decision tree classifier. The most general is the overall accuracy, defined as a percentage of correctly classified instances from all instances (correctly classified and not correctly classified). The accuracy is usually measured both for the training set and the testing set.

Attribute A feature that describes an aspect of an object (both training and testing) used for a decision tree. An object is typically represented as a vector of attribute values. There are two types of attributes: continuous attributes whose domain is numerical, and discrete attributes whose domain is a set of predetermined values. There is one distinguished attribute called decision class (a dependent attribute). The remaining attributes (the independent attributes) are used to determine the value of the decision class.

Attribute node Also called a test node. It is an internal node in the decision tree model that is used to determine a branch from this node based on the value of the corresponding attribute of an object being classified.

Classification A process of mapping instances (i. e. training or testing objects) represented by attribute-value vectors to decision classes. If the predicted decision class of an object is equal to the actual decision class of the object, then the classification of the object is accurate. The aim of classification methods is to classify objects with the highest possible accuracy.

Classifier A model built upon the training set used for classification. The input to a classifier is an object (a vector of known values of the attributes) and the output of the classifier is the predicted decision class for this object.

Decision node A leaf in a decision tree model (also called a decision) containing one of the possible decision classes. It is used to determine the predicted decision class of an object being classified that arrives to the leaf on its path through the decision tree model.

Instance Also called an object (training and testing), represented by attribute-value vectors. Instances are used to describe the domain data.

Induction Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to classify objects by analyzing a set of instances (already solved cases) whose classes are known. Instances are typically represented as attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known class, and the output consists of a mapping from attribute val-

Decision Trees

VILI PODGORELEC, MILAN ZORMAN
University of Maribor, Maribor, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
The Basics of Decision Trees
Induction of Decision Trees
Evaluation of Quality
Applications and Available Software
Future Directions
Bibliography

ues to classes. This mapping should accurately classify both the given instances (a training set) and other unseen instances (a testing set).

Split selection A method used in the process of decision tree induction for selecting the most appropriate attribute and its splits in each attribute (test) node of the tree. The split selection is usually based on some impurity measures and is considered the most important aspect of decision tree learning.

Training object An object that is used for the induction of a decision tree. In a training object both the values of the attributes and the decision class are known. All the training objects together constitute a training set, which is a source of the “domain knowledge” that the decision tree will try to represent.

Testing object An object that is used for the evaluation of a decision tree. In a testing object the values of the attributes are known and the decision class is unknown for the decision tree. All the testing objects together constitute a testing set, which is used to test an induced decision tree – to evaluate its quality (regarding the classification accuracy).

Training set A prepared set of training objects.

Testing set A prepared set of testing objects.

Definition of the Subject

The term decision trees (abbreviated, DT) has been used for two different purposes: in decision analysis as a decision support tool for modeling decisions and their possible consequences to select the best course of action in situations where one faces uncertainty, and in machine learning or data mining as a predictive model; that is, a mapping from observations about an item to conclusions about its target value. This article concentrates on the machine learning view of DT.

More descriptive names for DT models are classification tree (discrete outcome) or regression tree (continuous outcome). In DT structures, leafs represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a DT classifier from data (training objects) is called decision tree learning, or decision trees.

The main goal of classification (and regression) is to build a model that can be used for prediction [16]. In a classification problem, we are given a data set of training objects (a training set), each object having several attributes. There is one distinguished attribute called a decision class; it is a dependent attribute, whose value should be determined using the induced decision tree. The remaining attributes (the independent attributes, we will de-

note them just as attributes in further text) are used to determine the value of the decision class. Classification is thus a process of mapping instances (i. e. training or testing objects) represented by attribute-value vectors to decision classes.

The aim of DT learning is to induce such a DT model that is able to accurately predict the decision class of an object based on the values of its attributes. The classification of an object is accurate if the predicted decision class of the object is equal to the actual decision class of the object. The DT is induced using a training set (a set of objects where both the values of the attributes and decision class are known) and the resulting DT is used to determine decision classes for unseen objects (where the values of the attributes are known but the decision class is unknown). A good DT should accurately classify both the given instances (a training set) and other unseen instances (a testing set).

DT have a wide range of applications, but they are especially attractive in data mining [16]. Because of their intuitive representation, the resulting classification model is easy to understand, interpret and criticize [5]. DT can be constructed relatively fast compared to other methods [28]. And lastly, the accuracy of classification trees is comparable to other classification models [19,28]. Every major data mining tool includes some form of classification tree model construction component [17].

Introduction

In the early 1960s Feigenbaum and Simon presented EPAM (the Elementary Perceiver And Memorizer) [12] – a psychological theory of learning and memory implemented as a computer program. EPAM was the first system to use decision trees (then called discrimination nets). It memorized pairs of nonsense syllables by using discrimination nets in which it could find images of syllables it had seen. It stored as its cue only enough information to disambiguate the syllable from others seen at the time the association was formed. Thus, old cues might be inadequate to retrieve data later, so the system could “forget”. Originally designed to simulate phenomena in verbal learning, it has been later adapted to account for data on the psychology of expertise and concept formation.

Later in the 1960s Hunt et al. built further on this concept and introduced CLS (Concept Learning System) [23] that used heuristic lookahead to construct trees. CLS was a learning algorithm that learned concepts and used them to classify new cases. CLS was the precursor to decision trees; it lead to Quinlan’s ID3 system. ID3 [38] added the idea of using information content to choose the attribute

to split; it also initially chooses a window (a subset) of the training examples, and tries to learn a concept that will correctly classify all based on that window; if not, it increases window size. Quinlan later constructed C4.5 [41], an industrial version of ID3. Utgoff's ID5 [53] was an extension of ID3 that allowed many-valued classifications as well as incremental learning. From the early 1990s both the number of researchers and applications of DT have grown tremendously.

Objective and Scope of the Article

In this article an overview of DT is presented with the emphasis on a variety of different induction methods available today. Induction algorithms ranging from the traditional heuristic-based techniques to the most recent hybrids, such as evolutionary and neural network-based approaches, are described. Basic features, advantages and drawbacks of each method are presented. For the readers not very familiar with the field of DT this article should be a good introduction into this topic, whereas for more experienced readers it should broaden their perspective and deepen their knowledge.

The Basics of Decision Trees

Problem Definition

DT are a typical representative of a symbolic machine learning approach used for the classification of objects into decision classes, whereas an object is represented in a form of an attribute-value vector (*attribute*₁, *attribute*₂, ... *attribute*_N, *decision class*). The attribute values describe the features of an object. The attributes are usually identified and selected by the creators of the dataset. The decision class is one special attribute whose value is known for the objects in the learning set and which will be predicted based on the induced DT for all further unseen objects. Normally the decision class is a feature that could not be measured (for example some prediction for the future) or a feature whose measuring is unacceptably expensive, complex, or not known at all.

Examples of attributes-decision class objects are: patient's examination results and the diagnosis, raster of pixels and the recognized pattern, stock market data and business decision, past and present weather conditions and the weather forecast. The decision class is always a discrete-valued attribute and is represented by a set of possible values (in the case of regression trees a decision class is a continuous-valued attribute). If the decision should be made for a continuous-valued attribute, the values should be dis-

cretized first (by appropriately transforming continuous intervals into corresponding discrete values).

Formal Definition Let A_1, \dots, A_n, C be random variables where A_i has domain $\text{dom}(A_i)$ and C has domain $\text{dom}(C)$; we assume without loss of generality that $\text{dom}(C) = \{c_1, c_2, \dots, c_j\}$. A DT classifier is a function

$$\text{dt} : \text{dom}(A_1) \times \dots \times \text{dom}(A_n) \mapsto \text{dom}(C)$$

Let $P(A', C')$ be a probability distribution on $\text{dom}(A_1) \times \dots \times \text{dom}(A_n) \times \text{dom}(C)$ and let $t = \langle t.A_1, \dots, t.A_n, t.C \rangle$ be a record randomly drawn from P ; i.e., t has probability $P(A', C')$ that $\langle t.A_1, \dots, t.A_n \rangle \in A'$ and $t.C \in C'$. We define the misclassification rate R_{dt} of classifier dt to be $P(\text{dt}(\langle t.A_1, \dots, t.A_n \rangle) \neq t.C)$. In terms of the informal introduction to this article, the training database D is a random sample from P , the A_i correspond to the attributes, and C is the decision class.

A DT is a directed, acyclic graph T in the form of a tree. Each node in a tree has either zero or more outgoing edges. If a node has no outgoing edges, then it is called a decision node (a leaf node); otherwise a node is called a test node (or an attribute node).

Each decision node N is labeled with one of the possible decision classes $c \in \{c_1, \dots, c_j\}$. Each test node is labeled with one attribute $A_i \in \{A_1, \dots, A_n\}$, called the splitting attribute. Each splitting attribute A_i has a splitting function f_i associated with it. The splitting function f_i determines the outgoing edge from the test node, based on the attribute value A_i of an object O in question. It is in a form of $A_i \in Y_i$ where $Y_i \subset \text{dom}(A_i)$; if the value of the attribute A_i of the object O is within Y_i , then the corresponding outgoing edge from the test node is chosen.

The problem of DT construction is the following. Given a data set $D = \{t_1, \dots, t_d\}$ where the t_i are independent random samples from an unknown probability distribution P , find a decision tree classifier T such that the misclassification rate $R_T(P)$ is minimal.

Inducing the Decision Trees

Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to classify objects by analyzing a set of instances (already solved cases) whose decision classes are known. Instances are typically represented as attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known decision class, and the output consists of a mapping from attribute values to decision classes. This mapping should accurately classify both the given instances and other unseen instances.

A decision tree is a formalism for expressing such mappings [41] and consists of test nodes linked to two or more sub-trees and leafs or decision nodes labeled with a decision class. A test node computes some outcome based on the attribute values of an instance, where each possible outcome is associated with one of the sub-trees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate sub-tree. When a leaf is eventually encountered, its label gives the predicted decision class of the instance.

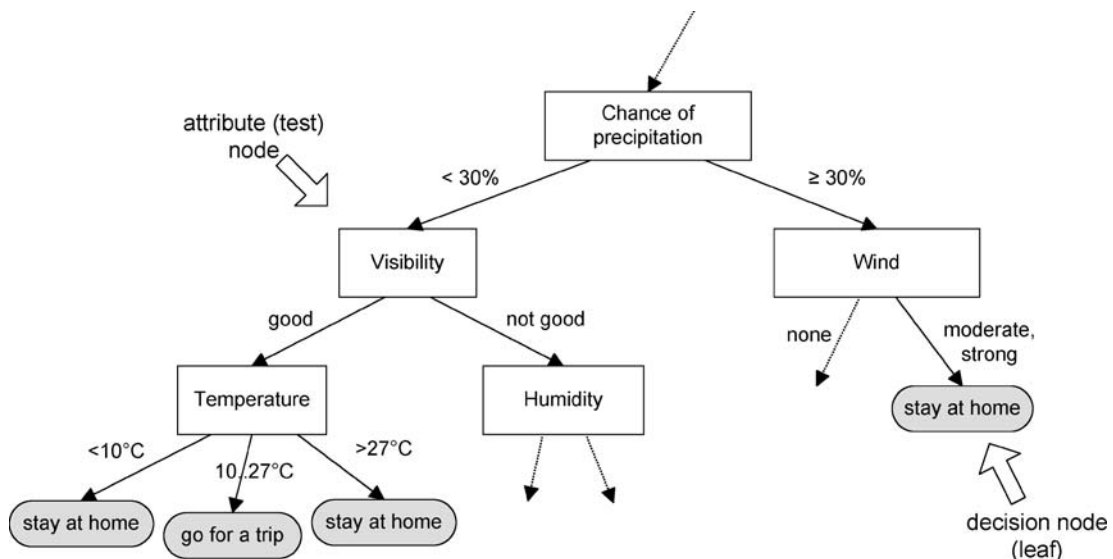
The finding of a solution with the help of DT starts by preparing a set of solved cases. The whole set is then divided into (1) a training set, which is used for the induction of a DT classifier, and (2) a testing set, which is used to check the accuracy of an obtained solution. First, all attributes defining each case are described (input data) and among them one attribute is selected that represents a decision class for the given problem (output data). For all input attributes specific value classes are defined. If an attribute can take only one of a few discrete values then each value takes its own class; if an attribute can take various numeric values then some characteristic intervals must be defined, which represent different classes. Each attribute can represent one internal node in a generated DT, also called a test node or an attribute node (Fig. 1). Such a test node has exactly as many outgoing edges as its number of different value classes. The leafs of a DT are decisions and represent the value classes of the de-

Decision Trees, Table 1

An example training set for object classification

#	Attributes			Decision class
	Color	Edge	Dot	
1	green	dotted	no	triangle
2	green	dotted	yes	triangle
3	yellow	dotted	no	square
4	red	dotted	no	square
5	red	solid	no	square
6	green	solid	yes	triangle
7	green	solid	yes	square
8	yellow	dotted	no	triangle
9	yellow	solid	no	square
10	red	solid	no	square
11	green	solid	yes	square
12	yellow	dotted	yes	square
13	yellow	solid	no	square
14	red	dotted	yes	triangle

cision attribute – decision classes (Fig. 1). When a decision has to be made for an unsolved case, we start with the root node of the DT classifier and moving along attribute nodes select outgoing edges where values of the appropriate attributes in the unsolved case matches the attribute values in the DT classifier until the leaf node is reached representing the decision class. An example training database is shown in Table 1 and a sample DT is shown in Fig. 1.



Decision Trees, Figure 1

An example of a (part of a) decision tree

The DT classifier is very easy to interpret. From the tree shown in Fig. 1 we can deduce for example the following rules:

- If the chance of precipitation is less than 30% and the visibility is good and the temperature is in the range of 10–27°C, then we should go for a trip,
- If the chance of precipitation is less than 30% and the visibility is good and the temperature is less than 10°C or more than 27°C, then we should stay at home,
- If the chance of precipitation is 30% or more and the wind is moderate or strong, then we should stay at home.

A DT can be built from a set of training objects with the “divide and conquer” principle. When all objects are of the same decision class (the value of the output attribute is the same) then a tree consists of a single node – a leaf with the appropriate decision. Otherwise an attribute is selected and a set of objects is divided according to the splitting function of the selected attribute. The selected attribute builds an attribute (test) node in a growing DT classifier, for each outgoing edge from that node the inducing procedure is repeated upon the remaining objects regarding the division until a leaf (a decision class) is encountered.

From a geometric point of view a set of n attributes defines a n -dimensional space, where each data object represents a point. A division of data objects regarding the attribute's class suits the definition of decision planes in the same space. Those planes are hyper-planes which are orthogonal to the selected attribute – DT divides a search space into hyper-rectangles, each of them represents one of the possible decision classes; of course more rectangles can also represent the same decision class.

Induction of Decision Trees

In 1986 Quinlan introduced an algorithm for inducing decision trees called ID3 [39,40]. In 1993 ID3 was upgraded with an improved algorithm C4.5 [41] that is still regarded as the reference model to build a DT based on the traditional statistical approach. Both algorithms ID3 and C4.5 use the statistical calculation of information gain from a single attribute to build a DT. In this manner an attribute that adds the most information about the decision upon a training set is selected first, followed by the next one that is the most informative from the remaining attributes, etc.

The method for constructing a DT as paraphrased from Quinlan [41], pp. 17–18, is as follows:

If there are j classes denoted $\{c_1, c_2, \dots, c_j\}$, and a training set D , then

- *If D contains one or more objects which all belong to a single class c_i , then the decision tree is a leaf identifying class c_i*
- *If D contains no objects, the decision tree is a leaf determined from information other than D*
- *If D contains objects that belong to a mixture of classes, then a test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes $\{o_1, o_2, \dots, o_n\}$. D is partitioned into subsets D_1, D_2, \dots, D_n , where D_i contains all the objects in D that have outcome o_i of the chosen test. The same method is applied recursively to each subset of training objects.*

Split Selection

The most important aspect of a traditional DT induction strategy is the way in which a set is split, i. e. how to select an attribute test that determines the distribution of training objects into sub-sets upon which sub-trees are built consequently. This process is called split selection. Its aim is to find an attribute and its associated splitting function for each test node in a DT. In the following text some of the most widely used split selection approaches will be presented.

Let D be the learning set of objects described with attributes A_1, \dots, A_n and a decision class C . Let n denote the number of training objects in D , n_i the number of training objects of decision class c_i , n_j the number of training objects with the j th value of splitting attribute, and n_{ij} the number of training objects of decision class c_i and j th value of the splitting attribute. The relative frequencies (probabilities) of training objects in D are as follows:

$p_{ij} = \frac{n_{ij}}{n}$ is the relative frequency of training objects of decision class c_i and j th value of splitting attribute within the training set D , $p_i = \frac{n_i}{n}$ is the relative frequency of training objects of decision class c_i within the training set D , $p_j = \frac{n_j}{n}$ is the relative frequency of training objects with the j th value of splitting attribute within the training set D , and $p_{i|j} = \frac{n_{ij}}{n_j}$ is the relative frequency of training objects of decision class c_i and j th value of splitting attribute regarding all the training objects with j th value of splitting attribute within the training set D .

Entropy, Information Gain, Information Gain Ratio

The majority of classical DT induction algorithms (such as ID3, C4.5, ...) are based on calculating entropy from the information theory [45] to evaluate splits. In information theory entropy is used to measure the unreliability of a message as the source of information; the more infor-

mation a message contains, the lower is the entropy. Two splitting criteria are implemented:

- Gain criterion, and
- Gain ratio criterion.

The gain criterion [41] is developed in the following way:

For any training set D , n_i is the number of training objects of decision class c_i within the training set D . Then consider the “message” that a randomly selected object belongs to decision class c_i . The “message” has probability $p_i = \frac{n_i}{n}$, where n is the total number of training objects in D . The information conveyed by the message (in bits) is given by

$$I = -\log_2 p_i = -\log_2 \frac{n_i}{n}.$$

Entropy E of an attribute A of a training object w with the possible attribute values a_1, \dots, a_m and the probability distribution $p(A(w) = a_i)$ is thus defined as

$$E_A = -\sum_j p_j \cdot \log_2 p_j.$$

Let E_C be entropy of decision class distribution, E_A be entropy of values of an attribute A , and E_{CA} be entropy of combined decision class distribution and attribute values as follows:

$$E_C = -\sum_i p_i \cdot \log_2 p_i$$

$$E_A = -\sum_j p_j \cdot \log_2 p_j$$

$$E_{CA} = -\sum_i \sum_j p_{ij} \cdot \log_2 p_{ij}.$$

The expected entropy of decision class distribution regarding the attribute A is thus defined as

$$E_{C|A} = E_{CA} - E_A.$$

This expected entropy $E_{C|A}$ measures the reduction of entropy that is to be expected in the case when A is selected as the splitting attribute. The information gain is thus defined as

$$I_{\text{gain}}(A) = E_C - E_{C|A}.$$

In each test node an attribute A is selected with the highest value $I_{\text{gain}}(A)$. The gain criterion [41] selects a test to maximize this information gain.

The gain criterion has one significant disadvantage in that it is biased towards tests with many outcomes. The

gain ratio criterion [41] was developed to avoid this bias; it is defined as

$$I_{\text{gain ratio}}(A) = \frac{I_{\text{gain}}(A)}{E_A}.$$

If the split is near trivial, split information will be small and this ratio will be unstable. Hence, the gain ratio criterion selects a test to maximize the gain ratio subject to the constraint that the information gain is large.

Gini Index The gain ratio criterion compares with CART’s impurity function approach [5], where impurity is a measure of the class mix of a subset and splits are chosen so that the decrease in impurity is maximized. This approach led to the development of the Gini index [5].

The impurity function approach considers the probability of misclassifying a new sample from the overall population, given that the sample was not part of the training sample, T . This probability is called the misclassification rate and is estimated using either the resubstitution estimate (or training set accuracy), or the test sample estimate (test set accuracy). The node assignment rule selects i to minimize this misclassification rate. In addition, the Gini index promotes splits that minimize the overall size of the tree.

$$\text{gini}(A) = -\sum_j p_j \sum_i p_{i|j}^2 - \sum_i p_i^2.$$

Chi-Square Various types of chi-square tests are often used in statistics for testing the significance. The Chi-square (chi-square goodness-of-fit) test compares the expected frequency e_{ij} with the actual frequency n_{ij} of training objects, which belong to decision class c_i and having j th value of the given attribute [55]. It is defined as

$$\chi^2(A) = \sum_i \sum_j \frac{(e_{ij} - n_{ij})^2}{e_{ij}}$$

where

$$e_{ij} = \frac{n_j n_i}{n}.$$

The higher value of χ^2 means a clearer split. Consequently, the attribute with the highest χ^2 value is selected.

J-Measure The J-measure heuristics has been introduced by Smyth and Goodman [48] as an information-theoretical model of measuring the information content of a rule.

The crossover entropy J_j is appropriate for selecting a single value of a given attribute A for constructing a rule; it is defined as

$$J_j(A) = p_j \sum_i p_{i|j} \log \frac{p_{i|j}}{p_i}.$$

The generalization over all possible values of a given attribute A measures the purity of the attribute

$$J(A) = \sum_j J_j(A).$$

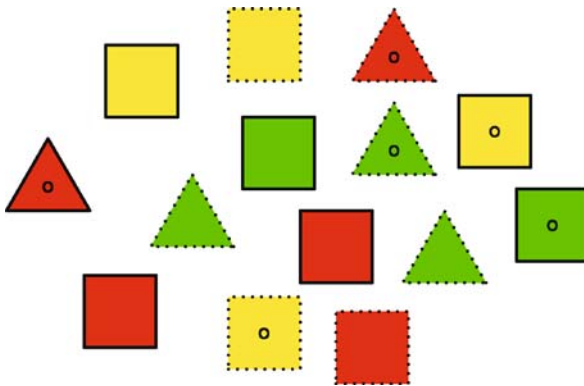
The J-measure is also used to reduce over-fitting in the process of pre-pruning (see the following sections).

DT Induction Example: Classifying Geometrical Objects

Let us have a number of geometrical objects: squares and triangles. Each geometrical object is described with three features: color describes the color of an object (it can be either green, yellow or red), edge describes the line type of an object's edge (it can be either solid or dotted), and dot describes whether there is a dot in an object. See Table 1 for details.

We want to induce a DT that will predict the shape of an unseen object based on the three features. The decision class will thus be **shape**, having two possible values: *square* or *triangle*. The three features represent the tree attributes: **color** (with possible values *green*, *yellow* and *red*), **edge** (with possible values *solid* and *dotted*), and **dot** (with possible values *yes* and *no*). Table 1 resembles the training set with 14 training objects; in Fig. 2 the training objects are also visually represented.

In the training set (Table 1, Fig. 2) there are 14 training objects: five triangles and nine squares. We can calculate



Decision Trees, Figure 2
Visual representation of training objects from Table 1

the class distributions:

$$p(\text{square}) = \frac{9}{14} \text{ and } p(\text{triangle}) = \frac{5}{14}.$$

The entropy of decision class (shape) distribution is thus

$$\begin{aligned} E_{\text{shape}} &= - \sum_i p_i \cdot \log_2 p_i \\ &= -\frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14} = 0.940. \end{aligned}$$

The entropy can be reduced by splitting the whole training set. For this purpose either of the attributes (color, edge, or dot) can be used as a splitting attribute. Regarding the information gain split selection approach, the entropy of each attribute is calculated first, then the information gain of each attribute is calculated based on these entropies, and finally the attribute is selected that has the highest information gain.

To demonstrate it, let us calculate the entropy for the attribute color. Using different possible values of color (green, yellow, and red) the whole training set is split into three subsets (Fig. 3). The entropies are as follows:

$$E_{\text{color=green}} = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 0.971$$

$$E_{\text{color=yellow}} = -\frac{4}{4} \cdot \log_2 \frac{4}{4} - \frac{0}{4} \cdot \log_2 \frac{0}{4} = 0.0$$

$$E_{\text{color=red}} = -\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} = 0.971.$$

Consequently, the entropy of the attribute color based on the whole training set is

$$\begin{aligned} E_{\text{color}} &= \sum_i p_i E_i = \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0.0 + \frac{5}{14} \cdot 0.971 \\ &= 0.694. \end{aligned}$$

The information gain of the attribute color is thus

$$\begin{aligned} I_{\text{gain}}(\text{color}) &= E_{\text{shape}} - E_{\text{shape|color}} \\ &= 0.940 - 0.694 = 0.246. \end{aligned}$$

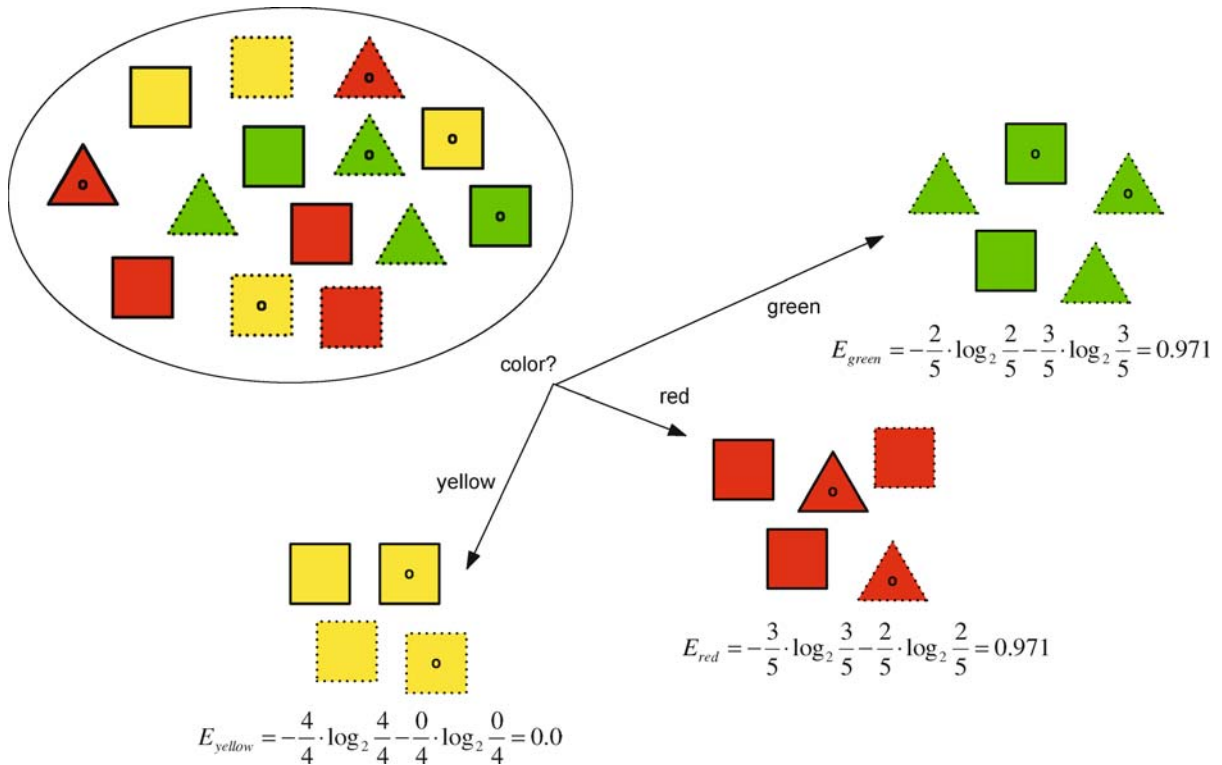
Similarly, the information gain can be calculated also for the remaining two attributes, edge and dot. The information gain for all three attributes are thus

$$I_{\text{gain}}(\text{color}) = 0.246$$

$$I_{\text{gain}}(\text{edge}) = 0.151$$

$$I_{\text{gain}}(\text{dot}) = 0.048.$$

As we can see the attribute color has the highest information gain and is therefore chosen as the splitting attribute



Decision Trees, Figure 3

Splitting the data set based on the attribute color

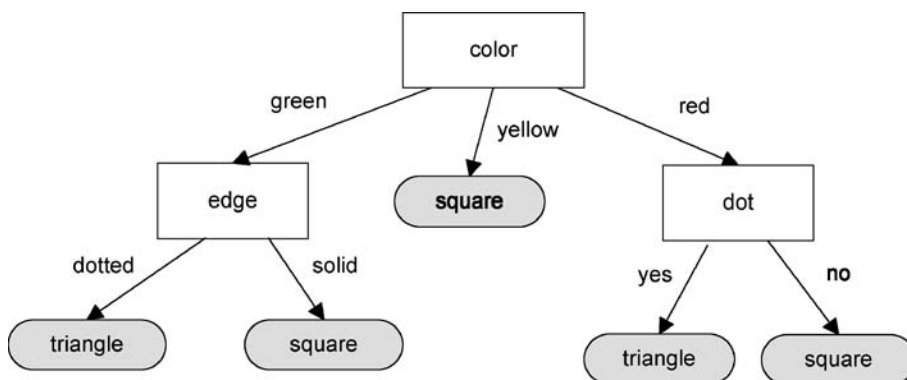
at the root node of the DT classifier. The whole process of splitting is recursively repeated on all subsets. The resulting DT classifier is shown in Fig. 4.

Tests on Continuous Attributes

In the above example all attributes were discrete-valued. For the continuous attributes a split threshold needs to

be determined whenever the attribute is to be used as a splitting attribute. The algorithm for finding appropriate thresholds for continuous attributes [5,33,41] is as follows:

The training objects are sorted on the values of the attribute. Denote them in order as $\{w_1, w_2, \dots, w_k\}$. Any threshold value lying between w_i and w_{i+1} will have the same effect, so there are only $k-1$ possible splits, all of which are examined.



Decision Trees, Figure 4

The resulting DT classifier for the classification of geometrical objects

Some approaches for building DT use discretization of continuous attributes. Two ways are used to select discretized classes:

- equidistant intervals; The Number Of Classes Is Selected First And Then Successive Equidistant Intervals Are Determined Between Absolute Lower And Upper Bounds, And
- percentiles; Again The Number Of Classes Is Selected First And Then Successive Intervals Are Determined Based On The Values Of The Appropriate Attribute In The Training Set So That All Intervals Contain The Same Number Of Training Objects.

Dynamic Discretization of Attributes In the MtDeciT 2.0 tool authors implemented an algorithm for finding subintervals [58], where the distribution of training objects is considered and there are more than two subintervals possible. The approach is called dynamic discretization of continuous attributes, since the subintervals are determined dynamically during the process of building a DT. This technique first splits the interval into many subintervals, so that every training object's value has its own subinterval. In the second step it merges together smaller subintervals that are labeled with the same outcome into

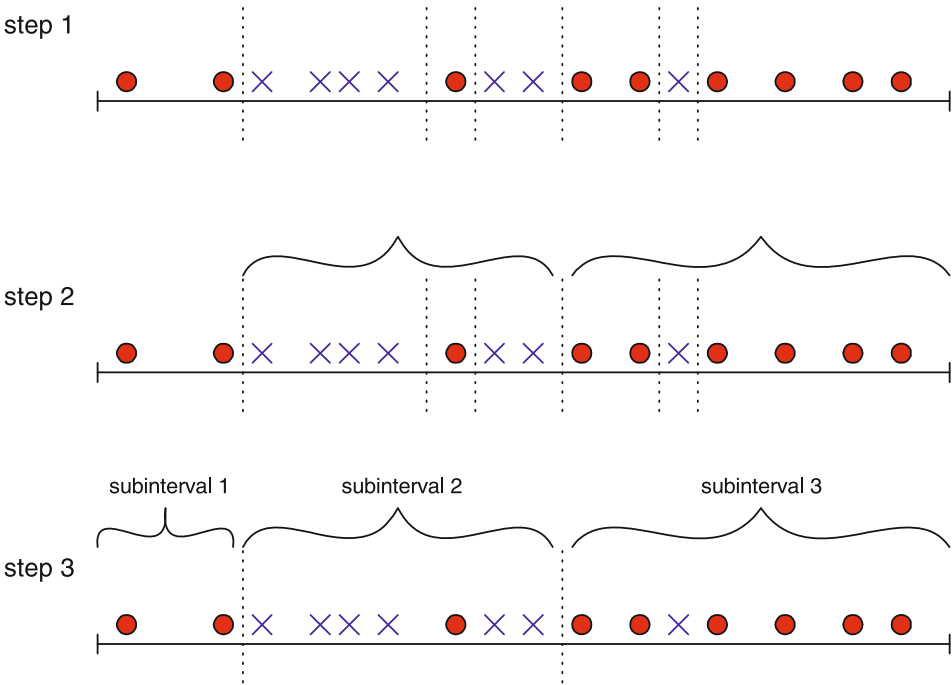
larger subintervals. In each of the following steps three subintervals are merged together: two “stronger” subintervals with one “weak” interval, where the “weak” interval lies between those two “strong” subintervals. Here strong and weak applies to the number of training objects in the subinterval tree (Fig. 5). In comparison to the previous two approaches the dynamic discretization returns more natural subintervals, which result in better and smaller DT classifiers.

In general we differentiate between two types of dynamic discretization:

- General dynamic discretization, and
- Nodal dynamic discretization.

General dynamic discretization uses all available training objects for the definition of subintervals. That is why the general dynamic discretization is performed before the start of building a DT. All the subintervals of all attributes are memorized in order to be used later in the process of building the DT.

Nodal dynamic discretization performs the definition of subintervals for all continuous attributes which are available in the current node of a DT. Only those training objects that came in the current node are used for setting the subintervals of the continuous attributes.



Decision Trees, Figure 5
Dynamic discretization of continuous attribute, which has values between 60 and 100. Shapes represent different attribute's values

In a series of tests the authors showed that nodal dynamic discretization produces smaller DT with higher accuracy than DT built with general dynamic discretization [56,58]. Nodal dynamic discretization also outperforms classical discretization techniques in the majority of cases.

Oblique Partitioning of Search Space

The so-far presented algorithms are using univariate partitioning methods, which are attractive because they are straightforward to implement (since only one feature is analyzed at a time) and the derived DT is relatively easy to understand. Beside univariate partitioning methods there are also some successful partitioning methods that do not partition the search space axis-parallel based on only one attribute at a time but are forming oblique partition boundaries based on a combination of attributes (Fig. 6).

Oblique partitioning provides a viable alternative to univariate methods. Unlike their univariate counterparts, oblique partitions are formed by combinations of attributes. The general form of an oblique partition is given by

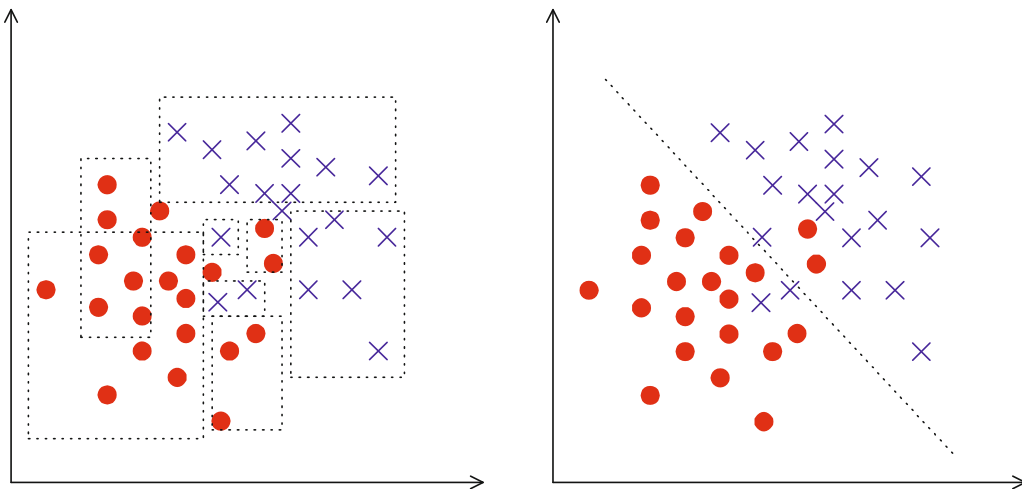
$$\sum_{i=1}^d \beta_i x_i \leq C$$

where β_i represents the coefficient of the i th attribute. Because of their multivariate nature, oblique methods offer far more flexibility in partitioning the search space; this flexibility comes at a price of higher complexity, however. Consider that given a data set containing n objects

described with d attributes, there can be $2 \cdot \sum_{i=0}^d \binom{n-1}{i}$ oblique splits if $n > d$ [50]; each split is a hyper-plane that divides the search space into two non-overlapping halves. For univariate splits, the number of potential partitions is much lower, but still significant, $n \cdot d$ [29]. In short, finding the right oblique partition is a difficult task.

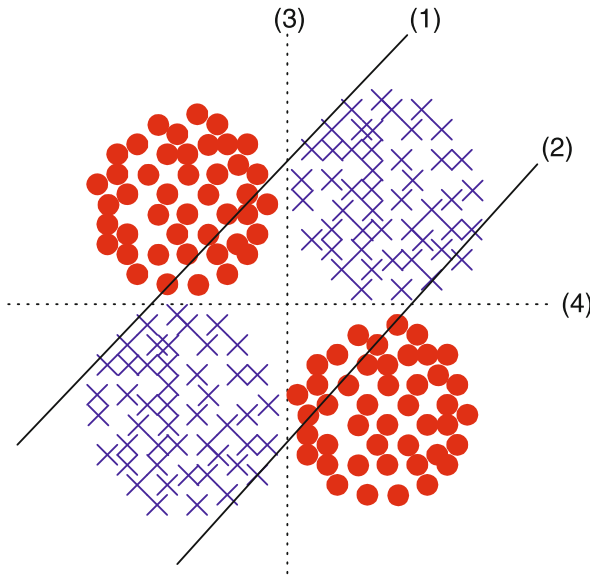
Given the size of the search space, choosing the right search method is of critical importance in finding good partitions. Perhaps the most comprehensive reference on this subject is [5] on classification and regression trees (CART). Globally, CART uses the same basic algorithm as Quinlan in C4.5. At the decision node level, however, the algorithm becomes extremely complex. CART starts out with the best univariate split. It then iteratively searches for perturbations in attribute values (one attribute at a time) which maximize some goodness metric. At the end of the procedure, the best oblique and axis-parallel splits found are compared and the better of these is selected.

Although CART provides a powerful and efficient solution to a very difficult problem, it is not without its disadvantages. Because the algorithm is fully deterministic, it has no inherent mechanism for escaping from local optima. As a result, CART has a tendency to terminate its partition search at a given node too early. The most fundamental disadvantage of CART (and of the traditional approach of inducing DT in general) is that the DT induction process can cause the metrics to produce misleading results. Because traditional DT induction algorithms choose what is locally optimal for each decision node, they inevitably ignore splits that score poorly alone, but yield better solution when used in combination. This problem is illustrated by Fig. 7. The solid lines indicate the splits



Decision Trees, Figure 6

Given axes that show the attribute values and shape corresponding to class labels: i axis-parallel and ii oblique decision boundaries



Decision Trees, Figure 7

CART generated splits (solid lines – 1,2) minimize impurity at each decision node in a way that is not necessarily optimal regarding the natural structure of data (denoted by dotted lines – 3,4)

found by CART. Although each split optimizes the impurity metric, the end product clearly does not reflect the best possible partitions (indicated by the dotted lines). However, when evaluated as individuals, the dotted lines register high impurities and are therefore not chosen. Given this, it is apparent that the sequential nature of DT can prevent the induction of trees that reflect the natural structure of the data.

Pruning Decision Trees

Most of the real-world data contains at least some amount of noise and outliers. Different machine-learning approaches elaborate different levels of sensitivity to noise in data. DT fall into the category of methods, sensitive to noise and outliers. In order to cure that, some additional methods must be applied to DT in order to reduce the complexity of the DT and increase its accuracy. In DTs we call this approach pruning.

Though pruning can be applied in different stages, it follows the basic idea, also called the Ockham's razor. William of Ockham (1285–1349), one of the most influential medieval philosophers has been given credit for the rule, which said that “one should not draw more conclusions about a certain matter than minimum necessary and that the redundant conclusions should be removed or shaved off”. That rule was a basis for latter interpretation

in the form of the following statement: “If two different solutions solve the same problem with the same accuracy, then the better of the two is the shorter solution.”

Mapping this rule to DT pruning allows us to prune or shave away those branches in the DT, which do not decrease the classification accuracy. By doing so, we end up with a DT, which is not only smaller, but also has higher (or in the worse case scenario at least the same) accuracy as the original DT.

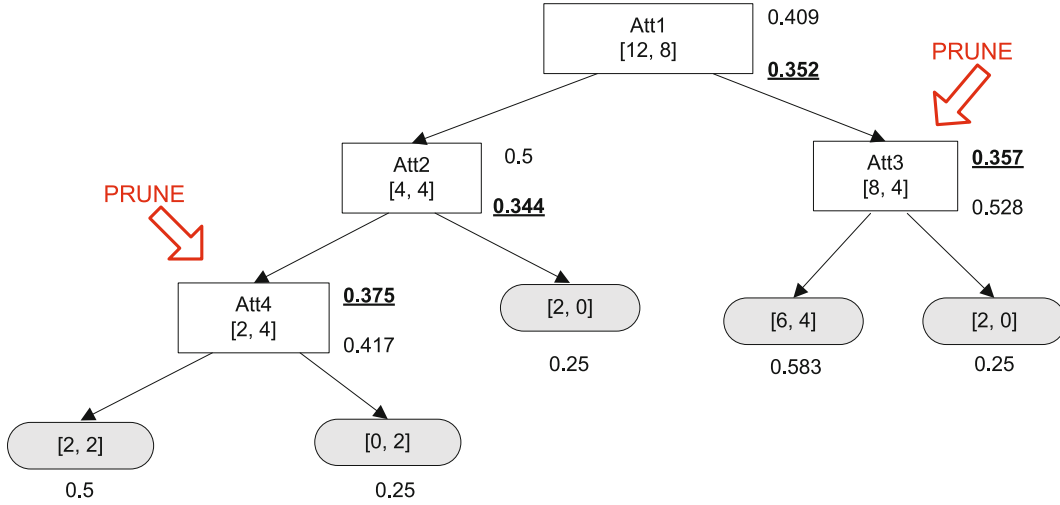
DT classifiers aim to refine the training sample T into subsets which have only a single class. However, training samples may not be representative of the population they are intended to represent. In most cases, fitting a DT until all leafs contain data for a single class causes over-fitting. That is, the DT is designed to classify the training sample rather than the overall population and accuracy on the overall population will be much lower than the accuracy on the training sample.

For this purpose most of the DT induction algorithms (C4.5, CART, OC1) use pruning. They all grow trees to maximum size, where each leaf contains single-class data or no test offers any improvement on the mix of classes at that leaf, and then prune the tree to avoid over-fitting. Pruning occurs within C4.5 when the predicted error rate is reduced by replacing a branch with a leaf. CART and OC1 use a proportion of the training sample to prune the tree. The tree is trained on the remainder of the training sample and then pruned until the accuracy on the pruning sample can not be further improved.

In general we differentiate between two pruning approaches: prepruning, which takes place during DT induction and postpruning, applied to already-induced DT in order to reduce complexity.

We will describe three representatives of pruning – one prepruning and two postpruning examples.

Prepruning Prepruning (called also stopping criteria) is a very simple procedure, performed during the first phase of induction. Early stopping of DT construction is based on criteria, which measures the percentage of the dominant class in the training subset in the node. If that percentage is higher than the preset threshold, then the internal node is transformed to a leaf and marked with the dominant class label. Such early stopping of DT construction reduces the size and complexity of DT, reduces the possibility of over-fitting, making it more general. However, there is also danger of oversimplification when the preset threshold is too low. This problem is especially present in training sets where frequencies of class labels are not balanced. An improper threshold can cause some special training objects, which are very important for solv-



Decision Trees, Figure 8

Unpruned decision tree with static error estimates

ing the classification problem, to be discarded during the prepruning process.

Postpruning The opposite to prepruning, postpruning operates on DT that is already constructed. Using simple frequencies or error estimation, postpruning approaches calculate whether or not substituting a subtree with a leaf (pruning or shaving-off according to Ockham's razor) would increase the DT's accuracy or at least reduce its size without negatively effecting classification accuracy.

Both the following approaches are practically the same up to a point, where we estimate error in a node or subtree.

Static Error Estimate Pruning Static error estimate pruning uses frequencies of training objects' class labels in nodes to estimate error, caused by replacing an internal node with a leaf. If the error estimate in the node is lower or equal to the error computed in the subtree, then we can prune the subtree in question.

Let C be a set of training objects in node V , and m the number of all possible class labels. N is the number of training objects in C and O is the most frequent class label in C . Let n be the number of training objects in C that belong to class O . The static error estimate (also called Laplace estimate) in node V is $E(V)$:

$$E(V) = \frac{N - n + m - 1}{N + m}.$$

Let us compute the subtree error of V using the next expression

$$\text{SubtreeError}(V) = \sum_{i=1}^{\text{NoOfChildNodes}} P_i * \text{Err}(V_i).$$

Since we do not have information about class distribution, we use relative frequencies in the place of probabilities P_i . $\text{Err}(V_i)$ is the error estimate of child node V_i .

Now we can compute the new error estimate $\text{Err}(V)$ in node V by using the next expression:

$$\text{Err}(V) = \min(E(V), \text{SubtreeError}(V)).$$

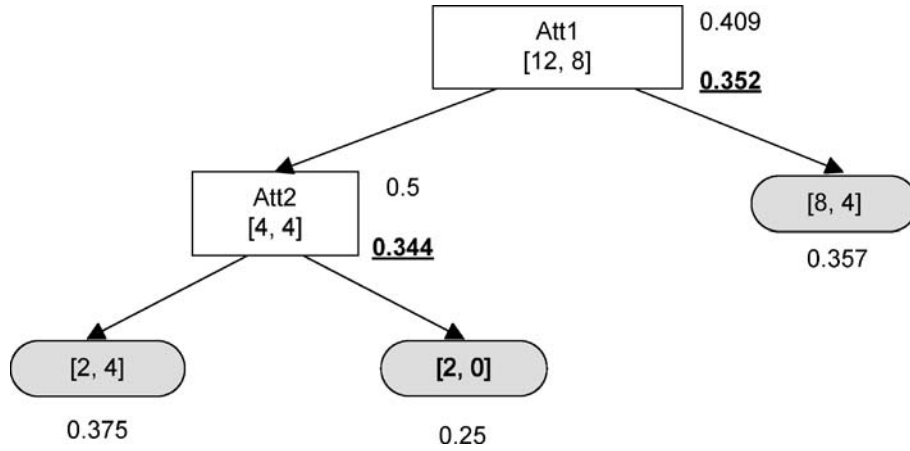
If $E(V)$ is less than or equal to $\text{SubtreeError}(V)$, then we can prune the subtree of V , replace it by a leaf, label it class O and assign the value of $E(V)$ to $\text{Err}(V)$. In the opposite case we only assign the value of $\text{SubtreeError}(V)$ to $\text{Err}(V)$.

Static error estimate pruning is a recursive bottom-up approach, where we start to calculate error estimates in leafs and perform error propagation and eventual pruning actions toward parent nodes until the root node is reached.

In Fig. 8 we have an unpruned DT with two possible decision classes, O_1 and O_2 . Inside each node there are frequencies of training objects according to their class label, written in brackets (first for O_1 and second for O_2). Beside each internal node (squares) there are two numbers. The top one is static error estimate in the node and the bottom one is error estimate in the subtree. The number below each leaf (ellipses) is its error estimate.

We can see that static error estimates are lower than subtree errors for two nodes: Att3 and Att4. Therefore, we prune those two subtrees and we get a smaller but more accurate DT (Fig. 9).

Reduced Error Pruning Reduced error pruning implements a very similar principle as static error estimate pruning. The difference is in error estimation. Reduced error



Decision Trees, Figure 9

Pruned decision tree with static error estimates

pruning uses the pruning set to estimate errors in nodes and leafs. The pruning set is very similar to training set, since it contains objects, described with attributes and class label. In the first stage we try to classify each pruning object. When a leaf is reached, we compare DT's classification with actual class label and mark correct/incorrect decision in the leaf. At the end there is a number of errors and number of correct classifications recorded in each leaf.

In a recursive bottom-up approach, we start to calculate error estimates in leafs, perform error propagation and eventual pruning actions toward parent nodes until the root node is reached.

Let $\text{SubtreeError}(V)$ be a sum of errors from all subtrees of node V .

$$\text{SubtreeError}(V) = \begin{cases} V \in \text{Leafs} \Rightarrow \text{NoOfErrorsInLeaf} \\ V \notin \text{Leafs} \Rightarrow \sum_{i=1}^{\text{NoOfChildNodes}} \text{Err}(V_i) \end{cases}$$

Let C be a set of training objects in node V and O the most frequent class label in C . Let N be the number of all pruning objects that reached the node V and n be the number of pruning objects in V that belong to class O . The error estimate in node V is defined by $E(V)$:

$$E(V) = N - n.$$

Now we can compare the error estimate in the node with the error estimate in the subtree and compute a new error estimate $\text{Err}(V)$ in node V :

$$\text{Err}(V) = \min(E(V), \text{SubtreeError}(V)).$$

If $E(V)$ is less than or equal to $\text{SubtreeError}(V)$, then we can prune the subtree of V , replace it by a leaf labeled with class O and assign the value of $E(V)$ to $\text{Err}(V)$. In the opposite case we only assign the value of $\text{SubtreeError}(V)$ to $\text{Err}(V)$.

The drawback of this pruning approach is the fact that it is very hard to construct a pruning set that is capable of "visiting" each node in a DT at least once. If we cannot assure that, the pruning procedure will not be as efficient as other similar approaches.

Deficiencies of Classical Induction

The DT has been shown to be a powerful tool for decision support in different areas. The effectiveness and accuracy of classification of DTs have been a surprise for many experts and their greatest advantage is in simultaneous suggestion of a decision and the straightforward and intuitive explanation of how the decision was made. Nevertheless, the classical induction approach also contains several deficiencies.

One of the most obvious drawbacks of classical DT induction algorithms is poor processing of incomplete, noisy data. If some attribute value is missing, classical algorithms do not perform well on processing of such an object. For example, in Quinlan's algorithms before C4.5 such data objects with missing data were left out of the training set – this fact of course resulted in decreased quality of obtained solutions (in this way the training set size and the information about the problem's domain were reduced). Algorithm C4.5 introduced a technique to overcome this problem, but this is still not very effective. In real world problems, especially in medicine, missing data is very common – therefore, effective processing of such data is of vital importance.

The next important drawback of classical induction methods is the fact that they can produce only one DT classifier for a problem (when the same training set is used). In many real-world situations it would be of great

benefit if more DT classifiers would be available and a user could choose the most appropriate one for a single case. As it is possible for training objects to miss some attribute value, the same goes also for a testing object – there can be a new case where some data is missing and it is not possible to obtain it (unavailability of some medical equipment, for example, or invasiveness of a specific test for a patient). In this way another DT classifier could be chosen that does not include a specific attribute test to make a decision.

Let us mention only one more disadvantage of classical induction methods, namely the importance of different errors. Between all decisions possible there are usually some that are more important than the others. Therefore, a goal is to build a DT in such a way that the accuracy of classification for those most important decisions is maximized. Once again, this problem is not solved adequately in classical DT induction methods.

Variant Methods

Although algorithms such as ID3, C4.5, and CART make up the foundation of traditional DT induction practice, there is always room for improvement of the accuracy, size, and generalization ability of the generated trees. As would be expected, many researchers have tried to build on the success of these techniques by developing better variations of them.

Alternatively, a lot of different methods for the induction of DT have been introduced by various researchers, which try to overcome the deficiencies noted above. Most of them are based on so-called soft methods, like evolutionary techniques or neural networks, sometimes several methods are combined in a hybrid algorithm.

A vast number of techniques have also been developed that help to improve only a part of the DT induction process. Such techniques include evolutionary algorithms for optimizing split functions in attribute nodes, dynamic discretization of attributes, dynamic subset selection of training objects, etc.

Split Selection Using Random Search Since random search techniques have proven extremely useful in finding solutions to non-deterministic polynomial complete (NP-complete) problems [30], naturally they have been applied to DT induction. Heath [20,21] developed a DT induction algorithm called SADT that uses a simulated annealing process to find oblique splits at each decision node. Simulated annealing is a variation of hill climbing which, at the beginning of the process, allows some random downhill moves to be made [42]. As a computational process,

simulated annealing is patterned after the physical process of annealing [25], in which metals are melted (at high temperatures) and then gradually cool until some solid state is reached.

Starting with an initial hyper-plane, SADT randomly perturbs the current solution and determines the goodness of the split by measuring the change in impurity (ΔE). If ΔE is negative (i. e. impurity decreases), the new hyper-plane becomes the current solution; otherwise, the new hyper-plane becomes the current split with probability $e^{-(\Delta E/T)}$ where T is the temperature of the system. Because simulated annealing mimics the cooling of metal, its initially high temperature falls with each perturbation. At the start of the process, the probability of replacing the current hyper-plane is nearly 1. As the temperature cools, it becomes increasingly unlikely that worse solutions are accepted. When processing a given data set, Heath typically grows hundreds of trees, performing several thousands perturbations per decision node. Thus, while SADT has been shown to find smaller trees than CART, it is very expensive from a computational standpoint [21].

A more elegant variation on Heath's approach is the OC1 system [29]. Like SADT, OC1 uses random search to find the best split at each decision node. The key difference is that OC1 rejects the brute force approach of SADT, using random search only to improve on an existing solution. In particular, it first finds a good split using a CART-like deterministic search routine. OC1 then randomly perturbs this hyper-plane in order to decrease its impurity. This step is a way of escaping the local optima in which deterministic search techniques can be trapped. If the perturbation results in a better split, OC1 resumes the deterministic search on the new hyper-plane; if not, it re-perturbs the partition a user-selectable number of times. When the current solution can be improved no further, it is stored for later reference. This procedure is repeated a fixed number of times (using a different initial hyper-plane in each trial). When all trials have been completed, the best split found is incorporated into the decision node.

Incremental Decision Tree Induction The DT induction algorithms discussed so far grow trees from a complete training set. For serial learning tasks, however, training instances may arrive in a stream over a given time period. In these situations, it may be necessary to continually update the tree in response to the newly acquired data. Rather than building a new DT classifier from scratch, the incremental DT induction approach revises the existing tree to be consistent with each new training instance. Utgoff implemented an incremental version of ID3, called

ID5R [53]. ID5R uses an *E-Score* criteria to estimate the amount of ambiguity in classifying instances that would result from placing a given attribute as a test in a decision node. Whenever the addition of new training instances does not fit the existing tree, the tree is recursively restructured such that attributes with the lowest E-Scores are moved higher in the tree hierarchy. In general, Utgoff's algorithm yields smaller trees compared to methods like ID3, which batch process all training data. Techniques similar to ID5R include an incremental version of CART [8]. Incremental DT induction techniques result in frequent tree restructuring when the amount of training data is small, with the tree structure maturing as the data pool becomes larger.

Decision Forests Regardless of the DT induction method utilized, subtle differences in the composition of the training set can produce significant variances in classification accuracy. This problem is especially acute when cross-validating small data sets with high dimensionality [11]. Researchers have reduced these high levels of variance by using decision forests, composed of multiple trees (rather than just one). Each tree in a forest is unique because it is grown from a different subset of the same data set. For example, Quinlan's windowing technique [41] induces multiple trees, each from a randomly selected subset of the training data (i.e., a window). Another approach was devised by Ho [22], who based each tree on a unique feature subset. Once a forest exists, the results from each tree must be combined to classify a given data instance. Such committee-type schemes for accomplishing this range from using majority rules voting [21,37] to different statistical methods [46].

A Combination of Decision Trees and Neural Networks

When DT and neural networks are compared, one can see that their advantages and drawbacks are almost complementary. For instance knowledge representation of DT is easily understood by humans, which is not the case for neural networks; DT have trouble dealing with noise in training data, which is again not the case for neural networks; DT learn very fast and neural networks learn relatively slow, etc. Therefore, the idea is to combine DT and neural networks in order to combine their advantages. In this manner, different hybrid approaches from both fields have been introduced [7,54,56].

Zorman in his MtDecit 2.0 approach first builds a DT that is then used to initialize a neural network [56,57]. Such a network is then trained using the same training objects as the DT. After that the neural network is again converted to a DT that is better than the original DT [2]. The

source DT classifier is converted to a disjunctive normal form – a set of normalized rules. Then the disjunctive normal form serves as the source for determining the neural network's topology and weights. The neural network has two hidden layers, the number of neurons on each hidden layer depends on rules in the disjunctive normal form. The number of neurons in the output layer depends on how many outcomes are possible in the training set. After the transformation, the neural networks is trained using back-propagation. The mean square error of such a network converges toward 0 much faster than it would in the case of randomly set weights in the network. Finally, the trained neural network has to be converted into a final DT. The neural network is examined in order to determine the most important attributes that influence the outcomes of the neural network. A list containing the most important attributes is then used to build the final DT that in the majority of cases performs better than the source DT. The last conversion usually causes a loss of some knowledge contained in the neural network, but even so most knowledge is transformed into the final DT. If the approach is successful, then the final DT has better classification capabilities than the source DT.

Using Evolutionary Algorithms to Build Decision Trees

Evolutionary algorithms are generally used for very complex optimization tasks [18], for which no efficient heuristic method exist. Construction of DT is a complex task, but heuristic methods exist that usually work efficiently and reliably. Nevertheless, there are some reasons justifying an evolutionary approach. Because of the robustness of evolutionary techniques they can be successfully used also on incomplete, noisy data (which often happens in real life data because of measurement errors, unavailability of proper instruments, risk to patients, etc.). Because of evolutionary principles used to evolve solutions, solutions can be found which can be easily overlooked otherwise. Also the possibility of optimizing the DT classifier's topology and the adaptation of split thresholds is an advantage.

There have been several attempts to build a DT with the use of evolutionary techniques [6,31,35], one the most recent and also very successful in various applications is genTrees, developed by Podgorelec and Kokol [35,36,37]. First an initial population of (about one hundred) semi-random DT is seeded. A random DT classifier is built by randomly choosing attributes and defining a split function. When a pre-selected number of test nodes is placed into a growing tree then a tree is finalized with decision nodes, which are defined by choosing the most fit decision class in each node regarding the training set. After the initialization the population of decision trees evolves through

many generations of selection, crossover and mutation in order to optimize the fitness function that determines the quality of the evolved DT.

According to the fitness function the best trees (the most fit ones) have the lowest function values – the aim of the evolutionary process is to minimize the value of local fitness function (LFF) for the best tree. With the combination of selection that prioritizes better solutions, crossover that works as a constructive operator towards local optimums, and mutation that works as a destructive operator in order to keep the needed genetic diversity, the searching for the solution tends to be directed toward the global optimal solution. The global optimal solution is the most appropriate DT regarding the specific needs (expressed in the form of fitness function). As the evolution repeats, more qualitative solutions are obtained regarding the chosen fitness function.

One step further from Podgorelec's approach was introduced by Sprogar with his evolutionary vector decision trees – VEDEC [49]. In his approach the evolution of DT is similar to the one by Podgorelec, but the functionality of DT is enhanced in the way that not only one possible decision class is predicted in the leaf, but several possible questions are answered with a vector of decisions. In this manner, for example a possible treatment for a patient is suggested together with the diagnosis, or several diagnoses are suggested at the same time, which is not possible in ordinary DT.

One of the most recent approaches to the evolutionary induction of DT-like methods is the AREX algorithm developed by Podgorelec and Kokol [34]. In this approach DT are extended by so-called decision programs that are evolved with the help of automatic programming. In this way a classical attribute test can be replaced by a simple computer program, which greatly improves the flexibility, at the cost of computational resources, however. The approach introduces a multi-level classification model, based on the difference between objects. In this manner "simple" objects are classified on the first level with simple rules, more "complex" objects are classified on the second level with more complex rules, etc.

Evolutionary Supported Decision Trees Even though DT induction isn't as complicated from the parameter settings' point of view as some other machine-learning approaches, getting the best out of the approach still requires some time and skill. Manual setting of parameters takes more time than induction of the tree itself and even so, we have no guarantee that parameter settings are optimal. With MtDeciT3.1Gen Zorman presented [59] an implementation of evolutionary supported DT. This approach

did not change the induction process itself, since the basic principle remained the same. The change was in evolutionary support, which is in charge of searching the space of induction parameters for the best possible combination. By using genetic operators like selection, crossover and mutation evolutionary approaches are often capable of finding optimal solutions even in the most complex of search spaces or at least they offer significant benefits over other search and optimization techniques.

Enhanced with the evolutionary option, DT don't offer us more generalizing power, but we gain on better coverage of method parameter's search space and significant decrease of time spent compared to manual parameter setting. Better exploitation of a method usually manifests in better results, and the goal is reached.

The rule of the evolutionary algorithm is to find a combination of parameters settings, which would enable induction of DT with high overall accuracy, high class accuracies, use as less attributes as possible, and be as small as possible.

Ensemble Methods An extension of individual classifier models are ensemble methods that use a set of induced classifiers and combine their outputs into a single classification. The purpose of combining several individual classifiers together is to achieve better classification results.

All kinds of individual classifiers can be used to construct an ensemble, however DT are used often as one of the most appropriate and effective models. DT are known to be very sensitive to changes in a learning set. Consequently, even small changes in a learning set can result in a very different DT classifier which is a basis for the construction of efficient ensembles. Two well-known representatives of ensemble methods are Random forests and Rotation forests, which use DT classifiers, although other classifiers (especially in the case of Rotation forests) could be used as well.

Dealing with Missing Values

Like most of the machine-learning approaches, the DT approach also assumes that all data is available at the time of induction and usage of DT. Real data bases rarely fit this assumption, so dealing with incomplete data presents an important task. In this section some aspects of dealing with missing values for description attributes will be given.

Reasons for incomplete data can be different: from cases where missing values could not be measured, human errors when writing measurement results, to usage of data bases, which were not gathered for the purpose of machine learning, etc.

Missing values can cause problems in three different stages of induction and usage of DT:

- Split selection in the phase of DT induction,
- Partition of training set in a test node during the phase of DT induction, and
- Selecting which edge to follow from a test node when classifying unseen cases during the phase of DT usage.

Though some similarities can be found between approaches for dealing with missing data for listed cases, some specific solutions exist, which are applicable only to one of the possible situations.

Let us first list possible approaches for the case of dealing with missing values during split selection in the phase of DT induction:

- Ignore – the training object is ignored during evaluation of the attribute with the missing value,
- Reduce – the evaluation (for instance entropy, gain or gain ratio) of the attribute with missing values is changed according to the percentage of training objects with missing values,
- Substitute with DT – a new DT with a reduced number of attributes is induced in order to predict the missing value for the training object in question; this approach is best suited for discrete attributes,
- Replace – the missing value is replaced with mean (continuous attributes) or most common (discrete attributes) value, and
- New value – missing values for discrete attributes are substituted with new value (“unknown”) which is then treated the same way as other possible values of the attribute.

The next seven approaches are suited for partitioning of the training set in a test node during the phase of DT induction:

- Ignore – the training object is ignored during partition of the test set,
- Substitute with DT – a new DT with a reduced number of attributes is induced in order to predict the missing value for the training object in question – according to the predicted value we can propagate the training object to one of the successor nodes; this approach is best suited for discrete attributes,
- Replace – the missing value is replaced with the mean (continuous attributes) or most common (discrete attributes) value,
- Probability – the training object with the missing value is propagated to one of the edges according to prob-

ability, proportional to strength of subsets of training objects,

- Fraction – a fraction of the training object with a missing value is propagated to each of the subsets; the size of the fraction is proportional to the strength of each subset of training objects,
- All – the training object with the missing value is propagated to all subsets, coming from the current node, and
- New value – training objects with new value (“unknown”) are assigned to a new subset and propagated further, following a new edge, marked with the previously mentioned new value.

The last case for dealing with missing attribute values is during the decision-making phase, when we are selecting which edge to follow from a test node when classifying unseen cases:

- Substitute with DT – a new DT with a reduced number of attributes is induced in order to predict the missing value for the object in question – according to predicted value we can propagate the unseen object to one of the successor nodes; this approach is best suited for discrete attributes,
- Replace – the missing value is replaced with the mean (continuous attributes) or most common (discrete attributes) value,
- All – the unseen object with a missing value is propagated to all subsets, coming from the current node; the final decision is subject to voting by all leaf nodes we ended in,
- Stop – when facing a missing value, the decision-making process is stopped and the decision with the highest probability according to the training subset in the current node is returned, and
- New value – if there exists a special edge marked with value “unknown”, we follow that edge to the next node.

Evaluation of Quality

Evaluation of an induced DT is essential for establishing the quality of the learned model. For the evaluation it is essential to use a set of unseen instances – a testing set. A testing set is a set of testing objects which have not been used in the process of inducing the DT and can therefore be used to objectively evaluate its quality.

A quality is a complex term that can include several measures of efficiency. For the evaluation of the DT's quality the most general measure is the overall accuracy, defined as a percentage of correctly classified objects from all objects (correctly classified and not correctly classified).

Accuracy can be thus calculated as:

$$ACC = \frac{T}{T + F}$$

where T stands for “true” cases (i. e. correctly classified objects) and F stands for “false” cases (i. e. not correctly classified objects).

The above measure is used to determine the overall classification accuracy. In many cases, the accuracy of each specific decision class is even more important than the overall accuracy. The separate class accuracy of i th single decision class is calculated as:

$$ACC_{k,i} = \frac{T_i}{T_i + F_i}$$

and the average accuracy over all decision classes is calculated as

$$ACC_k = \frac{1}{M} \cdot \sum_{i=1}^M \frac{T_i}{T_i + F_i}$$

where M represents the number of decision classes.

There are many situations in real-world data where there are exactly two decision classes possible (i. e. positive and negative cases). In such a case the common measures are sensitivity and specificity:

$$\text{Sens} = \frac{TP}{TP + FN}, \quad \text{Spec} = \frac{TN}{TN + FP}$$

where TP stands for “true positives” (i. e. correctly classified positive cases), TN stands for “true negatives” (i. e. correctly classified negative cases), FP stands for “false positives” (i. e. not correctly classified positive cases), and FN stands for “false negatives” (i. e. not correctly classified negative cases).

Besides accuracy, various other measures of efficiency can be used to determine the quality of a DT. Which measures are used depends greatly on the used dataset and the domain of a problem. They include complexity (i. e. the size and structure of an induced DT), generality (differences in accuracy between training objects and testing objects), a number of used attributes, etc.

Applications and Available Software

DT are one of the most popular data-mining models because they are easy to induce, understand, and interpret. High popularity means also many available commercial and free applications. Let us mention a few.

We have to start with Quinlan’s C4.5 and C5.0/See5 (MS Windows/XP/Vista version of C5.0), the classic version of DT tool. It is commercial and available at <http://www.rulequest.com/see5-info.html>.

Another popular commercial tool is CART 5, based on original regression trees algorithm. It is implemented in an

intuitive Windows-based environment. It is commercially available at <http://www.salford-systems.com/1112.php>.

Weka (Waikato Environment for Knowledge Analysis) environment contains a collection of visualization tools and algorithms for data analysis and predictive modeling. Among them there are also some variants of DTs. Beside its graphical user interfaces for easy access, one of its great advantages is the possibility of expanding the tool with own implementations of data analysis algorithms in Java. A free version is available at <http://www.cs.waikato.ac.nz/ml/weka/>.

OC1 (Oblique Classifier 1) is a decision tree induction system written in C and designed for applications where the instances have numeric attribute values. OC1 builds DT that contain linear combinations of one or more attributes at each internal node; these trees then partition the space of examples with both oblique and axis-parallel hyperplanes. It is available free at <http://www.cs.jhu.edu/~salzberg/announce-oc1.html>.

There exist many other implementations, some are available as a part of larger commercial packages like SPSS, Matlab, etc.

Future Directions

DT has reached the stage of being one of the fundamental classification tools used on a daily basis, both in academia and in industry. As the complexity of stored data is growing and information systems become more and more sophisticated, the necessity for an efficient, reliable and interpretable intelligent method is growing rapidly. DT has been around for a long time and has matured and proved its quality. Although the present state of research regarding DT gives reasons to think there is nothing left to explore, we can expect major developments in some trends of DT research that are still open. It is our belief that DT will have a great impact on the development of hybrid intelligent systems in the near future.

Bibliography

Primary Literature

1. Babic SH, Kokol P, Stiglic MM (2000) Fuzzy decision trees in the support of breastfeeding. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000, Houston, pp 7–11
2. Banerjee A (1994) Initializing neural networks using decision trees In: Proceedings of the International Workshop on Computational Learning Theory and Natural learning Systems, Cambridge, pp 3–15
3. Bonner G (2001) Decision making for health care professionals: use of decision trees within the community mental health setting. J Adv Nursing 35:349–356

4. Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140
5. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
6. Cantu-Paz E, Kamath C (2000) Using evolutionary algorithms to induce oblique decision trees. In: Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2000, Las Vegas, pp 1053–1060
7. Craven MW, Shavlik JW (1996) Extracting tree-structured representations of trained networks. In: Advances in Neural Information Processing Systems, vol 8. MIT Press, Cambridge
8. Crawford S (1989) Extensions to the CART algorithm. *Int J Man-Mach Stud* 31(2):197–217
9. Cremilleux B, Robert C (1997) A theoretical framework for decision trees in uncertain domains: Application to medical data sets. In: Lecture Notes in Artificial Intelligence, vol 1211. Springer, London, pp 145–156
10. Dantchev N (1996) Therapeutic decision trees in psychiatry. *Encephale-Revue Psychiatr Clinique Biol Therap* 22(3):205–214
11. Dietterich TG, Kong EB (1995) Machine learning bias, statistical bias and statistical variance of decision tree algorithms. *Mach Learn, Corvallis*
12. Feigenbaum EA, Simon HA (1962) A theory of the serial position effect. *Br J Psychol* 53:307–320
13. Freund Y (1995) Boosting a weak learning algorithm by majority. *Inf Comput* 121:256–285
14. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Machine Learning: Proc. Thirteenth International Conference. Morgan Kaufman, San Francisco, pp 148–156
15. Gambhir SS (1999) Decision analysis in nuclear medicine. *J Nucl Med* 40(9):1570–1581
16. Gehrke J (2003) Decision Trees. In: Nong Y (ed) *The Handbook of Data Mining*. Lawrence Erlbaum, Mahwah
17. Goebel M, Gruenwald L (1999) A survey of data mining software tools. *SIGKDD Explor* 1(1):20–33
18. Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, Reading
19. Hand D (1997) Construction and assessment of classification rules. Wiley, Chichester
20. Heath D et al (1993) k-DT: A multi-tree learning method. In: Proceedings of the Second International Workshop on Multistrategy Learning, Harpers Ferry, pp 138–149
21. Heath D et al (1993) Learning Oblique Decision Trees. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence IJCAI-93, pp 1002–1007
22. Ho TK (1998) The Random Subspace Method for Constructing Decision Forests. *IEEE Trans Pattern Anal Mach Intell* 20(8):832–844
23. Hunt EB, Marin J, Stone PT (1966) Experiments in Induction. Academic Press, New York, pp 45–69
24. Jones JK (2001) The role of data mining technology in the identification of signals of possible adverse drug reactions: Value and limitations. *Curr Ther Res-Clin Exp* 62(9):664–672
25. Kilpatrick S et al (1983) Optimization by Simulated Annealing. *Science* 220(4598):671–680
26. Kokol P, Zorman M, Stiglic MM, Malcic I (1998) The limitations of decision trees and automatic learning in real world medical decision making. In: Proceedings of the 9th World Congress on Medical Informatics MEDINFO'98, 52, pp 529–533
27. Letourneau S, Jensen L (1998) Impact of a decision tree on chronic wound care. *J Wound Ostomy Conti Nurs* 25:240–247
28. Lim T-S, Loh W-Y, Shih Y-S (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach Learn* 48:203–228
29. Murthy KVS (1997) On Growing Better Decision Trees from Data, Ph D dissertation. Johns Hopkins University, Baltimore
30. Neapolitan R, Naimipour K (1996) Foundations of Algorithms. DC Heath, Lexington
31. Nikolaev N, Slavov V (1998) Inductive genetic programming with decision trees. *Intell Data Anal Int J* 2(1):31–44
32. Ohno-Machado L, Lacson R, Massad E (2000) Decision trees and fuzzy logic: A comparison of models for the selection of measles vaccination strategies in Brazil. Proceedings of AMIA Symposium 2000, Los Angeles, CA, US, pp 625–629
33. Paterson A, Niblett TB (1982) ACLS Manual. Intelligent Terminals, Edinburgh
34. Podgorelec V (2001) Intelligent systems design and knowledge discovery with automatic programming. Ph D thesis, University of Maribor
35. Podgorelec V, Kokol P (1999) Induction of medical decision trees with genetic algorithms. In: Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications CIMA. Academic Press, Rochester
36. Podgorelec V, Kokol P (2001) Towards more optimal medical diagnosing with evolutionary algorithms. *J Med Syst* 25(3): 195–219
37. Podgorelec V, Kokol P (2001) Evolutionary decision forests – decision making with multiple evolutionary constructed decision trees. In: Problems in Applied Mathematics and Computational Intelligence. WSES Press, pp 97–103
38. Quinlan JR (1979) Discovering rules by induction from large collections of examples. In: Michie D (ed) *Expert Systems in the Micro Electronic Age*, University Press, Edinburgh, pp 168–201
39. Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
40. Quinlan JR (1987) Simplifying decision trees. *Int J Man-Mach Stud* 27:221–234
41. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco
42. Rich E, Knight K (1991) Artificial Intelligence, 2nd edn. McGraw Hill, New York
43. Sanders GD, Hagerty CG, Sonnenberg FA, Hlatky MA, Owens DK (2000) Distributed decision support using a web-based interface: prevention of sudden cardiac death. *Med Decis Making* 19(2):157–166
44. Schapire RE (1990) The strength of weak learnability. *Mach Learn* 5:197–227
45. Shannon C, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Champaign
46. Shlien S (1992) Multiple binary decision tree classifiers. *Pattern Recognit Lett* 23(7):757–763
47. Sims CJ, Meyn L, Caruana R, Rao RB, Mitchell T, Krohn M (2000) Predicting cesarean delivery with decision tree models. *Am J Obstet Gynecol* 183:1198–1206
48. Smyth P, Goodman RM (1991) Rule induction using information theory. In: Piatsky-Scharpiro G, Frawley WJ (eds) *Knowledge Discovery in Databases*, AAAI Press, Cambridge, pp 159–176
49. Sprogar M, Kokol P, Hleb S, Podgorelec V, Zorman M (2000) Vector decision trees. *Intell Data Anal* 4(3–4):305–321
50. Tou JT, Gonzalez RC (1974) Pattern Recognition Principles. Addison-Wesley, Reading

51. Tsien CL, Fraser HSF, Long WJ, Kennedy RL (1998) Using classification tree and logistic regression methods to diagnose myocardial infarction. In: Proceedings of the 9th World Congress on Medical Informatics MEDINFO'98, 52, pp 493–497
52. Tsien CL, Kohane IS, McIntosh N (2000) Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. *Artif Intell Med* 19(3):189–202
53. Utgoff PE (1989) Incremental induction of decision trees. *Mach Learn* 4(2):161–186
54. Utgoff PE (1989) Perceptron trees: a case study in hybrid concept representations. *Connect Sci* 1:377–391
55. White AP, Liu WZ (1994) Bias in information-based measures in decisions tree induction. *Mach Learn* 15:321–329
56. Zorman M, Hleb S, Sprogar M (1999) Advanced tool for building decision trees MtDecit 2.0. In: Arabnia HR (ed) Proceedings of the International Conference on Artificial Intelligence ICAI-99. Las Vegas
57. Zorman M, Kokol P, Podgorelec V (2000) Medical decision making supported by hybrid decision trees. In: Proceedings of the ICSC Symposia on Intelligent Systems & Applications ISA'2000, ICSC Academic Press, Wollongong
58. Zorman M, Podgorelec V, Kokol P, Peterson M, Lane J (2000) Decision tree's induction strategies evaluated on a hard real world problem. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000, Houston, pp 19–24
59. Zorman M, Sigut JF, de la Rosa SJL, Alayón S, Kokol P, Verliè M (2006) Evolutionary built decision trees for supervised segmentation of follicular lymphoma images. In: Proceedings of the 9th IASTED International conference on Intelligent systems and control, Honolulu, pp 182–187

Books and Reviews

- Breiman L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
- Han J, Kamber M (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco
- Hand D, Manilla H, Smyth P (2001) Principles of Data Mining. MIT Press, Cambridge
- Kantardzic M (2003) Data Mining: Concepts, Models, Methods, and Algorithms. Wiley, San Francisco
- Mitchell TM (1997) Machine Learning. McGraw-Hill, New York
- Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco
- Ye N (ed) (2003) The Handbook of Data Mining. Lawrence Erlbaum, Mahwah

Delay and Disruption in Complex Projects

SUSAN HOWICK¹, FRAN ACKERMANN¹, COLIN EDEN¹, TERRY WILLIAMS²

¹ Strathclyde Business School, University of Strathclyde, Glasgow, UK

² School of Management, Southampton University, Southampton, UK

Article Outline

Glossary

Definition of the Subject

Introduction

Disruption and Delay

Analyzing D&D and Project Behavior

Cascade Model Building Process

Implications for Development

Future Directions

Bibliography

Glossary

Cause map A cause map is similar to a cognitive map however it is not composed of an individuals perception but rather the views/statements from a number of participants. It follows the same formalisms as cognitive mapping but does not reflect cognition as it is composite.

Cognitive map A cognitive map is a representation of an individuals perception (cognition) of an issue. It is graphically depicted illustrating concepts/statements connected together with arrows representing causality. They are created using a set of established formalisms.

Complex project A complex project is a project in which the project behaviors and outcomes are difficult to predict and difficult to explain post-hoc.

Disruption and delay Disruption and delay (D&D) is primarily the consequence of interactions which feed on themselves as a result of an initial disruption or delay or portfolio of disruptions and delays.

Project A project is a temporary endeavor undertaken to create a unique product or service [1].

Definition of the Subject

There are many examples of complex projects suffering massive time and cost overruns. If a project has suffered such an overrun there may be a need to understand why it behaved the way it did. Two main reasons for this is (i) to gain learning for future projects or (ii) because one party of the project wishes to claim compensation from another party and thus is trying to explain what occurred during the project. In the latter case, system dynamics has been used for the last 30 years to help to understand why projects behave the way they do. Its success in this arena stems from its ability to model and unravel complex dynamic behavior that can result in project overruns. Starting from the first use of system dynamics in a claim situation in the late 1970's [2], it has directly influenced claim results worth millions of dollars. However, the number

of claims which system dynamics has been involved in is still small as it is not perceived by project management practitioners as a standard tool for analyzing projects. System dynamics has a lot to offer in understanding complex projects, not only in a post-mortem situation, but it could also add value in the pre-project analysis stage and during the operational stage of a project.

Introduction

In this chapter we discuss the role of system dynamics (SD) modeling in understanding, and planning, a complex project. In particular we are interested in understanding how and why projects can go awry in a manner that seems surprising and often very difficult to unravel.

When we refer to projects we mean “a temporary endeavor undertaken to create a unique product or service” [1]. Projects are a specific undertaking, which implies that they are “one-shot”, non-repetitive, time-limited, and, when complex, frequently bring about revolutionary (rather than evolutionary) improvements, start (to some extent) without precedent, and are risky with respect to customer, product, and project. If physical products are being created in a project, then the product is in some way significantly different to previous occasions of manufacturing (for example, in its engineering principles, or the expected operating conditions of the product, etc.), and it is this feature that means there is a need to take a project orientation.

Complex projects often suffer massive cost overruns. In recent decades those that have been publicized relate to large public construction projects, for example airports, bridges, and public buildings. Some examples include Denver’s US\$5 billion airport that was 200% overspent [3], the 800 million Danish Kroner Oresund bridge that was 68% overspent [4], and the UK’s Scottish Parliament, which was 10 times the first budget [5]. The Major Projects Association [6] talks of a calamitous history of cost overruns of very large projects in the public sector. Flyvberg et al., [7] describe 258 major transportation infrastructure projects showing 90% of projects overspent. Morris and Hough [8] conclude that “the track record of projects is fundamentally poor, particularly for the larger and more difficult ones. ... Projects are often completed late or over budget, do not perform in the way expected, involve severe strain on participating institutions or are canceled prior to their completion after the expenditure of considerable sums of money.” (p.7).

“Complex” projects are ones in which the project behaviors and outcomes are difficult to predict and difficult to explain post-hoc. Complex projects, by their nature,

comprise multiple interdependencies, and involve nonlinear relationships (which are themselves dynamic). For example, choices to accelerate might involve the use of additional overtime which can affect both learning curves and productivity as a result of fatigue – each of which are non-linear relationships. In addition many of the important features of complex projects are manifested through ‘soft’ relationships – for example managers will recognize deteriorating morale as projects become messy and look a failure, but assessing the impact of morale on levels of mistakes and rate of working has to be a matter of qualitative judgment. These characteristics are amenable particularly to SD modeling which specializes in working with qualitative relationships that are non-linear [9,10,11].

It is therefore surprising that simulation modeling has not been used more extensively to construct post-mortem analyses of failed projects, and even more surprising because of SD’s aptitude for dealing with feedback. Nevertheless the authors have been involved in the analysis of 10 projects that have incurred time and cost overruns and PA Consulting Group have claimed to have used SD to explain time and cost overruns for over 30 litigation cases [12]. Although in the mid-1990’s, attempts to integrate SD modeling with more typical approaches to project management were emerging, their use has never become established within the project management literature or practice [13,14,15]. In addition, recognition that the trend towards tighter project delivery and accelerated development times meant that parallelism in project tasks was becoming endemic, and the impact of increasing parallelism could result in complex feedback dynamics where vicious cycles exist [16]. These vicious cycles are often the consequence of actions taken to enforce feedback control designed to bring a project back on track.

As project managers describe their experiences of projects going wrong they will often talk of these “vicious cycles” occurring, particularly with respect to the way in which customer changes seem to generate much more rework than might be expected, and that the rework itself then generates even more rework. Consider a small part of a manager’s description of what he sees going on around him:

“For some time now we’ve been short of some really important information the customer was supposed to provide us. As a consequence we’ve been forced to progress the contract by making engineering assumptions, which, I fear, have led to more mistakes being made than usual. This started giving us more rework than we’d planned for. But, of course, rework on some parts of the project has meant re-

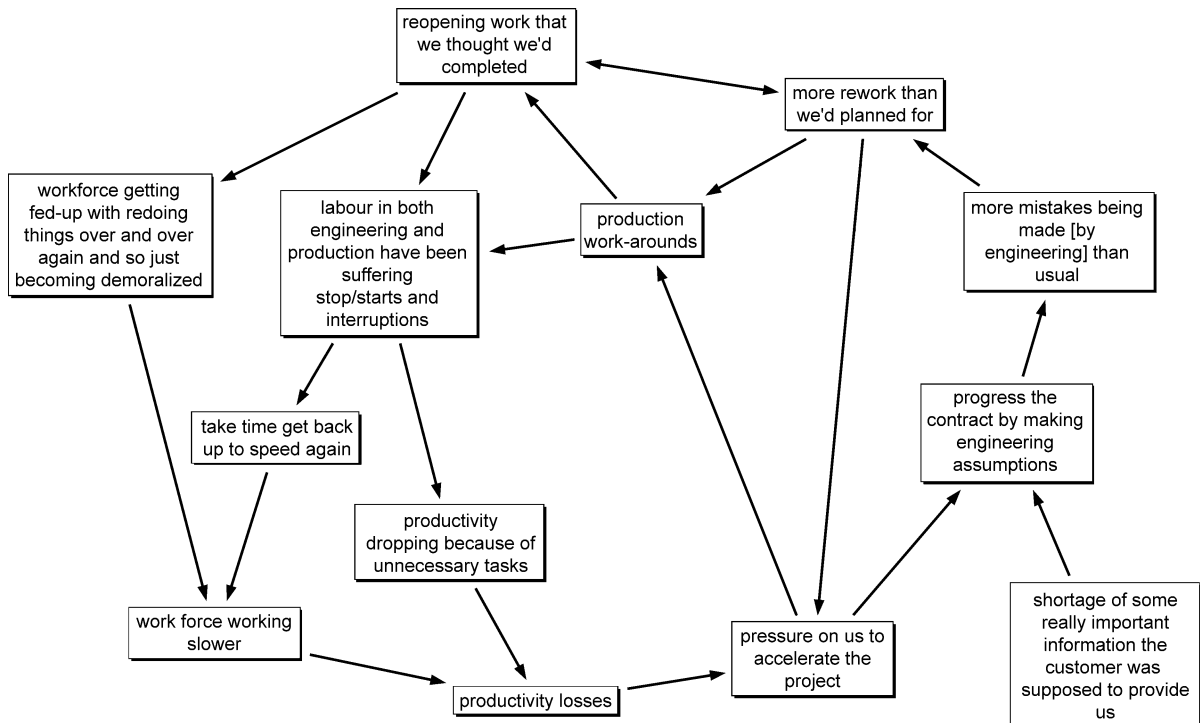
opening work that we thought we'd completed, and that, in turn has reopened even more past work. Engineering rework has led to the need for production work-arounds and so our labour in both engineering and production have been suffering stop/starts and interruptions – and each time this happens they take time to get back up to speed again. This has led to productivity dropping because of unnecessary tasks, let alone productivity losses from the workforce getting fed-up with redoing things over and over again and so just becoming demoralized and so working slower. Inevitably all the rework and consequential productivity losses have put pressure on us to accelerate the project forcing us to have to make more engineering assumptions and do work-arounds.”

Figure 1 shows a ‘cause map’ of the arguments presented by this project manager – the words used in the map are those used by the project manager and the arrows represent the causality described by the manager. This description is full of vicious cycles (indeed there are 35 vicious cycles discussed – see Fig. 1) all triggered by a shortage of customer furnished information and resulting in

the rework cycle [17,18,19,20,21] and the need to accelerate in order to keep the project on schedule. Using traditional project management models such as Critical Path Method/Network Analysis cannot capture any of the dynamics depicted in Fig. 1, but SD simulation modeling is absolutely appropriate [22].

So, why has SD modeling been so little used? Partly it is because in taking apart a failed project the purpose is usually associated with a contractor wishing to make a claim for cost-overruns. In these circumstances the traditions of successful claims and typical attitudes of courts tend to determine the approach used. A ‘measured-mile’ approach is common, where numerical simplicity replaces the need for a proper understanding [23].

It was not until the early 1980’s that the use of simulation modeling became apparent from publications in the public-domain. The settlement of a shipbuilding claim [2] prompted interest in SD modeling and [24], in the same year, reported on the use of management science modeling for the same purpose. It was not surprising that this modeling for litigation generated interest in modeling where the purpose was oriented to learning about failed projects (indeed the learning can follow from litigation modeling [25], although it rarely does).



Delay and Disruption in Complex Projects, Figure 1

Cause map showing the interactions described by a project manager and illustrating the feedback loops resulting from the complex dynamics behavior of a project under duress. The arrows represent causality

As Fig. 1 demonstrates, it is not easy to understand fully the complex dynamic behavior of a project under duress. Few would realize that 35 feedback loops are encompassed in the description that led to Fig. 1. Indeed one of the significant features of complex projects is the likelihood of underestimating the complexity due to the dynamics generated by disruptions. [6] has reported on the more specific difficulty of understanding feedback behavior and research in the field of managerial judgment reinforces the difficulties of biases unduly influencing judgment [27].

In the work presented here we presume that there is a customer and a contractor, and there is a bidding process usually involving considerations of liquidated damages for delays and possibly strategic reputational consequences for late delivery. Thus, we expect the project to have a clear beginning and an end when the customer (internal or external) signs off a contract. Finally, we do not explore the whole project business life cycle, but that part where major cost overruns occur: thus, we start our consideration when a bid is to be prepared, consider development and manufacturing or construction, but stop when the product of the project is handed over to the customer.

Thus, in this chapter we shall be concerned specifically with the use of SD to model the consequences of disruptions and delays. Often these disruptions are small changes to the project, for example design changes [28]. The work discussed here is the consequence of 12 years of constructing detailed SD simulation models of failed complex projects. The first significant case was reported in Ackermann et al. [10] and Bennett et al. [29]. In each case the prompt for the work was the reasonable prospect of the contractor making a successful claim for damages. In all the cases the claim was settled out of court and the simulation model played a key role in settling the dispute.

The chapter will firstly consider why modeling disruption and delay (D&D) is so difficult. It will discuss what is meant by the term D&D and the typical consequences of D&D. This will be examined using examples from real projects that have suffered D&D. The contribution of SD modeling to the analysis of D&D and thus to the explanation of project behavior will then be discussed. A process of modeling which has been developed over the last 12 years and one that provides a means of modeling and explaining project behavior will be introduced. This process involves constructing both qualitative cause maps and quantitative system dynamics models. The chapter will conclude by considering potential future developments for the use of SD in modeling complex projects.

Disruption and Delay

(The following contains excerpts from Eden et al. [22] which provides a full discussion on the nature of D&D).

The idea that small disruptions can cause serious consequences to the life of a major project, resulting in massive time and cost overruns, is well established. The terms 'disruption and delay' or 'delay and disruption' are also often used to describe what has happened on such projects. However, although justifying the direct impact of disruptions and delays is relatively easy, there has been considerable difficulty in justifying and quantifying the claim for the indirect consequences. Our experience from working on a series of such claims is that some of the difficulty derives from ambiguity about the nature of disruption and delay (D&D). We now consider what we mean by D&D before moving onto considering the types of consequences that can result from the impact of D&D.

What is a Disruption?

Disruptions are events that prevent the contractor completing the work as planned. Many disruptions to complex projects are planned for at the bid stage because they may be expected to unfold during the project. For example, some level of rework is usually expected, even when everything goes well, because there will always be 'normal' errors and mistakes made by both the contractor and client. The disruption and delay that follows would typically be taken to be a part of a risk factor encompassed in the base estimate, although this can be significantly underestimated [30]. However, our experience suggests that *there are other types of disruptions that can be significant in their impact and are rarely thought about during original estimating*. When these types of disruptions do occur, their consequences can be underestimated as they are often seen by the contractor as aberrations with an expectation that their consequences can be controlled and managed. The linkage between risk assessment and the risks as potential triggers of D&D is often missed [31]. Interferences with the flow of work in the project is a common disruption. For example, when a larger number of design comments than expected are made by the client an increased number of drawings need rework. However it also needs to be recognized that these comments could have been made by the contractors own methods engineering staff. In either case, the additional work needed to respond to these comments, increases the contractor's workload and thus requires management to take mitigating actions if they still want to deliver on time. These mitigating actions are usually regarded as routine and capable of easily bringing the contract back to plan,

even though they can have complex feedback ramifications.

Probably one of the most common disruptions to a project comes when a customer or contractor causes changes to the product (a Variation or Change Order). For example, the contractor may wish to alter the product after engineering work has commenced and so request a direct change. However, sometimes changes may be made unwittingly. For example, a significant part of cost overruns may arise where there have been what might be called 'give-aways'. These may occur because the contractor's engineers get excited about a unique and creative solution and rather than sticking to the original design, produce something better but with additional costs. Alternatively, when the contractor and customer have different interpretations of the contract requirements unanticipated changes can occur. For example, suppose the contract asks for a door to open and let out 50 passengers in 2 minutes, but the customer insists on this being assessed with respect to the unlikely event of dominantly large, slow passengers rather than the contractor's design assumptions of an average person. This is often known as 'preferential engineering'. In both instances there are contractor and/or customer requested changes that result in the final product being more extensive than originally intended.

The following example, taken from a real project and originally cited in Eden et al. [30], illustrates the impact of a client induced change to the product:

Project 1: The contract for a 'state of the art' train had just been awarded. Using well-established design principles – adopted from similar train systems – the contractor believed that the project was on track. However within a few months problems were beginning to emerge. The client team was behaving very differently from previous experience and using the contract specification to demand performance levels beyond that envisioned by the estimating team. One example of these performance levels emerged during initial testing, 6 months into the contract, and related to water tightness. It was discovered that the passenger doors were not sufficiently watertight. Under extreme test conditions a small (tiny puddle) amount of water appeared. The customer demanded that there must be no ingress of water, despite acknowledging that passengers experiencing such weather would bring in more water on themselves than the leakage.

The contractor argued that no train had ever met these demands, citing that most manufacturers and operators recognized that a small amount of wa-

ter would always ingress, and that all operators accepted this. Nevertheless the customer interpreted the contract such that new methods and materials had to be considered for sealing the openings. The dialog became extremely combative and the contractor was forced to redesign. An option was presented to the customer for their approval, one that would have ramifications for the production process. The customer, after many tests and after the verdict of many external experts in the field, agreed to the solution after several weeks. Not only were many designs revisited and changed, with an impact on other designs, but also the delays in resolution impacted the schedule well beyond any direct consequences that could be tracked by the schedule system (www.Primavera.com) or costs forecasting system.

What is a Delay?

Delays are any events that will have an impact on the final date for completion of the project. Delays in projects come from a variety of sources. One common source is that of the client-induced delay. Where there are contractual obligations to comment upon documents, make approvals, supply information or supply equipment, and the client is late in these contractually-defined duties, then there may be a client-induced delay to the expected delivery date (although in many instances the delay is presumed to be absorbed by slack). But also a delay could be self-inflicted: if the sub-assembly designed and built did not work, a delay might be expected.

The different types of client-induced delays (approvals, information, etc.) have different effects and implications. Delays in client approval, in particular, are often ambiguous contractually. A time to respond to approvals may not have been properly set, or the expectations of what was required within a set time may be ambiguous (for example, in one project analyzed by the authors the clients had to respond within n weeks – but this simply meant that they sent back a drawing after n weeks with comments, then after the drawing was modified, they sent back the same drawing after a further m weeks with more comments). Furthermore, excessive comments, or delays in comments can cause chains of problems, impacting, for example, on the document approval process with sub-contractors, or causing over-load to the client's document approval process.

If a delay occurs in a project, it is generally considered relatively straightforward to cost. However, ramifications resulting from delays are often not trivial either to under-

stand or to evaluate. Let us consider a delay only in terms of the CPM (Critical Path Method), the standard approach for considering the effects of delays on a project [32]. The consequences of the delay depend on whether the activities delayed are on the Critical Path. If they *are* on the Critical Path, or the delays are sufficient to cause the activities to become on the critical path, it is conceptually easy to compute the effect as an Extension Of Time (EOT) [33]. However, even in this case there are complicating issues. For example; what is the effect on other projects being undertaken by the contractor? When this is not the first delay, then to which schedule does the term “critical path” refer? To the original, planned programme, which has already been changed or disrupted, or to the “as built”, actual schedule? Opinions differ here. It is interesting to note that, “the established procedure in the USA [of using as-built CPM schedules for claims] is almost unheard of in the UK” [33].

If the delay is *not* on the Critical Path then, still thinking in CPM terms, there are only indirect costs. For example, the activities on the Critical Path are likely to be resource dependent, and it is rarely easy to hire and fire at will – so if non-critical activities are delayed, the project may need to work on tasks in a non-optimal sequence to keep the workforce occupied; this will usually imply making guesses in engineering or production, requiring later re-work, less productive work, stop/starts, workforce overcrowding, and so on.

The following example, taken from a real project, illustrates the impact of a delay in client furnished information to the project:

Project 2: A state of the art vessels project had been commissioned which demanded not only the contractor meeting a challenging design but additionally incorporating new sophisticated equipment. This equipment was being developed in another country by a third party. The client had originally guaranteed that the information on the equipment would be provided within the first few months of the contract – time enough for the information to be integrated within the entire design. However time passed and no detailed specifications were provided by the third party – despite continual requests from the contractor to the client.

As the project had an aggressive time penalty the contractor was forced to make a number of assumptions in order to keep the design process going. Further difficulties emerged as information from the third party trickled in demanding changes from the emerging design. Finally manufacturing which

had been geared up according to the schedule were forced to use whatever designs they could access in order to start building the vessel.

Portfolio Effect of many Disruptions

It is not just the extent of the disruption or delay but the number of them which may be of relevance. This is particularly the case when a large number of the disruptions and/or delays impact immediately upon one another thus causing a portfolio of changes. These portfolios of D&D impacts result in effects that would probably not occur if only one or two impacts had occurred. For example, the combination of a large number of impacts might result in overcrowding or having to work in poor weather conditions (see example below). In these instances it is possible to identify each individual item as a contributory cause of extra work and delay but not easy to identify the combined effect.

The following example, taken from another real project, illustrates the impact of a series of disruptions to the project:

Project 3: A large paper mill was to be extended and modernized. The extension was given extra urgency by new anti-pollution laws imposing a limit on emissions being enacted with a strict deadline.

Although the project had started well, costs seemed to be growing beyond anything that made sense given the apparent minor nature of the disruptions. Documents issued to the customer for ‘information only’ were changed late in the process. The customer insisted on benchmarking proven systems, involving visits to sites working with experimental installations or installations operating under different conditions in various different countries. In addition there were many changes of mind about where equipment should be positioned and how certain systems should work. Exacerbating these events was the circumstance of both the customer’s and contractor’s engineers being co-located, leading to ‘endless’ discussions and meetings slowing the rate of both design and (later) commissioning.

Relations with the customer, who was seen by the contractor to be continually interfering with progress of the project, were steadily deteriorating. In addition, and in order to keep the construction work going, drawings were released to the construction team before being fully agreed. This meant that construction was done in a piecemeal fashion, often inefficiently (for example, scaffolding would be

put up for a job, then taken down so other work could proceed, then put up in the same place to do another task for which drawings subsequently had been produced). As the construction timescale got tighter and tighter, many more men were put on the site than was efficient (considerable overcrowding ensued) and so each task took longer than estimated.

As a result the project was behind schedule, and, as it involved a considerable amount of external construction work, was vulnerable to being affected by the weather. In the original project plan (as used for the estimate) the outer shell (walls and roof) was due to be completed by mid-Autumn. However, the project manager now found himself undertaking the initial construction of the walls and roofing in the middle of winter! As chance would have it, the coldest winter for decades, which resulted in many days being lost while it was too cold to work. The combination of the particularly vicious winter and many interferences resulted in an unexpectedly huge increase in both labour hours and overall delay. Over-time payments (for design and construction workers) escalated. The final overspend was over 40% more than the original budget.

Consequences of Disruptions and Delays

Disruption and delay (D&D) is primarily the consequence of interactions which feed on themselves as a result of an initial disruption or delay or portfolio of disruptions and delays. If an unexpected variation (or disruption) occurs in a project then, if no intervention was to take place, a delivery delay would occur. In an attempt to avoid this situation, management may choose to take actions to prevent the delay (and possible penalties). In implementing these actions, side-effects can occur which cause further disruptions. These disruptions then cause further delays to the project. In order to avoid this situation, additional managerial action is required. Thus, an initial disruption has led to a delay, which has led to a disruption, which has led to a further delay. A positive feedback loop has been formed, where both disruption and delay feed back on themselves causing further disruptions and delays. Due to the nature of feedback loops, a powerful vicious cycle has been created which, if there is no alternative intervention, can escalate with the potential of getting 'out of control'. It is the dynamic behavior caused by these vicious cycles which can cause severe disruption and consequential delay in a project.

The dynamic behavior of the vicious cycles which are responsible for much of the D&D in a project make the costing of D&D very difficult. It is extremely difficult to separate each of the vicious cycles and evaluate their individual cost. Due to the dynamic behavior of the interactions between vicious cycles, the cost of two individual cycles will escalate when they interact with one another, thus disruptions have to be costed as part of a portfolio of disruptions.

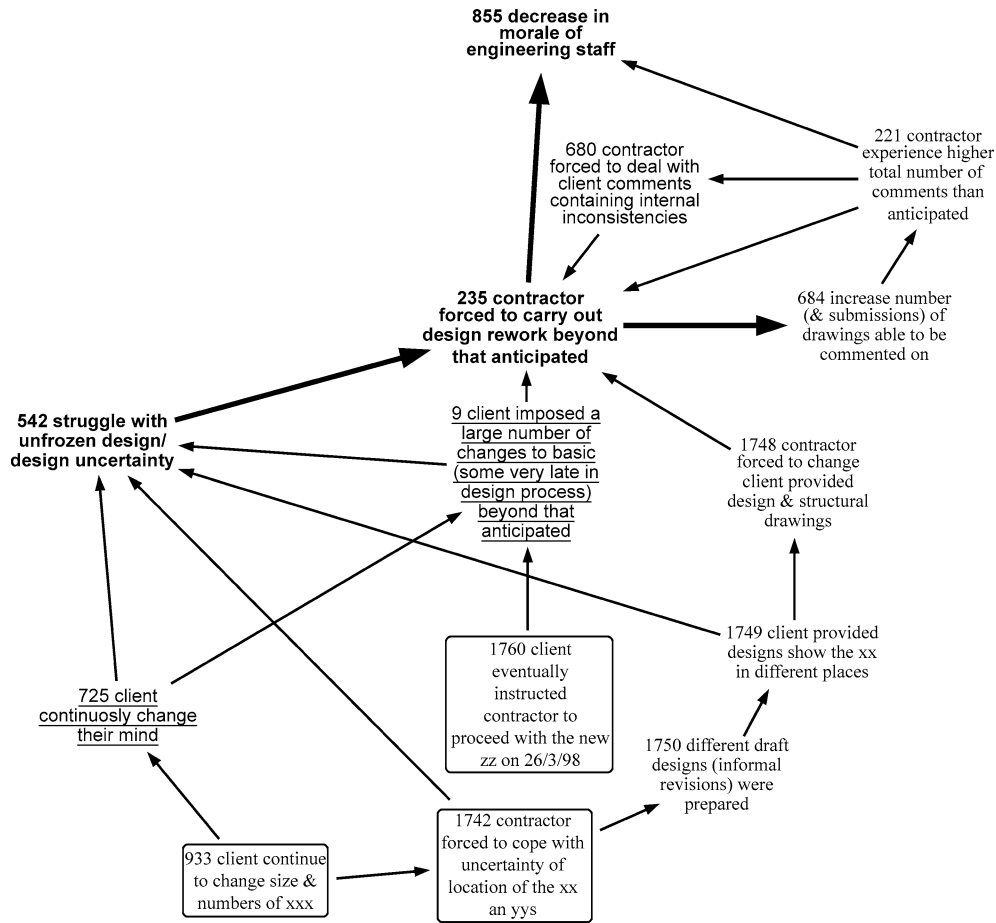
Returning to Project 2, the vessel case, as can be seen in Fig. 2, the client caused both disruptions (continuous changes of mind) and delays (late permission to use a particular product). Both of these caused the contractor to undertake rework, and struggle with achieving a frozen (fixed) design. These consequences in turn impacted upon staff morale and also developed as noted above dynamic behavior – where rework resulted in more submissions of designs, which led to further comments, some of which were inconsistent and therefore led to further rework. As mentioned in the introduction, the rework cycle [17,18,19,20,21] can be a major driver of escalating feedback within a complex project.

Managerial Actions and the Consequences of D&D

The acceleration of disrupted projects to avoid overall project delays is common practice by managers who are under pressure from the client and/or their own senior management to deliver on time. However, the belief that this action will always help avoid delays is naive as it does not take into account an appreciation of the future consequences that can be faced. For example, one typical action designed to accelerate a project is to hire new staff. In doing so, some of the difficulties which may follow are:

- New staff take time to become acquainted with both the project and thus their productivity is lower than that of an existing skilled worker.
- New staff require training on the project and this will have an impact on the productivity of existing staff.
- Rather than hiring new staff to the organization, staff may be moved from other parts of the organization. This action results in costs to other projects as the other project is short of staff and so may have to hire workers from elsewhere, thereby suffering many of the problems discussed above.

Many of the outcomes of this action and other similar actions can lead to a reduction in expected productivity levels. Low productivity is a further disruption to the project through a lack of expected progress. If management identifies this lack of progress, then further managerial actions



Delay and Disruption in Complex Projects, Figure 2

Excerpt from a cause map showing some of the consequences of disruption and delay in Project 2. *Boxed* statements are specific illustrations with statements *underlined* representing generic categories (e.g. changes of mind). Statements in *bold text* represent the SD variables with the remainder providing additional context. *All links* are causal however those in *bold* illustrate sections of a feedback loop. The numbers at the beginning of concept are used as reference numbers in the model

may be taken in an attempt to avoid a further delay in delivery. These actions often lead to more disruptions, reinforcing the feedback loop that had been set up by the first actions.

Two other common managerial actions taken to avoid the impact of a disruption on delivery are (i) the use of overtime and (ii) placing pressure on staff in an attempt to increase work rate. Both of these actions can also have detrimental effects on staff productivity once they have reached particular levels. Although these actions are used to increase productivity levels, effects on fatigue and morale can actually lead to a lowering of productivity via a slower rate of work and/or additional work to be completed due to increased levels of rework [21,34]. This lowering of productivity causes a delay through lack of expected progress on the project, causing a further delay

to delivery. Management may then attempt to avoid this by taking other actions which in turn cause a disruption which again reinforces the feedback loop that has been set up.

Analyzing D&D and Project Behavior

The above discussion has shown that whilst D&D is a serious aspect of project management, it is a complicated phenomenon to understand. A single or a series of disruptions or delays can lead to significant impacts on a project which cannot be easily thought through due to human difficulties in identifying and thinking through feedback loops [26,35]. This makes the analysis of D&D and the resulting project behavior particularly difficult to explain.

SD modeling has made a significant contribution to increasing our understanding of why projects behave in the way they do and in quantifying effects. There are two situations in which this is valuable: the claim situation, where one side of the party is trying to explain the project's behavior to the other (and, usually, why the actions of the other party has caused the project to behave in the way it has) and the post-project situation, where an organization is trying to learn lessons from the experience of a project. In the case of a claim situation, although it has been shown that SD modeling can meet criteria for admissibility to court [36], there are a number of objectives which SD, or any modeling method, would need to address [37]. These include the following:

1. Prove causality – show what events triggered the D&D and how the triggers of D&D caused time and cost overruns on the project.
2. Prove the 'quantum' – show that the events that caused D&D created a specific time and cost over-run in the project. Therefore, there is a need to replicate over time the hours of work due to D&D that were over-and-above those that were contracted, but were required to carry out the project.
3. Prove responsibility – show that the defendant was responsible for the outcomes of the project. Also to demonstrate the extent to which plaintiff's management of the project was reasonable and the extent that overruns could not have been reasonably avoided.
4. All of the above have to be proved in a way which will be convincing to the several stakeholders in a litigation audience.

Over the last 12 years the authors have developed a model building process that aims to meet each of these purposes. This process involves constructing qualitative models to aid the process of building the 'case' and thus help to prove causality and responsibility (purposes 1 and 3). In addition, quantitative system dynamics models are involved in order to help to prove the quantum (purpose 2). However, most importantly, the process provides a structured, transparent, formalized process from "real world" interviews to resulting output which enables multiple audiences, including multiple non-experts as well as scientific/expert audiences to appreciate the validity of the models and thus gain confidence in these models and the consulting process in which they are embedded (purpose 4). The process is called the 'Cascade Model Building Process'. The next section describes the different stages of the model building process and some of the advantages of using the process.

Cascade Model Building Process

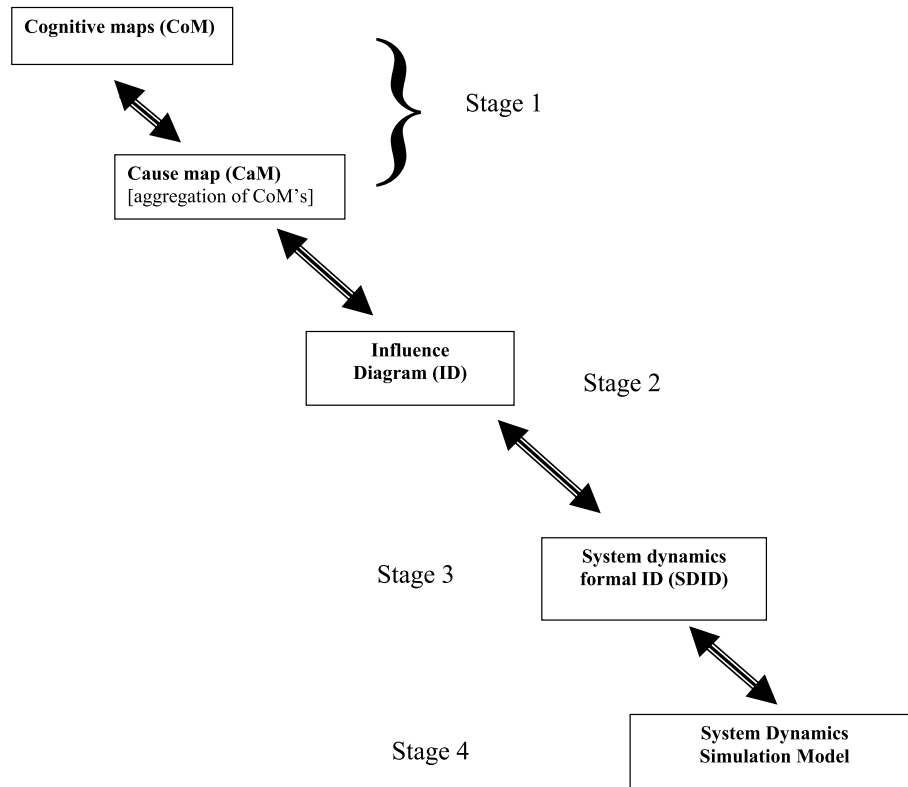
(The following contains excerpts from Howick et al. [38], which contains a full description of the Cascade Model Building process).

The 'Cascade Model Building Process' involves four stages (see Fig. 3) each of which are described below.

Stage 1: Qualitative Cognitive and Cause Map

The qualitative cognitive maps and /or project cause map aim to capture the key events that occurred on the project, for example a delay as noted above in the vessel example in Project 2. The process of initial elicitation of these events can be achieved in two ways. One option is to interview, and construct cognitive maps [39,40,41] for each participant's views. Here the aim is to gain a deep and rich understanding that taps the wealth of knowledge of each individual. These maps act as a preface to getting the group together to review and assess the total content represented as a merged cause map [42] in a workshop setting. The second option is to undertake group workshops where participants can contribute directly, anonymously and simultaneously, to the construction of a cause map. The participants are able to 'piggy back' off one another, triggering new memories, challenging views and developing together a comprehensive overview [43]. As contributions from one participant are captured and structured to form a causal chain, this process triggers thoughts from others and as a result a comprehensive view begins to unfold. In Project 1, this allowed the relevant design engineers (not just those whose responsibility was the water tight doors, but also those affected who were dealing with car-body structure, ventilation, etc.), methods personnel and construction managers to surface a comprehensive view of the different events and consequences that emerged.

The continual development of the qualitative model, sometimes over a number of group workshops, engenders clarity of thought predominantly through its adherence to the coding formalisms used for cause mapping [44]. Members of the group are able to debate and consider the impact of contributions on one another. Through bringing the different views together it is also possible to check for coherency – do all the views fit together or are there inconsistencies? This is not uncommon as different parts of the organizations (including different discipline groups within a division e. g. engineering) encounter particular effects. For example, during an engineering project, manufacturing can often find themselves bewildered by engineering processes – why are designs so late. However, the first stage of the cascade process enables the views from



Delay and Disruption in Complex Projects, Figure 3
The Cascade Model Building Process

engineering, methods, manufacturing, commissioning etc. to be integrated. Arguments are tightened as a result, inconsistencies identified and resolved and detailed audits (through analysis and features in the modeling software) undertaken to ensure consistency between both modeling team and model audience. In some instances the documents generated through reports about the organizational situation can be coded into a cause map and merged into the interview and workshop material [45].

The cause map developed at this stage is usually large – containing up to 1000 nodes. Computer supported analysis of the causal map can inform further discussion. For example, it can reveal those aspects of causality that are central to understanding what happened. Events that have multiple consequences for important outcomes can be detected. Feedback loops can be identified and examined. The use of software facilitates the identification of sometimes complex but important feedback loops that follow from the holistic view that arises from the merging of expertise and experience across many disciplines within the organization.

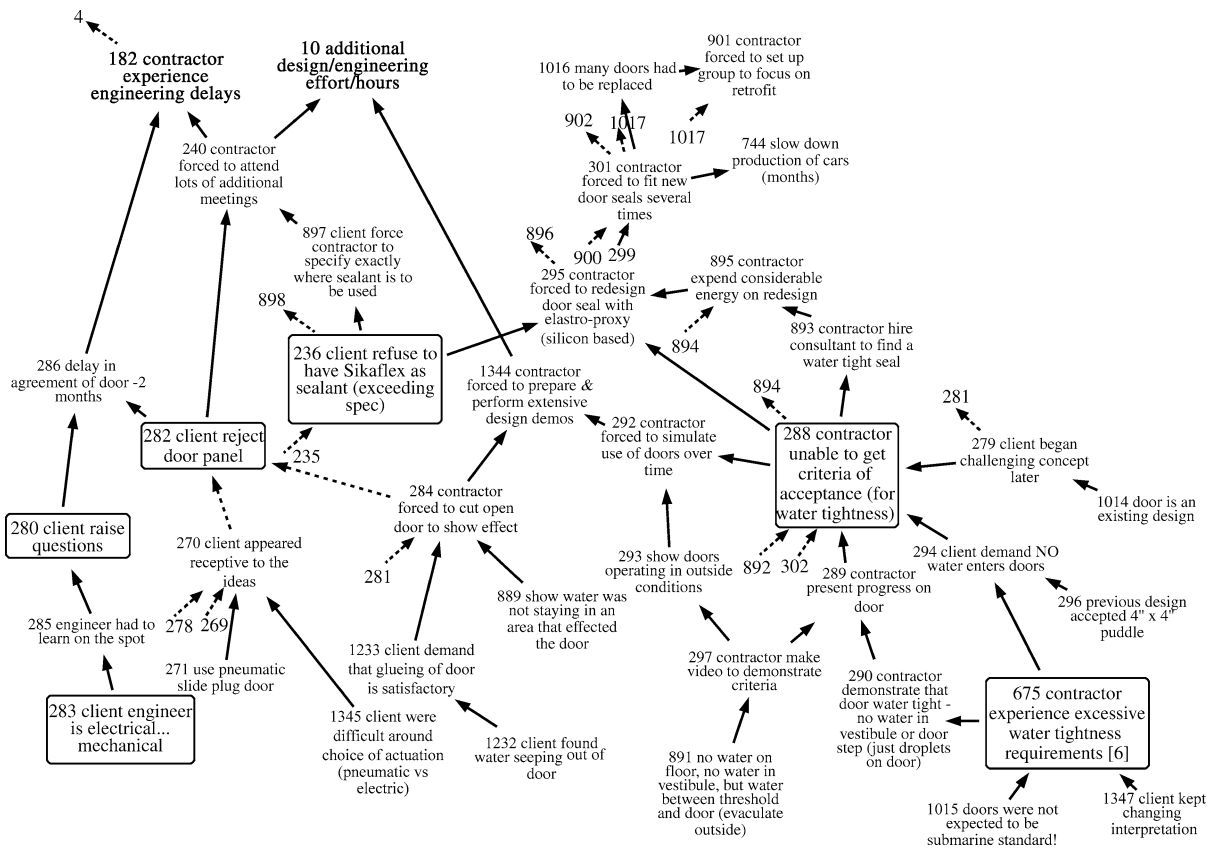
The resulting cause map from stage 1 can be of particular use in proving causality. For example, Fig. 4 repre-

sents some of the conversations made regarding the water ingress situation described in the above case. In this figure, consequences such as additional engineering effort and engineering delays can be traced back to events such as client found water seeping out of door.

Stage 2: Cause Map to Influence Diagram

The causal model produced from stage 1 is typically very extensive. This extensiveness requires that a process of ‘filtering’ or ‘reducing’ the content be undertaken – leading to the development of an Influence Diagram (ID) (the second step of the cascade process). Partly this is due to the fact that many of the statements captured whilst enabling a detailed and thorough understanding of the project, are not relevant when building the SD model in stage 4 (as a result of the statements being of a commentary like nature rather than a discrete variable). Another reason is that for the most part SD models comprise fewer variables/auxiliaries to help manage the complexity (necessary for good modeling as well as comprehension).

The steps involved in moving from a cause map to an ID are as follows:



Delay and Disruption in Complex Projects, Figure 4

Excerpt from a cause map showing some of the conversations regarding the water ingress situation in Project 1. As with Fig. 2, statements that have borders are the illustrations, those with **bold font** represent variables with the remainder detailing context. Dotted arrows denote the existence of further material which can be revealed at anytime

Step 1: Determining the core/endogenous variables of the ID

- Identification of feedback loops: It is not uncommon to find over 100 of these (many of these may contain a large percentage of common variables) when working on large projects with contributions from all phases of the project.
- Analysis of feedback loops: Once the feedback loops have been detected they are scrutinized to determine a) whether there are nested feedback 'bundles' and b) whether they traverse more than one stage of the project. Nested feedback loops comprise a number of feedback loops around a particular topic where a large number of the variables/statements are common but with variations in the formulation of the feedback loop. Once detected, those statements that appear in the most number of the nested feedback loops are identified as they provide core variables in the ID model.

Where feedback loops straddle different stages of the process for example from engineering to manufacturing note is taken. Particularly interesting is where a feedback loop appears in one of the later stages of the project e.g. commissioning which links back to engineering. Here care must be taken to avoid chronological inconsistencies – it is easy to link extra engineering hours into the existing engineering variable however by the time commissioning discover problems in engineering, the majority if not all engineering effort has been completed.

Step 2: Identifying the triggers/exogenous variables for the ID The next stage of the analysis is to look for triggers – those statements that form the exogenous variables in the ID. Two forms of analysis provide clues which can subsequently be confirmed by the group:

- The first analysis focuses on starting at the end of the chains of argument (the tails) and laddering up

(following the chain of argument) until a branch point appears (two or more consequences). Often statements at the bottom of a chain of argument are examples which when explored further lead to a particular behavior e.g. delay in information, which provides insights into the triggers.

- (ii) The initial set of triggers created by (i) can be confirmed through a second type of analysis – one which takes two different means of examining the model structure for those statements that are central or busy. Once these are identified they can be examined in more detail through creating hierarchical sets based upon them and thus “tear drops” of their content. Each of these teardrops is examined as possible triggers.

Step 3: Checking the ID Once the triggers and the feedback loops are identified care is taken to avoid double counting – where one trigger has multiple consequences some care must be exercised in case the multiple consequences are simple replications of one another.

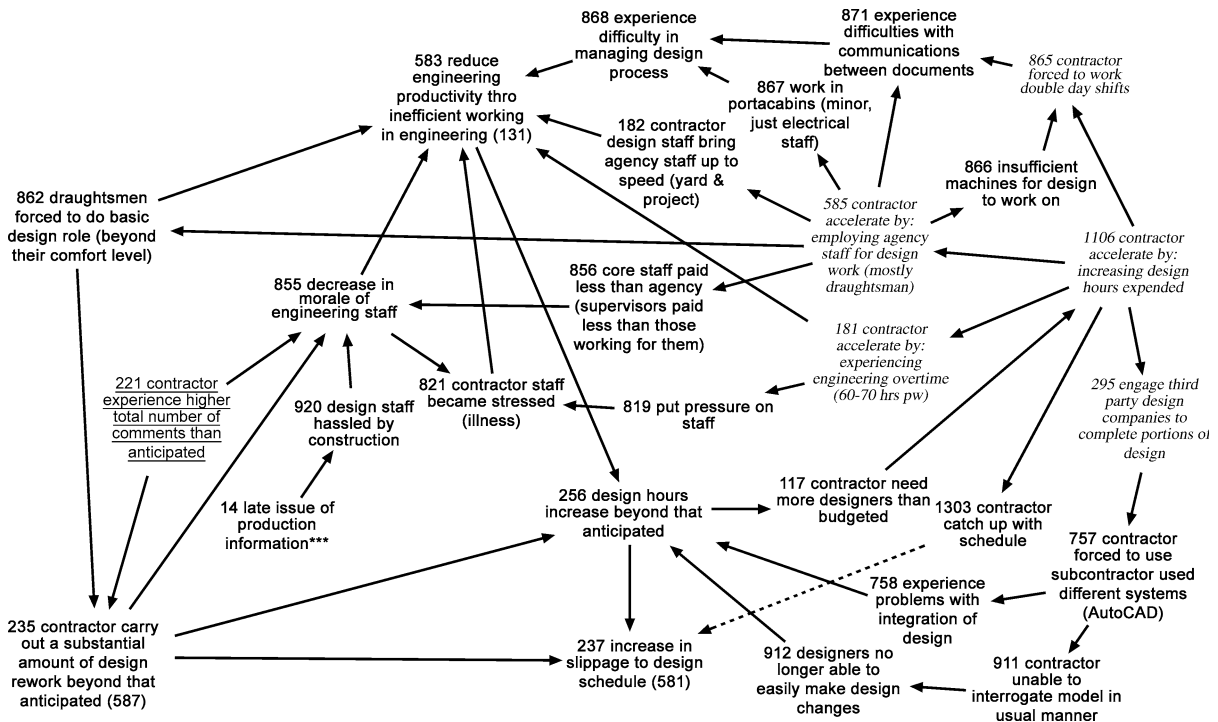
The resulting ID is comparable to a ‘causal loop diagram’ [46] which is often used as a pre-cursor to a SD

model. From the ID structure it is possible to create “stories” where a particular example triggers an endogenous variable which illustrates the dynamic behavior experienced.

Stage 3: Influence Diagram to System Dynamics Influence Diagram (SDID)

When a SD model is typically constructed after producing a qualitative model such as an ID (or causal loop diagram), the modeler determines which of the variables in the ID should form the stocks and flows in the SD model, then uses the rest of the ID to determine the main relationships that should be included in the SD model. However when building the SD model there will be additional variables/constants that will need to be included in order to make it ‘work’ that were not required when capturing the main dynamic relationships in the ID. The SDID is an influence diagram that includes all stocks, flows and variables that will appear in the SD model and is, therefore a qualitative version of the SD model. It provides a clear link between the ID and the SD model.

The SDID is therefore far more detailed than the ID and other qualitative models normally used as a pre-cursor to a SD model.



Delay and Disruption in Complex Projects, Figure 5

A small section of an ID from Project 2 showing mitigating actions (*italics*), triggers (underline) and some of the feedback cycles

Methods have been proposed to automate the formulation of a SD model from a qualitative model such as a causal loop diagram [47,48,49] and for understanding the underlying structure of a SD model [50]. However, these methods do not allow for the degree of transparency required to enable the range of audiences involved in a claim situation, or indeed as part of an organizational learning experience, to follow the transition from one model to the next. The SDID provides an intermediary step between an ID and a SD model to enhance the transparency of the transition from one model to another for the audiences. This supports an auditable trail from one model to the next.

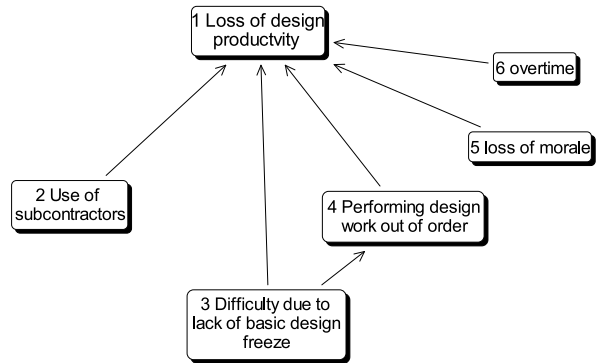
The approach used to construct the SDID is as follows: The SDID is initially created in parallel with the SD model. As a modeler considers how to translate an ID into a SD model, the SDID provides an intermediary step. For each variable in the ID, the modeler can do either of the following:

- (i) Create one variable in the SD & SDID: If the modeler wishes to include the variable as one variable in the SD model, then the variable is simply recorded in both the SDID and the SD model as it appears in the ID.
- (ii) Create multiple variables in the SD & SDID: To enable proper quantification of the variable, additional variables need to be created in the SD model. These variables are then recorded in both the SD model and SDID with appropriate links in the SDID which reflect the structure created in the SD model.

The SDID model forces all qualitative ideas to be placed in a format ready for quantification. However, if the ideas are not amenable to quantification or contradict one another, then this step is not possible. As a result of this process, a number of issues typically emerge including the need to add links and statements and the ability to assess the overall profile of the model though examining the impact of particular categories on the overall model structure. This process can also translate back into the causal model or ID model to reflect the increased understanding.

Stage 4: The System Dynamics Simulation Model

The process of quantifying SD model variables can be a challenge, particularly as it is difficult to justify subjective estimates of higher-level concepts such as “productivity” [51]. However, moving up the cascade reveals the causal structure behind such concepts and allows quantification at a level that is appropriate to the data-collection opportunities available. Figure 6, taken from the ID for Project 1, provides an example. The quantitative model

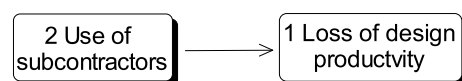


Delay and Disruption in Complex Projects, Figure 6
Section of an ID from Project 1 showing the factors affecting productivity

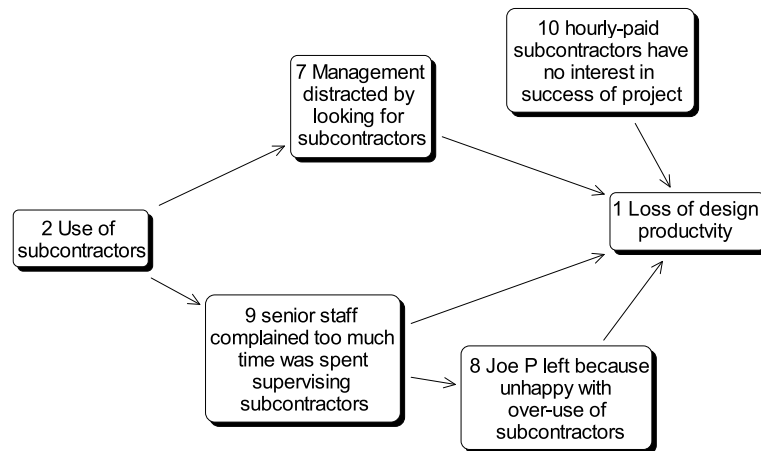
will require a variable “productivity” or “morale”, and the analyst will require estimation of the relationship between it and its exogenous and (particularly) endogenous causal factors. But while the higher-level concept is essential to the quantitative model, simply presenting it to the project team for estimation would not facilitate justifiable estimates of these relationships.

Reversing the Cascade

The approach of moving from stage 1 through to stage 4 can increase understanding and stimulate learning for all parties. However, the process of moving back up the cascade can also facilitate understanding between the parties. For example, in Fig. 7 the idea that a company was forced to use subcontractors and thus lost productivity might be a key part of a case for lawyers. The lawyers and the project team might have come at Fig. 7 as part of their construction of the case. Moving back up from the ID to the Cause Map (i. e. Fig. 7 to Fig. 8) as part of a facilitated discussion not only helps the parties to come to an agreed definition of the (often quite ill-defined) terms involved, it also helps the lawyers understand how the project team arrived at the estimate of the degree of the relationship. Having established the relationship, moving through the SDID (ensuring well-defined variables etc.) to the SD model enables the analysts to test the relationships to see whether



Delay and Disruption in Complex Projects, Figure 7
Section of an ID from Project 1 indicating the influence of the use of subcontractors on productivity



Delay and Disruption in Complex Projects, Figure 8

Section of a Cause Map from Project 1 explaining the relationship between the use of subcontractors and productivity

any contradictions arise, or if model behaviors are significantly different from actuality, and it enables comparison of the variables with data that might be collected by (say) cost accountants. Where there are differences or contradictions, the ID can be re-inspected and if necessary the team presented with the effect of the relationship within the SD model explained using the ID, so that the ID and the supporting cause maps can be re-examined to identify the flaws or gaps in the reasoning. Thus, in this example, as simulation modelers, cost accountants, lawyers and engineers approach the different levels of abstraction, the cascade process provides a unifying structure within which they can communicate, understand each other, and equate terms in each others discourse.

Advantages of the Cascade

The Cascade integrates a well-established method, cause mapping, with SD. This integration results in a number of important advantages for modeling to explain project behavior:

Achieving Comprehensiveness Our experience suggests that one of the principal benefits of using the cascade process derives from the added value gained through developing a rich and elaborated qualitative model that provides the structure (in a formalized manner) for the quantitative modeling. The cascade process immerses users in the richness and subtlety that surrounds their view of the projects and ensures involvement and ownership of all of the qualitative and quantitative models. The comprehensiveness leads to a better understanding of what occurred,

which is important due to the complex nature of D&D, and enables effective conversations to take place across different organizational disciplines.

The process triggers new contributions as memories are stimulated and both new material and new connections are revealed. The resultant models thus act as organizational memories providing useful insights into future project management (both in relation to bids and implementation). These models provide more richness and therefore an increased organizational memory when compared to the traditional methods used in group model building for system dynamics models (for example [52]). However this outcome is not untypical of other problem structuring methods [53].

Testing the Veracity of Multiple Perspectives The cascade's bi-directionality enabled the project team's understandings to be tested both numerically and from the perspective of the coherency of the systemic portrayal of logic. By populating the initial quantitative model with data [10] rigorous checks of the validity of assertions were possible.

In a claim situation, blame can be the fear of those participating in accounting for history and often restricts contributions [44]. When initiating the cascade process, the use of either interviews or group workshops increases the probability that the modeling team will uncover the rich story rather than partial explanations or as is often the case with highly politicized situations, 'sanitized' explanations. By starting with 'concrete' events that can be verified, and exploring their multiple consequences, the resultant model provides the means to reveal and explore the different experiences of various stakeholders in the project.

Modeling Transparency By concentrating the qualitative modeling efforts on the capture and structuring of multiple experiences and viewpoints the cascade process initially uses natural language and rich description as the medium which facilitates generation of views and enables a more transparent record to be attained.

There are often insightful moments as participants viewing the *whole* picture realize that the project is more complex than they thought. This realization results in two advantages. The first is a sense of relief that they did not act incompetently given the circumstances i. e. the consequences of D&D took over – which in turn instills an atmosphere more conducive to openness and comprehensiveness (see [44]). The second is learning – understanding the whole, the myriad and interacting consequences and in particular the dynamic effects that occurred on the project (that often acts in a counter-intuitive manner) provides lessons for future projects.

Common Understanding Across many Audiences

Claim situations involve numerous stakeholders, with varying backgrounds. The cascade process promotes ownership of the models from this mixed audience. For example, lawyers are more convinced by the detailed qualitative argument presented in the cause map (stage 1) and find this part of greatest utility and hence engage with this element of the cascade. However, engineers get more involved in the construction of the quantitative model and evaluating the data encompassed within it.

A large, detailed system dynamics model can be extremely difficult to understand for many of the stakeholders in a claim process [54]. However, the rich qualitative maps developed as part of the cascade method are presented in terms which are easier for people with no modeling experience to understand. In addition, by moving back up the cascade, the dynamic results that are output by the simulation model are given a grounding in the key events of the project, enabling the audience to be given fuller explanations and reasons for the D&D that occurred on the project.

Using the cascade method, any structure or parameters that are contained in the simulation model can be easily, and quickly, traced back to information gathered as a part of creating the cognitive maps or cause maps. Each contribution in these maps can then normally be traced to an individual witness who could defend that detail in the model. This auditable trail can aid the process of explaining the model and refuting any attacks made on the model.

Clarity The step-by-step process forces the modeler to be clear in what statements mean. Any illogical or incon-

sistent statements highlighted, require the previous stage to be revisited and meanings clarified, or inconsistencies cleared up. This results in clear, logical models.

Confidence Building As a part of gaining overall confidence in a model, any audience for the model will wish to have confidence in the structure of the model (for example [55,56,57,58]). When assessing confidence levels in a part of the structure of a SD model, the cascade process enables any member of the ‘client’ audience to clearly trace the structure of the SD model directly to the initial natural language views and beliefs provided from individual interviews or group sessions.

Scenarios are also an important test in which the confidence of the project team in the model can be considerably strengthened. Simulation is subject to the demands to reproduce scenarios that are recognizable to the managers capturing a portfolio of meaningful circumstances that occur at the same time, including many qualitative aspects such as morale levels. For example, if a particular time-point during the quantitative simulation is selected, clearly the simulated values of all the variables, and in particular the relative contributions of factors in each relationship, can be output from the model. If we consider Fig. 6, the simulation might show that at a particular point in a project, loss of productivity is 26% and that the loss due to:

“Use of subcontractors” is 5%.

“Difficulty due to lack of basic design freeze” is 9%.

“Performing design work out of order” is 3%.

“loss of morale” is 5%.

“overtime” is 4%.

Asking the project team their estimates of loss of productivity at this point in time, and their estimation of the relative contribution of these five factors, will help to validate the model. In most cases this loss level is best captured by plotting the relative levels of productivity against the time of critical incidents during the life the project. Discussion around this estimation might reveal unease with the simple model described in Fig. 6, which will enable discussion around the ID and the underlying cause map, either to validate the agreed model, or possibly to modify it and return up the cascade to further refine the model. In this scenario, validation of the cascade process provides a unifying structure within which the various audiences can communicate and understand each other.

The Cascade Model Building Process provides a rigorous approach to explaining why a project has behaved in

a certain way. The cascade uses rich, qualitative stories to give a grounding in the key events that drive the behavior of the project. In addition, it provides a quantifiable structure that allows the over time dynamics of the project to be described. The Cascade has therefore contributed significantly in understanding why projects behave in the way they do.

This chapter has focused on the role of SD modeling in explaining the behavior of complex projects. The final two sections will consider the implications of this work and will explore potential future directions for the use of SD modeling of projects.

Implications for Development

So what is the current status of SD modeling of projects? What is the research agenda for studying projects using SD? Below we consider each aspect of the project life-cycle in turn, to suggest areas where SD modeling may be applied, and to consider where further work is needed.

The first area is pre-project risk analysis. Risk analysis traditionally looks at risks individually, but looking at the systemicity in risks has clear advantages [59]. Firstly, the use of cause mapping techniques by an experienced facilitator, aided by software tools, is a powerful means of drawing out knowledge of project risk from an individual manager (or group of managers), enhancing clarity of thought, allowing investigation of the interactions between risks, and enhancing creativity. It is particularly valuable when used with groups, bringing out interactions between the managers and helping to surface cultural differences. And it clearly enables analysis of the systemicity, in particular identification of feedback dynamics, which can help explicate project dynamics in the ways discussed above. The influence of such work has led to the ideas of cause maps, influence diagrams and SD to be included into risk practice standard advice (the UK “PRAM” Guide, edition 2 [60] – absent from Edition 1). In one key example [31], the work described above enabled the team to develop a ‘Risk Filter’ in a large multinational project-based organization, for identifying areas of risk exposure on future projects and creating a framework for their investigation. The team reviewed the system after a few years; it had been used by 9 divisions, on over 60 major projects, and completed by 450 respondents; and it was used at several stages during the life of a project to aid in the risk assessment and contribute to a project database. The system allowed investigation of the interactions between risks, and so encouraged the management of the causality of relationships between risks, rather than just risks, thus focusing attention on those

risks and causality that create the most frightening ramifications on clusters of risks, as a system, rather than single items. This also encouraged conversations about risk mitigation across disciplines within the organization. Clearly cause mapping is useful in risk analysis, but there are a number of research questions that follow, for example:

- In looking at possible risk scenarios, what are appropriate methodologies to organize and facilitate heterogeneous groups of managers? And how technically can knowledge of systemicity and scenarios be gathered into one integrated SD model and enhance understanding? [61]
- How can SD models of possible scenarios be populated to identify key risks? How does the modeling cascade help in forward-looking analysis?
- There are many attempts to use Monte-Carlo simulation to model projects, without taking the systemic issues into account – leading to models which can be seriously misleading [62]. SD models can give a much more realistic account of the effect of risks – but how can essentially deterministic SD models as described above be integrated into a stochastic framework to undertake probabilistic risk analyses of projects which acknowledges the systemicity between the risks and the systemic effects of each risk?
- The use of SD is able to identify structures which give projects a propensity for the catastrophic systemic effects discussed in the Introduction. In particular, the three dimensions of structural complexity, uncertainty, and severe time-limitation in projects can combine together to cause significant positive feedback. However, defining metrics for such dimensions still remains an important open question. While a little work has been undertaken to give operational measures to the first of these (for example [63,64]), and de Meyer et al. [65] and Shenhar and Dvir [66] suggest selecting the management strategy based on such parameters, there has been little success so far in quantifying these attributes. The use of the SD models discussed above needs to be developed to a point where a project can be parametrized to give quantitatively its propensity for positive feedback.
- Finally, SD modeling shows that the effects of individual risks can be considerably greater than intuition would indicate, and the effects of clusters of risks particularly so. How can this be quantified so that risks or groups of risks can be ranked in importance to provide prioritization to managers? Again, Howick et al. [61] gives some initial indications here, but more work is needed.

The use of SD in operational control of projects has been less prevalent (Lyneis et al., [12] refers to and discusses examples of where it has been used). For a variety of reasons, SD and the traditional project management approach do not match well together. Traditional project-management tools look at the project in its decomposed pieces in a structured way (networks, work breakdown structures, etc.); they look at operational management problems at a detailed level; SD models aggregate into a higher strategic level and look at the underlying structure and feedback. Rodrigues and Williams [67] describe one attempt at an integrated methodology, but there is scope for research into how work with the SD paradigm can contribute to operational management of projects, and Williams [68] provides some suggestions for hybrid methods.

There is also a more fundamental reason why SD models do not fit in easily into conventional project management. Current project management practice and discourse is dominated by the “Bodies of Knowledge” or BoKs [69], which professional project management bodies consider to be the core knowledge of managing projects [1,70], presenting sets of normative procedures which appear to be self-evidently correct. However, there are three underlying assumptions to this discourse [71].

- Project Management is self-evidently correct: it is rationalist [72] and normative [73].
- The ontological stance is effectively positivist [74].
- Project management is particularly concerned with managing scope in individual parts [75].

These three assumptions lead to three particular *emphases* in current project management discourse and thus in the BoKs [71]:

- A heavy emphasis on planning [73,76].
- An implication of a very conventional control model [77].
- Project management is generally decoupled from the environment [78].

The SD modeling work provided explanations for why some projects severely over-run, which clash with these assumptions of the current dominant project management discourse.

- Unlike the third assumption, the SD models show behavior arising from the complex interactions of the various parts of the project, which would *not* be predicted from an analysis of the individual parts of the project [79].
- Against the first assumption, the SD models show project behavior which is complex and non-intuitive,

with feedback exacerbated through management response to project perturbations, conventional methods provide unhelpful or even disbeneficial advice and are not necessarily self-evidently correct.

- The second assumption is also challenged. Firstly, the models differ from the BoKs in their emphasis on, or inclusion of, “soft” factors, often important links in the chains of causality. Secondly, they show that the models need to incorporate not only “real” data but management perceptions of data and to capture the socially constructed nature of “reality” in a project.

The SD models tell us why failures occur in projects which exhibit complexity [63] – that is, when they combine *structural complexity* [80] – many parts in complex combinations – and *uncertainty*, in project goals and in the means to achieve those goals [81]. Goal uncertainty in particular is lacking in the conventional project management discourse [74,82], and it is when uncertainty affects a structurally complex traditionally-managed project that the systemic effects discussed above start to occur. But there is a third factor identified in the SD modeling. Frequently, events arise that compromise the plan at a faster rate than that at which it is practical to re-plan. When the project is heavily *time-constrained*, the project manager feels forced to take acceleration actions. A structurally complex project when perturbed by external uncertainties can become unstable and difficult to manage, and under time-constraints dictating acceleration actions when management has to make very fast and sometimes very many decisions, the catastrophic over-runs described above can occur. Work from different direction seeking to establish characteristics that cause complexity projects come up with similar characteristics (for example [66]). But the SD modeling explains *how* the tightness of the time-constraints strengthen the power of the feedback loops which means that small problems or uncertainties can cause unexpectedly large effects; it also shows how the type of under-specification identified by Flyvberg et al. [4] brings what is sometimes called “double jeopardy” – under-estimation (when the estimate is elevated to the status of a project control-budget) which leads to acceleration actions that then cause feedback which causes much greater over-spend than the degree of under-estimation.

Because of this, the greatest contribution that SD has made – and perhaps can make – is to increase our understanding of why projects behave in the way they do. There are two situations in which this is valuable: the claim situation, where one side of the party is trying to explain the project’s behavior to the others (and, usually, why the actions of the other party has caused the project to behave

in the way it has) and the post-project situation, where an organization is trying to learn lessons from the experience of a project.

The bulk of the work referred to in this chapter comes in the first of these, the claim situation. However, while these have proved popular amongst SD modelers, they have not necessarily found universal acceptance amongst the practicing project-management community. Work is needed therefore in a number of directions. These will be discussed in the next section.

Future Directions

We have already discussed the difficulty that various audiences can have in comprehending a large, detailed system dynamics model [54], and that gradual explanations that can be given by working down (and back up) the cascade to bring understanding to a heterogeneous group (which might include jurors, lawyers, engineers and so on) and so link the SD model to key events in the project. While this is clearly effective, more work is needed to investigate the use of the cascade. In particular, ways in which the cascade can be most effective in promoting understanding, in formalizing the methodology and the various techniques mentioned above to make it replicable, as well as how best to use SD here (Howick [54] outlines nine particular challenges the SD modeler faces in such situations). Having said this, it is still the case that many forums in which claims are made are very set in conventional project-management thinking, and we need to investigate more how the SD methods can be combined with more traditional methods synergistically, so that each supports the other (see for example [83]).

Significant unrealized potential of these methodologies are to be found in the post-project “lessons learned” situation. Research has shown many problems in learning generic lessons that can be extrapolated to other projects, such as getting to the root causes of problems in projects, seeing the underlying systemicity, and understanding the narratives around project events (Williams [84], which gives an extensive bibliography in the area). Clearly, the modeling cascade, working from the messiness of individual perceptions of the situation to an SD model, can help in these areas. The first part of the process (Fig. 3), working through to the cause map, has been shown to enhance understanding in many cases; for example, Robertson and Williams [85] describe a case in an insurance firm, and Williams [62] gives an example of a project in an electronics firm, where the methodology was used very “quick and dirty” but still gave increased understanding of why a (in that case successful) project turned out as it did, with

some pointers to lessons learned about the process. However, as well as formalization of this part of the methodology and research into the most effective ways of bringing groups together to form cause maps, more clarity is required as to how far down the cascade to go and the additional benefits that the SD modeling brings. “Stage 4” describes the need to look at quantification at a level that is appropriate to the data-collection opportunities available, and there might perhaps be scope for SD models of parts of the process explaining particular aspects of the outcomes. Attempts to describe the behavior of the whole project at a detailed level may only be suitable for the claims situation; there needs to be research into what is needed in terms of Stages 3 and 4 for gaining lessons from projects (or, if these Stages are not carried out, how the benefits such as enhanced clarity and validity using the cause maps, can be gained).

One idea for learning lessons from projects used by the authors, following the idea of simulation “learning labs”, was to incorporate learning from a number of projects undertaken by one particular large manufacturer into a simulation learning “game” [25]. Over a period of 7 years, several hundred Presidents, Vice-Presidents, Directors and Project Managers from around the company used the simulation tool as a part of a series of senior management seminars, where it promoted discussion around the experience and the effects encountered, and encouraged consideration of potential long-term consequences of decisions, enabling cause and effect relationships and feedback loops to be formed from participants’ experiences. More research is required here as to how such learning can be made most effective.

SD modeling has brought a new view to project management, enabling understanding of the behavior of complex projects that was not accessible with other methods. The chapter has described methodology for where SD has been used in this domain. This last part of the chapter has looked forward to a research agenda into how the SD work needs to be developed to bring greater benefits within the project-management community.

Bibliography

1. Project Management Institute (2000) A guide to the Project Management Body of Knowledge (PMBOK). Project Management Institute, Newtown Square
2. Cooper KG (1980) Naval ship production: a claim settled and a framework built. *Interfaces* 10:20–36
3. Szyliowicz JS, Goetz AR (1995) Getting realistic about megaproject planning: the case of the new Denver International Airport. *Policy Sci* 28:347–367
4. Flyvberg B, Bruzelius N, Rothengatter W (2003) Megaprojects

- and risk: an anatomy of ambition. Cambridge University Press, Cambridge
5. Scottish Parliament (2003) Corporate body issues August update on Holyrood. Parliamentary News Release 049/2003
 6. Major Projects Association (1994) Beyond 2000: A source book for major projects. Major Projects Association, Oxford
 7. Flyvberg B, Holm MK, Buhl SL (2002) Understanding costs in public works projects: error or lie? *J Am Plan Assoc* 68:279–295
 8. Morris PWG, Hough GH (1987) The anatomy of major projects. A study of the reality of project management. Wiley, Chichester
 9. Forrester J (1961) Industrial dynamics. Productivity Press, Portland
 10. Ackermann F, Eden C, Williams T (1997) Modeling for litigation: mixing qualitative and quantitative approaches. *Interfaces* 27:48–65
 11. Lyneis JM, Ford DN (2007) System dynamics applied to project management: a survey, assessment, and directions for future research. *Syst Dyn Rev* 23:157–189
 12. Lyneis JM, Cooper KG, Els SA (2001) Strategic management of complex projects: a case study using system dynamics. *Syst Dyn Rev* 17:237–260
 13. Ford DN (1995) The dynamics of project management: an investigation of the impacts of project process and coordination on performance. Massachusetts Institute of Technology, Boston
 14. Rodrigues A, Bowers J (1996) The role of system dynamics in project management. *Int J Proj Manag* 14:213–220
 15. Rodrigues A, Bowers J (1996) System dynamics in project management: a comparative analysis with traditional methods. *Syst Dyn Rev* 12:121–139
 16. Williams TM, Eden C, Ackermann F (1995) The vicious circles of parallelism. *Int J Proj Manag* 13:151–155
 17. Cooper KG (1993) The rework cycle: benchmarks for the project manager. *Proj Manag J* 24:17–21
 18. Cooper KG (1993) The rework cycle: How it really works.. and reworks... *PMNetwork* VII:25–28
 19. Cooper KG (1993) The rework cycle: why projects are mismanaged. *PMNetwork* VII:5–7
 20. Cooper KG (1993) The rework cycle: benchmarks for the project manager. *Proj Manag J* 24:17–21
 21. Cooper KG (1994) The \$2,000 hour: how managers influence project performance through the rework cycle. *Proj Manag J* 25:11–24
 22. Eden C, Williams TM, Ackermann F, Howick S (2000) On the nature of disruption and delay. *J Oper Res Soc* 51:291–300
 23. Eden C, Ackermann F, Williams T (2004) Analysing project cost overruns: comparing the measured mile analysis and system dynamics modelling. *Int J Proj Manag* 23:135–139
 24. Nahmias S (1980) The use of management science to support a multimillion dollar precedent-setting government contract litigation. *Interfaces* 10:1–11
 25. Williams TM, Ackermann F, Eden C, Howick S (2005) Learning from project failure. In: Love P, Irani Z, Fong P (eds) Knowledge management in project environments. Elsevier, Oxford
 26. Sterman JD (1989) Modelling of managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Manag Sci* 35:321–339
 27. Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
 28. Williams TM, Eden C, Ackermann F, Tait A (1995) The effects of design changes and delays on project costs. *J Oper Research Society* 46:809–818
 29. Bennett PG, Ackermann F, Eden C, Williams TM (1997) Analysing litigation and negotiation: using a combined methodology. In: Mingers J, Gill A (eds) Multimethodology: the theory and practice of combining management science methodologies. Wiley, Chichester, pp 59–88
 30. Eden C, Ackermann F, Williams T (2005) The amoebic growth of project costs. *Proj Manag J* 36(2):15–27
 31. Ackermann F, Eden C, Williams T, Howick S (2007) Systemic risk assessment: a case study. *J Oper Res Soc* 58(1):39–51
 32. Wickwire JM, Smith RF (1974) The use of critical path method techniques in contract claims. *Public Contract Law J* 7(1):1–45
 33. Scott S (1993) Dealing with delay claims: a survey. *Int J Proj Manag* 11(3):143–153
 34. Howick S, Eden C (2001) The impact of disruption and delay when compressing large projects: going for incentives? *J Oper Res Soc* 52:26–34
 35. Diehl E, Sterman JD (1995) Effects of feedback complexity on dynamic decision making. *Organ Behav Hum Decis Process* 62(2):198–215
 36. Stephens CA, Graham AK, Lyneis JM (2005) System dynamics modelling in the legal arena: meeting the challenges of expert witness admissibility. *Syst Dyn Rev* 21:95–122.35
 37. Howick S (2003) Using system dynamics to analyse disruption and delay in complex projects for litigation: Can the modelling purposes be met? *J Oper Res Soc* 54(3):222–229
 38. Howick S, Eden C, Ackermann F, Williams T (2007) Building confidence in models for multiple audiences: the modelling cascade. *Eur J Oper Res* 186:1068–1083
 39. Eden C (1988) Cognitive mapping: a review. *Eur J Oper Res* 36:1–13
 40. Ackermann F, Eden C (2004) Using causal mapping: individual and group: traditional and new. In: Pidd M (ed) Systems modelling: theory and practice. Wiley, Chichester, pp 127–145
 41. Bryson JM, Ackermann F, Eden C, Finn C (2004) Visible thinking: unlocking causal mapping for practical business results. Wiley, Chichester
 42. Shaw D, Ackermann F, Eden C (2003) Approaches to sharing knowledge in group problem structuring. *J Oper Res Soc* 54:936–948
 43. Ackermann F, Eden C (2001) Contrasting single user and networked group decision support systems. *Group Decis Negot* 10(1):47–66
 44. Ackermann F, Eden C, Brown I (2005) Using causal mapping with group support systems to elicit an understanding of failure in complex projects: some implications for organizational research. *Group Decis Negot* 14(5):355–376
 45. Eden C, Ackermann F (2004) Cognitive mapping expert views for policy analysis in the public sector. *Eur J Oper Res* 152:615–630
 46. Lane (2000) Diagramming conventions in system dynamics. *J Oper Res Soc* 51(2):241–245
 47. Burns JR (1977) Converting signed digraphs to Forrester schematics and converting Forrester schematics to differential equations. *IEEE Trans Syst Man Cybern SMC* 7(10):695–707
 48. Burns JR, Ulgen OM (1978) A sector approach to the formulation of system dynamics models. *Int J Syst Sci* 9(6):649–680
 49. Burns JR, Ulgen OM, Beights HW (1979) An algorithm for con-

- verting signed digraphs to Forrester's schematics. *IEEE Trans Syst Man Cybern SMC* 9(3):115–124
50. Oliva R (2004) Model structure analysis through graph theory: partition heuristics and feedback structure decomposition. *Syst Dyn Rev* 20(4):313–336
 51. Ford D, Sterman J (1998) Expert knowledge elicitation to improve formal and mental models. *Syst Dyn Rev* 14(4):309–340
 52. Vennix J (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
 53. Rosenhead J, Mingers J (2001) Rational analysis for a problematic world revisited. Wiley, Chichester
 54. Howick S (2005) Using system dynamics models with litigation audiences. *Eur J Oper Res* 162(1):239–250
 55. Ackoff RL, Sasieni MW (1968) Fundamentals of operations research. Wiley, New York
 56. Rivett P (1972) Principles of model building. Wiley, London
 57. Mitchell G (1993) The practice of operational research. Wiley, Chichester
 58. Pidd M (2003) Tools for thinking: modelling in management science. Wiley, Chichester
 59. Williams TM, Ackermann F, Eden C (1997) Project risk: systemicity, cause mapping and a scenario approach. In: Kahkonen K, Artto KA (eds) Managing risks in projects. E & FN Spon, London, pp 343–352
 60. APM Publishing Ltd (2004) Project risk analysis and management guide. APM Publishing Ltd, High Wycombe, Bucks
 61. Howick S, Ackermann F, Andersen D (2006) Linking event thinking with structural thinking: methods to improve client value in projects. *Syst Dyn Rev* 22(2):113–140
 62. Williams TM (2004) Learning the hard lessons from projects – easily. *Int J Proj Manag* 22(4):273–279
 63. Williams TM (1999) The need for new paradigms for complex projects. *Int J Proj Manag* 17:269–273
 64. Shenhar AJ (2001) One size does not fit all projects: exploring classical contingency domains. *Manag Sci* 47:394–414
 65. De Meyer A, Loch CH, Rich MT (2002) Managing project uncertainty: from variation to chaos. *MIT Sloan Mgmt Rev* 43(2):60–67
 66. Shenhar AJ, Dvir D (2004) How project differ and what to do about it. In: Pinto J, Morris P (eds) Handbook of managing projects. Wiley, New York, pp 1265–1286
 67. Rodrigues A, Williams TM (1997) Systems dynamics in software project management: towards the development of a formal integrated framework. *Eur J Inf Syst* 6:51–66
 68. Williams TM (2002) Modelling complex projects. Wiley, Chichester
 69. Dixon M (ed) (2000) The Association for Project Management (APM) Body of Knowledge (BoK), 4th edn. Association for Project Management, High Wycombe
 70. Stevens M (2002) Project management pathways. Association for Project Management, High Wycombe
 71. Williams TM (2005) Assessing and building on project management theory in the light of badly over-run projects. *IEEE Trans Eng Manag* 52(4):497–508
 72. Lundin RA (1995) Editorial: temporary organizations and project management. *Scand J Mgmt* 11:315–317
 73. Packendorff J (1995) Inquiring into the temporary organization: new directions for project management research. *Scand J Mgmt* 11:319–333
 74. Linehan C, Kavanagh D (2004) From project ontologies to communities of virtue. Paper presented at the 2nd International Workshop, Making projects critical, University of Western England, 13–14th December 2004
 75. Koskela L, Howell G (2002) The theory of project management: explanation to novel methods. In: Proceedings 10th Annual Conference on Lean Construction, IGLC-10, August 2002, Gramado, Brazil
 76. Koskela L, Howell G (2002) The underlying theory of project management is obsolete. In: Proc. PMI (Project Management Institute) Research Conference, Seattle 2002, pp 293–301
 77. Hodgson DE (2004) Project work: the legacy of bureaucratic control in the post-bureaucratic organization. *Organization* 11:81–100
 78. Malgrati A, Damiani M (2002) Rethinking the new project management framework: new epistemology, new insights. In: Proc. PMI (Project Management Institute) Research Conference, Seattle 2002, pp 371–380
 79. Lindkvist L, Soderlund J, Tell F (1998) Managing product development projects: on the significance of fountains and deadlines. *Org Stud* 19:931–951
 80. Baccarini D (1996) The concept of project complexity – a review. *Int J Proj Manag* 14:201–204
 81. Turner JR, Cochrane RA (1993) Goals-and-methods matrix: coping with projects with ill defined goals and/or methods of achieving them. *Int J Proj Manag* 11:93–102
 82. Engwall M (2002) The futile dream of the perfect goal. In: Sahil Andersson K, Soderholm A (eds) Beyond project management: new perspectives on the temporary-permanent dilemma. Libe Ekonomi, Copenhagen Business School Press, Malmö, pp 261–277
 83. Williams TM (2003) Assessing extension of time delays on major projects. *Int J Proj Manag* 21(1):19–26
 84. Williams TM (2007) Post-project reviews to gain effective lessons learned. Project Management Institute, Newtown Square
 85. Robertson S, Williams T (2006) Understanding project failure: using cognitive mapping in an insurance project. *Proj Manag J* 37(4):55–71

Dependency and Granularity in Data-Mining

SHUSAKU TSUMOTO, SHOJI HIRANO

Department of Medical Informatics, Shimane University, School of Medicine, Enya-cho Izumo City, Shimane, Japan

Article Outline

[Definition of the Subject](#)

[Introduction](#)

[Contingency Table from Rough Sets](#)

[Rank of Contingency Table \(\$2 \times 2\$ \)](#)

[Rank of Contingency Table \(\$m \times n\$ \)](#)

[Rank and Degree of Dependence](#)

[Degree of Granularity and Dependence](#)

[Conclusion](#)

Acknowledgment

Bibliography

Definition of the Subject

The degree of granularity of a contingency table is closely related with that of dependence of contingency tables. We investigate these relations from the viewpoints of determinantal divisors and determinants. From the results of determinantal divisors, it seems that the divisors provide information on the degree of dependencies between the matrix of all elements and its submatrices and that an increase in degree of granularity may lead to an increase in dependency. However, another approach shows that a constraint on the sample size of a contingency table is very strong, which leads to an evaluation formula in which an increase of degree of granularity gives a decrease of dependency.

Introduction

Independence (dependence) is a very important concept in data mining, especially for feature selection. In rough sets [2], if two attribute-value pairs, say $[c = 0]$ and $[d = 0]$ are independent, their supporting sets, denoted by C and D do not have an overlapping region ($C \cap D = \phi$), which means that an attribute independent to a given target concept may not appear in the classification rule for the concept.

This idea is also frequently used in other rule discovery methods: let us consider deterministic rules, described as *if-then* rules, which can be viewed as classic propositions ($C \rightarrow D$). From the set-theoretical point of view, a set of examples supporting the conditional part of a deterministic rule, denoted by C , is a subset of a set whose examples belong to the consequence part, denoted by D . That is, the relation $C \subseteq D$ holds and deterministic rules are supported only by positive examples in a dataset [4].

When such a subset relation is not satisfied, indeterministic rules can be defined as if-then rules with probabilistic information [6]. From the set-theoretical point of view, C is not a subset, but closely overlapped with D . That is, the relations $C \cap D \neq \phi$ and $|C \cap D|/|C| \geq \delta$ will hold in this case¹. Thus, probabilistic rules are supported by a large number of positive examples and a small number of negative examples.

On the other hand, in a probabilistic context, independence of two attributes means that one attribute (a_1) will

not influence the occurrence of the other attribute (a_2), which is formulated as $p(a_2|a_1) = p(a_2)$.

Although independence is a very important concept, it has not been fully and formally investigated as a relation between two attributes. Tsumoto introduces linear algebra into formal analysis of a contingency table [5]. The results give the following interesting results: First, a contingency table can be viewed as comparison between two attributes with respect to information granularity. Second, algebra is a key point of analysis of this table. A contingency table can be viewed as a matrix and several operations and ideas of matrix theory are introduced into the analysis of the contingency table. Especially for the degree of independence, rank plays a very important role in extracting a probabilistic model from a given contingency table.

This paper presents a further investigation into the degree of independence of a contingency matrix.

Intuitively and empirically, when two attributes have many values, the dependence between these two attributes becomes low. However, from the results of determinantal divisors, it seems that the divisors provide information on the degree of dependencies between the matrix of all elements and its submatrices and that an increase in degree of granularity may lead to an increase in dependency. The key of the resolution of these conflicts is to consider constraints on the sample size.

In this paper we show that a constraint on the sample size of a contingency table is very strong, which leads to the evaluation formula in which the increase of degree of granularity gives a decrease of dependency. The paper is organized as follows: Sect. "Contingency Table from Rough Sets" shows preliminaries. Sect. "Rank of Contingency Table (2×2)" discusses the former results. Sect. "Rank of Contingency Table ($m \times n$)" gives the relations between rank and submatrices of a matrix. Finally, Sect. "Degree of Granularity and Dependence" concludes this paper.

Contingency Table from Rough Sets

Notations

In the subsequent sections, the following notation, introduced in [3], is adopted: Let U denote a nonempty, finite set called the universe and A denote a nonempty, finite set of attributes, that is, $a: U \rightarrow V_a$ for $a \in A$, where V_a is called the domain of a . Then, a decision table is defined as an information system, $A = (U, A \cup \{D\})$, where $\{D\}$ is a set of given decision attributes. The atomic formulas over $B \subseteq A \cup \{D\}$ and V are expressions of the form $[a = v]$, called descriptors over B , where $a \in B$ and $v \in V_a$. The set $F(B, V)$ of formulas over B is the least set containing all atomic formulas over B and closed with respect to disjunc-

¹The threshold δ is the degree of the closeness of overlapping sets, which will be given by domain experts. For more information, please refer to Sect. "Rank of Contingency Table (2×2)".

Dependency and Granularity in Data-Mining, Table 1
Contingency table ($m \times n$)

	A_1	A_2	\dots	A_n	Sum
B_1	x_{11}	x_{12}	\dots	x_{1n}	$x_{1\cdot}$
B_2	x_{21}	x_{22}	\dots	x_{2n}	$x_{2\cdot}$
\dots	\dots	\dots	\dots	\dots	\dots
B_m	x_{m1}	x_{m2}	\dots	x_{mn}	$x_{m\cdot}$
Sum	$x_{\cdot 1}$	$x_{\cdot 2}$	\dots	$x_{\cdot n}$	$x_{\cdot\cdot} = U = N$

tion, conjunction and negation. For each $f \in F(B, V)$, f_A denotes the meaning of f in A , that is, the set of all objects in U with property f , defined inductively as follows:

1. If f is of the form $[a = v]$ then, $f_A = \{s \in U | a(s) = v\}$.
2. $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \cup g_A$; $(\neg f)_A = U - f_A$.

Contingency Table ($m \times n$)

Definition 1 Let R_1 and R_2 denote multinomial attributes in an attribute space A which have m and n values.

A contingency table is a table of a set described by the following formulas: $|[R_1 = A_j]_A|$, $|[R_2 = B_i]_A|$, $|[R_1 = A_j \wedge R_2 = B_i]_A|$, $|[R_1 = A_1 \wedge R_2 = A_2 \wedge \dots \wedge R_1 = A_m]_A|$, $|[R_2 = B_1 \wedge R_2 = A_2 \wedge \dots \wedge R_2 = A_n]_A|$ and $|U|$ ($i = 1, 2, 3, \dots, n$ and $j = 1, 2, 3, \dots, m$).

This table is arranged into the form shown in Table 1, where: $|[R_1 = A_j]_A| = \sum_{i=1}^m x_{ji} = x_{\cdot j}$, $|[R_2 = B_i]_A| = \sum_{j=1}^n x_{ji} = x_{i\cdot}$, $|[R_1 = A_j \wedge R_2 = B_i]_A| = x_{ij}$, $|U| = N = x_{\cdot\cdot}$ ($i = 1, 2, 3, \dots, n$ and $j = 1, 2, 3, \dots, m$).

Example 1 Let us consider an information table shown in Table 2.

The relationship between b and e can be examined by using the corresponding contingency table. First, the frequencies of four elementary relations, called *marginal distributions*, are counted: $[b = 0]$, $[b = 1]$, $[e = 0]$, and $[e = 1]$. Then, the frequencies of four kinds of conjunction are counted: $[b = 0] \wedge [e = 0]$, $[b = 0] \wedge [e = 1]$, $[b = 1] \wedge [e = 0]$, and $[b = 1] \wedge [e = 1]$.

Dependency and Granularity in Data-Mining, Table 2
A small dataset

a	b	c	d	e
1	0	0	0	1
0	0	1	1	1
0	1	2	2	0
1	1	1	2	1
0	0	2	1	0

Dependency and Granularity in Data-Mining, Table 3
Corresponding contingency table

	$b = 0$	$b = 1$	
$e = 0$	1	1	2
$e = 1$	2	1	3
	3	2	5

Then, the following contingency table is obtained (Table 3).

From this table, accuracy and coverage for $[b = 0] \rightarrow [e = 0]$ are obtained as $1/(1 + 2) = 1/3$ and $1/(1 + 1) = 1/2$.

One of the important observations from granular computing is that a contingency table shows the relations between two attributes with respect to intersection of their supporting sets. For example, in Table 3, both b and e have two different partitions of the universe and the table gives the relation between b and e with respect to the intersection of supporting sets. It is easy to see that this idea can be extended into an $n \times n$ contingency table, which can be viewed as an $n \times n$ -matrix. When two attributes have a different number of equivalence classes, the situation may be a little complicated. But, in this case, due to knowledge about linear algebra, we have only to consider the attribute which has a smaller number of equivalence classes and the surplus number of equivalence classes of the attributes with larger number of equivalence classes can be projected into other partitions. In other words, an $m \times n$ matrix or contingency table includes a projection from one attribute to the other one.

Rank of Contingency Table (2×2)

Preliminaries

Definition 2 A corresponding matrix $C_{T_{a,b}}$ is defined as a matrix the elements of which are equal to the value of the corresponding contingency table $T_{a,b}$ of two attributes a and b , except for marginal values.

Definition 3 The rank of a table is defined as the rank of its corresponding matrix. The maximum value of the rank is equal to the size of (square) matrix, denoted by r .

Example 2 Let the table given in Table 3 be defined as $T_{b,e}$. Then, $C_{T_{b,e}}$ is:

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}.$$

Since the determinant of $C_{T_{b,e}}$ $\det(C_{T_{b,e}})$ is not equal to 0, the rank of $C_{T_{b,e}}$ is equal to 2. It is the maximum value ($r = 2$), so b and e are statistically dependent.

Independence when the Table is 2×2

From the application of linear algebra, several results are obtained. (The proof is omitted). First, it is assumed that a contingency table is given as $m = 2$, $n = 2$ in Table 1. Then the corresponding matrix ($C_{T_{R_1, R_2}}$) is given as:

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}.$$

Proposition 1 *The determinant of $\det(C_{T_{R_1, R_2}})$ is equal to $|x_{11}x_{22} - x_{12}x_{21}|$.*

Proposition 2 *The rank will be:*

$$\text{rank} = \begin{cases} 2, & \text{if } \det(C_{T_{R_1, R_2}}) \neq 0 \\ 1, & \text{if } \det(C_{T_{R_1, R_2}}) = 0. \end{cases}$$

If the rank of $\det(C_{T_{b,e}})$ is equal to 1, according to the theorems of linear algebra, it is obtained that one row or column will be represented by the other column. That is,

Proposition 3 *Let r_1 and r_2 denote the rows of the corresponding matrix of a given 2×2 table, $C_{T_{b,e}}$. That is,*

$$r_1 = (x_{11}, x_{12}), \quad r_2 = (x_{21}, x_{22}).$$

Then, r_1 can be represented by r_2 : $r_1 = kr_2$, where k is given as:

$$k = \frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}} = \frac{x_1}{x_2}.$$

From this proposition, the following theorem is obtained.

Theorem 1 *If the rank of the corresponding matrix is 1, then two attributes in a given contingency table are statistically independent.*

Thus,

$$\text{rank} = \begin{cases} 2, & \text{dependent} \\ 1, & \text{statistical independent} \end{cases}.$$

Rank of Contingency Table ($m \times n$)

In the case of a general square matrix, the results in the 2×2 contingency table can be extended. It is especially important to observe that conventional statistical independence is supported only when the rank of the corresponding matrix is equal to 1. Let us consider the contingency table of c and a in Table 2, which is obtained as follows. Thus, the corresponding matrix of this table is:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

whose determinant is equal to 0. It is clear that its rank is 2.

Dependency and Granularity in Data-Mining, Table 4
Contingency table for a and c

	$a = 0$	$a = 1$	
$c = 0$	0	1	1
$c = 1$	1	1	2
$c = 2$	2	0	2
	3	2	5

It is interesting to see that if the case of $[d = 0]$ is removed, then the rank of the corresponding matrix is equal to 1 and two rows are equal. Thus, if the value space of d into $\{1, 2\}$ is restricted, then c and d are statistically independent. This relation is called *contextual independence* [1], which is related to conditional independence.

However, another type of weak independence is observed: let us consider the contingency table of a and c . The table is obtained as Table 4.

Its corresponding matrix is:

$$C_{T_{a,c}} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 0 \end{pmatrix},$$

Since the corresponding matrix is not square, the determinant is not defined. But it is easy to see that the rank of this matrix is two. In this case, even removing any attribute-value pair from the table will not generate statistical independence. Finally, the relation between rank and independence in a $n \times n$ contingency table is obtained.

Theorem 2 *Let the corresponding matrix of a given contingency table be a square $n \times n$ matrix. If the rank of the corresponding matrix is 1, then two attributes in a given contingency table are statistically independent. If the rank of the corresponding matrix is n , then two attributes in a given contingency table are dependent. Otherwise, two attributes are contextually dependent, which means that several conditional probabilities can be represented by a linear combination of conditional probabilities. Thus,*

$$\text{rank} = \begin{cases} n & \text{dependent} \\ 2, \dots, n-1 & \text{contextual independent} \\ 1 & \text{statistical independent} \end{cases}.$$

This theorem can be generalized into $m \times n$ matrix. If the corresponding matrix of a given contingency table is not square and of the form $m \times n$, then its rank is at most $\min(m, n)$.

For example, since $C_{T_{a,c}}$ is 3×2 , the rank is at most 2. Actually, from the calculation of subdeterminants shown in the next section, this matrix has a rank of 2.

Theorem 3 *Let the corresponding matrix of a given contingency table be an $m \times n$ matrix. The rank of this matrix*

is less than $\min(m, n)$. If the rank of the corresponding matrix is 1, then two attributes in a given contingency table are statistically independent. If the rank of the corresponding matrix is n , then two attributes in a given contingency table are dependent. Otherwise, two attributes are contextually dependent, which means that several conditional probabilities can be represented by a linear combination of conditional probabilities. Thus,

$$\text{rank} = \begin{cases} \min(m, n) & \text{dependent} \\ 2, \dots, \min(m, n) - 1 & \text{contextual independent} \\ 1 & \text{statistical independent} \end{cases}$$

In the cases of $m \neq n$, we need a discussion on submatrix and subdeterminant in the next section.

Rank and Degree of Dependence

Submatrix and Subdeterminant

The next interest is the structure of a corresponding matrix with $1 \leq \text{rank} \leq n - 1$. First, let us define a submatrix (a subtable) and subdeterminant.

Definition 4 Let A denote a corresponding matrix of a given contingency table ($m \times n$). A corresponding submatrix $A_{j_1 j_2 \dots j_s}^{i_1 i_2 \dots i_r}$ is defined as a matrix which is given by an intersection of r rows and s columns of A ($i_1 < i_2 < \dots < i_r$, $j_1 < j_2 < \dots < j_s$).

Definition 5 A subdeterminant of A is defined as a determinant of a submatrix $A_{j_1 j_2 \dots j_s}^{i_1 i_2 \dots i_r}$, which is denoted by $\det(A_{j_1 j_2 \dots j_s}^{i_1 i_2 \dots i_r})$.

Let us consider the contingency table given as Table 1. Then, a subtable for $A_{j_1 j_2 \dots j_s}^{i_1 i_2 \dots i_r}$ is given as Table 5.

Rank and Subdeterminant

Let δ_{ij} denote a co-factor of a_{ij} in a square corresponding matrix of A . Then,

$$\Delta_{ij} = (-1)^{i+j} \det(A_{1,2,\dots,i-1,i+1,\dots,n}^{1,2,\dots,j-1,j+1,\dots,n}).$$

Dependency and Granularity in Data-Mining, Table 5
A subtable ($r \times s$)

	A_{j_1}	A_{j_2}	...	A_{j_r}	Sum
B_{i_1}	$x_{i_1 j_1}$	$x_{i_1 j_2}$...	$x_{i_1 j_r}$	$x_{i_1 \cdot}$
B_{i_2}	$x_{i_2 j_1}$	$x_{i_2 j_2}$...	$x_{i_2 j_r}$	$x_{i_2 \cdot}$
...
B_{i_r}	$x_{i_r j_1}$	$x_{i_r j_2}$...	$x_{i_r j_r}$	$x_{i_r \cdot}$
Sum	$x_{\cdot 1}$	$x_{\cdot 2}$...	$x_{\cdot n}$	$x_{\cdot \cdot} = U = N$

It is notable that a co-factor is a special type of submatrix, where only the i th-row and j -column are removed from a original matrix.

By the use of co-factors, the determinant of A is defined as:

$$\det(A) = \sum_{j=1}^n a_{ij} \Delta_{ij},$$

which is called *Laplace expansion*.

From this representation, if $\det(A)$ is not equal to 0, then $\Delta_{ij} \neq 0$ for $\{a_{i_1}, a_{i_2}, \dots, a_{i_n}\}$ which are not equal to 0. Thus, the following proposition is obtained.

Proposition 4 $\det(A)$ is not equal to 0 if at least one co-factor of $a_{ij} (\neq 0)$, Δ_{ij} is not equal to 0.

It is notable that the above definition of a determinant gives the relation between an original matrix A and its submatrices (co-factors). Since cofactors give a square matrix of size $n - 1$, the above proposition gives the relation between a matrix of size n and submatrices of size $n - 1$. In the same way, we can discuss the relation between a corresponding matrix of size n and submatrices of size r ($1 \leq r < n - 1$).

Rank and Submatrix

Let us assume that corresponding matrix and submatrix are square ($n \times n$ and $r \times r$, respectively).

Theorem 4 If the rank of a corresponding matrix of size $n \times n$ is equal to r , the determinant of at least one submatrix of size $r \times r$ is not equal to 0. That is, there exists a submatrix $A_{j_1 j_2 \dots j_r}^{i_1 i_2 \dots i_r}$, which satisfies $\det(A_{j_1 j_2 \dots j_r}^{i_1 i_2 \dots i_r}) \neq 0$.

Corollary 1 If the rank of a corresponding matrix of size $n \times n$ is equal to r , all the determinants of the submatrices whose number of columns and rows are larger than $r + 1 (\leq n)$ are equal to 0.

Example 3 Let us consider the corresponding matrix mentioned in the above section, $C_{T_{a,c}}$.

The submatrices of this matrix are:

$$\begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Since all the subdeterminants are not equal to 0, the rank of this corresponding matrix is equal to 2.

Example 4 Let us consider the following corresponding matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

The determinant of A is:

$$\begin{aligned}\det(A) &= 1 \times (-1)^{1+1} \det \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} \\ &\quad + 2 \times (-1)^{1+2} \det \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} \\ &\quad + 3 \times (-1)^{1+3} \det \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} \\ &= 1 \times (-3) + 2 \times 6 + 3 \times (-3) = 0.\end{aligned}$$

Thus, the rank of A is smaller than 2. All the subdeterminants of A are:

$$\begin{aligned}\det \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} &= -3, & \det \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} &= -6, \\ \det \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} &= -3, & \det \begin{pmatrix} 1 & 2 \\ 7 & 8 \end{pmatrix} &= -6, \\ \det \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix} &= -12, & \det \begin{pmatrix} 2 & 3 \\ 8 & 9 \end{pmatrix} &= -6, \\ \det \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} &= -3, & \det \begin{pmatrix} 1 & 3 \\ 4 & 6 \end{pmatrix} &= -6, \\ \det \begin{pmatrix} 2 & 3 \\ 5 & 6 \end{pmatrix} &= -3.\end{aligned}$$

Since all the subdeterminants of A are not equal to 0, the rank of A is equal to 2.

Actually, since

$$\begin{pmatrix} 4 & 5 & 6 \end{pmatrix} = \frac{1}{2} \{ \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} + \begin{pmatrix} 7 & 8 & 9 \end{pmatrix} \},$$

and $\begin{pmatrix} 7 & 8 & 9 \end{pmatrix}$ cannot be represented by $k \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$ (k : integer), the rank of this matrix is equal to 2.

Thus, one attribute-value pair is statistically dependent on other two pairs, statistically independent of the other attribute. In other words, if two pairs are fixed, the one remaining attribute-value pair will be statistically independently determined.

Determinantal Divisors

From the subdeterminants of all submatrices of size 2, all the subdeterminants of a corresponding matrix has the greatest common divisor, equal to 3.

From the recursive definition of the determinants, it is show that the subdeterminants of size $r + 1$ will have the greatest common divisor of the subdeterminants of size r as a divisor. Thus,

Theorem 5 Let $d_k(A)$ denote the greatest common divisor of all the subdeterminants of size k , $\det(A_{j_1 j_2 \dots j_r}^{i_1 i_2 \dots i_k})$.

$d_1(A), d_2(A), \dots, d_n(A)$ are called *determinantal divisors*. From the definition of Laplace expansion,

$$d_k(A) | d_{k+1}(A).$$

In the example of the above subsection, $d_1(A) = 1$, $d_2(A) = 3$ and $d_3(A) = 0$.

Example 5 Let us consider $C_{T_{a,c}}$ as an example. $d_1(C_{T_{a,c}}) = 1$ and $d_2(C_{T_{a,c}}) = 1$.

Example 6 Let us consider the following corresponding matrix:

$$B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 11 & 9 \end{pmatrix}.$$

Calculation gives: $d_1(B) = 1$, $d_2(B) = 3$ and $d_3(B) = 18$.

It is notable that a simple change of a corresponding matrix gives a significant change to the determinant, which suggests a change of structure with regard to dependency/independence.

The relation between $d_k(A)$ gives an interesting constraint.

Proposition 5 Since $d_k(A) | d_{k+1}(A)$, the sequence of the divisors is a monotonically increasing one:

$$d_1(A) \leq d_2(A) \leq \dots \leq d_r(A),$$

where r denotes the rank of A .

The sequence of B illustrates this: $1 < 3 < 18$.

Let us define a ratio of $d_k(A)$ to $d_{k-1}(A)$, called *elementary divisors*, where C denotes a corresponding matrix and $k \leq \text{rank } A$:

$$e_k(C) = \frac{d_k(C)}{d_{k-1}(C)} (d_0(C) = 0).$$

Elementary divisors may give the increase of dependency between two attributes. For example, $e_1(B) = 1$, $e_2(B) = 3$, and $e_3(B) = 6$. Thus, a transition from 2×2 to 3×3 have a higher impact on the dependency of two attributes.

It is trivial to see that $\det(B) = e_1 e_2 e_3$, which can be viewed as a decomposition of the determinant of a corresponding matrix.

Divisors and Degree of Dependence

Since the determinant can be viewed as the degree of dependence, this result is very important. If values of all the

subdeterminants (size r) are very small (nearly equal to 0) and $d_r(A) \simeq 1$, then the values of the subdeterminants (size $r + 1$) are very small. This property may hold until the r reaches the rank of the corresponding matrix. Thus, the sequence of the divisors of a corresponding matrix gives a hidden structure of a contingency table.

Also, these results show that $d_1(A)$ and $d_2(A)$ are very important to estimate the rank of a corresponding matrix. Since $d_1(A)$ is given only by the greatest common divisor of all the elements of A , $d_2(A)$ are much more important components. This also intuitively suggests that the subdeterminants of A with size 2 are principal components of a corresponding matrix from the viewpoint of statistical dependence.

Recall that statistical independence of two attributes is equivalent to a corresponding matrix with a rank of 1. A matrix of rank 2 gives a context-dependent independence, which means three values of two attributes are independent, but two values of two attributes are dependent.

Subdeterminants and Degree of Dependence

Since the determinants give the degree of dependence, the degree of dependence can be evaluated by the values of subdeterminants.

For the above examples (A), since

$$\det \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix} = -12$$

gives the maximum value, the first and the third attribute-value pairs for two attributes are dependent on each other.

On the other hand, concerning B , since

$$\det \begin{pmatrix} 2 & 3 \\ 11 & 9 \end{pmatrix} = -15$$

gives the maximum value, the second and the third attribute-value pairs for two attributes are dependent on each other.

This discussion can be extended into the dependency between attribute-value pairs and a corresponding attribute. Let us consider 3×2 submatrices of A , which removes one column of the matrix

$$A_1 = \begin{pmatrix} 2 & 3 \\ 5 & 6 \\ 8 & 9 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 3 \\ 4 & 6 \\ 7 & 9 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{pmatrix}.$$

From the discussions in the above subsections, a set of the subdeterminants of 2×2 submatrices of A_j , denoted by

D_{A_j} are obtained as:

$$D_{A_1} = \{-3, -6, -3\}$$

$$D_{A_2} = \{-12, -6, -6\}$$

$$D_{A_3} = \{-3, -6, -3\}.$$

Thus, the first and third attribute value pairs are more dependent than the second value pairs, concerning the classification of attributes for the rows.

Elementary Divisors and Elementary Transformation

Let us define the following three elementary (row/column) transformations of a corresponding matrix:

1. Exchange two rows (columns), i_0 and j_0 ($P(i_0, j_0)$).
2. Multiply -1 to a row (column) i_0 ($T(i_0; -1)$).
3. Multiply t to a row (column) j_0 (i_0) and add it to a row i_0 (j_0). ($W(i_0, j_0, t)$).

Then, three transformations have several interesting characteristics.

Proposition 6 *Matrices corresponding to three elementary transformations are regular.*

Proposition 7 *Three elementary transformations do not change the rank of a corresponding matrix.*

Proposition 8 *Let \tilde{A} denote a matrix transformed by finite steps of three operations. Then,*

$$\text{rank } \tilde{A} = \text{rank } A, \quad d_r(\tilde{A}) = d_r(A),$$

where r denotes the rank of matrix A .

Then, from the application of linear algebra, the following interesting result is obtained.

Theorem 6 *With the finite steps of elementary transformations, a given corresponding matrix is transformed into*

$$\tilde{A} = \left(\begin{array}{ccc|c} e_1 & & & \\ & e_2 & & \\ & & \ddots & \\ & & & e_r \\ & & & & 0 \\ & & & & & 0 \end{array} \right)$$

where $e_j = \frac{d_j(A)}{d_{j-1}(A)}$ ($d_0(A) = 1$) and r denotes the rank of a corresponding matrix. Then, the determinant is decomposed into the product of e_j .

$$d_r(\tilde{A}) = d_r(A) = e_1 e_2 \dots e_r.$$

Degree of Granularity and Dependence

From Theorem 6, it seems that the increase of the degree of granularity gives that of the dependence between two attributes.

However, our empirical observations are different from the above intuitive analysis. Thus, there should be a strong constraint which suppresses the above effects on the degree of granularity.

Let us assume that the determinant of a given contingency matrix gives the degree of the dependence of the matrix. Then, from the application of linear algebra, we obtain the following theorem.

Theorem 7 *Let A denote an $n \times n$ contingency matrix, which includes N samples. If the rank of A is equal to n , then there exists a matrix B ($n \times n$) which satisfies*

$$BA = \begin{pmatrix} \rho_1 & & & O \\ & \rho_2 & & \\ & & \ddots & \\ O & & & \rho_n \end{pmatrix} = P,$$

where $\rho_1 + \rho_2 + \dots + \rho_n = N$.

It is notable that the value of determinants of P is larger than A :

$$\det A \leq \det P.$$

Example 7 Let us consider B as an example (Example 6). Let C denote the orthogonal matrix for transformation of B . Since the cardinality of B is equal to 48, the diagonal matrix which gives the maximum determinant is equal to:

$$\begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}.$$

On the other hand, the determinant of B is equal to 18. Thus, $\det B = 18 < 16^3 = 4096$. Then, C is obtained from the following equation:

$$C \times \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 11 & 9 \end{pmatrix} = \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}.$$

Thus,

$$C = \begin{pmatrix} -56/3 & 40/3 & -8/3 \\ 16/3 & -32/3 & 16/3 \\ 8 & 8/3 & -8/3 \end{pmatrix}.$$

It is notable that the determinant of C is equal to 2048/9. Also, since $\det B = 18$, we do not have any diagonal matrix

whose determinant is equal to 18 and whose sum of all the elements is equal to 48.

It is easy to see that the transformed matrix P has a very nice property for calculating the determinant.

Proposition 9 *The determinant of the transformed matrix P is equal to the multiplication of ρ_1 to ρ_n . That is,*

$$\det P = \rho_1 \rho_2 \dots \rho_n.$$

Then, the following constraint will be have the special meaning:

$$\rho_1 + \rho_2 + \dots + \rho_n = N, \quad (1)$$

because the following inequality holds in general:

$$\frac{\rho_1 + \rho_2 + \dots + \rho_n}{n} \geq \sqrt[n]{\rho_1 \rho_2 \dots \rho_n} \quad (2)$$

where the equality holds when $\rho_1 = \rho_2 = \dots = \rho_n$.

Since the above inequality can be transformed into:

$$\rho_1 \rho_2 \dots \rho_n \leq \left(\frac{\rho_1 + \rho_2 + \dots + \rho_n}{n} \right)^n,$$

the following inequality is obtained:

$$\det P = \rho_1 \rho_2 \dots \rho_n \leq \left(\frac{\rho_1 + \rho_2 + \dots + \rho_n}{n} \right)^n, \quad (3)$$

where the equality holds when $\rho_1 = \rho_2 = \dots = \rho_n$.

From the Theorem 7 and Eq. (1), the following theorem is obtained.

Theorem 8 *When a contingency matrix A holds $AB = P$, where P is a diagonal matrix, the following inequality holds:*

$$\det A \leq \left(\frac{N}{n} \right)^n.$$

Proof

$$\det A = \det(PB^{-1})$$

$$\leq \det P$$

$$= \rho_1 \rho_2 \dots \rho_n$$

$$\leq \left(\frac{\rho_1 + \rho_2 + \dots + \rho_n}{n} \right)^n = \left(\frac{N}{n} \right)^n, \quad (4)$$

where the former equality holds when $\det B^{-1} = \det B = 1$ and the latter equality holds when $\rho_1 = \rho_2 = \dots = \rho_n = \frac{N}{n}$. \square

Example 8 Let us consider the following contingency matrices D and E :

$$D = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 0 \\ 7 & 11 & 9 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$E = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 0 \\ 7 & 10 & 9 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The numbers of examples of D and E are 49 and 48, respectively, which can be comparable to that of B . Then, from Theorem 8,

$$\det D = 18 < (49/4)^4 = \frac{5764801}{256} \sim 22\,518$$

$$\det E = 12 < (48/4)^4 = 20\,736.$$

Thus, the maximum value of the determinant of A is at most $\left(\frac{N}{n}\right)^n$. Since N is constant for the given matrix A , the degree of dependence will decrease very rapidly when n becomes very large. That is,

$$\det A \sim n^{-n}.$$

Thus,

Corollary 2 *The determinant of A will converge to 0 when n increases to infinity*

$$\lim_{n \rightarrow \infty} \det A = 0.$$

This result suggests that when the degree of granularity becomes higher, the degree of dependence will become lower due to constraints on the sample size.

However, it is notable that N/n is very important. If N is very large, a rapid decrease will be observed when N is close to n . Even when N is 48 as shown in Example 8, $n = 3, 4$ may give a strong dependency between two attributes. For the behavior of $(N/n)^n$, we can apply the technique of real analysis, which will be our future work.

Conclusion

In this paper, a contingency table is interpreted from the viewpoint of granular computing and statistical independence. Matrix algebra is a key point of the analysis of a contingency table and the degree of independence, and rank plays a very important role in extracting a probabilistic model. From the correspondence between contingency table and matrix, the following results are obtained: First, the

value of determinants gives the degree of dependency between attribute-value pairs for a set of submatrices with the same size. Second, from the characteristics of the determinants, the larger the rank a corresponding matrix has, the more the two attributes are dependent. This result is shown by a monotonicity of a sequence of determinantal divisors. Third, elementary divisors give a decomposition of the determinant of a corresponding matrix. Finally, the constraint on the sample size of a contingency table is very strong, which leads to an evaluation formula in which an increase of degree of granularity gives the decrease of dependency.

Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research (13131208) on Priority Areas (No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Science, Culture, Sports, Science and Technology of Japan.

Bibliography

1. Butz C (2002) Exploiting contextual independencies in web search and user profiling. In: Proceedings of World Congress on Computational Intelligence (WCCI'2002) (CD-ROM)
2. Pawlak Z (1991) Rough sets. Kluwer, Dordrecht
3. Skowron A, Grzymala-Busse J (1994) From rough set theory to evidence theory. In: Yager R, Fedrizzi M, Kacprzyk J (eds) Advances in the Dempster-Shafer Theory of Evidence. Wiley, New York, pp 193–236
4. Tsumoto S (2000) Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Inf Sci 124:125–137
5. Tsumoto S (2003) Statistical independence as linear independence. In: Skowron A, Szczuka M (eds) Electronic Notes in Theoretical Computer Science, vol 82. Elsevier
6. Tsumoto S, Tanaka H (1996) Automated discovery of medical expert system rules from clinical databases based on rough sets. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 96. AAAI Press, Palo Alto, pp 63–69

Development, Complex Dynamic Systems of

PAUL VAN GEERT

The Heymans Institute, Groningen, The Netherlands

Article Outline

Glossary

Definition of the Subject

Introduction

Adaptation and Adaptive Agent Systems

Main Theories in the Field

and Short Historical Background

The Study of Individual Development in Light of Complexity Theory

An Overview of the Human Life Span in Light of the Theory of Complex Dynamic Systems

Future Directions

Bibliography

Glossary

Development In the context of developmental psychology, development implies the process of increasing knowledge, skill, capacity and so forth across the life span, in an ordered and directional process, leading to a stable state of maturity. Development implies an increase in complexity of the developing person or system.

Education A co-adaptive process involving asymmetrical relationships between educators (parents, teachers, ...) and young persons (children, pupils, ...); the process of the upbringing of children, by means of teaching, providing resources, models, teaching, guidance and so forth

Learning and teaching The process of gaining knowledge or skills, often in the context of help by a more competent person, who enables the learning through teaching, consisting of guidance, transmission of knowledge, structuring, promoting and confining the learner's zones of action, often in the context of explicit learning and teaching goals (e.g. teaching and learning how to write, teaching and learning how to weld iron on a construction site, etc.)

Complex developmental system A developmental system consisting of components such as the persons involved, material and cultural artifacts, properties attributable to individuals, that shows development as defined earlier, through the interactions between and interdependence of the components

Dynamic systems (developmental) A way of describing how one state of a developmental system changes into another state, as defined by a developmental state space (the whole of possible states distinguishable in the system, described by the set of dimensions or variables needed to specify the system as being developmental)

Developmental states, stages and ages A developmental state is any possible state in the developmental state space, which is defined by the dimensions used to describe development; with continuous dimensions, the number of possible states is continuous and infinite;

stages are states characterized by a stability that lasts over a sufficiently long time span (a few years) and by a pattern of mutually dependent properties, i.e. values of the developmental dimensions, stages are characteristic of the founding theories, such as Piaget's theory; ages are periods in the lifespan characterized by sufficient stability of the properties to qualify as properties characteristic of that period, ages often coincide with stages as distinguished in classical theories but often comprise additional properties required for coherent description.

Developmental domains Physical, neurological and sensorimotor development; cognitive development; language development; emotional development, social and personality development; self-, gender- and identity development, moral development; domains follow characteristic paths of development, can be distinguished from one another on the basis of their components and laws of change, but closely interact with one another and form interdependencies on the level of action and developmental time scales.

Developmental time scales Development takes place across various time scales, characterized by their characteristic event duration, the laws or principles that govern change on that particular time scale and interdependencies with other time scales; in order of descending duration the time scales relevant to human development are the scale of biological evolution; the time scale of socio-cultural historical development; the time scale of development across the human life span; the time scale of action and real-time experience and the time scale of underlying neurobiological processes.

Definition of the Subject

Developmental psychology concerns the study of developmental changes in human beings across the life span. Developmental changes are broadly defined as changes in the organism that are mostly progressive – in terms of increasing complexity, adaptation to the environment, efficiency of actions and operations and so forth. The subject of developmental psychology is the individual person, embedded in a particular social, cultural and material context. Although it is now widely accepted that development is a life-long process, the main developmental events take place during the first part of the life span, which co-occur with physical growth and development and coincide roughly with the age from birth to adulthood. Developmental changes through adulthood and old-age are often referred to as processes of aging and imply processes of

loss, for instancing due to aging of the nervous system, and processes compensating for such losses.

Developmental psychology is also concerned with problematic development, for instance in the form of developmental psychopathology, studying the life-span development of conditions such as autism, attention deficit disorders, oppositional behavior and so forth.

The aim of studying development from the perspective of complex dynamic systems is to apply the concepts of complexity and dynamic systems to the phenomena, theories and explanations currently found in developmental psychology, including the educational sciences, in order to arrive at a comprehensive theoretical approach on the subject that focuses on the mechanisms and forms of change.

Introduction

Terms of Change

Developmental psychology deals with various terms of change, some of which have already been defined under the Glossary terms: Development; education; learning. Other terms referring to developmental change that are worth considering are the life span (the period between conception and death in a single individual); socio-cultural evolution (the historical process of changes in human cultures and societies as they pertain to the life span and development of individuals); and variability and fluctuation. The latter refer to non-permanent changes in a developmentally relevant property or variable that are in principle occurring over the short-term (as compared to the long-term of the preceding forms of change), variability and fluctuation typically occur around one (or several) central tendencies or pivotal points.

Etymologically, development means unwrapping or unfolding, as in the unwrapping or unfolding of a book roll, or the unwrapping or unfolding of a flower bud [317,330,341,345]. Development thus carries an implicit notion of an inner logic in the sequence of the unfolding, a notion of potentiality (what is in there must come out) and a notion of finality (the unfolding comes to an end when the folded object is spread out). Although the historical meaning of development can of course not determine how we see or define development in scientific discourse, the deliberate application of this term in a particular context – instead of words like maturation, learning etc. – implies that we wish to refer to a phenomenon that is characterized to a more than a trivial extent, by these notions of inner logic, potentiality and finality. Development implies a directed process of change towards or unfolding of a mature state. It is a directed process, from an immature to a mature state, implying increasing complex-

ity in terms of a system that differentiates (incorporates more and more elements, features, knowledge ...) and at the same time integrates (constructs connections between the components).

Readers familiar with dynamic systems will immediately recognize these notions as metaphorical representations of self-organizing dynamics. The inner logic corresponds with the evolution term or the change function that governs the dynamics, and the potentiality and finality refer to self-organization or the systems tendency to move towards a particular attractor state. The notion of increasing developmental complexity refers to theories of complexity and emergence [60,146,147,362]. In short, given its core assumptions, developmental psychology is a natural domain of application for the approach of nonlinear, complex dynamic systems. Development, moreover, concerns a complex dynamic system characterized by adaptation in various senses, namely adaptation of the developing individual to the environment, adaptation of the environment to the developing individual (in the sense of education, but also in the sense of long-lasting historical and cultural adaptations as a result of intergenerational effects on development). In short, in order to describe human development, we need three grounding notions: complexity, dynamic systems and adaptation.

Complexity

A Working Definition of Complexity In the context of our discussion of the complex dynamics of developmental processes, we shall use the following working definition of complexity.

- A complex system consists of many components or elements; the magnitude of “many” typically depends on the nature of the system
- The components are interacting; the interactions occur on the basis of a few, simple interaction principles, with a system-characteristic degree of connectedness among the components
- The components change because of their interactions with other components
- And are thus interdependent
- The complex system shows characteristic higher-order properties (exceeding the properties or behaviors of the single components, implying characteristic patterns of related behaviors among many components)
 - Examples are sub-systems, trajectories of long-term change or development, and events at various time scales (see further)
- That are emergent on the interactions, i. e. they occur through self-organization

- Complex systems naturally divide and organize into sub-systems
 - Sub-systems are also complex in the sense of the current definition
 - Sub-systems are defined by the strength of the connections between the components of the sub-system, and these strengths may vary over a broad range, thus allowing the possibility of hard- as well as soft-assembled subsystems
- Complex systems are characterized by patterns and mechanisms of change that occur on various interdependent time scales
 - A characteristic distinction is that between short-term versus long-term processes, that are interdependent, with long-term processes determining the constraints and parameters of the short-term processes, and short-term processes determining the constraints and parameters of the long-term change
 - Patterns of change are characterized by non-linear phenomena such as the emergence of attractors, phase transitions, “tipping point” or “domino” effects, slow and gradual change, surge or peak phenomena, bimodal states and fluctuations, and so forth
- Complex social or human systems are characterized by the embeddedness of the observer
 - The observer is a member of the system he or she studies, and thus makes self-referring statements when explaining and describing the complex system
 - To the embedded observer, complexity often corresponds with various cognitive states, that relate to properties of the complex system, such as ambiguity, fuzziness, contradictions, superposition and entanglement, surprise and so forth, that are not necessarily in principle reducible to unambiguous, crisp, non-contradictory and independent statements or beliefs.

An Application of Complexity to Development

Networks of Interacting and Interdependent Components and Subsystems They depend on the level of organization (and corresponding time scale) or system component chosen. For instance, for a student of language development, the major component chosen is that of language. To understand its development, one must reckon with the fact that language must be subdivided into various subsystems, e. g. the lexicon, syntax, meaning and so forth. Each subsystem, for instance the lexicon, can be further subdivided into components, such as the lemma’s (words) in the lex-

icon. The subsystems interact, but they also interact with non-linguistic systems, such as the members of the community of language users (e. g. the parents, siblings), the perceptual-motor system of the child itself, and so forth. A theory of the development of such collection of elements or components will need to specify the dynamic relationships between them. For instance, across development, the development of language is dynamically related to the development of social understanding (e. g. theory-of-mind) and vice versa. Hence, the developing language of a child is a network of interactions among components. Interdependency means that connected components cannot be treated as independent variables, or independent components. For instance, a child’s current linguistic skill is dependent on its effective environment, in terms of learning opportunities, and its the effective environment is interdependent on the child’s mind (in terms of the language addressed to the child).

In addition to the notion of subsystems, we can also invoke the term “levels of organization”, which refers to levels of particular types of patterns or structures that are stable at their characteristic time scale. An example is the life-span history of a person with a number of characteristic properties, including those of the person’s characteristic life spaces; another example at a shorter time scale is the example of a counting strategy in which a child uses his fingers to count and make simple arithmetic calculations.

In a complex dynamic system like development, all phenomena are interconnected. A major theoretical insight from dynamic systems is that the patterns of action, thinking, or development in the long term, result not from any single factor (plus some additional “noise”), but from the local and temporal confluence of many factors, operating on many time scales [19,271]. In order to study a phenomenon, for instance the development of abstract thinking, or of language, that phenomenon must be isolated for study. However, although it is possible to take a phenomenon out of a complex system, it is not possible to take the complex system out of the phenomenon. That is to say, while isolating a phenomenon for study, the model must explicitly account for the complex system properties from which the phenomenon is taken.

Time Scales Time scales are characteristic durations of processes or phenomena, corresponding with characteristic levels of organization. An example most directly pertaining to development is the time scale of developmental phenomena, spanning developmental events taking place over years or decades and limited by the duration of the human life span

These time scales are

- Scale of biological evolution: practically speaking, this time scale accounts for static constraints, i.e. constraints that do not change across the number of generations that developmental theories customarily address; exceptionally, rapid evolutionary changes can occur based on intergenerational links between developmental processes, such as certain food tolerances under high survival pressure (famine)
- Scale of socio-cultural historical development
 - The constraints and affordances at this time scale change relatively rapidly, it is virtually impossible to formulate an a-historical developmental psychology, i.e. a developmental psychology that takes the historical conditions as a constant; Baltes, Reese and Lippsit [13] made a distinction between normative-historical and normative age-related influences, for instance, the influence of war (normative historical) on a generation of adolescents (normative age-related); other examples are the influence of computers and the internet on young children growing up with them, in contrast with parents who did not use computers at that age.
 - Socio-cultural development can be conceived of as as a Complex Dynamic System in the following way. Its components are agents (many, of different ages, forming intergenerational networks), cultural tools, social organizations and channels of communication and interaction and so forth. The emergent phenomena emerging from the interactions between the agents are: the historical production of the life space, historical production of tools, products for human action, continuous innovation and elaboration of the human life space of human environment or Umwelt, which are emergent phenomena of mass interaction in the social networks of human culture and human society.
 - Scale of development across the human life span: this time scale encompasses changes describable only at the level of the human life span and involves processes such as development in the sense of increasing complexity, skill, and so forth, and processes of aging which refer to losses and negative changes, e.g. decrease of information-processing speed as a consequence of aging. These changes form the topic of the current article.
 - Scale of action and real-time experience: this time scale encompasses actions as goal-driven or intentional behaviors and require the duration characteristic of actions, lasting from seconds to minutes to

hours; the dynamics are explained by means of the theories of adaptive agents (see further).

- Scale of underlying fast neurobiological processes: the time scale of rapid processes in the brain, nervous, motor and visceral system.

Theories of human development differ with respect to the time scale(s) they wish to emphasize. For instance, evolutionary developmental theory explains development from the perspective of innate properties that served the fitness of humans in evolutionary times; socio-cultural and Vygotskian theory of development emphasized the cultural tools available to individuals and that depend on the historical evolution of their society and the actual socio-economic position they occupy, enabling them to use or not use those tools to various extents; Piaget's theory which emphasizes development as a partial life-span trajectory (from birth to about twelve years of age), dependent on processes taking place at the time scale of action.

Self-Organization, Emergence Self-organization is the spontaneous increase in order, complexity or structure, i.e. structure increases not because it is imposed or transferred by an external source, as in transmission (e.g. transmission of certain knowledge items from one person to another through communication). An example is increasing differentiation and integration in cognitive skills and performance across development that occurs spontaneously out of the interactions among all the components involved. It is not imposed on the human life span by the unfolding of a genetic blueprint or by transmission of knowledge and skill from a teacher to an apprentice. Genes, environment and the person's actions are all interdependent components, the interactions of which lead to a self-organization in terms of developmental levels [314,340].

A common theme among proponents of (complex) dynamic systems theory in development is that they view the developmental system as a self-organizing system, showing attractor states, non-linearity in its behavior, emergence and so forth [185,186,188,190,287,314,337,345,348].

Emergence is the spontaneous appearance or evolving of a new property of a system, in the form of a coherence or pattern on a global or macro level of organization. Emergence is related to the viewpoint of the observer (e.g. [76]) and implies a certain degree of surprise from the side of the observer of the system [121,122]. In the field of developmental psychology, emergence relates to the fundamental issue of whether the mature state of human skills, knowledge and so forth is an emergent phenomenon or not. Developmental viewpoints have tra-

ditionally emphasized the emergent nature and have focused on the appearance of (subjective) novelty in a system. A prime example of an emergentist view on development is Piaget's theory. Vygotsky's theory emphasized the emergence of novel forms or innovations on the time scale of cultural-historical evolution through the contributions of individuals and collectives of people. Vygotsky primarily defined ontogenesis as an appropriation by the individual of tools and symbols generated through cultural-historical innovations. The notion of a zone of proximal development may imply emergence, if viewed as an instance of a proper dynamic system [341,352]. Modern theories, such as nativism, assign emergence to the level of phylogeny, i. e. the evolution of the human species leading to heritable biological preconditions for development (as in the concept of core knowledge). Traditional theories of learning equally denied the primacy of emergence for development by emphasizing the role of transmission and appropriation, more particularly transmission of knowledge, skills etc. by instructors [348]. As such, the question of emergence is at the heart of developmental psychology. Notwithstanding the centrality of this question, development originally (see etymology of the term) implied the unfolding of intrinsic properties, i. e. the unfolding or uncovering of what is already there [330,345] whereas emergence implies the coming about of something truly new.

Development as an Increase in Complexity of the Developing System A developing system, for instance a child in a particular familial, cultural and historical context, is characterized by an increase in the system's complexity, often with an asymptotic level of complexity implied as the "final state" of the developmental trajectory. The description of development thus requires a descriptive framework or manifold specifying the space of complexity. Imagine such a space as a multi-dimensional space consisting of all the descriptive dimensions or features needed to specify a distinction between any possible developmental states or levels [330,331]. The simplest possible description entails a single developmental scale or "ruler" specifying the relevant developmental order. For instance, the complexity or structure of a child's thinking and problem solving is given by its position on an ordered scale of cognitive accomplishments, which are often inferred from a relevant content theory. An example is Fischer's theory of iterative embeddings of components, such as single representations of objects or properties and the relations between them, describing an orderly structure of increasingly complex levels [102,103]. A child's developmental level is assessed by letting the child perform a number of actions that map on

the developmental scale at issue, which is mostly done in the form of a standardized task setting, i. e. a test.

A typical and enduring problem of developmental psychology concerns the relationship between the structure of the developmental scale with the structure of the child's "mind". The identification of one with the other – the structure of the test and the associations between test results on the one hand with the structure of the mind on the other hand – is a tacit but common stance for many (developmental) psychologists and is an example of an essentialist and primarily Aristotelian view on the nature of the human mind. It does not reckon with the fact that the relationships involved – those between child, observer, theory and observation instrument – are in themselves an example of a complex adaptive system, and not an instance of a straightforward measurement problem, where a property of an object (a child) is measured by a measurement operation that has no effect on the property being measured.

The expression of knowledge in an activity (nonverbal, verbal, symbolic) is a matter of particular stabilities and patternings of the actions (or expressions). For instance the infant's searching for a hidden object in the correct hiding place in the case of "object permanence", or the older child's verbal communications about perceived events in the case of causal understanding or Theory of Mind, reveal the knowledge in the form of certain stabilities of the pattern of reaching (it is not perturbed by a replacement of the hiding object for instance) or in the pattern of certain verbal explanations, which can take various concrete forms. Because the pattern is a temporal, self-sustaining pattern over a state space of possibilities (e. g. reaching and grasping possibilities; the space of words and grammatical relationships), the knowledge is said to be soft-assembled, i. e. existing for the moment and context in which it is actually expressed [314]. This model is very different from a standard mentalistic one, in which the mind entertains a symbolic or conceptual entity – such as the object concept or a Theory-of-Mind – which is then expressed in the form of or linked to overt activities, which borrow their meaning from the mental contents to which they are attached. In this sense, knowledge has no existence outside its expression in the form of real-time action. The stable patterns that express a particular form of knowledge (e. g. the object concept) emerge through coordinations of many components, part of which are "internal" or linked to the individual, part of which are linked to the world in which the action at issue takes place and are given through perception ([271]). Many of the internal components are non-cognitive in the classical sense, and include goals, concerns and emotions [19,303]. Through

development, the rules and components of the coordinations change, giving rise to actions, reasoning, and so forth that can be described in terms of formal structures that increase in complexity. Such formal descriptions (e. g. that children's thinking involves systems of relationships between components), however, do not refer to the underlying causes of the developing knowledge expressions and skills, but only to their abstract form. They are ways or frameworks for comparing knowledge and skills at various levels of development.

Development as Increasing Complexity Applied to Language Theory and Theory of Language Development As explained above, a theory of a domain of development – for instance language, cognition, ... – serves as a model, eventually as an implicit model, of the state space in which the developmental trajectory is situated. Thus, in order to define what counts as development, the researcher must take a definition or description of the developmental domain and specify changes along developmental time as changes in the quality or quantity of the features that figure in the definition. However, a theory of, for instance, language, need not in itself contain the elements for a complete developmental state space description, i. e. a description of all the possible developmental states relevant to language development. Simply stated, a theory of language can contain a specific description of a state of development, i. e. a mature or ideal state, but be entirely undetermined in terms of the possible paths that lead to this final state. For instance, assume that using passive constructions is a feature of mature language. Given that passive constructions form part of the description of mature language or language per se, the developmental route towards passive construction use is logically confined to a two-step process, namely no-passive-construction followed by passive-construction. This simple from-nothing-to-all switching process vastly underestimates the variety of observable developmental steps leading to passive constructions. That is, it falls short in specifying the potential developmental state space.

Transformational generative grammar, originating in Chomsky's work, made a major contribution to solving this problem by advancing a theory of language that logically entailed a theory of the possible developmental steps towards mature language [6]. By so doing, transformational generative grammar included a description of the developmental state space. Such a description should not be identified with a description of the developmental process per se, since that process can be any of the possible trajectories through the state space. However, if all states of the state space are conditionally ordered, such that for any

state there is only one possible state it can emerge from, the state space description trivially becomes a description of the developmental trajectory, since it is the only one possible. A comparable situation occurs if we take Piaget's theory of cognition, which logically entailed a description of the developmental state space. The possible states are conditionally ordered, and only one trajectory is logically possible (see [330,332,333,334,335,336] also for discussion of developmental models based on Galperin's and Erikson's work). However, if the formal theory of a domain, e. g. linguistic theory, defines – or even simply constrains – the state space for developmental processes, the empirical and theoretical validity of developmental findings becomes conditional on the validity of the formal theory at issue. In short, the developmental findings answer the question of how a developing system reaches a particular developmental outcome, as defined by the domain-theory at issue. However, by studying development as such, it is possible to directly contribute to the formal domain theory itself. For instance, according to genetic epistemology, by studying cognitive development one obtains a better understanding of the nature of human knowledge in general, i. e. of what uniquely defines human knowledge [241].

In a similar vein, by addressing the logical problem of learnability of language as a human competence, Chomsky contributed to the formal definition and specification of that competence [62,63]. The earlier notion of transformational generative grammar has now been replaced by a linguistic theory entailed in the so-called minimalist program, which "... proposes that the computational system central to human language is a perfect solution to the task of relating sound and meaning" [64,181]. A major question is of course what linguistic theory has to say about meaning, how it is structured and what defines its developmental state space. An interesting development in this regard is the work relating catastrophe theory to semantics, thus defining semantics in dynamic systems terms [275,316,369,370]. Meanings are represented in terms of attractor states in morpho-dynamic fields, and such type of representation might be linked to dynamic field theories that were developed in the context of developmental research by Schöner and others [273,273,274,296]. If meaning can indeed be described as a morpho-dynamic field, language development amounts, at least in part, to the dynamic unfolding of this field. Formal theories of how such unfolding can take place can thus form a theoretical basis for a complex dynamic systems theory of language development. Relating meaning and sound, the solution of which is formally described by the minimalist program, leaves the question of why or for what purpose a person would want to relate meaning

and sound in a concrete situation. Transformational generative linguistics, which has dominated the field of linguistics in the second part of the 20th century, sees this question as related to the actual use of language, the “performance”, that is not part of fundamental linguistic theory. This stance relates to the essentially anti-functionalist view of transformational generative linguistics. However, from an evolutionary and developmental point of view, language is the outcome of complex dynamic and adaptive processes, and it is hard to avoid the conclusion that this adaptive evolution has not fundamentally shaped language down to its deepest layers.

Simulation studies and mathematical models of iterative processes in language evolution and acquisition provide evidence that the formal structure of a language is shaped by the dynamics of language transmission and appropriation by individuals, and more particularly, by its use in social interaction [48,65,66,168,169,229,231,290].

An important feature of these models is the very close dynamic interaction they propose between learning, culture and biological evolution. Over the course of language evolution, these three components transform each other in a process generally known as co-evolution. In that sense, human biology is deeply transformed by human culture and vice versa. For instance, the biological pre-adaptation for language acquisition – in whatever form one wishes to specify it – is the result of a dynamic systems process occurring over the intergenerational (i. e. evolutionary and historical) and intra-generational (i. e. ontogenetic) timescales. In that sense, the dynamic systems approach can help explain – in principle – how language structure emerged through self-organization over the course of generations [165,228]. The evolutionary processes that have shaped language were modified by the fact that they had to pass through the highly specific constraints and opportunities of transmission and appropriation actions in individual agents. They have resulted in language being an essential aspect of the psychological life space of individuals and being appropriated in an extremely rapid and robust way, given the complexities of the task of language development. The question is how language functions in the psychological life space of individuals, and thus, under which constraints and opportunities language develops.

A classical theory in that regard is Vygotsky's, whose work is still of theoretical importance to the field of developmental psychology. Vygotsky saw language as a complex set of evolutionarily and historically developed tools, that individuals use to act with and solve problems [328,361]. Hence, language development can at least partly be understood in accordance with the dynamics of tool use and its development [200]. This view leads to the

idea that language is a cognitive niche or a material scaffolding structure that the child and its environment construct during the developmental process, as an additional dimension and partition of the psychological life space (e. g. the name of an object as a feature of the object, relating the object through the linguistic relationships of the word; linguistic forms as objects of action in themselves, related in complex ways to other features of the complex world, an issue which relates to the so-called grounding problem [24,68,69,70,71,75,357,358,375]).

Development and the Embeddedness of the Observer As explained earlier (in the section on developmental level, order and structure), the assessment of a child's developmental level is not a simple measurement issue, of a determinate property (the developmental level) being tapped by an otherwise neutral measuring instrument (a test or observation). For instance, Elbers [92] showed that children bring specific expectations of answers to particular questioning situations and use the (non-)reactions of the test administrators as information, thus turning the alleged test or measurement situation into a social dynamics. For a young child, the test is an educational situation, and the child will react to the situation on the basis of the expected consensual frames [107]. From a measurement-theoretical point of view, this is not a trivial issue. The adult's intended act of measurement co-determines the measured content in a direct and objective sense, and there is no measurement outside the context in which the adult's probing changes the probe. Exceptions are observations in natural, free contexts, e. g. observations of spontaneous behavior as in language developmental studies, which are observations of relatively unconstrained social interactions. However, the result of these observations should also not be seen as measurements of the “true” level of some developmental property, but as observations of the dynamics of social interactions, involving the dynamics of what children already know and are able to do in an environment that is adapted and reacts to the child's possibilities and, equally important, anticipates on the child's growing potentials.

Another fundamental issue, which is typical of complex human systems, is that in the field of human development, ontological and epistemological issues become entangled. The researcher's view on the cognitive development of a child to an adult, for instance, requires an ontology (a theory of being, in this case of the mental contents, knowledge etc. of the developing persons under study) of a process that ends with an epistemology, i. e. a theory of how humans know, including the researcher. The researcher's level of understanding of the cognitive develop-

ment of children towards adults is in itself the endpoint (roughly speaking) of his own, personal developmental trajectory.

For most developmental scholars, there is no problem with this entanglement, if any such entanglement is observed at all. For instance, a cognitive developmental theorist, working in the spirit of Piaget's stage theory and using his own mature, formal operational thinking as a tool for understanding the world, can fruitfully and without any internal inconsistency, study the emergence of formal-operational thought starting from its roots in the baby's non-operational, actional and sensorimotor way of understanding the world. However, the understanding of the developmental process will thus be determined by its endpoint, also because that is the tool with which this understanding is accomplished [330]. If we assume that development continues, also because scholars invest in making historical change processes come true and contribute to transforming systems of understanding, the representation of the process of development by developmental scholars will shift, as their own developmental levels (forms of scientific understanding) shift over the course of their lifetime, or over the course of historical time.

Dynamic Systems

A Definition of "Dynamic System" Dynamic systems theory is an approach to the description and explanation of *change*. A simple definition is "a means of describing how one state develops into another state over the course of time" [365], which can be expressed mathematically as

$$y_{t+1} = f(y_t) \quad (1)$$

expressing that the next state (at time $t + 1$) is a function f of the preceding state, at time t . In a slightly different notation

$$\Delta y / \Delta t = f(y). \quad (2)$$

The equation states that the change of a system, denoted by y , over some amount of time, denoted by Δt , is a function f of the state of y . The function f is also referred to as the evolution term or evolution "law". That is, it is important that f specifies some causal principle of change.

An important property of the current equation is that it represents recursive relationships. Thus, y_t leads to y_{t+1} , and according to the same principle, y_{t+1} generates y_{t+2} and so on.

A system can be described as a set of entities that are related to one another and that influence one another, and a state of the system is the set of properties of its compo-

nents at any particular moment in time. The properties of the system are expressed in terms of dimensions or variables, for instance the variable y from the preceding equations. Dynamic systems can consist of any number of such variables. For instance, if y represents a child's current lexical knowledge, and z represents the child's knowledge of syntactic rules, the dynamic system consists of two dimensions, and the child's current developmental state is a point in this two-dimensional space. The space of developmental dimensions is the developmental state space, and development can then be defined as a trajectory across the developmental state space.

Properties of Dynamic Systems Applied to Development

Iterativeness The iterative or recursive nature of a dynamic system refers to the map or flow that the system instantiates, and which, in qualitative and metaphorical terms can be rephrased as "explaining after by before" [352]. The application to development is – apparently – trivial, in the sense that a developmental process is a process that transforms a current state to arrive at a new state, the "state" being any point or region in the developmental state or phase space, as defined earlier (or alternatively, the symbol string corresponding with a set of properties describing the current properties of a developing person in a co-developing environment). Examples in classical developmental theories are Piaget's notion of assimilation, implying that the representation of given information is a function of the current forms of understanding and representing, or Vygotsky's notion of the zone of proximal development, in which the next attainable level of development is a function of the level already consolidated.

Dynamic Rule/Principle/Function The dynamic rule or function describes the way a current state of the system is mapped onto, or transformed into, another state at some later time. In principle, this dynamic rule corresponds with the basic mechanism of development that a developmental theory entails. In principle, this basic mechanism is any causal mechanism that operates on the current developmental state and that brings about a particular change of that state, including a 0-change, which occurs if the developmental state has reached stability.

Classical developmental theories were usually explicit about the mechanisms operating on developmental states. For instance, for Piaget, the mechanism is one of adaptation, with constituent mechanisms of assimilation and accommodation. The working of these mechanisms is ex-

plicitly determined by the properties of the current developmental state. For instance, the child assimilates the information it obtains from its actions to the cognitive structures it currently maintains. The same holds true for accommodation, which is the driving force behind developmental change. A similar logic applies to the fundamental Vygotskian notion of the zone of proximal development, which implies that a more competent person adapts, in terms of help given, to the current level of development of an apprentice, and by giving this level-adapted help, stimulates a process of interiorization in the apprentice that leads to a new and higher developmental level. Other examples are Werner's notions of differentiation and integration, which are mechanisms explicitly operating on the cognitive, behavioral and emotional structures that are present in a child (or in a child-environment system).

Providing such models of mechanisms boils down to specifying an implicit function for development, i. e. given a state so-and-so, application of the mechanisms or rules will result in a different state, and through iteration or recurrence, to a developmental process. In principle, the mechanism or developmental function itself implies not only the present developmental state of the system, but also any influences – coupled to or independent of the state – that are incorporated by the mechanism and that moderate development. For instance, a mechanism explaining the growth of a child's lexicon not only operates on the current state of the lexicon, but also implies external influences, such as the environment's lexicon, teaching activities and so forth.

The majority of modern developmental studies have replaced these implicit developmental functions by explicit ones, which are based on samples of independent subjects. For instance, the time dimension, which is a fundamental kinematical variable for specifying change, can be used in an explicit developmental function, assigning a developmental level (of whatever kind is required) to a particular value of developmental time (most particularly age, but also duration of experience, for instance). Although these explicit models claim to achieve generality (or as it is often called, generalization), they achieve this result at a devastating cost. The cost is that they bypass the actual process, and are in fact completely ignorant as to which causal mechanisms explain actual developmental processes, which are processes applying to concrete, individual systems. For instance, lexical growth is a process occurring with a particular child in a particular environment, which is to a considerable extent dependent on the child's actions. For instance, the linguistic environment tends to adapt itself to the language use and understanding of the child. An explicit model provides a model

of the lexicon as a function of a given time (age), other variables such as maternal language proficiency and an additional stochastic influence. The model is achieved by averaging over many individuals, and thus risks to lose all information about actual processes (unless the processes are virtually uniform over all individuals, which is rarely, if ever, the case). This problem, which refers to the fact that such explicit models of associations between independent and dependent variables are not in any way logically related to implicit models describing mechanisms operating on developmental states, has recently come under the attention of statisticians and methodologists working on individual developmental trajectories [136,216] (e. g. Molenaar, Hamaker, ...). The reason why such implicit functions are central to developmental science is that they attempt to specify the causal processes that operate in real time, and thus provide models or prescriptions of actions by practitioners.

Attractor An attractor is a set of points in the phase or state space towards which a system will evolve, given its dynamic function, if it is in the basin of attraction.

The application of the notion of attractor to a developmental system implies that the system will asymptotically evolve towards a particular state if it is under a particular set of conditions (the basin of attraction). More precisely, if the system is in an attractor state and gets perturbed, it will spontaneously return to the attractor state, unless effectively counteracted by some external force. For instance, a particular child, in a particular development, will tend to show certain stabilities or stable patterns in its behavior. The criterion of return after perturbation, under certain limits concerning the strengths of the perturbation, is an important criterion for distinguishing attractor states from accidental states for which the system has no particular "preference" (if any such states exist). Many developmentally relevant attractor states are likely to be rather idiosyncratic, i. e. dependent on individual and local circumstances. Other states may be relatively general and predictable on the basis of broad criteria such as age. An example of the latter type of state are the classical developmental states or stages, for instance the states distinguished by Piaget, by neo-Piagetian scholars, stages distinguished in the progression towards mature language and so forth. An example of a clinical application of the concept of attractor is the notion of resilience, which refers to a child's capacity to spontaneously move back to a healthy psychological state after having experienced highly stressful or adverse experiences. Although resilience can easily be defined as a personality property of the child or as an individual capacity, it is probably more realistic to view

it as a an attractor state of the developmental system in which the child is embedded [209]. A comparable example is the emergence and self-sustainment of highly problematic teaching-learning patterns in children with developmental disorders such as ADHD [353].

Developmental State Space and State Space Grids The use of dynamic systems theory to development applies a fundamentally geometric way of thinking to the study, description and explanation of development. Developmentalists are used to thinking in terms of variables that they identify with psychologically real properties of the mind, which implies that their frame of reference is a model of the mind or the brain, i.e. a model with a topology that is similar to the topology of the mind or the brain (since the topology of the mind is difficult to imagine, and the topology of the brain is at least seemingly given by modern brain imaging studies, developmental psychologists are increasingly turning towards the brain as a model; however, see [50] for a critical discussion). From a dynamic systems point of view, however, a developing system is a geometric manifold or space, consisting of all the dimensions used to describe the system (and this number can eventually be very small), including the evolution laws or rules that specify the change of positions (developmental states) in this space. There is no implicit reference to topological similarity with the mind or the brain. The use of such geometric, state space descriptions can free the researcher from “unsolicited” ontological claims, i.e. implicit claims about the nature and composition of the human mind that are relatively standard in mainstream developmental psychological investigations [350].

The notion of state spaces, in particular categorical state spaces, has been promoted by several researchers, mainly working in the field of social interaction and social development [127,128,129,148,149,191,192].

Static Versus Dynamic Models in Development Developmental psychology, has almost exclusively focused on *static* models and has implicitly assumed that change, for instance developmental change in an individual, could be approximated by stretching static relationships over the time axis [352]. A characteristic expression of a static relationship takes the form

$$y_i = f(x_i) \quad (3)$$

with y a dependent variable and x an independent variable, which, for any possible value x_i generates a corresponding value for the dependent variable y . The variable x can also be time, but the use of time as such does not turn the model into a dynamic model.

A difficulty arising with this definition of a static model is that any dynamic model that is expressed as a function of time

$$y_i = f(t_i) \quad (4)$$

must strictly speaking be qualified as a static model [300]. Hence, we should confine static models to those where the x -variable is not time. However, in Sect. “A Definition of ‘Dynamic System’” it was claimed that the f in the dynamic equation must specify some causal principle of change, with the implicit assumption that this causal principle, however general, is theoretically justified and based on what we know about how things change. Hence, a model of the form of Eq. (4), $y_i = f(t_i)$, that applies to an empirical sample (e.g. a regression model of time applying to a cross-sectional sample, or a sample with two or a few consecutive measurements), can easily be transformed into a model of change by taking the first derivative of the model. By doing so, however, one does not automatically arrive at a meaningful dynamic model, since the function term f featuring in a descriptive time-serial model is not necessarily *descriptively adequate*. A dynamic model can be characterized as developmentally descriptively adequate if the mechanism implied in the dynamic model (1) corresponds with a developmentally plausible mechanism, (2) in principle applies to the whole developmental time scale of the developmental phenomenon in question [193].

A static system describes a particular value of the variable of interest as a function of the value of another variable (or set of such variables). For instance, for any possible age, or for any level of the mother’s lexical knowledge, or for a combination of age and maternal lexicon, the static system or model will generate a predicted or expected size of the lexicon, without any reference to recursiveness.

This distinction between static and dynamic type models has considerable consequences [152,300]. Whereas a dynamic model recursively generates a time series (a state and the next state and the next ...), a static model generates a sample or population of individuals that are in principle independent of one another (an individual with age i and lexicon i , an individual with age j and lexicon j , and so forth). Statements about associations between variables across populations do not necessarily apply to the mechanisms that apply to change in the individuals in the population. However, the behavioral sciences, including developmental psychology, often implicitly take a relationship between variables that holds across a sample as a representation of some dynamic rule or principle (also known as the homology or ergodicity error) [136,216,224]. For example, a study showed that early math skills in 5

to 6 year olds have the greatest predictive power for later school achievement, whereas socio-emotional behaviors, on the other hand, had little or no predictive power, irrespective of gender and socioeconomic background [90]. From such findings, it is easy to infer that increasing early math achievement, e. g. through preschool teaching programs, will thus lead to better school achievement at a later age, implying also that attempts to increase socio-emotional skills should be reduced since they do not relate to academic achievement. However, there exists no logical or direct relationship between the static relationship (how is it associated across a population) and the dynamic relationship (how can something be increased or decreased in individuals).

Adaptation and Adaptive Agent Systems

Adaptation is the process of adapting something to something else, usually in the context of an organism adapting to its environment. An important question concerns the relationship between adaptation and development. There is no doubt that adaptation and complex adaptive systems play a major role in the process of development (for further discussion, see Sects. “Theory Of Complex Adaptive Systems (CAS): Developmental Agent Models” and “Theory Of Complex Adaptive Systems (CAS): Epigenetic Robotics”). A classical theory such as Piaget’s conceived of development as a process of adaptation, which shows a clear pattern of increasing complexity. However, adaptation, especially in the sense of organism-environment adaptation, need not be a process of increasing complexity, adaptation can also mean loss of specialization, decreasing complexity etc., if the latter is better adapted to the organism’s current environment. In developmental processes, there are also processes of loss of knowledge and of complexity, depending on changes in the environmental circumstances (e. g. language loss [176,234]). An encompassing theory might claim that an adaptive process as it applies to a growing organism (literally as in embryogenesis, metaphorically as in cognitive or language development), must by necessity show an increase in complexity (and size, etc.), given the constraints on reproduction (reproductive activities produce an offspring that is less complex (of smaller size etc.) than the progenitor).

The theory of complex systems refers to complex adaptive systems, which are collections of interacting components (agents) that adapt to each other. A developmental system, e. g. a child in his or her environment, can be conceived of as a complex adaptive system, with the agents or components adapting to each other. Examples of mutual adaptations are given in dynamic reinterpretations of Vy-

gotsky’s theory [341,352] or in the theory of transactional development [263,265].

An adaptive system is not necessarily a goal-driven or teleological system. Human beings and organisms in general, however, are complex adaptive systems that are goal-driven or teleological, and in addition to adapting to their environment, they also wish to control their environment [77,167]. Developmental theory that explains the mutuality between the long-term time scale of development and the short-term time scale of action, thus needs a theory of adaptive, goal-driven or concern-driven agents in order to explain the level of action, and the developmental level of changes in goals and concerns [302,303] (for further explanation see Sects “Theory Of Complex Adaptive Systems (CAS): Developmental Agent Models” and “Theory Of Complex Adaptive Systems (CAS): Epigenetic Robotics”).

Main Theories in the Field and Short Historical Background

Founding Historical Theories

A brief look into any arbitrary collection of handbooks on developmental psychology illustrates the field’s historical concern with the question of whether development implies the unfolding of what is already given at birth (which refers to the original meaning of development as unwrapping) or whether development implies a start from zero and a process of appropriating of whatever is necessary to become a mature person. The main figures embodying these standpoints are John Locke (1632–1704) and Jean-Jacques Rousseau (1712–1778) respectively. Although it is no longer stated in this naïve form, theorists and researchers struggle with this issue even at the present day (see for instance the discussion on gene-environment relationships; [261]). Biologically-inspired theories – seeing development as the unfolding of a biologically given program – are associated with the work of Charles Darwin (1809–1882) and G. Stanley Hall (1844–1924). Historically, however, developmental psychology is based on the confluence of many theoretical strands.

Main Theoretical Viewpoints

Biologically Inspired Explanations The notion that the important components or aspects of the human condition in its mature form are in fact innate and not appropriated thanks to external influences, or constructed on the basis of one’s own action, has received a major impetus by the work of Noam Chomsky in generative linguistics. Chomsky argued that language – which is obvi-

ously learned from the input received from the environment, in that no child will for instance learn French if confronted with a Dutch language environment and input – cannot be learned or acquired without an innate language acquisition device, which defines the major properties or degrees of freedom, of human language (see the section on Language theory and theory of language development nativism and modern nativism for further discussion [62,63]). The major argument was that language is in fact underdetermined by the input if the learning device has no preset clues about what the input means or implies in terms of structural relationships among identifiable components. A similar line of thought is followed by proponents of *core knowledge theory*, such as Elizabeth Spelke, who claim that the major components of human knowledge about the world – such as the notion of space, number, causality and so forth – must be innately given (see [295] for an overview).

Evolutionary developmental psychology [32,235] is a theory that applies Darwinian principles of evolutionary adaptation to explain the evolutionary emergence of epigenetic programs that evolved under specific selection pressures [33,96,115,194,195]. Examples include early fantasy play, parental investment and cognitive development. Since the epigenetic programs evolved under historical environmental conditions that are no longer present in contemporary environments, mismatches between such programs and the requirements of contemporary life may lead to perturbed development. Evolutionary developmental psychology can be seen as an offspring of ethological theory, a biological theory that tries to understand the adaptive functions of behavior of an organism. Ethological theory, primarily through the work of Konrad Lorenz, has given rise to theories about critical periods, i. e. specific ages at which the development of a particular category of action or skill, such as language, is particularly stimulated and beyond which that development cannot take place. Modern theories of development tend to speak about sensitive periods or windows of opportunity, and eschew the notion of critical periods and the impossibility of development if the period, for some reason or another, is missed [8,49,159,319].

Developmental behavioral genetics attempts to explain development and particular developmental trajectories on the basis of the person's specific genetic endowment [244,245]. Recently, major progress has been made in the study of the effects of “generalist genes” on development [246,275,276]. In the context of learning disabilities, Plomin and Kovas describe “generalist genes” as genes that affect not only the disability but also the normal variation in the behavior at issue (e. g. reading), that affect all aspects

of the disability and related disabilities (or normal behavior) and not just a particular aspect.

The influence of genes is not unidirectional: genes and environments are to a considerable extent linked with transactional relationships. Specific genes act so as to make the person more sensitive to or selective to particular environments, whereas environments have an influence on the activation of particular genes or moderate the effect of genetic influence [5,261].

Psycho-dynamically Inspired Theories Psychodynamical theories originated in the work of Sigmund Freud (1856–1939). According to Freud, human behavior and action are determined by energetic forces resulting from basic human drives, which, for Freud, were primarily sexual in origin. The actions required for fulfilling these drives conflict with the exigencies of reality, and actions serve to resolve this perennial conflict, leading to particular psychological structures (such as the unconscious, the Id or superego, etc.). The principle of drive-determined conflicting actions operates from birth on, and leads to a series of psycho-social stages (the oral, anal and genital phases, from birth to about 7 years). Although Freud can be criticized for his narrow view on the importance of sexual drives for the explanation of human behavior and development, his theory is one of the few that actually tries to understand human action and development from the goals and concerns that human beings try to accomplish, and from the problems they encounter in doing so.

In modern developmental psychology, psychodynamic theories mainly survive through the work of Erik Erikson (1902–1994) who emphasized the social, cultural and environmental aspects of the conflicts between the person's drives and the challenges of the environment. Erikson saw the human life cycle as a sequence of basic conflicts, arising out of the demands that culture and society make on the growing individual and that depend on that individual's psychological and biological maturation. These conflicts, for instance the conflict between identity versus confusion which is typical of adolescence form a conditional sequence, in that the solution of an earlier conflict is a condition for the way in which later conflicts will be manifested and solved [115].

Socio-culturally Inspired Theories These theories find their origin in the work of Lev Semenovich Vygotsky (1896–1934) who developed his theories during the early stages of the development of the Soviet state, which may help account for the particular properties of this theory [251,328,359,360,366]. For Vygotsky, development is a process that builds upon the biological predispositions

given by biological evolution and that consists of an appropriation of cultural tools for action, including mental action. This appropriation of cultural tools – skills relating to tools in the literal sense as well as symbolic tools such as language and historically developed concepts – leads to the involvement of the person in his culture and is a requirement for his ability to contribute to the further development of his culture and society. Vygotsky emphasized the intertwining of processes on distinct time-scales, such as the historical time scale of societal evolution, the developmental time scale of the human life span and the short-term time scale of human action and tool use. Vygotsky's primary developmental mechanism consists of a combination of principles. One is that of interiorization, which states that children can interiorize or appropriate the actions and skills they first perform with the help of more competent others, such as parents, teachers or more competent peers. The second is that of the zone of proximal development, which refers to the distance between what a child can accomplish on his own and what it can accomplish with the help of others, such that interiorization is likely to take place and thus, development is further advanced. Development of an individual child takes place in direct and intensive interactions and transactions with other people, and is a deeply socially embedded and social process.

To Vygotsky, development of higher levels of thinking, including abstract thinking, is made possible by the appropriation of language-grounded concepts, such as time or causality, or names for numerical digits, and by the availability and use of physical symbolic systems such as writing.

The principle of social embeddedness in the culture has been used by various authors who have extended Vygotsky's work (examples are [255,324,367]). The theory is now more widely known as the socio-cultural theory of development, thus emphasizing the two main forces acting on and shaping human development.

Because of the emphasis on the mechanism of development and the direct causes of change, Vygotsky's major developmental principles lend themselves relatively easily to assimilation in a dynamic systems or agent-based framework [300,340,343,344].

Piaget and Cognitive Approaches to Development For Jean Piaget (1896–1980), cognition is a particular biological adaptation that allows the human organism to control and predict the environment through understanding and mental models. However, being a biological adaptation, it is not innately given but must develop, and it will do so on the basis of primarily biological mechanisms

that operate on the child's understanding of the world. These mechanisms are summarized by the term adaptation and involve processes of assimilation and processes of accommodation. It is through these processes that operate under the form of human activities that the child actually constructs his understanding of the world, in a series of stage-like steps, leading to a level of understanding that is self-sustaining under its interaction with the environment. Assimilation involves the transformation of information to the form of the cognitive structures already present, whereas accommodation refers to changes in the cognitive structures to accommodate the information given. Assimilation and accommodation form a dialectical pair of forces, allowing the child to move beyond the direct sensory and motor givenness of experiences, and to come to abstraction. Cognitive structures are interrelated structures of schemes, concepts, skills and so forth, which, through their interrelationships are self-sustaining, at least for a given period of time, after which the working of the adaptive mechanism transforms them in new structures. These sequential structures correspond with Piaget's main stages, the sensorimotor, pre-operational, concrete-operational and formal operational stage. Cognitive structures are constructed, and are thus characterized by structural properties that are not or cannot be inductively inferred from the information given. A core example of such a structural property is the property of reversibility, which develops around the age of about six years and which marks the transition from pre-operational to operational structures. Reversibility is the property that assigns a reversible operation to any operation the cognitive system may have or develop, and it does so by implication. With reversibility, cognitive systems obtain properties of mathematical groups of operations, which greatly increases their power, enabling them, among others, to take their own operations as topic of reflection.

As with Vygotsky, Piaget was mainly interested in the nature of the developmental mechanism and considerably less so in the actual trajectories of development (although the textbook-representation of Piaget focuses mainly on his theory of stages and only superficially on the developmental mechanisms; the reason for this skewed focus is that modern developmental psychology is more a collection of facts about age differences in phenomena and associations of variables within samples than a science of development that primarily addresses the developmental mechanisms).

Piaget's theory was not so much inspired by the wish to learn about the developmental trajectories followed by children than by his interest in the nature of human

thought and understanding, which he tried to capture by studying their development. Piaget was first of all a genetic epistemologist, interested in the nature of human knowledge. For Piaget, developmental science is an intellectually reflective science, i. e. a reflection on the nature of the thinking that makes that reflection possible. This interest puts Piaget in the ranks of complexity theorists who view complex systems as systems in which the observer and knower of the system himself is embedded [304].

Because of his focus on the mechanism of development, that is on the process of change itself, Piaget's basic theoretical notions lend themselves naturally to dynamic systems modeling [343,344]. Piaget's notion of development as construction of cognitive systems is closely related to dynamic systems notions of self-organization. Given the constraints and degrees of freedom present in the organism and in the environment, cognitive systems corresponding to the major developmental stages are self-organizing structures that develop towards equilibrium, i. e. they are self-sustaining, until, through the accumulation of experiences and most notably cognitive conflicts, a critical point is reached at which the structure tends to change rapidly towards another equilibrium, through a cascading process of inter-related events of change.

Neo-Piagetian theorists started from the major assumptions of Piaget regarding structural relationships in cognitive systems and the constructive nature of development, and added principles from modern cognitive science, such as biological maturation and brain development, limited working memory and other constraints on information processing [58,102]. Neo-Piagetian theorists have also employed dynamic systems modeling to explain processes of cognitive development, including stepwise growth and temporary regressions preceding developmental accelerations [103].

In addition to Piaget's theory, the theory of information processing has made an important contribution to developmental science. Information processing theory and its recent offspring in cognitive science and neurocognitive science starts with a basic model of human information processing, containing components such as input and effector components, short- and long-term memories and so forth, and studies development as the changes in those components. Changes concern quantitative changes in the operation of the components, for instance processing speed or size of working memory, but also qualitative, i. e. content-related changes, which amount to changes in the information-processing rules [171,210,279]. Information processing has been dynamically modeled by simulation architectures such as ACT-R (adaptive control of thought-rational [3,4,161]).

Learning-Theoretically Inspired Approaches Learning theory is concerned with how experiences shape future behavior and focuses on the addition of new knowledge and skills to what is already present. Classical learning theory remains close to the observable properties of behavior, e. g. under which environmental conditions particular behaviors occur. Learning and behavior are seen as being under the control of the environment. For the explanation of development, two main principles of learning are important. The first is the principle of contiguity and stems from the field of *respondent (or classical) conditioning*, which is mainly associated with the pioneering works of I.P. Pavlov (1849–1936). A stimulus that is contiguous (immediately precedes and overlaps) with a stimulus that evolves a particular response, will automatically obtain the response-retrieving properties of the latter. The second is the principle of functional or operant learning and is mainly associated with the work of B.F. Skinner (1904–1990). Behaviors have different consequences, and if the consequence of a behavior is positively evaluated by the organism, its future frequency will increase. Under the influence of the particular context, every behavior shows a spontaneous variation, and this variation allows for selective consequences, which may alter the frequency of the behavioral variant at issue, which on its turn will also vary spontaneously and be selectively reinforced. Through this principle of *operant conditioning*, behaviors can be shaped towards entirely new forms by applying reinforcements to each variant that comes closer to the goal behavior than other variants [283,285]. The principle is very similar to “survival-of-the-fittest” principles acting in evolutionary biology. If developmental processes must be seen as a result of many such learning processes over the long-term, it is likely that such processes will interfere, for instance in terms of contexts and reinforcements, and thus that complexity principles, involving networks of many interacting components, will apply. The question is to what extent non-linearities, attractor states and phase shifts may arise under these conditions, although principles of operant learning can be very easily transformed in a dynamic systems format [338]. The interaction of many levels of reinforcement is addressed in learning theory with relation to the matching law [142,143]. The *Matching Law* holds that the proportional distribution of behaviors (e. g. how much on-task versus off-task behavior in a child during a math lesson) will evolve towards a value that reflects the reinforcements provided by these behaviors (e. g. how pleasurable or goal-effective, on average, is on-task versus off-task behavior). The Matching Law explains the emergence of what in dynamic systems terms is called a specific attractor state and can be used to explain developmental

changes (see [85,301] for an application in the field of social interaction; [30,278,286] for an application in teaching-learning processes and [38,223] for applications with developmental disorders).

Finally *social learning theory*, which originated with the work of Bandura, shows that children learn by imitating other people. Imitation depends for instance on the effect that the imitated behavior has for other people, on the social power and status of the person imitated and on beliefs on whether the imitating person will be able to accomplish the imitation or not (self-efficacy) [1,14,16,52,124,131,132,260]. Imitation tends to be a process that is far more complex than mere mimicry, and even in very young children extends to relatively abstract properties of the imitated behaviors, such as the intentions of the person who is imitated [53,57,320]. Thus, even with a seemingly simple act as imitation, one needs to conceive of the imitating person as a complex system, operating on many levels at the same time. The principle of modeling and imitation, which is maybe better referred to as behavioral contagion, can be fruitfully used in dynamic systems models of social behavior in children, in a developmental context [302,303].

Systems and Dynamically Inspired Theories Developmental systems theory, as it is applied to developmental psychology, explains development as a transactional process, implying transactions (that is, mutual transformations) of individuals and their contexts or life spaces, or of biological and environmental influences [108,184,264]. The theory has its roots in developmental systems theory as a specific approach to evolutionary biology [125,128,194,195], and the theory of nonlinear, complex dynamic systems, which is the focus of the current chapter.

Dynamic systems theory of development is a general approach, relying on the basic definition of a dynamic system as a system explaining the change of a variable, or why a next state follows from a preceding state, for “state” defined as any value on one or many variables describing the developmental state space (see [265,348,352,365] for general introductions in the field of development). In developmental psychology, the term “dynamic systems theory” has a somewhat confusing meaning. It refers to a theory of development based on processes of physically and socially embedded and embodied action, which originates from the work of Thelen and Smith [314]. Another approach is inspired by ecological models from biology and applies principles of resource-dependent change and mutual relationships of competition and support between components of the developing system, and is based on the work of van Geert [337,339]. Finally, dynamic systems theory refers to a variety of applications of general concepts from

dynamic systems theory, such as self-organization, attractors, state space, etc. to developmental phenomena, for instance in the work of Lewis, Granic, Hollenstein, Fogel, Dishion and others (these theories will be discussed later).

Content-Driven Approaches A quick glance through the scholarly journals in developmental psychology shows that the field as a whole is theory-poor, in the sense that well-established, non-trivial and fundamental theories of developmental change that lay the groundwork for our understanding of developmental processes are virtually absent from most of the empirical work. Developmental psychology is primarily a science of relatively isolated fields or themes of development. Examples are attachment, theory-of-mind, aggression and bullying, and so forth. The standard research approach is to measure a collection of variables over samples of children of various ages, describe the change of the dependent variable based on averages for the distinct age groups and show how the dependent variable is associated with a host of independent variables (for instance, changes in average levels of theory-of-mind scores or reactions in experimental situations, which are associated with measures on language, intelligence, socio-economic status and so forth).

Conclusion

In its present state the field of developmental psychology is a scattered collection of approaches and fields that do not unite into a common framework or emerge from a common framework explaining the fundamental mechanisms of developmental change. The field lacks a generally accepted notion and theory of development in general, except for relatively trivial collections of principles taken from various (historical) theoretical approaches, such as Piaget’s theory, learning theory, knowledge about biological underpinnings of development and so forth, that form the inevitable introductory part of the field’s basic textbooks. Theoretical unification may take place by liberally employing principles from dynamic systems theory, theory of complex systems and of complex adaptive systems.

The Study of Individual Development in Light of Complexity Theory

Theory of Development and Complex Dynamic Systems Models

Development and the Theory of Embodied Action According to Thelen and Smith [314], current psycholog-

ical theories tend to invoke “ghostly” things to explain behavior, namely internal representations and concepts. The representationalist stance, also known as the computational-representational understanding of mind, or information-processing theory, states that thinking can best be understood in terms of representational structures in the mind and of computational procedures that operate on these structures [311]. That is, in order to explain thinking you need internal entities. For instance, if a baby watches an object being hidden by an adult, and then watches how the object is moved to another hiding place, the baby will still look in the first hiding place in an attempt to retrieve the hidden object. This so-called A-not-B error is traditionally explained by the absence of a fully developed object concept, i. e. an internal representation of the object concept [287,292]. According to Thelen and Smith, invoking the notion of “concept” to explain a particular behavior or action related to that concept, is a categorical error, as if one explains the color of the red traffic light by the workings of its inherent redness. It is this categorical mistake that Thelen and others seek to repair by explaining phenomena such as the A-not-B-error by a theory of situated action, in which an embodied subject (not an epistemic subject, such as in Piaget’s case) acts with the help of and under the constraints of a physical world that includes the external environment and the physical properties of the body [292].

In essence, cognition, thinking and action are explained as dynamic patterns unfolding from the continuous, “here-and-now” interaction between the person and the immediate environment. A particularly clear description, in the context of cognition and intelligence comes from Linda Smith [287]:

The embodiment hypothesis is the idea that intelligence emerges in the interaction of an organism with an environment and as a result of sensory-motor activity. The continual coupling of cognition to the world through the body both adapts cognition to the idiosyncrasies of the here and now, makes it relevant, and provides the mechanism for developmental change. (p. 205)

The dynamic system at issue is the continuous coupling between the organism and its environment. This system shows a short-term time-evolution that takes the form of intelligent action, which changes the body and the brain through processes of learning and adaptation, thus giving rise to a long-term evolution we call “development”.

In order to explain the short-term dynamics of action, thinking and knowledge in concrete contexts and the long-term dynamics of development, including the inter-

dependencies between these two time scales, Thelen and Smith invoke the notion of dynamic fields, which will be explained in a later section.

Development and Resource-Dependent Competition-Support Systems

As for any complex system, a developmental system can be viewed as a collection of elements or components. These components are related through functional relationships, implying that one component can change another, and vice versa. The system is embedded in an environment with which it is also functionally related (it can affect the environment and can be affected by it). The notions of “system” or “environment” refer to distinctions arbitrarily made by the describer of the system. For instance, a developmental system can be defined as a single variable or component, for instance a particular child’s social cognition, or a child’s lexical knowledge. This single-variable system then defines an environment consisting of any other component that functionally affects the component at issue. For instance, for the social-cognition system of a child, the child’s emotional repertoire, intelligence, linguistic knowledge all form part of that system’s environment, in addition to components that are not internal to the child, e. g. The child’s peers, family, culture etc. In this sense, the notion of environment does not refer to “environment” in the usual sense, namely the child’s current life space, family environment etc. (although such components do belong to the system-dynamic definition of environment). A system consisting of two variables, for instance a child’s lexical knowledge as one component and the language addressed to the child by the mother as the second component, is embedded in an environment that consists of all sorts of internal and external phenomena that are related to the two variables at issue (such as the cognitive systems of the child and of the mother, their emotions, but also the material and cultural artifacts of their homes, other family members, etc.). Such developmental systems can be easily extended towards any number of explicitly defined and related components, which then define a complementary system set, which is the environment of the system in the abstract sense just introduced.

If the components through which a developmental system is defined can be described as (roughly) numerical variables, the system can be treated as and modeled as a dynamic ecological system. That is, if the system’s components can be described by means of variables that specify the level of the component, the system dynamics amount to relationships affecting those levels. For instance, a child’s lexical knowledge can be described as a particular number of words actually known or under-

stood by the child. The child's social cognition can be described by means of a variable that distinguishes possible levels of development of social cognition (see Fischer's notion of developmental rulers, capturing the same idea [105,257,350]). This representation by means of numerical variables forms part of the abstract description of the system, and does not necessarily need to correspond with a homologous empirical measure of the variable in question. It suffices that the abstract numerical dimension and the empirical measure of the variable at issue are sufficiently analogous to warrant an eventual empirical test of the resulting dynamic model describing the system. The descriptive use of abstract numerical variables to describe components of the system also makes no implicit ontological claims with regard to the nature of those variables. For instance, describing a child's social cognitive knowledge as an underlying numerical variable does not entail any claim about that knowledge being localizable (e. g. in the child's brain), as internal or symbolic contents. The variable in question can equally well refer to a distributed, soft-assembled property of a child's actions that depends on causal loops between the child and its environment (see [350] for discussion).

The description of the developmental system as a dynamic ecological system makes use of the following general assumptions (for thorough discussions of these principles [334,337,345]). First, development is defined as the growth or increase in level of more developmentally advanced or complex variables and the decline or decrease of less developmentally advanced variables. The growth of a variable (e. g. a child's lexicon, a child's level of social cognitive understanding, and so on) is an auto-catalytic process, in that it depends on the level already attained. Thus, if l is the current level of some developmental variable (e. g. a child's level of lexical knowledge, of social cognitive skill, etc.), the growth or change of l over some time duration is expressed as

$$\Delta l = rl$$

for r any rate or change parameter. In an ecological system, of which developmental systems are examples, growth or change depends on the availability of resources, which are limited. For instance, a resource factor for lexical growth is the language spoken in the environment, but also the child's auditory system that helps it pick up acoustic signals that form the physical basis of spoken and heard language. Given that resources are limited, the growth parameter r approaches a zero limit as l approaches the level that is sustained by the available resources (a simple example is the resource "lexicon of the environment" which limits the number of lexical items that can be learned by a child

living in that environment). The effect of the limited resources is expressed mathematically as follows

$$\Delta l = r(1 - l/K)l$$

for K the limit level of l under the given resource conditions. This equation thus corresponds to the well-known and basic logistic or Verhulst equation (see [337] for further explanation). If applied to knowledge-related variables, it basically states that the growth of knowledge depends on what one already knows, and on what one does not know yet, given what there is to know in the current context of the particular person in a particular environment. In the simplest possible case of a single-dimensional developmental system, the resource component corresponds with the system's environment.

For a system description to be really explanatory interesting, it should contain various coupled components, corresponding with the major dimensions of a particular developmental system. For instance, a model of the child's entire cognitive system should consist of variables referring to the major components of the cognitive system, e. g. language, conceptual knowledge, emotions and appraisal components and so forth. Another example concerns a model of an educational system, which should at least consist of a variable that changes as a result of environmental stimulation and interaction on the one hand, and a variable that describes the educational influence on the other hand. These two variables should be coupled, since it is relatively obvious that, as the child's growth (at least partially) depends on educational influences, the educational influences themselves will depend on the developmental or growth level that the child has already attained (one is not going to teach higher algebra to a child who has not even mastered elementary arithmetics, for instance). Coupled variables, each of them dependent on background resource factors, form the heart of a dynamic ecological system. To explain the principle of coupled variables, let us take an example from the field of social and personality development during adolescence (the example is taken from [193]).

The adolescent's main conflicts with the parents are related to the adolescent's wish for growing autonomy and the parents' reluctance to grant this autonomy too easily or too rapidly. The tendency to increase autonomy depends on resources such as physical maturation and cognitive and social skills, which are clearly increasing during the stage of adolescence. However, the tendency to increase autonomy is coupled to a complementary variable, which is the adolescent's connectedness with the parents. We shall represent autonomy and connectedness as two coupled numerical variables, which can thus increase or

decrease in level. In psychologically healthy families, the level (including the quality) of the connectedness between parents and adolescent can have a positive effect on the adolescent's wish for autonomy, i. e. increase the growth of autonomy. The other way round, however, autonomy of the adolescent – and later adult – will have a positive effect on the quality of the connectedness and thus – given the principle that differences in quality are implicitly quantified – on the level of the connectedness (it is easy to see that if parents cannot accept the increasing autonomy of their children growing to adulthood, the level of connectedness that exists between parents and children is in fact going back). However, the actual striving towards autonomy (including participation in risk behaviors, staying out late at night, visiting places the parents do not like etc.) causes conflicts that tend to temporally reduce the good relationships between parents and children, i. e. temporally reduce the levels of connectedness they feel towards each other. Hence, the growth in autonomy (through conflicts) is related to the decrease in connectedness, whereas the level of connectedness (in a healthy family situation) is positively related to the growth of autonomy. With both autonomy and connectedness related to their own resource factors, K_A and K_C , the dynamic relationships between Autonomy A and Connectedness C are described as follows:

$$\begin{aligned}\Delta C &= r_C(1 - C/K_C) - a\Delta A + bAC \\ \Delta A &= r_A(1 - A/K_A) + cCA.\end{aligned}$$

The first part of the equation describing change in connectedness (ΔC) relates to its resource-dependent growth, the second part relates to its being negatively affected by the change or increase in autonomy, which takes place mainly through conflicts; the third part relates to the posi-

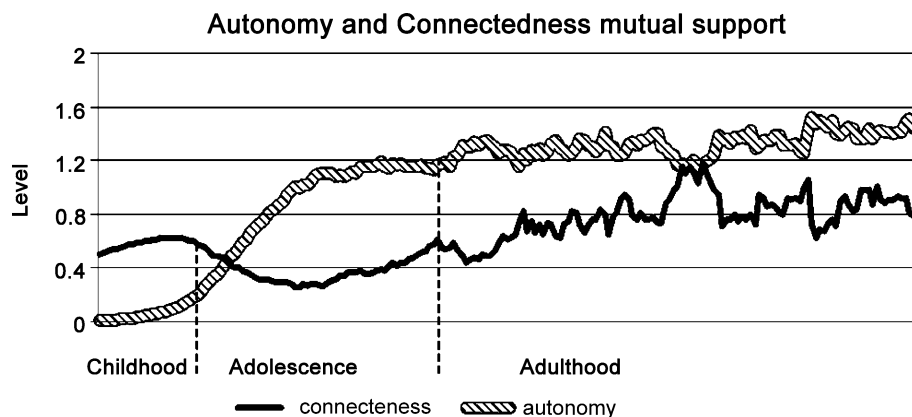
tive effect of autonomy on the level of connectedness. The first part in the equation describing the growth of Autonomy (ΔA) describes its resource-dependent growth, the second part describes the growth due to support from Connectedness.

A stochastic version of this model produces developmental trajectories as described in the literature on adolescent development. As autonomy grows, connectedness shows a temporary decline, from which it restores and then shows a gradual increase, more or less parallel with the gradual increase in connectedness. The system then stabilizes around an attractor level, with stochastic fluctuations (see Fig. 1). The local regression and restoration is characteristic of U-shaped growth, which is a typical developmental phenomenon [59,117,226,305,337].

The model is an example of the ecological relationships that can hold for any couple of “growers”, i. e. relatively autonomous components of a developmental system. These relationships can be symmetrical supportive, symmetrical competitive, asymmetrical competitive and supportive (as in a predator-prey relationship in a biological model) and, finally, conditional (if a particular, minimal level of a component is a necessary precondition for another component to start growing). The relationships are represented in Fig. 2.

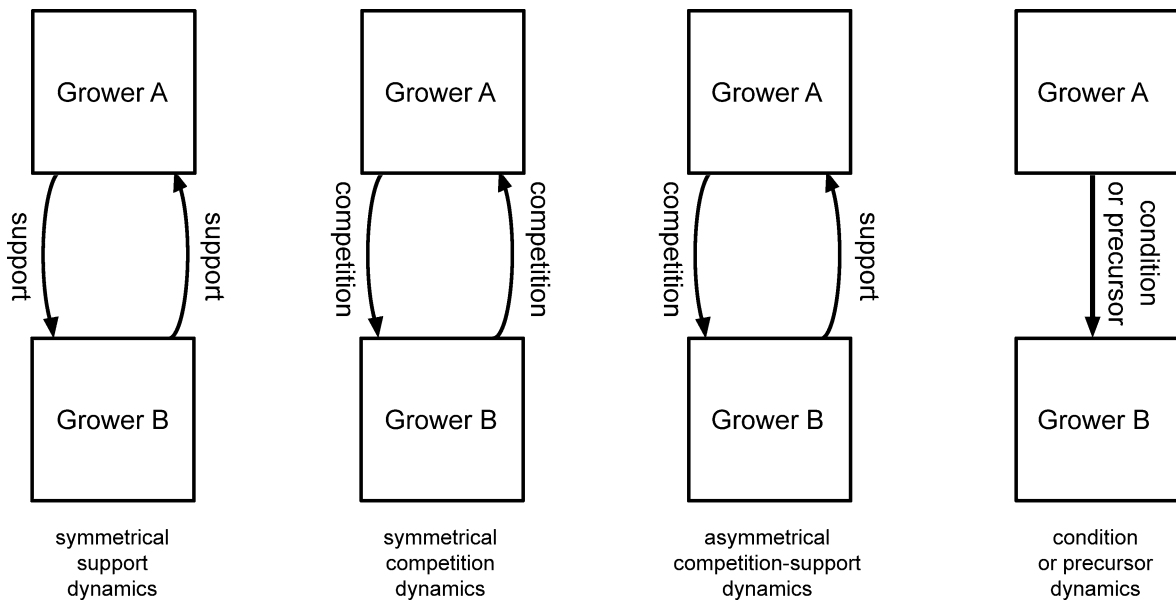
A developmental system is characterized by relationships between any of its components and forms a web of relationships, formally similar to the foodwebs described in biological models [252]. Figure 3 represents an imaginary web of relationships between components in a developmental system.

A major difference between ecological web models used in biology and those used in developmental psychology is that the latter are considerably less supported by em-



Development, Complex Dynamic Systems of, Figure 1

Developmental trajectories generated by a stochastic version of the autonomy-connectedness dynamics



Development, Complex Dynamic Systems of, Figure 2
 Four types of dynamic relationships between “growers”

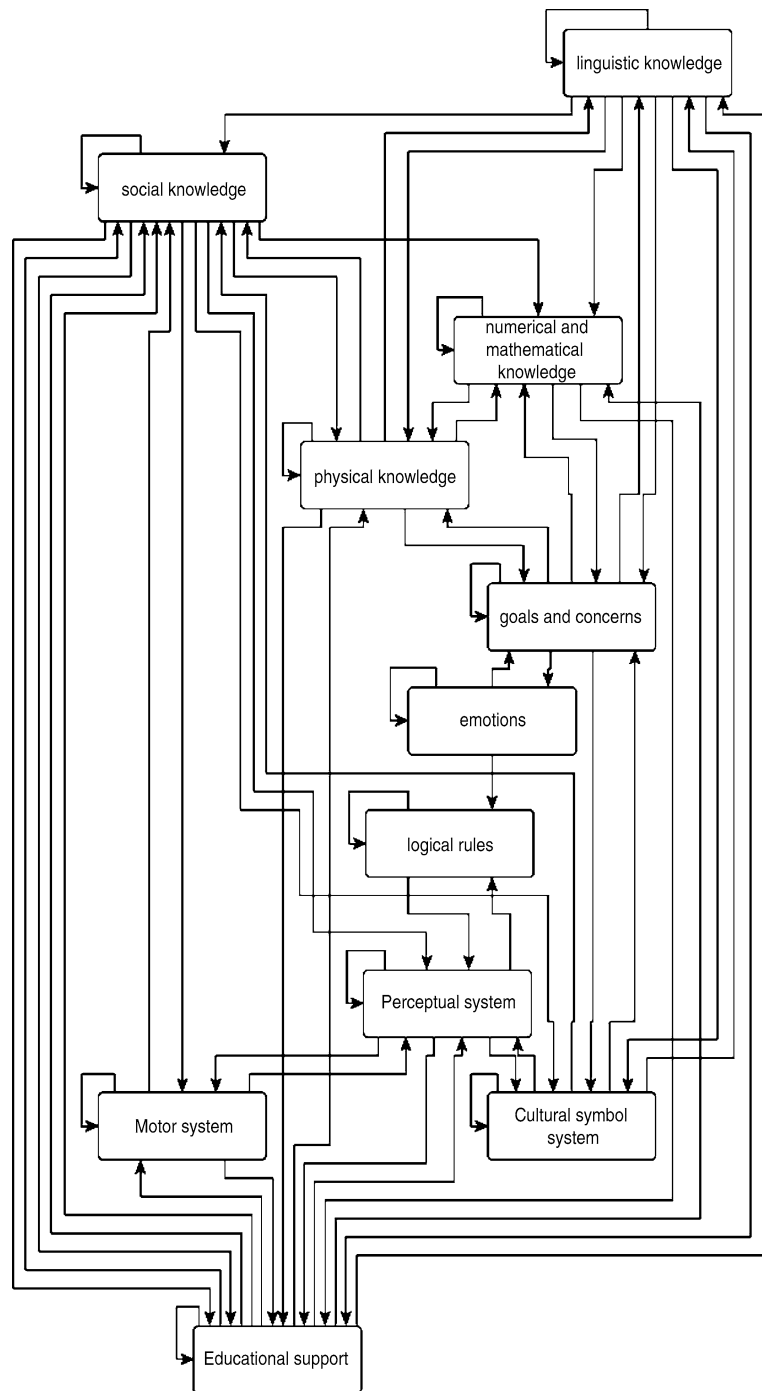
pirical data than the first. The lack of empirical support is due to two factors mainly: the first is that the components distinguished in developmental models are considerably more fuzzy and less tangible than those used in biological models, with the associated difficulty of measurement precision. The second reason is that developmental psychologists are considerably less accustomed than biologists to study real developmental systems over sufficient time spans. Developmentalists mainly focus on statistical relationships between distributions of variables in samples, which tell very little if anything about the dynamics of the developmental system. Exceptions to this rule are the studies carried out in the field of language development, where single-case studies are the rule rather than the exception. By conceiving of language as a developmental system, consisting of components such as semantics, syntax, phonetics etc., or components on a lower level of organization, such as prepositions, adjectives, verbs, nouns etc., ecological network models can be specified simulating the growth of linguistic variables in a single child (see for instance [254,337,340] on lexical development in relation to the growth of plurals; [259] on the growth of closed-class words; and [20] on the pattern of growth and decline of sentences of various sentence length).

A study by Bassano and van Geert [259] illustrates the process of the emergence of three, developmentally successive syntactic generators, the holophrastic, combinatorial and syntactic generators. The holophrastic generator is basically a “one-word grammar”, i. e. the set of early gram-

matical principles that generate utterances with a characteristic word length of one. The combinatorial generator is the developmentally more advanced set of principles that generate combinations of words, typically two per utterance. The syntactic generator is the set of principles that use the syntactic rules of sentence formation typical of mature language. Bassano and van Geert assume a series of asymmetric relationships between a less and a more advanced developmental structure, for instance, the holophrastic and the combinatorial generator. The less advanced structure has a conditional and supportive relationship with the more advanced structure. For instance, a minimum level of one-word productions is needed for the combinatorial generator to emerge, and, in addition, the level of one-word production supports the growth of the combinatorial generator, i. e. the production of two-to-three-word sentences based on a simple combinatorial principle. The more advanced structure on the other hand has a competitive relationship with the less advanced structure. For instance, the use of two-to-three-word sentences negatively affects the use of one-word sentences (see Fig. 4)

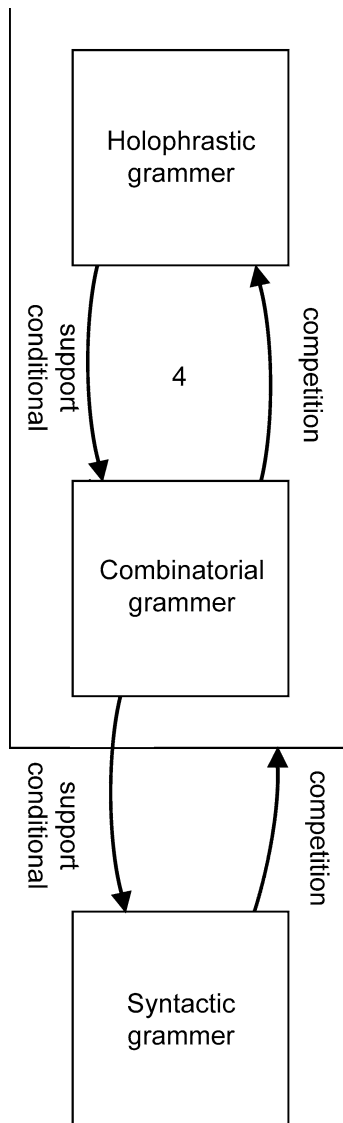
A mathematical model of these relationships with three connected growers (the holophrastic, combinatorial and syntactic grower) provides a good fit of the empirical data (model is shown in comparison with the smoothed data; see Fig. 5)

Another example of a model based on growth relationships between components of the developing system



Development, Complex Dynamic Systems of, Figure 3

An imaginary developmental growth web or network. *Arrows* represent either positive (supportive) or negative (competitive) relationships, in addition to conditional relationships. The *reflexive arrows* refer to the nodes' autocatalytic and resource-dependent growth



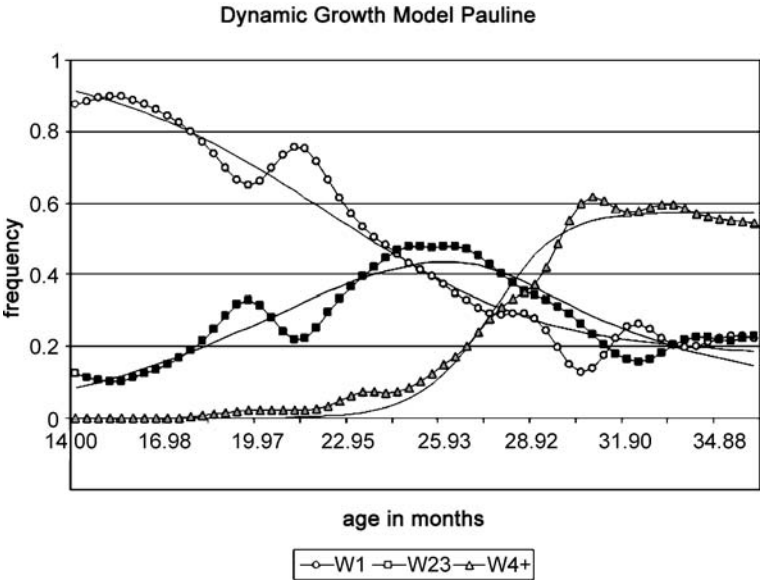
Development, Complex Dynamic Systems of, Figure 4
 Asymmetric growth relationships between three developmental levels of grammar (after [259])

is Fischer's model of development through tiers and levels [103]. The model describes development as the emergence of skills, which are context- and content-specific. Skills are general formats of behavioral control and perception, and they are characterized by general structural properties that develop over the life span through a series of more or less discontinuous changes. Discontinuity is primarily observable in what Fischer calls optimal performance, which is a subject's skill level under optimal conditions, including support from other persons (for instance, more competent persons, as in a context of teach-

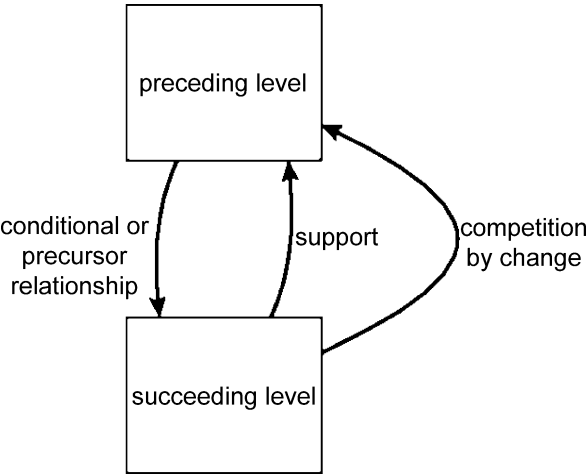
ing and learning). Fischer distinguishes three major structural types of skills, which he calls tiers: the tier of action (roughly from birth to two years), the tier of representations (from two to about 12 years), and the tier of abstractions (from 12 years up). Tiers are further subdivided into levels, for instance the tier of abstractions goes through a sequence of single abstractions, then mappings, then systems, and finally, principles. Fischer's model has an additive structure, in that earlier levels do not disappear – as in Bassano and van Geert's model of language development – but are conserved and in fact transformed through the emergence of a more advanced level. The postulated structure of relationships between a preceding and succeeding level of development is as follows (see Fig. 6). First, the preceding level acts as a precursor or condition for the emergence of the succeeding level, e. g. single abstractions are necessary precursors of mappings, because a mapping is a relationship between single abstractions. Second, the succeeding level competes with the preceding level, in that the change (growth) in the succeeding level has a negative influence on the preceding level. Third, the succeeding level supports the preceding level: its positive effect on the preceding level is proportional to its magnitude of occurrence ("level" in the quantitative sense).

Dynamic models constructed according to these growth relationships predict typical developmental patterns, for instance in reflective judgment [103,170] (see Fig. 7). In addition, such growth models have predicted developmentally discontinuous trajectories at high growth rates and smooth trajectories at low growth rates, consistent with the data [104].

Development and Dynamic Field Theory Thelen, Smith and co-workers explain the short-term dynamics of action, thinking and knowledge and the long-term dynamics of development, by means of dynamic fields. A dynamic field is defined by an abstract metric dimension (or a space consisting of various such dimensions) that describes the main variable (or variables) of an action or thinking process. For instance, a major variable in the aforementioned object search task in which babies make the A-not-B error is the spatial position of the hiding objects and hiding places, which also defines the major variable of the child's action, which is the place toward which the infant will reach in order to retrieve the hidden object. For each point of this metric variable, there exists a particular activation value, which, in the case of the current example, would mean a likelihood that the child reaches at a particular location. The whole of activation values forms an activation field, the form of which changes on the basis of "inputs" from various sources. In the present example,



Development, Complex Dynamic Systems of, Figure 5
Smoothed data of the development of one-word, two-to-three-word and four-plus-word sentences in a French girl, Pauline, with fitted growth model based on relationships of competition and support (after [259])



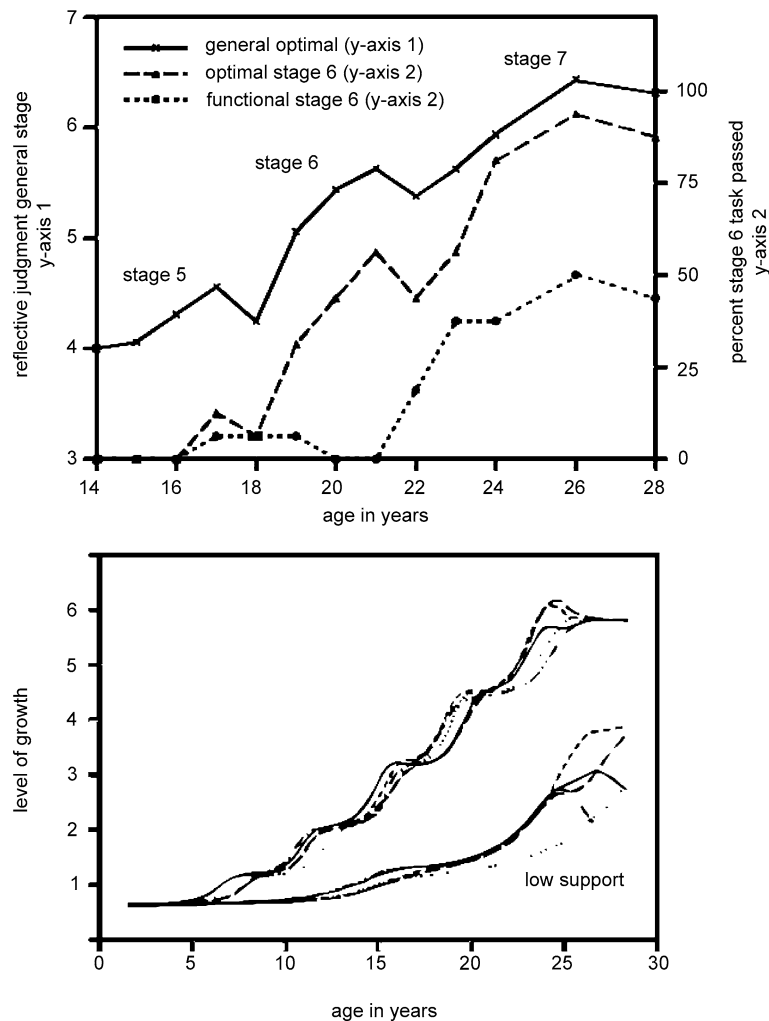
Development, Complex Dynamic Systems of, Figure 6
Growth relationships in Fischer's model of tiers and levels of development

inputs come from the child's perception of the environment, for instance the position of the hiding places, from events taking place in the environment, for instance the adult moving a hidden object to another place; and finally, from the child's memory [273,274,313], which is carried by the neural network that constitutes the brain and that changes as a result of experiences, maturation and self-organizing processes [188,189].

The inputs to a dynamic field are not just linearly superposed: they show cooperative and competitive interactions, leading to self-stabilization of activation patterns [99]. The mathematical properties of that field can be defined rigorously and allow for a dynamic systems model that is no longer just metaphorical. The dynamic field theory that describes the dynamics of this field thus bridges the "representational gap" that exists in current dynamic systems models [296]. This "representational gap" refers to the fact that developmental dynamic systems models of embodied and embedded action have no use for concepts and representations as mental entities that act as mental causes of behavior [74].

In addition to the development of the object concept, dynamic field theory has been applied to development of habituation [273] and development of working memory [274].

Dynamic fields can also be specified for abstract properties of the developmental state space in order to model long-term changes and mechanisms of development [343]. The starting point is the geometric notion of development as specified earlier, i.e. the developing system defined as a manifold of dimensions or variables, describing all of its relevant developmental properties. Since all those dimensions can be ordered along a scale of developmental progress (a developmental "ruler"), the developmental state space is thus characterized by a principal component (in the statistical sense) that can be used to specify any kind



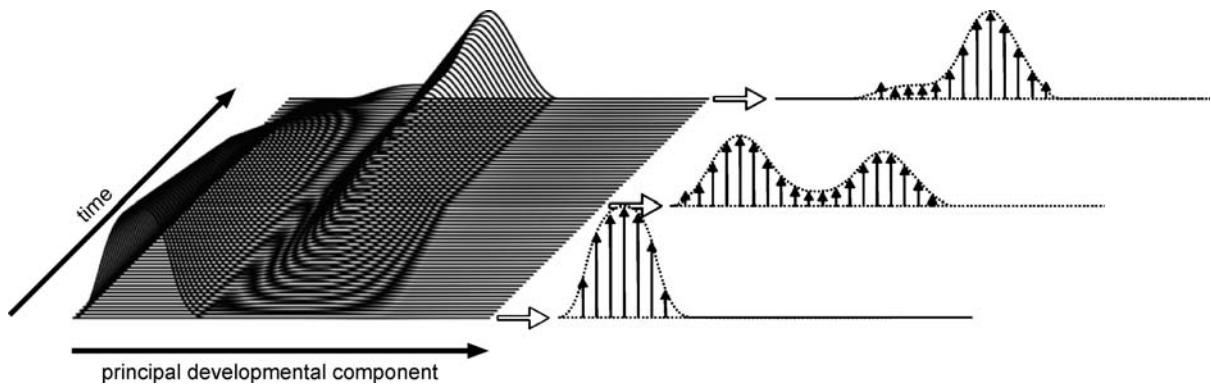
Development, Complex Dynamic Systems of, Figure 7

Fischer's model of structural stages in cognitive development, with empirical data from reflective judgment. The dynamic model based on growth relationships generates the pattern of stepwise change with intermittent regressions (after [103])

of developmental progress or succession. Any point or region in the developmental state space can be mapped onto the principal component of the space, i. e. the general developmental distance introduced above. Any point on this metric distance dimension has a certain likelihood of being "visited" by the developing system. The actual likelihood, i. e. the actual activation field, is determined by inputs from the child's momentary experience of the context, retrieved memories, actions from other persons, and so forth. These likelihoods can be represented as a vector field, with an activation vector for each point in the developmental principal component or distance dimension. The vector field can specify a single peak, in which case the developmental state of the individual is crisp and uni-modal (the classical ideal), or by a landscape of peaks,

in which case the developmental state of the individual is multi-modal, fluctuating and fuzzy (which is more like reality; see Fig. 8). For instance, if a child alternates between solving a problem either in a less or in a more developmentally advanced way, it occupies two regions in the developmental space between which it shifts randomly.

Development can then be represented as the change of the vector field over time, beginning with a dominant mode in the lower and ending with a dominant mode in the upper regions. The short-term dynamics of development consist of the individual's actions, experiences and interactions in real time. These real-time actions have a lasting effect on the structure of the vector field. The effect is moderated through two mechanisms that have already been described by the founding fathers of develop-



Development, Complex Dynamic Systems of, Figure 8

Probability functions of developmental levels assigned to potential actions or experiences of a child across time. The probability wave moves from a dominant mode on the left to a bimodal mode in the middle to a dominant mode on the right (see the three probability functions with vectors at the right)

mental psychology, Piaget and Vygotsky. They see development as the result of what I have freely termed *conservative* and *progressive* forces. These forces operate as follows (see [343,344] for an explanation of the model). It is assumed that any activation of components of the developing system in the form of a particular action, experience or event, have a consolidating effect on those components, and hence on the developmental level(s) that they represent. The consolidation depends on the functional success of the action or experience in question, i. e. on its short-term dynamics. The consolidation takes place in the form of increasing the vector values at the levels corresponding with the action or experience in question. The consolidation function spreads out to nearby regions and becomes negative (reducing vector values) for regions farther away on the developmental distance dimension. The consolidation function amounts to some sort of familiarity effect, which decreases with increasing distance from the actual, i. e. familiar level.

The developing system is also driven by a second force, namely novelty, which is a general term for novelty (new things) per se, curiosity, interest, goal-related activity and so forth. Novelty is a function that increases with increasing distance from the familiar. Assuming that familiarity and novelty are governed by their own characteristic parameters, there is a point on the developmental distance dimension where the combination of both has a maximal value or optimum. The vector values corresponding with this point are also upgraded, with an upgrade function the form of which is in principle similar to the conservative upgrade function. The updated vector field will generate new short-term actions and experiences, which will cause the vector field to update again, and so forth.

Simulations based on this model of development show that, depending on the values of the main parameters (fa-

miliarity and novelty parameters, rate of vector field upgrading, nature of information activating vector loadings and so forth), a rich landscape of developmental phenomena can be achieved, ranging from stepwise growth as described in the Piagetian and neo-Piagetian theories, to models of overlapping waves of strategies (Siegler), micro-genetic fluctuations in performance, and so forth. An interesting effect of these dynamic field models is that they are able to explain discontinuous changes in development.

Theory of Complex Adaptive Systems (CAS): Developmental Agent Models Although development is a prime example of a long-term adaptive process – and according to Piaget, for instance, adaptation is the central developmental mechanism – there is very little work on complex adaptive systems in the field of development. An adaptive system can be defined as a collection of agents and artifacts that interact with one another and through that interaction are aiming at reaching their goals, concerns or interests [7,17,94,288]. Agents pursue their goals by means of their action repertoires, their knowledge of the world and the information they obtain through acting. In a complex adaptive system, agents are interdependent, yet autonomous in achieving their goals. From a developmental viewpoint, agents learn from their experiences and show long-term change in their goals, knowledge, action repertoire and skills, a long-term process we call development [269,303,352]. In a situated and embodied agent, development not only concerns the change in the person, but also changes in the person's niches or preferred environments [68,69].

A characteristic property of agent-based models is that they conceive of agents as being equipped with relatively simple rules, instead of complex internal representational worlds, carrying out extensive computations before act-

ing [288]. Developmental psychology, on the other hand, emphasizes the complexity and the richness of a child's emerging knowledge and skills. However, the two approaches need not be in conflict with one another, in that very simple rules of agent interaction can in fact emerge on the basis of complex, self-organizing processes requiring the interplay of the environment, experience, knowledge and so forth. An example of how agent-based modeling and development come together is Steenbeek and van Geert's work on the development of social status, social power and social skills in young children [301,302,303]. The model describes the dyadic interaction of children by conceiving of children as agents with particular interests or goals with regard to playing alone or playing with other children. The agents in the model have a representation of the social value (popularity) of the other agents in their group, and try to optimize their positive appraisals of the interaction situation by drawing actions from a repertoire of either solitary action or actions aimed at other children. The principles governing the actions – for instance the choice among actions – are very simple at the level of the agent-based model, but such simple principles are in fact emerging out of a complex multitude of components and influences. An example of such emergent simplicity in social interaction concerns the emergence of goals and interests of the agent through the interactions themselves (see for instance [66]).

The agent-model described by Steenbeek and van Geert [299,301,302] simulates patterns of interactions that are empirically validated by data on play interactions between children from various sociometric statuses. The model can also be linked with a model of long-term changes in the peer selection preferences of children and the emergence of social statuses such as popularity or rejectedness. A similar model is used to explain the spreading of risk behaviors among groups of adolescents and the formation of friendship groups or cliques [12].

One possible drawback of agent models for development is that the nonlinear and complex patterns that typically result from interactions between agents do so if the number of interacting agents is high (see [11] for many examples). The typical number of agents in developmental models is small, e. g. two in dyadic interaction, or ten to twenty in small group interactions (e. g. a group of friends, a peer group). In spite of this limitation, agent models are typically suited for simulating short-term temporal patterns of interaction among persons of various developmental levels, and it is from these patterns that the interacting persons learn, adapt and develop.

Agent models have been used to study and simulate language development in connection with language

evolution, by modeling generations of language learning, teaching and communicating agents. Such models explain, among others, why language tends to evolve towards a capacity that is optimally adapted to biological and social learning principles, and thus tends to become an increasingly “innate” type of capacity (examples of such studies are [66,165,168,169,290]).

Theory of Complex Adaptive Systems (CAS): Epigenetic Robotics

A different type of agent models in developmental studies concerns the epigenetic robotics models. The embodied and embedded aspects of developing agents are literally implemented in the form of robotic models, i. e. artificial autonomous agents that wander about, act in real environments and interact with other agents, including humans, and must learn from their experiences. Metta and Berthouze [211] define epigenetic robotics as the study of how a realistically embedded and embodied robot-model of a person, including a brain, sensory and effector organs, changes and develops in interaction with a real world (for introductions and reviews, see [26,27,28,182,202,211,267]). In fact, “... Beyond a certain level, it becomes extremely difficult to study realistic interactions between the agent and the environment without including a real body and real people in it” (p. 130 in [267]). Epigenetic robotics serves two purposes. The first is of a technical nature, and is aimed at designing self-organizing and learning robots that serve practical goals, in cases where direct programming of the robot is too complicated (e. g. [376]). The second is of more concern to the present article, which is to understand developmental processes in humans by studying simulated but embodied (robotic) agents in real environments. Epigenetic robotics approaches have so far dealt with the following aspects of development.

The first refers to the observation that human learning and development very strongly depends on social scaffolding and socially situated processes of cognition and perception. An important issue is joint attention, which occurs at a very early age (approximately around nine months) and involves the infants capacity to infer an object or event of attention to an adult by using the adult's gaze, pointing, etc., and to share that topic of attention with the adult. Joint attention is a typical human capacity, and greatly facilitates the process of cognitive and social learning. A host of robotics studies have been carried out, showing that processes of joint attention or closely related to it can be implemented in a robotic system and greatly enhance the processes of cognitive, linguistic and social learning [163,225,268,306,307]. Joint attention is closely related to empathy, i. e. the ability to in-

tuitively understand the minds of others by picking up their intentions, an ability which is related to specific neurophysiological structures, the so-called mirror-neurons, in human and primate brains (for studies using principles of epigenetic robotics, see [39,212]; and for applications to impaired attentional processes, see [239]). Socially-situated learning rests heavily on imitation learning, or learning through emulation, which means transforming the perceived actions into one's own action repertoires, aiming at the inferred goal or intention of the perceived action (for studies on imitation learning in epigenetic robots, see [47,83,116]). An important aspect of perceptual development concerns the infant's very early capacity to integrate information from the various senses and perceive the world in a multi-modal way. Multi-modal perception can also be accomplished by epigenetic robots through associations of information from various sensory organs [106,248]. Epigenetic robot studies have tested embodied and socially-situated processes of cognitive and language development [86,87,197,200,377] and emotional development and communication [44,45,46,56].

The Forms of Development: Fluctuations, Variability, Continuity, Discontinuity and Critical States

Development and the Notion of Stages Classical developmental theories (e. g. Piaget, Erikson) typically view development as occurring in stages, that is, the course from the initial developmental state to some sort of end state is seen as a stepwise path, or a path moving across various qualitatively distinct states. Piaget's theory, for instance, describes a first stage as a sensorimotor level of thought, after which children proceed to a second level called pre-operational, then to concrete-operational thinking. Development finally stabilizes at a formal operational level, which is characteristic of adult thinking.

Recent stage-oriented theorists, in particular the neo-Piagetians, occupy a considerably more sophisticated standpoint [103]. They see "stages" as in fact qualitatively different forms of thought, or skill in general, that are developmentally ordered but are also context- and domain-specific [58,82,103]. A child may function on stage (or level) 1 in domain A (e. g. simple mathematical operations) and on level 2 in domain B (e. g. social relationships). Within a domain, such stages – or one should say levels – can fluctuate with varying context, because context is a part of a person's skill (e. g. a child who faces a particular problem context may function on level 2 with help and on level 1 without help). The levels or stages may fluctuate strongly over the short-term time scale of a problem-solving event, in a process that Fischer has

called "scalloping" [103,130]. Overall, however, there is also a fuzzy but nevertheless convincing ordering in the level or stages. Two-year olds, for instance will show a very different mixture and frequency of context- and domain-specific levels than adults, and are thus characterized by a different major-stage category than adults are (see [80] for an example). The notion of stages advocated by neo-Piagetian theorists reflects the complexity of the developmental system by viewing stages all the way down, in a complex, hierarchical and dynamic organization.

The notion of stage (level, phase, ...) entails an idea of internal coherence, a relatively stable structure of interdependent elements such as skills, habits, processes and so forth. The notion of stage is thus highly reminiscent of a basic notion from dynamic systems, namely the notion of attractor. Starting from the theory of complex systems, we can follow the assumption that such systems tend to self-organize into islands of relative stability rather than remain unconnected collections of features where any combination of such features is as likely and (in)stable as any other. From this, we can reach the conclusion that stages, defined in the dynamic and complex way explained above, should be the default option for a system as complex as human development.

Developmental Stages and the Theory of Bifurcations

Developmental attractor states do not need to be overall states in the sense of stages, but can amount to any stable pattern of coordinated knowledge or skills at any level of aggregation. The transition from one such state to another represents a discontinuity, since none of the intermediary states, if there are any, is stable. Developmental researchers have used the framework of catastrophe theory to answer their questions about developmental discontinuity (for an overview see [327,351]). By testing for empirical indicators of the so-called catastrophe flags (structural properties of discontinuities in general), they have tried to show that developmental transitions are instances of the so-called cusp catastrophe and thus entail a clear form of discontinuity. Examples of phenomena studied are the transition between non-conservation and conservation understanding in young children [139,150,151,157,158,326], reaching and grasping in infants [373,374] and syntactic development [259,329]. The results show that rapid, jump-wise development takes place in a variety of domains. However it remains unclear whether these changes are real discontinuities in the bifurcation sense. In addition, they seem to occur in some children but not in all. A problem with discontinuities is that the empirical detection depends on the definition given by the researcher (see [329] for discussion). Finally, different attractor patterns can occur si-

multaneously and in that sense show a form of superposition [343,352]. Children can act according to a less advanced skill pattern and only little later act in accordance with a more developmentally advanced pattern. If such patterns co-occur, they will lead to increased intra-individual variability in performance, which has been shown in the above-mentioned catastrophe-theoretical studies and in studies of language development [20]. As development proceeds, the less advanced pattern will disappear in some cases, but remain unchanged in others. In language development, for instance, less advanced grammatical principles disappear, but on the level of cognition as a whole, earlier patterns – one should not particularly call them less advanced – can survive and be used in contexts where the more advanced patterns are not directly applicable (see the scalloping principle mentioned earlier). Finally, developmentally divergent levels of skill or knowledge can even occur in the same action pattern, for instance if the child's verbal explanation refers to a less advanced level and his non-verbal gestures to a more advanced level of understanding (or vice versa, see [119,156]).

Because behavior and action are in themselves highly variable phenomena and because this variability is intrinsic to behavior and not a matter of added measurement error or noise, the continuity issue amounts to the question whether bands of fluctuation or variability in behavior develop continuously or discontinuously [329]. The growth models described above represent development or growth in a variable as a single point on a dimension, but this point should be seen as an estimation of a central point of what in reality amounts to a bandwidth of fluctuation.

Are Developmental Stage Transitions Like Physical Phase Transitions? One possibility is that what has traditionally been called “stages” are comparable, in that sense, to the phases of physical matter (gaseous, liquid, solid) and depend, in essence, on a single parameter or a confluence of parameters. Developmental stages form attractor states because they rely on network-like structures, i. e. on structures of relationships between the components of the system. A developmental attractor state is represented by habitual, coherent patterns of performance, skill or action that self-organize spontaneously in the person's habitual contexts, niches or living spaces. These patterns consist of mutually-supportive and sustaining features. To give a simple example, Piaget's sensorimotor stage defines thought in the form of external action on objects. For instance, reaching to and grasping an object requires the coordination in real-time of a myriad of components or aspects, including the coordination of the muscles in the arm and hand, the coordination of vision and move-

ment, the coordination of vision of the object and vision of the own arm and hand, and so forth. These patterns are not innately given, but self-organize through processes that eventually amount to discontinuous changes (e. g. Wimmer's studies of early prehension development, Wimmers, Beek, and Savelsbergh, 1998a; Wimmers et al., 1998b; [372]). The characteristic feature of these sensorimotor patterns is that their contextual self-organization (e. g. in the form of reaching to and grasping a particular object) emerges on the basis of dominant driving forces or control parameters that are of a sensory and motor nature (see for instance dynamic field theory described earlier). In addition, the sensory and motor control parameters of infant action are likely to be biologically pre-adapted to important features of the environment, such as object-person distinctions, numerosity, etc. (see for instance [295,372,373,374]).

Are Developmental Transitions Caused by Self-organized Criticality? Irrespective of the stage theory under consideration, the durations of sequential stages tend to increase in a logarithmic manner [340]. One might ask if the distribution of stage durations relates to the power law distribution characteristic of self-organizing phenomena [242,354], and more particularly, to self-organized criticality [10]. The phenomenon of self-organized criticality emerges in complex systems, consisting of many components that entertain local relationships. We have already seen that the embodied-embedded brain is such a system, consisting of many components (perceptions, thoughts, actions, memories, tools, environments) that are temporally and functionally connected. This complex system is under a certain external “tension”: there are problems to solve, goals to achieve. The person does so by means of the complex system of skills, knowledge, sensory and motor systems. As not every action is successful, the person will adapt and learn from his experiences and from being taught by other people. This complex, interconnected system exchanging information with the world is a good example of a system that shows self-organized criticality. Its attractor states are critical states, i. e. states for which any external influence can cause patterns of change with a wide variety of magnitude and duration, dissipating the stress that has been build up in the system. Note the major difference with a phase transition model: in a phase transition model the attractor states are the phases, whereas in a critical transition model the attractor states are those where a transition might occur.

The magnitudes and durations of changes are statistically distributed according to a power law distribution, with very few large-scale changes and increasing numbers

of smaller scale changes. It is tempting to see development as an example of such a self-organized criticality: a succession of meta-stable states punctuated by changes of various magnitude (e.g. a relatively small change in a relatively context-specific problem solving strategy, versus an avalanche of changes in many aspects and domains of cognitive performance, the latter characteristic of what would count as a stage transition).

If for some reason something changes in one skill (or knowledge, ability, action pattern, habit) it is likely to affect other skills (habits, etc.) to the extent that these two developmental components are interrelated. However, the second component, affected by the first, can eventually affect a third one to which it is connected, and so forth. In principle, such changes can remain quite limited, but they can also grow into an avalanche of changes that affects the whole developmental system. If we assume that in a developing system the “weakest”, i.e. the least adapted or effective skills (habits, knowledge) are eliminated (or altered) more easily than better adapted or more effective skills, we wind up with a system that closely resembles the Bak-Sneppen model of biological evolution through punctuated equilibria [9,35]. This model of evolution changes through many events of extinction and speciation, interspersed by periods of stasis.

The pattern of evolution with many small and only a few major extinction-speciation events is clearly reminiscent of the course of human development, with many small and a few major changes. The principle of eliminating or altering the weakest component is also applied in a routine for solving hard optimization problems, called extremal optimization [36]. The solution patterns are characterized by shifts following the power law distribution. In a certain sense, (cognitive) development is like solving a hard optimization problem, an adaptation of knowledge and skills to the complexities of reality. It would thus not be surprising that the general dynamic structure of cognitive development follows a pattern very close to that of the extremal optimization process, including the power law distribution of the changes.

An Overview of the Human Life Span in Light of the Theory of Complex Dynamic Systems

Preliminary Remark

A discussion of the issues of dynamics, self-organization, complexity and so forth in the context of human development requires that the reader has some basic knowledge of how current developmental psychology describes the human life span from the viewpoint of developmental processes. In order to provide such knowledge, I will

present an overview of selected themes and topics discussed in mainstream handbooks on developmental psychology [25,51,73,162,227,276,280,356]. The themes are chosen for several reasons, mostly because they are considered of great importance to development, but also because they provide interesting possibilities for applying a complex dynamic systems framework on issues that are usually not seen in this light. The overview presented is therefore, by necessity, only fragmentary and exemplary. Interested readers are referred to a host of introductory handbooks (see above for some suggestions). Because of their importance for starting and guiding development, I shall concentrate more on early processes of development than on the later ones, and confine myself to the period between birth and adolescence.

Most handbooks give a review in terms of stages or phases in the life span, with subdivisions based on major topics or fields of development such as physical and motor development, social and personality development, cognitive and language development as the main sections.

Handbooks often start with a theoretical introduction, discussing major theories such as Piaget, Vygotsky, Freud and so on. These theories mainly or exclusively refer to the work of historical figures who laid the foundations of the field. In addition to such theoretical perspectives, handbooks also discuss basic questions such as the nature-nurture relationship.

Prenatal Development

Prenatal development covers the period from conception to birth. Normal fetal development occurs through three stages. The germinal stage covering the first two weeks after conception is a period of cell division and implantation in the uterus. The second or embryonic period is a period of emergence of essential organ systems, such as the central nervous system, the heart etc. and lasts from week 3 to week 8. The third or fetal stage lasts from week 9 to week 38 in case of full term birth. During the embryonic and fetal stages, the embryo or fetus is sensitive to teratogenic influences, i.e. influences of substances in the mother's body that negatively affect the growth of specific organic systems, due to disease or intake of substances such as alcohol. The influence of teratogenic substances depends on the level of development of the embryo or fetus and is thus related to periods of greater or lesser sensitivity to such influences. During the fetal stage, the child becomes increasingly sensitive to sensory stimulation, which is limited and modified by the child's position inside the body in the amniotic fluid. Of particular importance is the communication with the mother's men-

tal and physical state through the exchange of hormonal and chemical substances that reflect the mother's current psychological state. Such influences can affect the child's later neurophysiological reaction to stress, for instance. Already before birth, the unborn child and the pregnant mother entertain a transactional relationship. The child is affected by the mother through the sensory and neurochemical links described above, whereas the unborn child affects the mother's behavior, moods and evaluations. The effect of the child on the mother ranges from direct effects, for instance the physical stress of pregnancy, to indirect effects via individual, family and cultural evaluations of the mother's current pregnancy. Of particular importance for later development are eventual problems during the birth process, such as hypoxia of the fetus during birth, or premature birth. Most of such relatively minor birth problems are related to diffuse and (very) mild effects on performance, for instance cognitive and academic performance at school age.

Infancy and the Preschool Years (Birth to 4 Years)

General Introduction Infancy is the period that lasts from birth to 24 months. It is the age of sensorimotor functioning, more precisely action that is mainly limited to a direct coupling between sensory and motor systems in an action context. Action itself emerges out of the newborn's reflexes and reflexive actions. The basic sensory capacities are relatively well developed, in that they suffice to allow the infant to make sense of the environment and perceive the environment "as it is". That is, the infant is capable of perceiving the environment as a structure of functional affordances, e.g. as a three-dimensional space with identifiable objects and events that the infant can relate to in its actions. For example, very soon after birth, the infant is able to locate objects in space, for instance by following moving targets with the eyes and turning to objects identified by their sound. Given the early and seemingly automatic adaptation of action of the infant to the core properties of the physical and social world, various developmentalists have endowed the infant with innate core knowledge of the main features of the world. These core knowledge systems comprise objects, actions, number, space and social partners [295]. The notion of identifiable core knowledge systems has been criticised from a dynamic systems point of view as not presenting a model of the real time mechanisms creating the developmental expression of such knowledge in real, physical situations, and which, in the words of Smith [291] are "general, probabilistic, emergent and distributed across several levels of analyzes".

Cognition and Intelligence Infant *cognition and intelligence* is deeply sensorimotor, that is, intelligent action takes place in the form of real physical action. Infant action, involving looking, grasping, repetitive actions, quitting the action and so forth, must be understood as dynamic processes, the patterning of which unfolds in real time, thanks to the continuous feedback loops between the perceiving and acting infant and the affordances of the infant's proximal environment. By "patterning" one can understand correlational regularities in the action sequences, for instance regularities such as following a moving object with the head and eyes, where the object gets temporally occluded by other objects, for instance. The patterning is a relatively high level patterning, however, in the sense that the regularities are interpretable as expressions of or as being consistent with real-world properties such as the permanence of objects or the existence of causal relations among events. In the mentalistic framework, interpretive patterns such as object permanence or causality were viewed as internally represented structures, generating or producing the observable actions. According to dynamic systems theorists, these patterns are not the cause of, but emerge through the real-time interaction between the infant and the proximal environment (see for instance the studies on the infant's understanding of object permanence in the A-not-B-error experiment [292,296]. As development proceeds, the dynamics of this real-time unfolding of understanding and acting on the world will become increasingly complex, e.g. by incorporating verbal actions, by enriching memory and by enriching and changing the perceptual organization of the environment.

The infancy-to-toddler phase is characterized by a succession of two major modes of thought. The first is the sensorimotor form, in which thought is entirely expressed through sensory and motor action as described above. In the second phase, thought incorporates forms of representation and symbolization, thanks to the blending of thought and sensorimotor action. In addition to using language for and in thought, for instance in the form of private speech [361], symbolization also typically entails symbolic or make-believe play, which is related to skills needed for thinking about other people's minds (so-called Theory-of-Mind [112]; ToM is further discussed in the section on childhood).

The distinction between perception and symbolization should not be taken too strictly. According to the tradition of Ecological psychology [118], perception means picking up invariants and properties of the environment that "resonate" with the functional abilities of the perceiving organism (this view is also closely related to Thelen and Smith's dynamic systems theory of development [314]).

A typical example of the abstractness of direct perception in infants, in addition to the examples already given on early core knowledge of space, time, causality and so forth, is the perception of the goal-directedness of actions of other persons by infants younger than 1 year of age [31].

Motor Development In the *motor domain*, the infant develops capacities such as sitting alone, crawling, standing and walking. Although such milestones are usually associated with age averages, such as walking around the age of 12 months, individual differences in onset of such motor skills are considerable, for instance between 9 and 18 months for walking. In the motor domain, the disappearance of the early stepping reflex and its replacement by real stepping at a later age provides an illustration of dynamic systems thinking. Thelen and Fisher [314] argued that the disappearance is not due to central, cortical processes, and that it is in fact not really disappearing. Due to more “peripheral” biodynamic processes, namely differential growth in body mass and muscle strength, the reflex is inhibited biomechanically, and “reappears” as soon as the baby’s legs are held under water for instance.

Language Development *Language* begins as vocalization, which around the age of six months changes into babbling, i. e. combining vocalizations into larger clusters. Around the 12th month, the first recognizable words begin to appear, notably words referring to caretakers, the meaning of which is probably assigned by the adults instead of actually intended by the children, who form such sound patterns (mama, babab etc.) more or less automatically. Meaning assignment forms an interesting illustration of the transactional nature of early developmental processes, with the adult automatically assigning a meaning to sound patterns uttered automatically by infants, which then sets out an iterative pattern of meaning assignments and language production, resulting in the child’s understanding of referential meaning and linguistic significance (around 18 months). Language production in infancy typically consists of one-word sentences expressing various semantic features and nuances (holophrastic speech). Word combinations, mostly in the form of two-word sentences, become abundant around the age of 24 months. Meanwhile, the child’s language production and comprehension gradually assimilates the syntactic features of the child’s ambient language. The process is typically dynamical, in that it consists of an iterative trajectory of assimilations of syntactic features given the child’s current state of language production. Various quantitative features of language development can be described by means of growth processes of linguistic and non-linguistic components con-

nected into a web of mutually supportive or competitive relationships, under the constraints of limited and specific resources [337,340,346]. In addition to continuous growth processes, language development is also characterized by discontinuities, corresponding to the emergence of new strategies of language production [21] or new linguistic forms or categories [259,329].

Our understanding of the dynamics of language development is crucially dependent on how we understand language per se, although – in the spirit of genetic epistemology – our understanding of the nature of language can be greatly enhanced by the study of language development. For a further discussion of this relationship, which is a typical issue of developmental psychological theory formation, see the Sect. “Development as Increasing Complexity Applied to Language Theory and Theory of Language Development.”

Brain Development As regards *brain development*, after the prenatal period, productions of new neurons virtually stops or is greatly reduced. Synaptic connections between neurons are abundant in the beginning, and will be selectively lost in a process called synaptic pruning, which relies on experience and practice. During infancy, the two hemispheres become increasingly specialized (lateralization process). Meanwhile, the brain, or more precisely the cortex, remains highly adaptive, a property known as brain plasticity. Brain specialization occurs through complex, non-linear dynamic processes involving interactions inside the brain and interactions between brain and body and body and environment [187,188,189]. Brain plasticity decreases as the person grows older, but does not disappear [175].

Brain plasticity however, is a typical dynamic property which is nonlinearly dependent on the brain’s developmental history. The change in plasticity is not linear or curvilinear, as the notion of a gradual decline in plasticity suggests. Rather, there are nonlinear peaks of plasticity, known as critical periods or sensitive periods. These critical or sensitive periods, in which the brain is particularly sensitive to particular experiences, are in themselves also self-organizing and dynamic phenomena [49]. They emerge epigenetically from the brain’s development and are thus co-dependent on biological brain growth and the unfolding of experiences, including teaching and learning over developmental time [173,319]. According to a now obsolete view on sensitive periods – critical periods – the sensitive period is like a time window of opportunity that, if missed, will never come back and will leave the person with an irreparable gap in development. This view relates sensitive periods to relatively isolated processes of growth

in the brain that unilaterally govern the developmental process. It is incompatible with the view, for which there is now abundant evidence, that sensitive periods are self-organizational states integrating interdependent phenomena of brain growth, experience and environment [159].

A dramatic illustration of how brain plasticity – and development as a whole, for that matter – always passes through the short-term dynamics of action, is the development of children after hemispherectomy. Hemispherectomy is the surgical removal of a brain hemisphere, mostly as a last possibility for curing major and highly frequent epileptic insults that cannot be treated pharmacologically [22,154,355].

Social and Emotional Development Basic *emotions* such as anger, sadness and happiness, are already present during early infancy, which suggests that basic emotions are innate patterns. However, dynamic systems theorists have challenged this interpretation and view basic emotions as the earliest patterns of emotional expression that arise through self-organization of components from various sources (motor, contextual, ...) and become stable attractor patterns [54,55]. Infants are able to interpret emotions of other persons early in infancy. The ability to understand the relation between an other person's emotional expressions and that person's goals and actions is already developed at the end of the first year [218,240].

Temperament is a person's habitual pattern of emotional reaction, activity level, attention and self-regulation. Temperament in the sense of such stable characteristics occurs from early infancy on, with about 2/3rds of the infants falling in the categories "easy child", "difficult child", "slow starter". Temperament, as an invariant property overarching contextual variability in reaction patterns, is a typical short-term form of stability. Temperament changes over the long-term process of development, but it does so to varying degrees, depending on the person (e. g. extreme patterns are less likely to change) and age (e. g. early temperament tends to change more easily than temperament at a later age). In that sense, temperament is a higher-order short-term attractor pattern of behavior and emotion, which over the long-term tends to shift across the temperamental state space with context-, person- and age-specific velocities. The classical study of Thomas and Chess on temperament development [318] provides an example of a dynamic person-context transaction, known as the goodness-of-fit hypothesis.

A third theme that is of particular importance for early social and emotional development is the development of *attachment*. Around 6 to 8 months, infants begin to develop a strong affectional bond with familiar people, the

mother or primary caretaker in the first place. This strong tie serves as a source of emotions, such as comfort and joy while the object of attachment is present and sadness or discomfort when he or she is absent. Attachment thus serves as a basic model for the emotionally close relationships that will develop and last throughout life. As with temperament, attachment shows a certain stability across short-term fluctuations in contexts, and a partitioning into characteristic patterns. About 2/3ds of the infants show a pattern of secure attachment; other infants show avoidant, resistant or disorganized patterns of attachment. In a dynamic systems framework, these patterns should be interpreted as short-term attractor states, showing a gradual long-term development, with considerable individual differences in the amount of displacement over the developmental state space [342].

Childhood (5 to 12 Years)

Cognitive Development A characteristic feature of cognitive development during this phase is that it makes a transition to a new mode of thought, concrete operational thinking, a term stemming from the work of Piaget (as do most of the basic terms used for stages or phases as represented in the major handbooks). A major feature of this type of cognition is that it shows reversibility, a notion that is similar to the notion of inverse operations in mathematical groups. Hence, for every action the child can think of, it automatically knows there is an imaginable operation that undoes the effect of the first. The emergence of a property such as reversibility amounts to an increase in the complexity of the cognitive system.

From a dynamic point of view, such formal properties must be given a concrete temporal meaning, for instance in the form of a pattern of reasoning about a possible inverse operation that is not explicitly given in perception.

Neo-Piagetian theorists such as Fischer and Case have postulated comparable qualitative extensions of the child's cognitive system. In Fischer's theory, for instance, the cognitive system which is able to represent relations between elements (e. g. a relationship is-a-brother between me and my brother) is transformed into one which can represent relationships between relationships; see for instance [103].

As regards the theoretical interpretation of what such increases in or extensions of complexity of the cognitive system actually mean, various considerations should be taken into account. The first is that properties such as reversibility and comparable formal properties of the cognitive system refer to action potentialities of a system that is based on or consists of an embodied neural network in a concrete, spatiotemporal world. Hence, the gener-

ative cause of child's actions and reasoning is the system consisting of a concrete context or niche on the one hand and the child – as embodied and organic neural-network on the other hand [350]. This generative cause generates a stream of action to which certain formal properties can be ascribed, such as reversibility (and comparable properties). Second, a major feature of dynamic theories is that they view this generative cause as a dynamic interdependence of many factors (memory, perception, action, changes brought about by action in the context, recent experiences, long-term effects of experiences and so forth) on many levels (from the micro-level of neuronal activity via the meso-level of bodily activity to the macro-level of physical-cultural environments). There is no single factor that can be held responsible for an emergent phenomenon such as reversibility of cognitive representations [19,103,271,293,368].

Examples of developmental acquisitions typical of this phase are conservation, classification, and seriation. Conservation, for instance, is the child's ability to understand that physical properties, such as the amount of liquid in a glass, are conserved under certain operations, such as pouring the liquid into another glass. The emergence of conservation understanding is a typical example of a discontinuous development, i. e. it tends to occur in an all-or-none fashion and can be described as a bifurcation or cusp catastrophe [325,327]. However, the developmental emergence of a phenomenon such as conservation is characterized by a considerable intra- and inter-individual variability. That is, children tend to dramatically fluctuate in their type of conservation answer (yes or no) over test occasions during the period of transition. Moreover, children also tend to differ considerably from each other in terms of the path they take towards conservation understanding. The emergence of conservation forms an interesting example of a phenomenon of development that resembles the phenomenon of phase transition in physics.

During childhood, an important aspect of the development of information processing concerns the development of executive functions, functions that critically depend on the frontal lobe. Important skills related to executive functions are inhibitory control (resisting habits, temptations, or distractions), working memory (mentally holding and using information), and cognitive flexibility (adjusting to change). These skills are of crucial importance for and to a considerable extent dependent on scholarly learning during childhood [79,84]. They are closely related to issues of time management (when to do what, how long to retain information), adaptability and goal-and-desire structures (how to focus on which goals and when). The dynamic manipulation of goals and tools (physical and sym-

bolic) and the patterning that emerges as a consequence of this process is a characteristic feature of dynamic accounts of cognition and cognitive development [19,322]. By "patterning" I understand the unique structure of components and relationships in a child's ability repertoire as witnessed through context-specific and context-supported actions. This ability repertoire greatly extends during childhood, as the child goes to school and acquires skills in major domains such as math, reading and writing, conceptual knowledge and so forth. These new domains of thought and symbolic action were prepared and prefigured during the preceding developmental stage (see for instance the discussion on core knowledge), but show an explosive development in (schoolgoing) children during this developmental stage. The acquisition of these fundamental domains of skill and knowledge requires massive recruitments of brain regions [298].

Social Cognition In this section I shall concentrate on an interesting aspect of social cognition which nicely illustrates the complexities of development and which is also increasingly accepted as being of applied and clinical importance, namely Theory-of-Mind. Theory-of-Mind (ToM) is the ability to attribute mental states to oneself and others and to use these attributions in understanding, predicting and explaining behavior of oneself and others [19,214] or the book 'Autism: mind and brain' [111]. A typical aspect of ToM is the ability to entertain first-order beliefs, i. e. beliefs about the beliefs (thoughts, knowledge, ...) of other persons, which typically develops around the age of 4.

There is supportive but indirect evidence of two 'approaches' to ToM: an intuitive (or automatic) and a reflective (or controlled) route [196]. Indirect evidence for an intuitive, neurophysiologically-based understanding of ToM-related properties of other persons comes from the rapidly growing literature on the neuronal systems that underlie the spontaneous understanding of human actions and psychological states of others. An example of such systems is the mirror neuron system (for the relationship between the mirror neuron system and autism, see for instance [113,135,153,183,232,371] but see [81] for critical remarks). There is neuropsychological evidence that specific parts of the brain, such as the medial prefrontal cortex and the temporal-parietal junction are involved in the processing of ToM-related information [110,174,196,266].

In a dynamic systems framework, the ability as described under the definition of ToM above does not emerge from an internal symbolic structure, generating prescriptions for acting under certain conditions (e. g. a false belief experiment). ToM, as it is actually expressed,

in real time, emerges out of a coordination of many components, including perceived similarities in bodily appearance and action between the child and other persons, goals and emotions, inhibitions of associations (e.g. between what I know and what another person eventually knows), automatic simulations of actions and emotions perceived in others, language and linguistic terms for expressions of states of mind and so forth. This ability grows throughout childhood, until it entails abilities such as the ability to think about how other people think you think (second-order beliefs).

Other aspects of social and socio-cognitive development that are worth mentioning in this regard are the views children develop on their own person in relation to others. These views concern the child's self-concept, the sense of competence and self efficacy, gender- and gender-role concepts and so forth. From a dynamic standpoint, these concepts are relatively stable and adaptive patterns of action and emotional evaluation featuring in situations or events that elicit self-referential aspects are in general self-sustaining (e.g. when confronted with a math assignment, a child must ask himself "will I be able to solve that assignment", "will it require little or great effort", etc.). In a social interaction the child may react with intense negative emotion when confronted with statements about his capacities, lack of fit with the gender stereotype and so forth. For examples of the interdependence between short- and long-term time scales in this particular domain, see [301,303].

Self-referential emotions and concerns play an important role in the dynamics of group relationships among peers. Children in this particular age range tend to form patterns of social interaction, creating particular interaction positions such as popular versus rejected children. These so-called sociometric structures (because they formed the subject of an approach to children's social interactions known as sociometry) tend to self-organize as a result of interactions driven by concerns, emotions and perceptions of actions of other children in the group (for an explicit dynamic model of this process, see [208,300,302,303]).

Adolescence

Biological Maturation and the Problem of Timing Puberty and adolescence are ages of accelerated change and development. A major aspect concerns sexual maturity during puberty, which involves major changes in primary sexual development (the person becomes biologically capable of reproduction) and secondary sexual development (the person develops bodily features characteristic of one's sex). These biological changes are related to a spurt in bod-

ily growth, the functioning of the endocrine system and changes in brain development [29,78]. Biological changes are directly related to social changes in terms of interaction, interaction forms and preferences, and changes in cultural expectations.

The structure of interactions between biological, social and cultural changes provides an interesting example of the importance of timing (the temporal ordering of related events). Sexual maturation is an extremely important event. From the point of view of the evolutionary time scale, it must occur at the "right" time, which means that it must run parallel with the timing, i.e. the temporal sequencing, of additional events, required for successful reproduction. The timing is a result of events that preceded it, and on the other hand timing provides a condition for later events. For instance, pubertal timing is responsive to ecological conditions earlier in life, which might be beneficiary as well as adverse. Such conditions may lead to either inhibiting or accelerating onset of puberty. The effects of earlier or later puberty, on the other hand, are non-linear, being most notable in the extremes (both positive and negative [43,95,97]). Differences in the timing of biological maturation are gender specific and occur mostly with girls [207]. The problematic or beneficiary effects of early or late timing of onset or puberty are not a matter of a linear ordering of events. For instance, it is not the eventually early menarche itself that leads girls to experience adverse effects on other variables, such as psychosocial adaptation or birth weight of the first offspring. What matters is that such timing issues cause a problematic (or beneficiary) coordination, that is, mutual ordering, of different time lines or event sequences. Examples of such event sequences are timing of romantic dating and first sexual intercourse, the nature of the partners and potential providers who may or may not be able to support the offspring, and so forth. From the point of view of the participants, i.e. the people who are actually involved in these issues, timing relates to the subjective order of meaningful life-events in their own life and in the life of other persons with whom they are intimately connected (educators, parents, peers, romantic friends,...; see for instance [206]). In short, biological maturation during puberty provides an excellent example of timing problems in a complex dynamic system.

Autonomy, Identity and Connectedness The processes of biological and social development described in the preceding section are related to a renegotiation of the person's place in the familial, social and peer network. According to the historically important and psycho-analytically inspired theory of Erik Erikson (1902–1994), the main chal-

lenge during adolescence is to develop a personal identity (see [41,42] for an overview). For Erikson, “At times, identity refers to a structure or a configuration, at other points it refers to a process. Still on other occasions identity is viewed as both a conscious subjective experience as well as an unconscious entity” [179]. Erikson himself saw the conscious feeling of having a personal identity as “... based on two simultaneous observations: the perception of the selfsameness and continuity of one’s own existence in time and space and the perception of the fact that others recognize one’s sameness and continuity” [100]. Self-sameness means the existence of invariant properties over change that is both developmental (long-term) and actional (short-term). Erikson’s view of identity is an example of an approach that respects the complex nature of the phenomenon under study and by doing so accepts the superposition of its features (it is both a process and an entity, it is an invariant defined by change, it is of the person but only insofar as it is also of other persons, and so on; see Sect. “A Working Definition of Complexity”). Although the concept of identity is an essential feature of the study of adolescent development, its complexity and in particular the acceptance of its complexity is a constant source of unease among researchers. Lichtwarck-Aschoff et al. [193] suggested a description – and the associated empirical study – of identity as a concept distributed across a two dimensional, categorical state space. One dimension concerns the distinction between dynamic and static approaches to phenomena, the other concerns the distinction between short- and long-term processes.

Identity development is closely related to the tension between autonomy and connectedness, both in terms of the short-term dynamics of action and social interaction and the long-term dynamics of development. As children become sexually mature they renegotiate their relationship with their parents by claiming a greater level of personal autonomy and self-determination. Meanwhile, if the relationship with the parents is positive, they do not want to lose the connectedness they feel with their parents, and these opposite tendencies create an interesting developmental dynamics. Lichtwarck-Aschoff et al. [193] have used this tension as the basic component of a long-term dynamic growth model of autonomy and connectedness, in order to account for the classical phenomenon of parent-child conflict during adolescence and for the considerable inter-individual variation in the form and outcome of this process (see also Sect. Development and Resource-Dependent Competition-Support Systems” for an explanation of the model; for comparable models see [180]).

Future Directions

Although human development is a prime example of a complex dynamic system, the theory of complex dynamic systems is not the mainstream approach to studying development. Human development is an example of one such system in which the observer – the researcher, the parent, the educator, ... – are part and parcel of the complex system itself, thus leading to the epistemological and conceptual complexities and entanglements that are characteristic of systems in which the observer is an essential agent, in one way or another. Maybe it is because of this entanglement and epistemological complexity that psychology, and the study of development for that matter, still adheres to an approach of linearizing processes and disentangling causal contributions from identifiable factors. Changes are observable, for instance in the form of theories and approaches that focus on developmental systems, holistic interactions, dynamic systems, and so forth. Unfortunately, these approaches are still mainly theoretical: their contribution to changing the habits of empirical research and theory formation are at best very modest. The statistical sophistication that is required from most manuscripts that are submitted for publication in developmental journals stands in a glaring contrast to the almost complete lack of requirements regarding theoretical depth and reflection. In a book review, the philosopher and psychologist Harré referred to the situation as follows:

It is a remarkable feature of mainstream academic psychology that, alone among the sciences, it should be almost wholly immune to critical appraisal as an enterprise. Methods that have long been shown to be ineffective or worse are still used on a routine basis by hundreds, perhaps thousands of people. Conceptual muddles long exposed to view are evident in almost every issue of standard psychology journals. ([138], p. 1303)

The blessing of our being so closely involved in the processes of development and education (all persons living have gone through a process of bio-psycho-social development themselves and cannot be but daily witnesses of ongoing developmental processes in others) is maybe also its curse. In order to study the process scientifically we probably feel obliged to discard all the intuitive knowledge and models we use in daily life and by doing so surrender to approaches that are critically incompatible with the complexity and dynamics of the phenomenon that we wish to understand. This problem is not easy to solve and will continue to be a very hard problem for the coming decades.

Meanwhile, it is the firm belief of the current author that progress can be made by reformulating the questions and methods on the basis of an approach of complex adaptive systems. The first steps that scholars in this field will have to take will most likely appear utterly naïve and limited to scholars studying other fields where conceptual and mathematical theory building has gone hand in hand with the development of rigorous empirical methods. Students of development will have to invest much more time and intellectual energy in the descriptive and exploratory study of individual developmental trajectories, studied with an intensity that is sufficient to capture the nature of the underlying process and with a thorough understanding of the fact that their focusing on the individual alone is in itself an important factor in the development of that individual. In addition, researchers should not be afraid – and in fact should promote the use – of so-called toy models that are highly simplified representations of the assumed, basic dynamics of some sort of phenomenon. It should not be expected that these toy models will provide us with stunning empirical fits with great data sets (but model fitting in itself is not the primary goal of scientific research, see [126]). However, if we ever wish to understand the intricacies of even relatively simple dynamic processes, we will have to study them in ways that can be made conceptually transparent, and this means, among others, taking toy models serious, as long as they are based on good theoretical considerations. An important consideration for a model is that it should be descriptively adequate, that is, that it captures the “essential” features of the phenomenon it addresses (the term descriptive adequacy stems from Chomskyan linguistics [61]). Since it is difficult to agree on what the essential features of some sort of phenomenon are, it is easier to define a model as not descriptively adequate if it leaves out or is incompatible with at least one feature of the phenomenon that is generally accepted as “essential”. A descriptively adequate approach to human development and education should be one that is deeply compatible with the major features of development, namely its complexity and its dynamic nature.

Bibliography

Primary Literature

1. Abramovitch R, Grusec JE (1978) Peer imitation in a natural setting. *Child Dev* 49:60–65
2. Allen P, Strathern M (2003) Evolution, emergence, and learning in complex systems. *Emerg* 5(4):8–33
3. Anderson JR (1996) ACT: A simple theory of complex cognition. *Am Psychol* 51:355–365
4. Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. *Psychol Rev* 111(4):1036–1060
5. Asbury K, Wachs T, Plomin R (2005) Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intell* 33(6):643–661
6. Atkinson M (1986) Learning and models of development. In: van Geert P (ed) *Theory building in developmental psychology*. North Holland, Amsterdam, pp 105–139
7. Axelrod R (1997) *The complexity of cooperation: agent-based models of competition and collaboration*. Princeton University Press, Princeton
8. Bailey DB, Bruer JT, Symons FJ, Lichtman JW (2001) *Critical thinking about critical periods*. Brookes Publishing, Baltimore
9. Bak P, Sneppen K (1993) Punctuated equilibrium and criticality in a simple model of evolution. *Phys Rev Lett* 71(24):4083–4086
10. Bak P (1996) *How Nature works: the science of self-organized criticality*. Springer, New York
11. Ball P (2005) *Critical mass: how one thing leads to another*. Arrow Books/Random House, London
12. Ballato L, van Geert P, Schumacher N (2007) A dynamic systems model of friendship interaction and risk behavior in adolescence: short-and-long term aspects of influence and selection mechanisms. Internal paper. Groningen: Department of Developmental Psychology
13. Baltes PB, Reese H, Lipsitt L (1980) Lifespan developmental psychology. *Ann Rev Psychol* 31:65–110
14. Bandura A, Bussey K (2004) On broadening the cognitive, motivational, and sociostructural scope of theorizing about gender development and functioning: comment on Martin, Ruble, and Szkrybalo (2002). *Psychol Bull* 130:691–701
15. Bandura A, Walters R (1977) *Social learning and personality development*. Holt, Rinehart and Winston, London
16. Bandura A, Ross D, Ross SA (1963) A comparative test of the status envy, social power, and secondary reinforcement theories of identificatory learning. *J Abnorm Soc Psychol* 67:527–534
17. Bankes S, Lempert R (2004) Robust reasoning with agent-based modeling. *Nonlin Dyn Psychol Life Sci* 8(2):259–277
18. Bankes S, Lempert R, Popper S (2002) Making computational social science effective: epistemology, methodology, and technology. *Soc Sci Comput Rev* 20(4):377–388
19. Barsalou L, Breazeal C, Smith L (2007) Cognition as coordinated non-cognition. *Cogn Process* 8(2):79–91
20. Bassano D, van Geert P (2007) Modeling continuity and discontinuity in utterance length: a quantitative approach to changes, transitions and intra-individual variability in early grammatical development. *Dev Sci* 10(5):588–612
21. Baron-Cohen S, Tager-Flusberg H, Cohen DJ (1993) The impairment of ToMM: some issues. In: Baron-Cohen S, Tager-Flusberg H, Cohen DJ (eds) *Understanding other minds. Perspectives from autism*. Oxford University Press, pp 102–105
22. Battro AM (2001) *Half a brain is enough: the story of Nico*. Cambridge University Press, New York
23. Baum WM, Davison M (2004) Choice in a variable environment: visit patterns in the dynamics of choice. *J Exp Anal Behav* 81:85–127
24. Belpaeme T, Cowley S (2007) Foreword: extending symbol grounding. *Interact Stud* 8(1):1–6

25. Berger KS (2003) The developing person through childhood and adolescence, 6th edn. Worth Publishers, New York
26. Berthouze L, Metta G (2005) Epigenetic robotics: modelling cognitive development in robotic systems. *Cogn Syst Res* 6(3):189–192
27. Berthouze L, Ziemke T (2003) Epigenetic robotics-modelling cognitive development in robotic systems. *Connect Sci* 15(4):147–150
28. Berthouze L, Metta G, Sun R (2005) Editorial: Epigenetic robotics: Modelling cognitive development in robotic systems. *Cogn Syst Res* 6(3):189–192
29. Bianchi-Berthouze N, Kleinsmith A (2003) A categorical approach to affective gesture recognition. *Connect Sci* 15(4):259–269
30. Billington E, DiTommaso N (2003) Demonstrations and applications of the matching law in education. *J Behav Educ* 12(2):91–104
31. Biro S, Leslie A (2007) Infants' perception of goal-directed actions: development through cue-based bootstrapping. *Dev Sci* 10(3):379–398
32. Bjorklund D, Pellegrini A (2000) Child development and evolutionary psychology. *Child Dev* 71(6):1687–1708
33. Bjorklund D, Smith P (2003) Evolutionary developmental psychology: Introduction to the special issue. *J Exp Child Psychol* 85(3):195–198
34. Boccia M, Pedersen C (2001) Animal models of critical and sensitive periods in social and emotional development. In: *Critical thinking about critical periods*. Brookes Publishing, Baltimore, pp 107–127
35. Boettcher S, Paczusi M (1996) Exact Results for Spatiotemporal Correlations in a Self-Organized Critical Model of Punctuated Equilibrium. *Phys Rev Lett* 76(3):348–351
36. Boettcher S, Percus A (2000) Nature's way of optimizing. *Artif Intell* 119:275–286
37. Bongers R, Smitsman A, Michaels C (2003) Geometrics and dynamics of a rod determine how it is used for reaching. *J Motor Behav* 35(1):4–22
38. Borrero J, Vollmer T (2002) An application of the matching law to severe problem behavior. *J Appl Behav Anal* 35(1):13–27
39. Borenstein E, Ruppini E (2005) The evolution of imitation and mirror neurons in adaptive agents. *Cogn Syst Res* 6(3):229–242
40. Bortfeld H, Whitehurst G (2001) Sensitive periods in first language acquisition. In: *Critical thinking about critical periods*. Brookes Publishing, Baltimore, pp 173–192
41. Bosma H, Kunnen E (2001) Determinants and mechanisms in ego identity development: a review and synthesis. *Dev Rev* 21(1):39–66
42. Bosma H, Kunnen E (2001) Identity and emotion: Development through self-organization. Cambridge University Press, New York
43. Boyce WT, Ellis BJ (2005) Biological sensitivity to context: I. An evolutionary-developmental theory of the origins and functions of stress reactivity. *Dev Psychopathol* 17:271–301
44. Breazeal C (2003) Emotion and sociable humanoid robots. *Int J Human-Comput Stud* 59(1):119–155
45. Breazeal C, Brooks R (2005) Robot emotion: a functional perspective. In: *Who needs emotions: the brain meets the robot*. Oxford University Press, New York, pp 271–310
46. Breazeal C, Scassellati B (2000) Infant-like social interactions between a robot and a human caregiver. *Adapt Behav* 8(1):49–74
47. Breazeal C, Scassellati B (2002) Robots that imitate humans. *Trends Cogn Sci* 6(11):481–487
48. Brighton H (2002) Compositional syntax from cultural transmission. *Artif Life* 8:25–54
49. Bruer J (2001) A critical and sensitive period primer. In: *Critical thinking about critical periods*. Brookes Publishing, Baltimore, pp 3–26
50. Bruer J (2002) Avoiding the pediatrician's error: how neuroscientists can help educators (and themselves). *Nature Neurosci* 5:1031–1033
51. Bukatko D, Daehler MW (2001) Child development: a thematic approach, 4th edn. Houghton, Mifflin and Company, Boston
52. Bussey K, Bandura A (1984) Influence of gender constancy and social power on sex-linked modeling. *J Personal Soc Psychol* 47:1292–1302
53. Call J, Carpenter M (2002) Three sources of information in social learning. In: Dautenhahn K, Nehaniv CL (eds) *Imitation in animals and artifacts*. MIT Press, Cambridge, pp 211–2228
54. Camras L, Witherington D (2005) Dynamical systems approaches to emotional development. *Dev Rev* 25(3):328–350
55. Camras L, Oster H, Bakeman R, Meng Z, Ujiie T, Campos J (2007) Do infants show distinct negative facial expressions for fear and anger? Emotional expression in 11-month-old European American, Chinese, and Japanese infants. *Infancy* 11(2):131–155
56. Cañamero L, Gaussier P (2005) Emotion understanding: robots as tools and models. In: *Emotional development: recent research advances*. Oxford University Press, New York, pp 235–258
57. Carpenter M, Call J, Tomasello M (2005) Twelve- and 18-month-olds copy actions in terms of goals. *Dev Sci* 8:13–20
58. Case R (1992) The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge. Lawrence Earlbaum, Hillsdale
59. Cashon CH, Cohen LB (2004) Beyond U-shaped development in infants' processing of faces: an information-processing account. *J Cogn Dev* 5:59–80
60. Casti JL (1994) Complexification: explaining a paradoxical world through the science of surprise. Harper Collins, New York
61. Chomsky N (1957) Syntactic structures. The Hague, Mouton
62. Chomsky N (1959) A review of Bf Skinner's verbal behavior. *Lang* 35(1):26–58
63. Chomsky N (1965) Aspects of the theory of syntax. MIT Press, Oxford
64. Chomsky N (1995) The minimalist program. MIT Press, Cambridge MA
65. Christiansen MH, Dale R (2003) Language evolution and change. In: Arbib MA (ed) *Handbook of brain theory and neural networks*, 2nd edn. MIT Press, Cambridge, pp 604–606
66. Christiansen MH, Kirby S (2003) Language evolution: consensus and controversies. *Trends Cogn Sci* 7(7):300–307
67. Cilliers P (2002) Why we cannot know complex things completely. *Emerg* 4(1/2):77–84
68. Clark A (1997) Being there: putting brain, body and world together again. MIT Press, Cambridge
69. Clark A (2006) Language, embodiment, and the cognitive niche. *Trends Cogn Sci* 10:370–374

70. Clark A (2006) Material symbols. *Philos Psychol* 19(3):291–307
71. Clark A, Chalmers D (1998) The extended mind. *Anal* 58:7–19
72. Clowes R (2007) Semiotic symbols and the missing theory of thinking. *Interact Stud* 8(1):105–124
73. Cole M, Cole SR (1993) *The development of children*, 2nd edn. Scientific American Books, New York
74. Colunga E, Smith L (2005, April) From the Lexicon to Expectations About Kinds: A Role for Associative Learning. *Psychol Rev* 112(2):347–382
75. Cowley S (2007) How human infants deal with symbol grounding. *Interact Stud* 8(1):83–104
76. Crutchfield JP (1994) The calculi of emergence: computation, dynamics and induction. *Physica D* 75:11–54
77. Cziko G (2000) *The things we do using the lessons of Bernard and Darwin to understand the what, how, and why of our behavior*. MIT Press, Cambridge
78. Dahl R, Spear L (2004) *Adolescent brain development: Vulnerabilities and opportunities*. New York Academy of Sciences, New York
79. Davidson M, Amso D, Anderson L, Diamond A (2006) Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. *Neuropsychol* 44(11):2037–2078
80. Dawson-Tunik T, Commons M, Wilson M, Fischer K (2005, June) The shape of development. *Eur J Dev Psychol* 2(2):163–195
81. de Hamilton CA, Brindley R, Frith U (2007) Imitation and action understanding in autistic spectrum disorders: How valid is the hypothesis of a deficit in the mirror neuron system? *Neuropsychologia* 45(8):1859–1868
82. Demetriou A, Kyriakides L (2006, June) The functional and developmental organization of cognitive developmental sequences. *Br J Educational Psychol* 76(2):209–242
83. Demiris Y, Johnson M (2003) Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning. *Connect Sci* 15(4):231–243
84. Diamond A, Barnett WS, Thomas J, Munro S (2007) The early years: preschool program improves cognitive control. *Sci* 318:1387–1388
85. Dishion T, Bullock B, Granic I (2002, September) Pragmatism in modeling peer influence: Dynamics, outcomes and change processes. *Dev Psychopathol* 14(4):969–981
86. Dominey PF (2007) Towards a construction-based framework for development of language, event perception and social cognition: insights from grounded robotics and simulation. *Neurocomput* 70(13):2288–2302
87. Dominey PF, Boucher J-D (2005) Developmental stages of perception and language acquisition in a perceptually grounded robot. *Cogn Syst Res* 6(3):243–259
88. Dooley K (2002) Organizational complexity In: Warner M (ed) *International encyclopedia of business and management*. Thompson Learning, London, pp 5013–5022
89. Dooley K, Corman S (2002) Agent-based, genetic, and emergent computational models of complex systems. In: Kiel LD (ed) *Encyclopedia of life support systems (EOLSS)*. UNESCO/EOLSS, Oxford UK, <http://www.eolss.net>
90. Duncan G, Dowsett C, Claessens A, Magnuson K, Huston A, Klebanov P et al (2007) School readiness and later achievement. *Dev Psychol* 43(6):1428–1446
91. Eggleston R, McCreight K, Young M (2005) Distributed cognition and situated behavior. Modeling human behavior with integrated cognitive architectures: comparison, evaluation, and validation. Lawrence Erlbaum, Mahwah, pp 177–235
92. Elbers E (2004) Conversational asymmetry and the child's perspective in developmental and educational research. *Int J Disabil Dev Educ* 51(2):201–215
93. Eley T, Sugden K, Corsico A, Gregory A, Sham P, McGuffin P et al (2004) Gene-environment interaction analysis of serotonin system markers with adolescent depression. *Mol Psychiatry* 9(10):908–915
94. Elliott E, Kiel L (2004) Agent-based modeling in the social and behavioral sciences. *Nonlin Dyn Psychol Life Sci* 8(2):121–130
95. Ellis BJ (2004) Timing of pubertal maturation in girls: An integrated life history approach. *Psychol Bull* 130:920–958
96. Ellis B, Bjorklund D (2005) *Origins of the social mind: evolutionary psychology and child development*. Guilford Press, New York
97. Ellis BJ, Essex MJ, Boyce WT (2005) Biological sensitivity to context: II. Empirical explorations of an evolutionary-developmental theory. *Dev Psychopathol* 17:303–328
98. Eoyang G (2004) The practitioner's landscape. *Emerg: Complex Organ* 6(1/2):55–60
99. Erhlagen W, Schöner G (2002, July) Dynamic field theory of movement preparation. *Psychol Rev* 109(3):545–572
100. Erikson EH (1968) *Identity: youth and crisis*. Norton, New York
101. Fioretti G, Visser B (2004) A cognitive interpretation of organizational complexity. *Emerg: Complex Organ* 6(1/2):11–23
102. Fischer KW (1980) A theory of cognitive development: the control and construction of hierarchies of skills. *Psychol Rev* 87:477–531
103. Fischer KW, Bidell TR (2006) Dynamic development of action, thought, and emotion. In: Damon W, Lerner RM (eds) *Theoretical models of human development. Handbook of child psychology*, vol 1, 6th edn. Wiley, New York, pp 313–399
104. Fischer KW, Kennedy B (1997) Tools for analyzing the many shapes of development: the case of self-in-rerelationships in Korea. In: *Change and development: issues of theory, method, and application*. Lawrence Erlbaum, Mahwah, pp 117–152
105. Fischer KW, Rose SP (1999) Rulers, models, and nonlinear dynamics: measurement and method in developmental research. In: Savelsbergh G, van der Maas H, van Geert P (eds) *Nonlinear developmental processes*. Royal Netherlands Academy of Arts and Sciences, Amsterdam, pp 197–212
106. Fitzpatrick P, Arsenio A, Torres-Jara E (2006) Reinforcing robot perception of multi-modal events through repetition and redundancy and repetition and redundancy. *Interact Stud* 7(2):171–196
107. Fogel A (1993) *Developing through relationships: Origins of communication, self, and culture*. University of Chicago Press, Chicago
108. Ford DH, Lerner RM (1992) *Developmental systems theory: an integrative approach*. Sage Publications, Thousand Oaks
109. Förster J, Liberman N, Friedman R (2007) Seven principles of goal activation: a systematic approach to distinguishing goal priming from priming of non-goal constructs. *Personal Soc Psychol Rev* 11(3):211–233
110. Frith U, Frith CD (2003) Development and neurophysiology of mentalizing. *Philos Trans Royal Soc Lond Biol Sci* 358:459–473
111. Frith U, Hill E (2003) *Autism: Mind and brain*. Oxford University Press, New York

112. Frith U, Morton J, Leslie A (1991) The cognitive basis of a biological disorder: autism. *Trends Neurosci* 14(10):433–438
113. Gallese V (2006) Intentional attunement: A neurophysiological perspective on social cognition and its disruption in autism. *Brain Res* 1079:15–24
114. Gallistel CR, Gibbon J (2000) Time, rate, and conditioning. *Psychol Rev* 107:289–344
115. Geary D (2006) Evolutionary developmental psychology: current status and future directions. *Dev Rev* 26(2):113–119
116. Gergely G (2003) What should a robot learn from an infant? Mechanisms of action interpretation and observational learning in infancy. *Connect Sci* 15(4):191–209
117. Gershkoff-Stowe L, Thelen E (2004) U-shaped changes in behavior: a dynamic systems perspective. *J Cogn Dev* 5:11–36
118. Gibson J (1966) The senses considered as perceptual systems. Houghton Mifflin, Oxford
119. Goldin-Meadow S, Wagner S (2005) How our hands help us learn. *Trends Cogn Sci* 9(5):234–241
120. Goldspink C (2007) Transforming education: evidential support for a complex systems approach. *Emerg: Complex Organ* 9(1/2):77–92
121. Goldstein J (1999) Emergence as a construct: History and issues. *Emergence* 1:49–72
122. Goldstein J (2002, October) The Singular Nature of Emergent Levels: Suggestions for a Theory of Emergence. *Nonlinear Dyn Psychol Life Sci* 6(4):293–309
123. Goldstone R, Janssen M (2005) Computational models of collective behavior. *Trends Cogn Sci* 9(9):424–430
124. Gottfried AW, Gottfried AE (1974) Influence of social power vs status envy modeled behaviors on children's preferences for models. *Psycholog Rep* 34:1147–1150
125. Gottlieb G (2001) A developmental psychobiological systems view: early formulation and current status. MIT Press, Cambridge
126. Gottman JM, Murray JD, Swanson CC, Tyson R, Swanson KR (2002) The mathematics of marriage: dynamic nonlinear models. MIT Press, Cambridge
127. Granic I, Hollenstein T (2003) Dynamic systems methods for models of developmental psychopathology. *Dev Psychopathol* 15(3):641–669
128. Granic I, Lamey A (2002) Combining dynamic systems and multivariate analyses to compare the mother-child interactions of externalizing subtypes. *J Abnorm Child Psychol* 30(3):265–283
129. Granic I, Hollenstein T, Dishion T, Patterson G (2003) Longitudinal analysis of flexibility and reorganization in early adolescence: a dynamic systems study of family interactions. *Dev Psychol* 39(3):606–617
130. Granott N (2002) How microdevelopment creates macrodevelopment: Reiterated sequences, backward transitions, and the Zone of Current Development. *Microdevelopment: Transition processes in development and learning*. Cambridge University Press, New York, pp 213–242
131. Grusec JE (1992) Social learning theory and developmental psychology: the legacies of Robert Sears and Albert Bandura. *Dev Psychol* 28:776–786
132. Grusec JE, Abramovitch R (1982) Imitation of peers and adults in a natural setting: a functional analysis. *Child Dev* 53:636–642
133. Guerin S, Kunkle D (2004) Emergence of constraint in self-organizing systems. *Nonlin Dyn Psychol Life Sci* 8(2):131–146
134. Gulyás L (2002) On the transition to agent-based modeling: Implementation strategies from variables to agents. *Soc Sci Comput Rev* 20(4):389–399
135. Hadjikhani N, Joseph RM, Snyder J, Tager-Flusberg H (2006) Anatomical differences in the mirror neuron system and social cognition network in autism. *Cereb Cortex* 16:1276–1282
136. Hamaker EL, Dolan CV, Molenaar PCM (2005) Statistical modeling of the individual: rationale and application of multivariate stationary time series analysis. *Multivar Behav Res* 40:207–233
137. Hammond S, Sanders M (2002) Dialogue as social self-organization: an introduction. *Emerg* 4(4):7–24
138. Harré R (2000) Acts of living. *Sci* 289(25):1303–1304
139. Hartelman PA, van der Maas HLJ, Molenaar PCM (1998) Detecting and modelling developmental transitions. *Br J Dev Psychol* 16:97–122
140. Henrickson L (2002) Old wine in a new wineskin: college choice, college access using agent-based modeling. *Soc Sci Comput Rev* 20(4):400–419
141. Hernández Blasi C, Bjorklund D (2003) Evolutionary developmental psychology: a new tool for better understanding human ontogeny. *Human Dev* 46(5):259–281
142. Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. *J Exp Anal Behav* 4:267–272
143. Herrnstein RJ (1970) On the law of effect. *J Exp Anal Behav* 13:243–266
144. Herrnstein RJ, Prelec D (1991) Melioration: A theory of distributed choice. *J Econom Perspect* 5:137–156
145. Hodgson G (2000) The concept of emergence in social sciences: its history and importance. *Emerg* 2(4):65–77
146. Holland JH (1995) Hidden order: How adaptation builds complexity. Addison-Wesley, Reading
147. Holland JH (1998) Emergence. From chaos to order. Addison-Wesley, Redwood City
148. Hollenstein T (2007) State space grids: analyzing dynamics across development. *Int J Beh Dev* 31(4):384–396
149. Hollenstein T, Lewis M (2006) A state space analysis of emotion and flexibility in parent-child interactions. *Emot* 6(4):656–662
150. Hosenfeld B, van der Maas HLJ, van den Boom DC (1997a) Detecting bimodality in the analogical reasoning performance of elementary schoolchildren. *Int J Behav Dev* 20:529–547
151. Hosenfeld B, van der Maas HLJ, van den Boom DC (1997b) Indicators of discontinuous change in the development of analogical reasoning. *J Exp Child Psychol* 64:367–395
152. Howe ML, Lewis MD (2005) The importance of dynamic systems approaches for understanding development. *Dev Rev* 25:247–251
153. Iacoboni M, Dapretto M (2006) The mirror neuron system and the consequences of its dysfunction. *Nat Rev Neurosci* 7:942–951
154. Immordino-Yang MH (2005) A tale of two cases: emotion and affective prosody after hemispherectomy. *ProQuest Information Learning*
155. Immordino-Yang MH (2007) A tale of two cases. Lessons for education from the study of two boys living with half their brains. *Mind Brain Educ* 1(2):66–83
156. Iverson J, Goldin-Meadow S (2005) Gesture paves the way for language development. *Psychol Sci* 16(5):367–371
157. Jansen BRJ, van der Maas HLJ (2001b) Evidence for the Phase

- transition from rule I to rule II on the balance scale task. *Developmental Review*, 21, 450–494
158. Jansen BRJ, van der Maas HLJ (2002b) The development of children's rule use on the balance scale task. *J Exp Child Psychol* 81:383–416
 159. Johnson M (2005) Sensitive periods in functional brain development: problems and prospects. *Dev Psychobiol* 46(3):287–292
 160. Jones G (2007) Agent-based modeling: Use with necessary caution. *Am J Public Health* 97(5):780–781
 161. Jones G, Ritter FE, Wood DJ (2000) Using a cognitive architecture to examine what develops. *Psychol Sci* 11(2):1–8
 162. Kail RV (2001) *Children and their development*, 2nd edn. Prentice-Hall, Englewood Cliffs
 163. Kaplan F, Hafner V (2006) The challenges of joint attention. *Interact Stud* 7(2):135–169
 164. Ke J, Holland J (2006) Language origin from an emergentist perspective. *Appl Linguist* 27(4):691–716
 165. Kello CT (2004) Characterizing the evolutionary dynamics of language. *Trends Cogn Sci* 8:392–394
 166. Kendal J (2006) Review of: Galef BG Jr, Heyes CM (2004) Social learning and imitation. *Learn Behav* 32(1):1–140. *Interact Stud* 7(2):273–288
 167. Kerpelman J, Pittman J, Lamke L (1997) Toward a micro-process perspective on adolescent identity development: an identity control theory approach. *J Adolesc Res* 12(3):325–346
 168. Kirby S (2002) Natural language from artificial life. *Artif Life* 8:185–215
 169. Kirby S, Smith K, Brighton H (2004) From UG to universals: linguistic adaptation through iterated learning. *Stud Lang* 28:587–607
 170. Kitchener K, Lynch C, Fischer K, Wood P (1993) Developmental range of reflective judgment: the effect of contextual support and practice on developmental stage. *Dev Psychol* 29(5):893–906
 171. Klahr D, Wallace JG (1976) Cognitive development: an information processing approach. Lawrence Erlbaum, Hillsdale
 172. Knafo A, Plomin R (2006) Prosocial behavior from early to middle childhood: genetic and environmental influences on stability and change. *Dev Psychol* 42(5):771–786
 173. Knudsen E (2004) Sensitive periods in the development of the brain and behavior. *J Cogn Neurosci* 16(8):1412–1425
 174. Kobayashi C, Glover GH, Temple E (2007) Children's and adults' neural bases of verbal and nonverbal 'theory of mind'. *Neuropsychologia*, 45, 1522–1532
 175. Kolb B, Whishaw IQ (1998) Brain plasticity and behavior. *Ann Rev Psychol* 49:43–64
 176. Kouritzin S (1999) Face[t]s of first language loss. Lawrence Erlbaum, Mahwah
 177. Kovas Y, Plomin R (2007) Learning abilities and disabilities: generalist genes, specialist environments. *Current Dir Psychol Sci* 16(5):284–288
 178. Kovas Y, Petrill S, Plomin R (2007) The origins of diverse domains of mathematics: generalist genes but specialist environments. *J Educ Psychol* 99(1):128–139
 179. Kroger J (2004) *Identity in adolescence. The balance between self and other*, 3rd edn. Routledge, East Sussex
 180. Kunnen E, Bosma H (2000) Development of meaning making: a dynamic systems approach. *New Ideas Psychol* 18(1):57–82
 181. Lasnik H (2002) The minimalist program in syntax. *Trends Cogn Sci* 6(10):432–437
 182. Lee K, Park N, Song H (2005) Can a robot be perceived as a developing creature? Effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it. *Human Commun Res* 31(4):538–563
 183. Lepage JF, Théoret H (2006) EEG evidence for the presence of an action observation-execution matching system in children. *Eur J Neurosci* 23:2505–2510
 184. Lerner RM (2006) *Developmental science, developmental systems, and contemporary theories of human development*. Wiley, Hoboken
 185. Lewis M (1995, March) Cognition-emotion feedback and the self-organization of developmental paths. *Hum Dev* 38(2):71–102
 186. Lewis M (2000, January) The promise of dynamic systems approaches for an integrated account of human development. *Child Dev* 71(1):36–43
 187. Lewis MD (2005) Bridging emotion theory and neurobiology through dynamic systems modeling. *Behav Brain Sci* 28:169–245
 188. Lewis MD (2005) Self-organizing individual differences in brain development. *Dev Rev* 25(3):252–277
 189. Lewis MD, Todd R (2007) The self-regulating brain: cortical-subcortical feedback and the development of intelligent action. *Cogn Dev* 22(4):406–430
 190. Lewis MD, Lamey A, Douglas L (1999) A new dynamic systems method for the analysis of early socioemotional development. *Dev Sci* 2(4):457–475
 191. Lewis MD, Zimmerman S, Hollenstein T, Lamey A (2004) Reorganization in coping behavior at 1½ years: Dynamic systems and normative change. *Dev Sci* 7(1):56–73
 192. Lewis T, Maurer D (2005) Multiple sensitive periods in human visual development: evidence from visually deprived children. *Dev Psychobiol* 46(3):163–183
 193. Lichtwarck-Aschoff A, van Geert P, Bosma H, Kunnen S (2008) Time and identity: a framework for research and theory formation. *Dev Rev* 28(3):370–400
 194. Lickliter R, Honeycutt H (2003) Developmental dynamics: toward a biologically plausible evolutionary psychology. *Psychol Bull* 129:819–835
 195. Lickliter R, Honeycutt H (2003) Evolutionary approaches to cognitive development: status and strategy. *J Cogn Dev* 4(4):459–473
 196. Lieberman MD (2007) Social cognitive neuroscience: A review of core processes. *Ann Rev Psychol* 58:259–289
 197. Lindblom J, Ziemke T (2006) The social body in motion: cognitive development in infants and androids. *Connect Sci* 18(4):333–346
 198. Locke J (2007) Bimodal signaling in infancy: Motor behavior, reference, and the evolution of spoken language. *Interact Stud* 8(1):159–175
 199. Lockman JJ (2000) A perception-action perspective on tool use development. *Child Dev* 71(1):137–144
 200. Lopes L, Chauhan A (2007) How many words can my robot learn? An approach and experiments with one-class learning. *Interact Stud* 8(1):53–81
 201. Louro M, Pieters R, Zeelenberg M (2007) Dynamics of multiple-goal pursuit. *J Personal Soc Psychol* 93(2):174–193
 202. Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. *Connect Sci* 15(4):151–190

203. MacDorman K (2007) Afterword: life after the symbol system metaphor. *Interact Stud* 8(1):1–999
204. MacDorman K, Ishiguro H (2006) Toward social mechanisms of android science: a CogSci Workshop, 25 and 26 July 2005, Stresa, Italy. *Interact Stud* 7(2):289–296
205. Maestripieri D, Roney J (2006). Evolutionary developmental psychology: contributions from comparative research with nonhuman primates. *Dev Rev* 26(2):120–137
206. Magnusson D (1999) On the individual: a person-oriented approach to developmental research. *Eur Psychol* 4(4):205–218
207. Magnusson D (2001) The holistic-interactionistic paradigm: some directions for empirical developmental research. *Eur Psychol* 6(3):153–162
208. Martin LC, Fabes RA, Hanish LD, Hollenstein T (2005) Social dynamics in the preschool. *Dev Rev* 25:299–327
209. Masten A (2007) Resilience in developing systems: progress and promise as the fourth wave rises. *Dev Psychopath* 19(3):921–930
210. McShane J (1991) Cognitive development: an information processing approach. Blackwell, Oxford
211. Metta G, Berthouze L (2006) Epigenetic robotics: modelling cognitive development in robotic systems. *Interact Stud* 7(2):129–134
212. Metta G, Sandini G, Natale L, Craighero L, Fadiga L (2006) Understanding mirror neurons: a bio-robotic approach. *Interact Stud* 7(2):197–232
213. Mezic I, Gruenewald P, Gorman D, Mezic J (2007) Agent-based modeling: use with necessary caution: reply. *Am J Public Health* 97(5):781–782
214. Mitchell P (1997) Introduction to theory of mind. Children, autism and apes. Arnold, London
215. Moldoveanu M (2004) An intersubjective measure of organizational complexity: a new approach to the study of complexity in organizations. *Emerg: Complex Organ* 6(3):9–26
216. Molenaar P (2004) A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Meas: Interdiscipl Res Perspect* 2(4):201–218
217. Molenaar P, Huizenga H, Nesselroade J (2003) The relationship between the structure of interindividual and intraindividual variability: a theoretical and empirical vindication of developmental systems theory. In: *Understanding human development: dialogues with lifespan psychology*. Kluwer, Dordrecht
218. Montague DPF, Walker-Andrews AS (2001) Peekaboo: a new look at infants' perception of emotion expressions. *Dev Psychol* 37:826–838
219. Moore D (2003) Trying to fix the development in evolutionary developmental psychology. *Am J Psychol* 116(2):299–307
220. Mosekilde E (1998) Topics in nonlinear dynamics: applications to physics biology and economic systems. World Scientific, Singapore
221. Moss S, Edmonds B (2005) Sociology and simulation: statistical and qualitative cross-validation. *Am J Sociol* 110(4):1095–1131
222. Munakata Y, Casey BJ, Diamond A (2004) Developmental cognitive neuroscience: progress and potential. *Trends Cogn Sci* 8(3): 122–128
223. Murray L, Kollins S (2000) Effects of methylphenidate on sensitivity to reinforcement in children diagnosed with attention deficit hyperactivity disorder: an application of the matching law. *J Appl Behav Anal* 33(4):573–591
224. Musher-Eizenman DR, Nesselroade JR, Schmitz B (2002) Perceived control and academic performance: a comparison of high- and low-performing children on within-person change patterns. *Int J Behav Dev* 26:540–547
225. Nagai Y, Hosoda K, Morita A, Asada M (2003) A constructive model for the development of joint attention. *Connect Sci* 15(4):211–229
226. Namy LL, Campbell AL, Tomasello M (2004) The changing role of iconicity in non-verbal symbol learning: a U-shaped trajectory in the acquisition of arbitrary gestures. *J Cogn Dev* 5:37–57
227. Newman BM, Newman PR (2006) Development through life: a psychosocial approach. Thomson Wadsworth, Belmont
228. Niyogi P, Berwick RC (2001) A dynamical systems model for language change. *Complex Syst* 11(3):161–204
229. Nowak MA, Komarova NL (2001) Towards an evolutionary theory of language. *Trends Cogn Sci* 5(7):288–295
230. Nowak MA, Komarova NL, Niyogi P (2001) Evolution of universal grammar. *Sci* 291:114–118
231. Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of language. *Nature* 417:611–617
232. Oberman LM, Ramachandran VS (2007) The simulating social mind: The role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychol Bull* 133:310–327
233. Oyama S, Griffiths PE, Gray RD (2001) Cycles of contingency: developmental systems and evolution. MIT Press, Cambridge
234. Pan B, Gleason J (1986) The study of language loss: models and hypotheses for an emerging discipline. *Appl Psycholinguist* 7(3):193–206
235. Pellegrini A, Bjorklund D (2004) The ontogeny and phylogeny of children's object and fantasy play. *Human Nature* 15(1):23–43
236. Pepper SC (1926) Emergence, *The J Philos* 23(9):241–245
237. Pepper SC (2004) Emergence. *Emerg: Complex Organ* 6(4):66–245
238. Pessa E (2004) Quantum connectionism and the emergence of cognition. In: Globus GG, Pribram KH, Vitiello G (eds) *Brain and being: at the boundary between science, philosophy, language and arts*. Benjamins, Amsterdam, pp 129–148
239. Bjorne P, Balkenius C (2005) A model of attentional impairments in autism: first steps toward a computational theory. *Cogn Syst Res* 6(3):193–204
240. Phillips AT, Wellman HM, Spelke ES (2002) Infants' ability to connect gaze and emotional expression to intentional action. *Cogn* 85:53–78
241. Piaget J (1950) Introduction à l'épistémologie génétique; I La pensée mathématique. – II La pensée physique. – III La pensée biologique, la pensée psychologique et la pensée sociologique. Presses Universitaires de France, Paris
242. Pincus D, Guastello SJ (2005) Nonlinear Dynamics and Interpersonal Correlates of Verbal Turn-Taking Patterns in a Group Therapy Session. *Small Group Res* 36:635–677
243. Plomin R (2004) Genetics and developmental psychology. *Merrill-Palmer Quarterly* 50(3):341–352
244. Plomin R (2007) Genetics and developmental psychology. Wayne State University Press, Detroit
245. Plomin R, Asbury K (2005) Nature and nurture: genetic and environmental influences on behavior. *Ann Am Acad Political Soc Sci* 600:86–98

246. Plomin R, Kovas Y (2005) Generalist genes and learning disabilities. *Psycholog Bull* 131(4):592–617
247. Popper KR (1976) *Unended quest: an intellectual autobiography*. Fontana, London
248. Prince CG, Hollich GJ (2005) Synching models with infants: a perceptual-level model of infant audio-visual synchrony detection. *Cogn Syst Res* 6(3):205–228
249. Raczynski S (2006) A self-destruction game. *Nonlin Dyn Psychol Life Sci* 10(4):471–483
250. Richardson K (2004) Systems theory and complexity: Part 1. *Emerg: Complex Organ* 6(3):75–79
251. Rieber RW, Carton AS (eds) (1987) *The collected works of LS Vygotsky: vol I Problems of general psychology*. Plenum, New York, pp 37–285
252. Ripa J, Ives AR (2003) Food web dynamics in correlated and autocorrelated environments. *Theor Popul Biol* 64(3):369–384
253. Robertson D (2005) Review of agent-based modeling toolkits. *Acad Manag Learn Educ* 4(4):525–527
254. Robinson BF, Mervis CB (1998) Disentangling early language development: modeling lexical and grammatical acquisition using and extension of case-study methodology. *Dev Psychol* 34:363–375
255. Rogoff B (1990) *Apprenticeship in thinking*. Oxford University Press, New York
256. Ronald A, Happé F, Plomin R (2006) Genetic research into autism. *Sci* 311(5763):952–952
257. Rose SP, Fischer KW (1998) Models and rulers in dynamical development. *Brit J Dev Psychol* 16:123–131
258. Rubino C (2002) The consolations of uncertainty: time, change, and complexity. *Emerg* 4(1/2):200–206
259. Ruhland R, van Geert P (1998) Jumping into syntax. *Brit J Dev Psychol* 16:65–95
260. Rushton JP (1975) Generosity in children: immediate and long-term effects of modeling, preaching, and moral judgment. *J Personal Soc Psychol* 31:459–466
261. Rutter M (2007) Gene-environment interdependence. *Dev Sci* 10:1–12
262. Sallach D (2003) Social theory and agent architectures: prospective issues in rapid-discovery social science. *Soc Sci Comput Rev* 21:179–195
263. Sameroff AJ (1975) Transactional models in early social relations. *Human Dev* 18(1):65–79
264. Sameroff AJ (2000) Developmental systems and psychopathology. *Dev Psychopathol* 12:297–312
265. Sameroff AJ, MacKenzie M (2003) Research strategies for capturing transactional models of development: the limits of the possible. *Dev Psychopathol* 15(3):613–640
266. Saxe R, Wexler A (2005) Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* 43:1391–1399
267. Schlesinger M (2003) A lesson from robotics: modeling infants as autonomous agents. *Adapt Behav* 11(2):97–107
268. Schlesinger M, Casey P (2003) Where infants look when impossible things happen: simulating and testing a gaze-direction model. *Conn Sci* 15(4):271–280
269. Schlesinger M, Parisi D (2001) The agent-based approach: a new direction for computational models of development. *Dev Rev* 21(1):121–146
270. Schlindwein S, Ison R (2004) Human knowing and perceived complexity: implications for systems practice. *Emerg: Complex Organ* 6(3):27–32
271. Schöner G, Dineva E (2007) Dynamic instabilities as mechanisms for emergence. *Dev Sci* 10(1):69–74
272. Schöner G, Thelen E (2006) Using dynamic field theory to rethink infant habituation. *Psychol Rev* 113:273–299
273. Schutte AR, Spencer JP (2002) Generalizing the dynamic field theory of the A-not-B error beyond infancy: three-year-olds' delay- and experience-dependent location memory biases. *Child Dev* 73:377–404
274. Schutte AR, Spencer JP, Schöner G (2003) Testing the dynamic field theory: working memory for locations becomes more spatially precise over development. *Child Dev* 74(5):1393–1417
275. Sebeok TA (2001) Biosemiotics: its roots, proliferation, and prospects. *Semiot* 134:61–78
276. Seifert KL, Hoffnung RJ (1991) *Child and adolescent development*, 2nd edn. Houghton, Mifflin and Company, Boston
277. Shahan T, Podlesnik C (2007) Divided attention and the matching law: sample duration affects sensitivity to reinforcement allocation. *Learn Behav* 35(3):141–148
278. Shriver M, Kramer J (1997) Application of the generalized matching law for description of student behavior in the classroom. *J Behav Educ* 7(2):131–149
279. Siegler RS (1983) Information processing approaches to development. In: Mussen PH (ed) *Recent advances in cognitive developmental theory*. Wiley, New York
280. Sigelman CK, Rider EA (2006) *Life-span human development*, 5th edn. Thomson Wadsworth, Belmont
281. Simon H (1962) The architecture of complexity. *Proc Am Philos Soc* ISSN 0003–049X 106(6):467–482
282. Simon H, Cilliers P (2005) The architecture of complexity. *Emerg: Complex Organ*, 7(3/4), 138–154
283. Skinner BF (1938) *The behavior of organisms: an experimental analysis*. Appleton, New York
284. Skinner BF (1945) The operational analysis of psychological terms. *Psychol Rev* :52:270–77
285. Skinner BF (1950) Are theories of learning necessary? *Psychol Rev* 57:193–216
286. Skinner C, Robinson S, Johns G, Logan P, Belfiore P (1996) Applying Herrnstein's matching law to influence students' choice to complete difficult academic tasks. *J Exp Educ* 65(1):5–17
287. Smith L (2005, September) Cognition as a dynamic system: Principles from embodiment. *Dev Rev* 25(3):278–298
288. Smith E, Conrey F (2007) Agent-based modeling: a new approach for theory building in social psychology. *Personal Soc Psychol Rev* 11(1):1–18
289. Smith K (2002) Natural selection and cultural selection in the evolution of communication. *Adapt Behav* 10(1):25–44
290. Smith K, Kirby S, Brighton H (2003) Iterated learning: a framework for the emergence of language. *Artif Life* 9:371–386
291. Smith LB (1999) Not 'either', not 'or', not 'both'. *Dev Sci* 2(2):162–163
292. Smith LB, Thelen E, Titzer R, McLin D (1999) Knowing in the context of acting: the task dynamics of the A-not-B error. *Psychol Rev* 106:235–260
293. Smith LB, Breazeal C (2007) The dynamic lift of developmental process. *Dev Sci* 10(1):61–68

294. Spear L (2000, August) Neurobehavioral changes in adolescence. *Curr Dir Psychol Sci* 9(4):111–114
295. Spelke E, Kinzler K (2007) Core knowledge. *Dev Sci* 10(1):89–96
296. Spencer JP, Schöner G (2003) Bridging the representational gap in the dynamic systems approach to development. *Dev Sci* 4(6):392–412
297. Spencer JP, Smith LB, Thelen E (2001) Tests of a dynamic systems account of the A-not-B error: The influence of prior experience on the spatial memory abilities of two-year-olds. *Child Dev* 72(5):1327–1346
298. Stanislas D (2007) A few steps toward a science of mental life. *Mind Brain Educ* 1(1):28–47
299. Steenbeek H (2006) Modeling dyadic child-peer interactions: sociometric status, emotional expressions and instrumental actions during play. Doctoral Dissertation, University of Groningen
300. Steenbeek H, van Geert P (2005) A dynamic systems model of dyadic interaction during play of two children. *Eur J Dev Psychol* 2(2):105–145
301. Steenbeek H, van Geert P (2007) Do you still like to play with him? Variability and the dynamic nature of children's sociometric ratings. *Neth J Psychol* 63(3):86–101
302. Steenbeek H, van Geert P (2007) The empirical validation of a dynamic systems model of interaction; do children of different sociometric statuses differ in their dyadic play interactions? *Dev Sci* 11(2):253–281
303. Steenbeek H, van Geert P (2008) A dynamic systems approach to dyadic interaction in children: emotional expression, action, dyadic play, and sociometric status. *Dev Rev* 27(1):1–40
304. Serman J (1994) Learning in and about complex systems. *Syst Dyn Rev* 10(2):291–330
305. Strauss S, Stavy R (1982) U-shaped behavioral growth. Academic, New York
306. Striano T, Henning A, Stahl D (2006) Sensitivity to interpersonal timing at 3 and 6 months of age. *Interact Stud* 7(2):251–271
307. Striano T, Henning A, Vaish A (2006) Selective looking by 12-month-olds to a temporally contingent partner. *Interact Stud* 7(2):233–250
308. Sulis W (2004) Archetypal dynamical systems and semantic frames in vertical and horizontal emergence. *Emerg: Complex Organ* 6(3):52–64
309. Sun R (2001) Cognitive science meets multi-agent systems: a prolegomenon. *Philos Psychol* 14:5–28
310. Taptiklis T (2005) After managerialism. *Emerg: Complex Organ* 7(3/4):2–14
311. Thagard P (1996) *Mind: introduction to cognitive science*. MIT Press, Cambridge
312. Thelen E, Fisher D (1982) Newborn stepping: an explanation for a 'disappearing' reflex. *Dev Psychol* 18(5):760–775
313. Thelen E, Schöner G, Scheier C, Smith L (2001) The dynamics of embodiment: A field theory of infant perseverative reaching. *Behav Brain Sci* 24(1):1–86
314. Thelen E, Smith L (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge
315. Thelen E, Whitmyer V (2005) Using dynamic field theory to conceptualize the interface of perception, cognition, and action. In: *Action as an organizer of learning and development*, vol 33 in the Minnesota Symposia on Child Psychology. Lawrence Erlbaum, Mahwah, pp 243–280
316. Thom R (1990) *Semiophysics: a sketch*, Addison-Wesley, Redwood City
317. Thomae H (1959) Entwicklungsbegriff und Entwicklungstheorie. In: Bergius R, Thomae H (eds) *Handbuch der Psychologie*, Band 3: Entwicklungspsychologie. Verlag für Psychologie, Göttingen
318. Thomas A, Chess S (1977) *Temperament and development*. Brunner/Mazel, Oxford
319. Thomas M, Johnson M (2006) The computational modeling of sensitive periods. *Dev Psychobiol* 48(4):337–344
320. Thompson DE, Russell J (2004) The ghost condition: imitation versus emulation in young children's observational learning. *Dev Psychol* 40:882–889
321. Thompson R (2001) Sensitive periods in attachment? In: *Critical thinking about critical periods*. Brookes Publishing, Baltimore, pp 83–106
322. Tschacher W, Haken H (2007) Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas Psychol* 25:1–15
323. Uylings H (2006) Development of the human cortex and the concept of 'critical' or 'sensitive' periods. *Lang Learn* 56(1):59–90
324. Valsiner J (1987) *Culture and the development of children's action: A cultural-historical theory of developmental psychology*. Wiley, Oxford
325. van der Maas HL (1993) *Catastrophe analysis of stagewise cognitive development*. Faculty of Psychology, University of Amsterdam, Amsterdam (Academic dissertation)
326. van der Maas H, Jansen B, Raijmakers M (2004) *Developmental patterns in proportional reasoning*. Cambridge University Press, New York
327. van der Maas HL, Molenaar PC (1992) Stagewise cognitive development: An application of catastrophe theory. *Psychol Rev* 99:395–417
328. van der Veer R, Valsiner J (1991) *Understanding Vygotsky: a quest for synthesis*. Blackwell, Oxford
329. van Dijk M, van Geert P (2007) Wobbles, humps and sudden jumps: a case study of continuity, discontinuity and variability in early language development. *Infant Child Dev* 16(1):7–33
330. van Geert P (1986) The concept of development. In: van Geert P (ed) *Theory building in developmental psychology*. North Holland, Amsterdam, pp 3–50
331. van Geert P (1987) Developmental theories as cognitive systems. *Cogn Syst* 2:5–39
332. van Geert P (1987) The concept of development and the structure of developmental theories. In: Baker WJ, Hyland ME (eds) *Current issues in theoretical psychology: selected/edited proceedings of the founding conference of the International Society for Theoretical Psychology held in Plymouth UK, 30 August–2 September 1985*. North-Holland, Oxford, pp 379–392
333. van Geert P (1987) The structure of developmental theories. A generative approach. *Hum Dev* 30(3):160–177
334. van Geert P (1987) The structure of Erikson's model of the Eight Ages of Man. A generative approach. *Hum Dev* 30(5):236–254
335. van Geert P (1987) The structure of Galperin's theory of the formation of mental acts: a generative approach. *Human Dev* 30(6):355–381
336. van Geert P (1988) A graph theoretical approach to the structure of developmental models. *Human Dev* 31(2):107–135

337. van Geert P (1991) A dynamic systems model of cognitive and language growth. *Psychol Rev* 98:3–53
338. van Geert P (1993) A dynamic systems model of cognitive growth: competition and support under limited resource conditions. In: Smith LB, Thelen E (eds) *A dynamic systems approach to development: applications*. MIT Press, Cambridge, pp 265–331
339. van Geert P (1994) *Dynamic systems of development. Change between complexity and chaos*. Harvester, New York
340. van Geert P (1994) Vygotskian dynamics of development. *Human Dev* 37:346–365
341. van Geert P (1995) Dimensions of change: a semantic and mathematical analysis of learning and development. *Human Dev* 38:322–331
342. van Geert PLC (1996) The development of attachment and attachment-related competences. A dynamic model. In: Koops W, Hoelsma J, van den Boom J (eds) *New developments in attachment research*. North Holland, pp 181–199
343. van Geert P (1998) A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky and beyond. *Psychol Rev* 105(4):634–677
344. van Geert P (2000) The dynamics of general developmental mechanisms: from Piaget and Vygotsky to dynamic systems models. *Current Direct Psychol Sci* 9(2):64–68
345. van Geert P (2003) Dynamic systems approaches and modeling of developmental processes. In: Valsiner J, Conolly KJ (eds) *Handbook of developmental Psychology*. Sage, London, pp 640–672
346. van Geert P (2007) The Dynamic Systems approach in the study of L1 and L2 acquisition: an introduction. *Mod Lang J* 91:179–199
347. van Geert P (2007) Transmission, self-organization and the emergence of language: a dynamic systems point of view. In: Schönpluf U (ed) *Cultural transmission: developmental, psychological, social, and methodological perspectives*. Cambridge University Press, Cambridge UK (in press)
348. van Geert P (2008) Nonlinear-Complex-Dynamic-Systems in developmental psychology. In: Guastello S, Koopmans M, PinCUS D (eds) *Chaos and Complexity Now and in the Future: Recent Advances in the Theory of Nonlinear Dynamical Systems Psychology*. Cambridge University Press, Cambridge
349. van Geert P (2008) The Dynamic Systems approach in the study of L1 and L2 acquisition: an introduction. *Mod Lang J* 92(2):179–199
350. van Geert P, Fischer KW (2008). Dynamic systems and the quest for individual-based models of change and development. In: Spencer JP, Thomas MSC, McClelland J (eds) *Toward a new grand theory of development? Connectionism and dynamic systems theory re-considered*. Oxford University Press, Oxford (in press)
351. van Geert PLC, Savelsbergh G, van der Maas H (1999) Transitions and non-linear dynamics in developmental psychology. In: Savelsbergh G, Maas H, van Geert PLC (eds) *Non-linear developmental processes*. Elsevier Science Publishers, New York
352. van Geert P, Steenbeek H (2005) The dynamics of scaffolding. *New Ideas Psychol* 23(3):115–128
353. van Geert P, Steenbeek H (2008) Brains and the dynamics of 'wants' and 'cans': a commentary on Immordino-Yang's "A tale of two cases". *Mind Brain Educ* 2(2):62–66
354. van Orden GC, Holden JG, Turvey MT (2005) Human Cognition and 1/f Scaling. *J Exp Psychol Gen* 134:117–123
355. Vargha-Khadem F, Carr LJ, Isaacs E, Brett E (1997) Onset of speech after left hemispherectomy in a nine-year-old boy. *Brain: J Neurol* 120:159–182
356. Vasta R, Haith MM, Miller SA (1995) *Child psychology: The modern science*, 2nd edn. Wiley, Oxford
357. Viger C (2007) The acquired language of thought hypothesis: a theory of symbol grounding. *Interact Stud* 8(1):125–142
358. Vogt P, Divina F (2007) Social symbol grounding and language evolution. *Interact Stud* 8(1):31–52
359. Vygotsky LS (1978) Cole M, John V-Steiner, Scribner S, Souberman E (eds) *Mind in society: the development of higher psychological processes*. Harvard University Press, Cambridge
360. Vygotsky LS (1986) Kozulin A (ed) *Thought and language*. MIT Press, Cambridge
361. Vygotsky LS (1987) Thinking and speech. In: Rieber RW, Carton AS (eds) *The collected works of LS Vygotsky: vol I Problems of general psychology*. Plenum, New York, pp 37–285
362. Waldrop MM (1992) *Complexity: The emerging science at the edge of order and chaos*. Penguin, London
363. Weaver W (1948) Science and complexity, *Am Sci* 36:536–544
364. Weber-Fox C, Neville H (2001) Sensitive periods differentiate processing of open- and closed-class words: an ERP study of bilinguals. *J Speech Lang Hear Res* 44(6):1338–1353
365. Weisstein EW (1999) *CRC concise encyclopedia of mathematics*. Chapman Hall/CRC, Boca Raton
366. Wertsch JV (1985) *Vygotsky and the social formation of mind*. Harvard University Press, Cambridge
367. Wertsch JV (1991) *Voices of the mind: a sociocultural approach to mediated action*. Harvard University Press, Cambridge
368. Westermann G, Mareschal D, Johnson M, Sirois S, Spratling M, Thomas M (2007) Neuroconstructivism. *Dev Sci* 10(1):75–83
369. Wildgen W (1986) Processual semantics of the verb. *J Semant* 5(4):321–344
370. Wildgen W (1982) Catastrophe theoretic semantics: an elaboration and application of René Thom's theory. Benjamins, Amsterdam
371. Williams JHG, Waiter GD, Gilchrist A, Perrett DI, Murray AD, Whiten A (2006) Neural mechanisms of imitation and 'mirror neuron' functioning in autistic spectrum disorder. *Neuropsychologia* 44:610–621
372. Wimmers RH, Beek PJ, Van Wieringen PC (1992) Phase transitions in rhythmic tracking movements: A case of unilateral coupling. *Hum Mov Sci* 11:217–226
373. Wimmers RH, Savelsbergh GJP, Beek PJ, Hopkins B (1998) Evidence for a phase transition in the early development of prehension. *Dev Psychobiol* 32:235–248
374. Wimmers RH, Savelsbergh GJP, van der Kamp J, Hartelman P (1998b) A developmental transition in prehension modeled as a cusp catastrophe. *Dev Psychobiol* 32:23–35
375. Worgan S, Damper R (2007) Grounding symbols in the physics of speech communication. *Interact Stud* 8(1):7–30
376. Yavuz H (2007) An integrated approach to the conceptual design and development of an intelligent autonomous mobile robot. *Robot Autonom Syst* 55(6):498–512
377. Yoshikawa Y, Asada M, Hosoda K, Koga J (2003) A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connect Sci* 15(4):245–258

Books and Reviews

- Jansen BRJ, van der Maas HLJ (2001a) Evidence for the Phase transition from rule I to rule II on the balance scale task. *Dev Rev* 21:450–494
- Jansen BRJ, van der Maas HLJ (2002a) The development of children's rule use on the balance scale task. *J Exp Child Psychol* 81:383–416
- van der Maas HLJ, Jansen BRJ (2003) What response times tell of children's behavior on the balance scale task. *J Exp Child Psychol* 85:141–177

Development, Evolution, and the Emergence of Novel Behavior

AMY K. GARDINER, DAVID F. BJORKLUND
Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Influence of Developmental Timing on Evolution:
Heterochrony
Epigenesis, Developmental Systems, and Plasticity
Gene \times Environment \times Development Interactions
Developmental Systems and Evolutionary Change
Epigenetic Theories of Evolution
Conclusion
Bibliography

Glossary

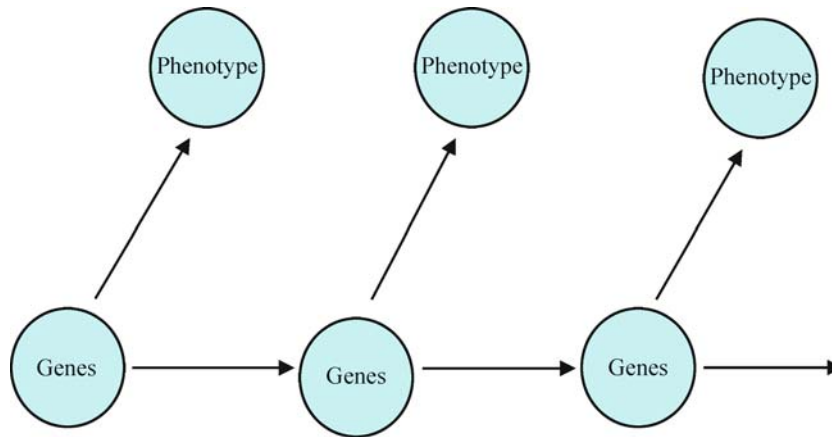
- Heterochrony** an evolutionary change in the timing of individual developmental events.
- Plasticity** the ability of individuals to respond flexibly in biologically or behaviorally adaptive ways to changes in the environment.
- Developmental Systems Theory** the idea that development unfolds via the bidirectional interaction of genes and environment at all levels of the developmental system, including genetic, cellular, structural, behavioral, and cultural.
- Epigenesis** an emergent process by which an organism's structure and function change from relatively undifferentiated states to increasingly specialized, differentiated forms throughout ontogeny.
- Epigenetic inheritance** the non-genetic transfer of information from one generation to another.

Definition of the Subject

Ontogeny, or individual development, results from the bidirectional interactions of genes and environment. It is this interaction that allows inherited traits to become expressed in the phenotypes of adult organisms. While each individual will develop along its own unique trajectory, most members of a species are very much the same because they all inherit a species-typical genotype and a species-typical environment. When this environment changes, individuals must adapt or they will fail to survive. Individuals with enough plasticity to respond to new environments by developing novel phenotypes will be more likely to survive than those without such resilience. In this way, developmental change can have substantial impact on evolution by providing the grist upon which natural selection acts. Successful developmental systems will be selected and inherited, and evolution may thus be seen as a series of ontogenies.

Introduction

There has been a resurgence of interest in evolution among psychologists and other students of behavior. Darwin's great insight that not only morphology but also behavior and "mind" have evolutionary histories has captured the attention of a myriad of scholars in disciplines from anthropology through zoology. An important issue in evolutionary psychology concerns how evolved, inherited characteristics become expressed in the phenotypes of adults. It seems obvious to some that such characteristics do not appear fully formed in the adult but emerge during *ontogeny* (the development of the individual), requiring a developmental analysis (e.g., [11,40,99]). On the flip side, developmental analyses can be used to provide insights into the processes of evolution. *Phylogeny* (the development of the species) can be viewed as a series of ontogenies: The many ancestors of extant creatures each themselves developed, and it was changes during the course of these ontogenies that produced evolutionary change. In the words of West-Eberhard (p. 89 in [124]), "The evolution of the phenotype is synonymous with the evolution of development". Natural selection has had as much or more of an impact on the early stages of development as it has had on the adult, and as a result modifications during the fetal, infant, or juvenile periods establish new contexts for further selection and thus the evolution of the species. From this perspective, the influences of both genetic and environmental mechanisms are expressed through the process of development, yielding phenotypic variation that selection might then act upon.



Development, Evolution, and the Emergence of Novel Behavior, Figure 1
 Genocentric view of inheritance (see [123] and canonical view of Modern Synthesis)

Arguing that an animal's ancestors developed and that a species' ontogeny had an influence on its phylogeny may seem self-evident; however, a closer examination reveals that this perspective was contrary to conventional wisdom in mainstream biology throughout most of the 20th century, and continues to be controversial for some today. Even before the rediscovery of Mendel's work, biologists realized that organisms could only inherit parental characteristics through the germ line (see Fig. 1), making what happened during the course of development irrelevant for the evolution of the species [123]. This viewpoint became dogma with the establishment of the Modern Synthesis [32,76,105], with any claims about the potential role of ontogeny on phylogeny being described as "Lamarckian", a label anathema to any biologist. There were a few theorists over the century who attempted to integrate development with evolution (e. g., [32,39,45,51,88]), but for the most part, development was seen by evolutionary biologists as an epiphenomenon – of critical importance to the individual but irrelevant for evolution.

In the waning years of the 20th century, with the increasing awareness that genes are always expressed in a developmental context, some biologists began to propose a significant role of ontogeny on phylogeny, albeit one quite different from that proposed by Lamarck. The new field of evolutionary developmental biology, or *Evo-Devo*, examines the role of ontogeny (especially developmental genetics) on phylogeny across a wide range of species (e. g., [24,102,124]). Evolutionary-oriented psychologists also took note, proposing that changes in behavior over the course of development could contribute to evolutionary change (e. g., [12,44,46,48,55,72,95,97,99]). These theorists recognized that the connection between genes and behavior (or genes and morphology, for that matter) was not

a direct one. Rather, new forms and function emerge via a dynamic interaction between genes and environments, with alterations in environment, especially during early development, having potentially profound effects on the adult phenotype. With respect to behavior, this perspective implies that there is a great deal of *plasticity*, defined as the ability to modify behavior as a result of environmental input. This plasticity permits an organism to change its behavior to adapt to novel environments, or perhaps to enter new environments, where they will experience new selection pressures. Despite this plasticity, most members of a species end up resembling one another to a significant degree, an outcome that is due to the fact that individuals inherit not only a species-typical genome but also a species-typical environment.

In this chapter, we examine the influence of development on evolution. We first briefly examine the possible role of developmental timing (heterochrony) on evolution. The bulk of our chapter deals with how an epigenetic account of gene-environment interactions describes the course of development and evolution. We specifically explore the developmental systems perspective of ontogeny (e. g., [46,49,50,95] and theories of *epigenetic inheritance*, or *epigenetic theories of evolution* (e. g., [57,61,62,78])). We focus toward the end of our chapter on the possible role that developmental plasticity may have had on the evolution of human social cognition, particularly through maternal effects (e. g., [7,10]).

The Influence of Developmental Timing on Evolution: Heterochrony

Prior to the Modern Synthesis, development was seen as playing a critical role in evolution. This can be seen in

the popularity of Ernst Haeckel's *biogenetic law*, which is best captured by the phrase *ontogeny recapitulates phylogeny* (see [45,51,80,106] for historical reviews). Briefly, Haeckel proposed that the development of the individual goes through, or repeats, the same stages as the evolution of the species, which is most clearly seen during embryonic development. Haeckel proposed that new stages are added to the adult stage of ancestral organisms, making evolution a progressive phenomenon. Findings in the early 20th century seriously questioned Haeckel's assumptions, however. For example, although teeth are seen in evolution prior to tongues, they develop in the reverse order in modern-day mammals [32]. Biologists soon came to recognize that the timing of some aspects of individual development could vary in a number of ways from ancestral development, a phenomenon termed *heterochrony* [32,51,82,83,107,108]. In addition to accelerations of developmental timing relative to an ancestor (implicit in Haeckel's recapitulation theory), development could also be slowed, or retarded (termed *paedomorphosis*). Modifications of developmental timing result in different phenotypes, which, if they prove adaptive, eventually result in evolutionary change [32]. "Evolution by heterochrony" implies that factors that influence the onset or offset of a developmental process can have cascading effects on an organism, resulting in substantial changes in form or function without requiring substantial changes in the genome.

One particular type of developmental retardation that caught the attention of many evolutionary biologists is termed *neoteny* (literally, "holding youth") and refers to a reduced rate of development and the retention of juvenile characteristics into adulthood [51,56,89,106]. For instance, Gould ([51], p. 375) stated "the early stages of ontogeny are a storehouse of potential adaptations, for they contain countless shapes and structures that are lost through later allometries. When development is retarded, a mechanism is provided (via retention of fetal growth rates and proportions) for bringing these features forward to later ontogenetic stages."

One oft-cited example of neoteny is the salamander species axolotl (see [51]). Axolotls begin life in water as tadpoles, their larval state. They metamorphose into adults as land-dwelling, air-breathing newts. However, when conditions in the water are favorable, the tadpoles will become sexually mature. The development of their reproductive system follows the species-typical pattern, while the development of their "gills-to-lungs" transformation is delayed, resulting in a sexually mature larval organism. Some of the offspring of these larval parents may then progress through the typical developmental se-

quence, becoming air-breathing newts, while their parents remain tadpoles (albeit sexually mature ones).

Changing the timing of certain developmental events can result in modifications in the juvenile and adult organism, some of which may be adaptive and lead, eventually, to evolutionary changes. However, changes in developmental rates *permit* evolutionary innovations rather than cause them. The origins of those changes must be found in genetics (through mutations, for example) and/or pressures in the environment, with heterochrony being a response to these pressures. But development should not be viewed as a simple unfolding of maturationally paced events that once set in motion (perhaps earlier or later than in an ancestor) have an inevitable outcome. The key to understanding evolutionary change, we argue, is to understand developmental change, and it is to this topic we now turn.

Epigenesis, Developmental Systems, and Plasticity

The phenotypic expression of genetically inherited traits involves many components, including DNA, RNA, proteins, and multiple levels of environmental surroundings. The traditional biological view of development described the genotype as a blueprint for the phenotypes of organisms. This "central dogma" asserted that genetic information flowed in one direction only, from DNA to RNA to proteins. Subsequent evidence revealed that the process of development is extensively complex and that environmental influences on gene expression play a critical role in how inherited traits emerge within organisms (see [48]). This is expressed in *developmental systems theory*, which posits that development progresses via a system of interacting levels, from the genetic through the cultural.

Epigenesis and Development

Central to developmental systems theory is the concept of *epigenesis*, which can be defined as "an emergent process by which an organism's structure and function change from relatively undifferentiated states to increasingly specialized, differentiated forms throughout ontogeny" ([86], p. 105). An epigenetic view of development describes ontogeny as a process of continuous, bidirectional interaction between components at all levels of the developmental system, including the genetic, cellular, phenotypic, behavioral, and cultural. Components of the system include genetic activity, structural maturation, activity emanating from structure, known as function, and the larger environment (see e. g., [48,49,95,124]). Genes are not given a privileged role in development, but are considered as one integral part of the developmental system that require in-

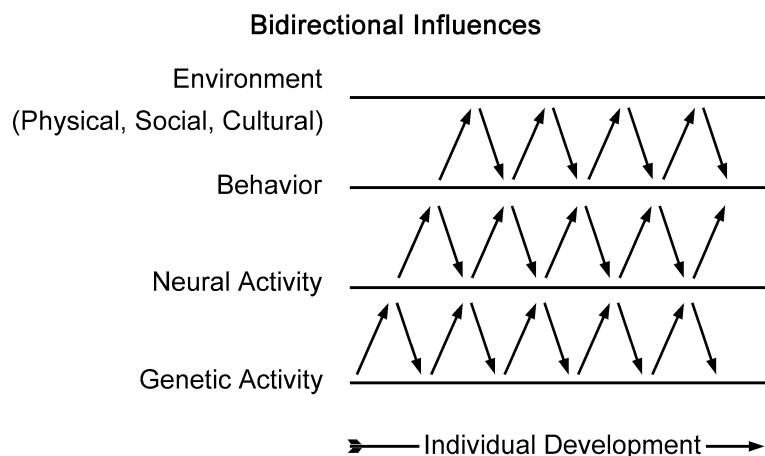
put from and interaction with other components of the system to function appropriately. From this perspective, no dichotomy is drawn between genetic and environmental influences on particular traits. Rather than distinguishing specific, individual contributions of genes and environment, or nature and nurture, the epigenetic perspective emphasizes the interaction between genes and environment as they jointly contribute to the emergence of phenotypic characteristics. Development is not construed as simply a combination of nature and nurture, but as the emergent process of nature *via* nurture.

Gottlieb [46,47] differentiated two types of epigenesis: predetermined and probabilistic. *Predetermined epigenesis* is reminiscent of the central dogma of biology, describing a unidirectional relationship between structure and function: genes \rightarrow structure \rightarrow function. In contrast, *probabilistic epigenesis* describes a bidirectional relationship between structure and function, with interaction among adjacent levels of the system: genes \leftrightarrow structural maturation \leftrightarrow function. This produces increasingly complex organization over the course of development as new structures and functions emerge from interactions between genes and environment.

These interactions occur throughout development, beginning at conception, when the first somatic cells of an organism emerge. Each somatic cell contains the entire inherited genotype. However, different parts of the genotype are expressed in different parts of the organism's body, and only a handful of genes are expressed at any single point in

development. Genetic activity requires environmental interaction and is therefore determined by the environment in which genes exist. Gene \times Environment interactions begin as the genotype of a newly conceived organism interacts with the cytoplasmic environment of the mother's ovum. These interactions continue throughout development as genes find themselves in increasingly complex biological and social environments. This bidirectional interaction between different levels of organization that occurs continuously over time, hypothesized by developmental systems theory, is illustrated in Fig. 2.

Central to the concept of probabilistic epigenesis is *developmental plasticity*, the flexibility that organisms show in response to their environments during development. While a genotype may be inherited as a replication of parental genes, individuals are unlikely to experience the precise environment of previous generations or even of members within their own generation. Therefore, the developmental of each individual follows a slightly different course than that of its parents and other conspecifics. The environment may also change within a single generation, requiring organisms to adapt or fail to survive. Demands of a shifting environment can be met by a genotype that has the potential to respond to environmental novelties by interacting with the environment in ways that facilitate successful development. Plasticity thus ensures individual survival by allowing organisms to adapt to environmental changes encountered during ontogeny and ensures species survival by allowing individuals to develop successfully in



Development, Evolution, and the Emergence of Novel Behavior, Figure 2

A simplified schematic of the developmental systems approach, showing a hierarchy of four mutually interacting components (from [46]). Activity at one level influences activity at adjacent levels. For instance, genetic activity may affect the generation or activity of neurons, which in turn influences behavior. That behavior then has some outcome on the environment. The response of environment, in turn, influences behavior, which affects brain activity, which itself alters genetic activity (turns on or off a particular gene, for example)

environments that may be different from those of their ancestors.

Not all aspects of the environment are equally likely to change. Certain environmental features, such as gravity and patterned light, will remain reliably constant across generations, whereas other environmental characteristics, such as climate, are more apt to vary. Likewise, developmental plasticity does not exist to the same degree across all traits, as some traits show higher levels of plasticity than others. Traits that evolved within reliable environments show low levels of plasticity and vary little across generations. They are considered highly *canalized* and require extremely atypical environments to develop abnormally (e. g., for humans, language). In contrast, traits that evolved within variable environments show high levels of plasticity and may vary considerably across populations or generations (e. g., literacy).

Because of the dynamic nature of the developmental system, the specific ontogeny of any individual is impossible to predict with complete accuracy. Development can be predicted to a certain degree, however, as implied by the concept of probabilistic epigenesis. While organisms may encounter atypical environments and experience environmental change that lead to atypical development, most individuals inherit a species-typical environment in addition to a species-typical genome. Organisms can expect to develop in a species-typical way so long as environmental conditions do not depart radically from those experienced by their ancestors. When an organism is faced with drastic changes, however, atypical development is expected. The fact that many organisms exhibit developmental plasticity that allows them to survive when confronted with species-atypical environments provides evidence that environmental change was a recurrent problem faced by many species during their evolution, a problem that plasticity evolved to solve.

Developmental Systems

The entire system of interacting gene and environment components constitutes a developmental system. The ontogeny of intersensory integration in bobwhite quails provides an excellent illustration of a developmental system in action. The auditory system of bobwhite quail chicks that are hatched and reared in a species-typical environment begins functioning before hatching, while the visual system begins functioning after hatching. Normally hatched chicks show an auditory preference for the maternal call of their species at both 24 and 48 hours after birth, and at 72 hours require simultaneous auditory and visual information (seeing the mother while hearing her call) to show

a preference. This typical pattern of development can be altered when the initial functioning of these sensory systems is manipulated.

Lickliter [69] removed a portion of the eggshell of bobwhite quail chicks and exposed the birds to patterned light 24–36 hours prior to hatching, thus shifting the initial stimulation and functioning of the visual system to an earlier point in development, corresponding to the time when the auditory system normally begins to function. These chicks showed no preference for the maternal call of their own species over a chicken call and no preference for a stuffed bobwhite hen over a stuffed quail hen at either 24 or 48 hours after hatching. This suggests interference of normal auditory perception because of the early exposure to visual information. However, when tested with both auditory and visual information in the form of a stuffed adult bobwhite hen with the bobwhite maternal call versus a stuffed quail hen with the chicken call, chicks showed a preference for the bobwhite hen at both 24 and 48 hours after hatching. As normally hatched chicks do not require both sources of information until 72 hours of age to show preference, this early activation of the visual system effectively accelerated the chicks' integration of the auditory and visual sensory systems. These findings indicate that the sequence of functional onset of these systems is an important factor in the early perceptual organization of this species. Genes were surely involved in this "imprinting" behavior, but the young bird's tendency to approach the maternal call was mediated by the timing of perceptual experience and not the simple outcome of a genetically programmed "instinct".

Further work with bobwhite quail hatchlings provided more details about their developmental system in regard to prenatal auditory learning. Lickliter and Hellewell [70] found that bobwhite quail chicks that were repeatedly exposed to a particular maternal call during the 24 hours before hatching showed a preference for this call over an unfamiliar variant until at least 24 hours after hatching, while unexposed chicks showed no preference. The addition of visual information appeared to interfere with this auditory learning when the maternal call was presented concurrently with patterned light. Chicks exposed to light at the same time that they heard the maternal call showed no preference for this call when tested after hatching. However, when the visual information was presented non-concurrently, chicks were able to learn the maternal call and showed a preference after hatching, suggesting that the timing of the visual information might be important.

Another factor to consider is the social environment in which the chicks were incubated. All chicks in the Lickliter and Hellewell [70] study were incubated communally, sur-

rounded by their broodmates during the pre-hatching period, and thus received sensory stimulation in the form of vocalizations uttered by their broodmates. To investigate if the social environment had an effect on bobwhite prenatal auditory learning, Lickliter and Lewkowicz [71] incubated some chicks in isolation and others communally. All chicks received auditory information of the particular maternal call. None of the chicks incubated communally received visual information, but some of the chicks incubated in isolation were exposed to light at the same time that they heard the call. Chicks incubated communally and chicks that were isolated but exposed to visual information preferred this call to an unfamiliar variant. Chicks incubated in isolation without visual information did not show a preference. Based on findings from these studies, Lickliter and Lewkowicz [71] concluded that bobwhite quail chicks need an optimal level of background stimulation to learn auditory information prenatally, but too much (broodmates, light, and sound) or too little (just sound) can prevent learning. The source of the sensory stimulation, broodmates or light, is less important than the stimulation itself. These studies provide a clear example of a developmental system and demonstrate how small changes to the system can produce noticeable perturbations in development.

Plasticity

Plasticity is essential for development. Especially for traits that develop in unpredictable environments, a degree of plasticity is necessary for the interactions between genes and environments to produce viable phenotypes.

Plasticity and Morphological Change Many animals display truly remarkable degrees of plasticity, at least in comparison to mammals. For instance, beginning in high-school biology class, we are told that sex is determined genetically, associated with genes on the X and Y chromosomes. This is true for all mammals and birds, and sex is similarly determined for most other animal species. However, whether one is male or female for some lizards, turtles, and crocodilians is not determined by specific genes but by the temperature at which eggs are incubated. In most turtles in which sex is determined by temperature, lower temperatures produce males and higher temperatures females, with this pattern being reversed for crocodilians [17,30]. Temperature also determines the camouflaging pattern on the wings of some butterflies. For instance, in the warm, wet season, the wings of the butterfly *Bicyclus anynana* develop many eyespots, which helps it blend into the foliage of the summer environment. In

cooler, dry weather, *Bicyclus anynana* develop fewer or no spots, which helps them blend into the mostly brown leaf litter [16]. Temperatures seem to affect a gene (Distal-less) associated with the production proteins, which in turn affects the generation of wing spots (described in [23]).

Early diet can also play a pivotal role in the development of some organisms. The moth *Nemoria arizonaria* lays its eggs on oak trees during different times of the year, when floral environments and available diets are different. Based on the diet ingested during the first few days after hatching, the caterpillars that emerge will develop one of two morphologies. The spring brood feeds on oak catkins and comes to resemble the catkins, while the summer brood feeds on oak leaves and resembles oak twigs. These morphologies allow the caterpillars to blend into the surrounding environment, protecting them from predators [52].

In each of these examples, important aspects of an animal's morphology are determined by environmental conditions, not by commonly-shared genes. A male and female alligator may be genetic twins, something that is impossible in mammals. Genes evolved to be sensitive to differences in some environmental conditions during particular times in development, which resulted in adaptive outcomes during the species' natural selective history, what Boyce and Ellis [14] refer to as *conditional adaptations*. Although the action of genes still determines an animal's morphology, variation in the environment, not in the genome, is responsible for important aspects (e. g., sex) of an animal's physique.

Plasticity and Behavioral Development Behavioral plasticity early in development has important consequences for animals. As we saw in the examples of bobwhite quails (e. g., [69]), important aspects of an animal's behavior (and thus presumably their brains) can be altered as a result of species-atypical perceptual experiences. The plasticity of brain and behavior can also serve a protective function, in that the effects of deleterious early environments can be reversed with a radical change of circumstances. For example, six months of social isolation in rhesus monkeys, which typically results in permanent social and sexual dysfunction, can be reversed if the animals are provided with appropriate "therapy" (in this case, six months of daily interaction beginning at 6-months of age with a younger conspecific [115]). Similarly, the placement of children into foster or adoptive homes who had spent their early years in stultifying institutions, results in significant improvements of intellectual and social functioning (e. g., [63,91,92,110]). The relatively slow development of the human brain over the preschool years provides a de-

gree of plasticity that permits a reversal of the effects of an early harmful environment, with the degree to which such effects can be reversed declining with age (e. g., [91,92]).

Plasticity also appears to play an important role in the ontogeny of children's social competencies. Huether [60] proposed a framework for understanding the interaction between early environment and neural plasticity and how it affects social development. According to this framework, individuals can experience at least two types of responses to environmental stress: controllable and uncontrollable. The type of response experienced depends on the degree of control an individual subjectively perceives he or she can assert on a situation through his or her own actions. A controllable stress response is in reaction to a situation one feels he or she can control, whereas an uncontrollable stress response is in reaction to a situation one feels he or she cannot control. Both responses begin as nonspecific patterns of arousal of cortical and limbic structures in the brain. Controllable responses direct arousal to specific neuronal pathways involved in behavioral response to the given stressor and increase noradrenergic output to stabilize these pathways. In contrast, uncontrollable responses escalate arousal and activate the hypothalamic-pituitary-adrenal (HPA) system and adrenal glucocorticoid secretion, which can destabilize previously established pathways and provide opportunity for neural reorganization through the acquisition of new coping strategies. Behavioral strategies that succeed in making a stressor subjectively controllable are reinforced and facilitated, whereas strategies that repeatedly fail and make a stressor subjectively uncontrollable will be eliminated or become a source of constant dysregulation. Both controllable and uncontrollable stress experiences are needed to establish individual patterns of coping that represent the sum of previous experience and to contribute to normal brain development, but it is repeated experience of controllable stress that allows children to develop appropriate, successful social strategies.

Flinn [37] notes that stress hormones such as glucocorticoids divert bodily resources from important health functions such as immune response and that chronic stress, especially during childhood, can have negative long-term effects on health. Therefore, uncontrollable stress responses, which activate glucocorticoid secretion, appear to come at a cost. Evolutionarily, if these responses were adaptive, it is expected that they would provide benefits that outweigh such costs. Flinn argues that the benefits of psychosocial stress during childhood exceed the costs because navigating relationships is of paramount importance in the socially dynamic human environment. From this perspective, elevated levels of stress hormones in re-

sponse to social situations can be advantageous if they facilitate neural reorganization that allows individuals to cope with a variable, unpredictable environment.

The relationships between social experiences, stress responses, and neural organization are extensively complex. In humans especially, the effects of early stress experiences on neural organization and social ontogeny are difficult to study because of methodological considerations. To understand the relationship between social experiences and stress response, Flinn measures the stress hormone cortisol, a glucocorticoid, through collection of children's saliva samples. In a long-term study, Flinn [37] collected saliva from the children of a small rural community on the Caribbean island of Dominica at least twice a day over a period of more than 10 years. He found that cortisol levels were generally highest in response to social challenges, particularly those involving conflict or change within children's families. Children living in families with high levels of marital conflict were more likely to show abnormal, elevated levels of cortisol than children living in more harmonious family environments. Longitudinal results support the conclusion that interactions within the family are a critical source of psychosocial stress in childhood. Children who experienced early family trauma in the form of parental divorce, death, or abuse have higher cortisol levels at age 10 and higher morbidity than children who do not experience early family trauma.

Flinn suggests that early experiences of family trauma result in permanent dysregulation of the HPA system, which leads to impaired regulation of cortisol levels in response to social challenges, as well as deleterious effects on brain regions involved in important neural, metabolic, and immune system health. However, temporary cortisol response to psychosocial challenges is important because it facilitates neural reorganization in response to the constantly shifting social landscapes of human communities. Because human relationships are of such great importance, the benefits of the stress response system outweigh the costs.

Both Huether [60] and Flinn [37] note that not all children will develop identically given the same social circumstances. Genetically-based individual differences between children in how they respond to social challenges, such as temperament, will have a significant effect on how they interact with their environments, as well as the development of their social competencies. These interactions between genes and environment are seen in studies of rhesus monkeys. For instance, in one study rhesus monkeys were classified as being either highly reactive to novel social experiences and displaying elevated cortisol levels or classified as less reactive [15]. When living in a spacious,

low-stress environment, low-reactive animals had slightly fewer injuries over the course of the year than high-reactive individuals. When the monkeys were moved to a confined, high-stress environment, the relation between stress reactivity and injuries reversed. Now, the high-reactive monkeys showed substantial increases in injuries, whereas there was no change in the rate of injuries for the low-reactive monkeys.

Other research with monkeys (e. g., see [112,114]) and children (e. g., see [5,42,111]) has shown that individual differences in temperament or reaction to stress interacts with the quality of the rearing environment to produce differential reaction to parental influence or to specific psychosocial outcomes (e. g., depression, antisocial behavior). In the past decade, researchers have identified specific genes associated with certain behavioral outcomes under some environmental contexts, providing greater insight into the nature of gene \times environment \times development interactions, and we examine briefly several examples of this research in the next section.

Gene \times Environment \times Development Interactions

Gene \times environment \times development interactions are a prominent feature of the developmental systems approach. Throughout development, an individual's inherited genotype interacts with the environment in the epigenetic process that leads to the emergence of phenotypic characteristics (see [9]). Several studies of the interaction between genes and environment in relation to human and non-human primate parenting behavior illustrate this effect.

Caspi et al. [25] examined the relationship between maltreatment during childhood and possession of a particular allele related to the production of Monoamine oxidase A (MAOA), a compound that metabolizes several neurotransmitters. Levels of MAOA are governed by different alleles of a gene located on the X chromosome and have been shown to be associated with aggression. Specifically, low levels of MAOA are associated with high levels of aggression. Caspi et al. [25] found that boys with low MAOA levels showed elevated amounts of antisocial behavior, but only if they had experienced maltreatment during childhood. In the group that had not experienced maltreatment, boys with low MAOA levels actually showed lower amounts of antisocial behavior than boys with high MAOA levels. Another study by Caspi and colleagues [26] found that different alleles of the 5-HTT serotonin transporter gene and childhood maltreatment were related to depression in young adulthood. Individuals with the LL (long-long) variant of the gene showed no relationship be-

tween maltreatment and depression, but individuals with the LS (long-short) or SS (short-short) variants who had experienced maltreatment were more likely to report depressive episodes during young adulthood than those who had not experienced maltreatment. In other research, adolescents who experienced high degrees of maternal rejection were more apt to be classified as clinically depressed than adolescents whose mothers did not reject them, but only if they had one combination of a particular gene that influences dopamine transfer [53].

These and other studies (e. g., [27,87]) suggest that early rearing environments may moderate genetic effects on behavior through interaction with particular gene variants. This conclusion is tentatively drawn, however, because these findings are correlational and because parenting behavior (and consequently early rearing environment) may naturally vary with genotype, thereby making true gene-environment interactions difficult to discern.

These difficulties can be overcome in animal studies. Suomi [114] investigated the relationship between the LL and LS alleles of the 5-HTT gene and parenting behavior in rhesus monkeys by examining the differences in aggression between peer-raised and mother-raised individuals. Peer-raised monkeys with the LS variant showed high levels of aggression, while those with the LL variant did not. There was no relation between allelic variation and aggression in mother-raised monkeys, suggesting a maternal buffering effect. Similar maternal-buffering effects for individuals with the LS allele were found for hypothalamic-pituitary-adrenal responses to social separation [3] and for neonatal neurobehavioral development [29].

Suomi [113] discussed additional findings from cross-fostering studies that investigated gene \times environment \times development interactions in regard to the temperament of rhesus monkey infants and the type of maternal parenting style received during the first months of life. Monkeys were selectively bred to fall within the normal range of temperamental reactivity or within the high reactive range. High-reactive infants tend to have fearful and anxious reactions to stressful situations, begin to use their mothers as a secure base from which to explore relatively late in development, and are initially shy in peer relationships. The monkeys were cross-fostered with mothers who were either in the normal range of maternal style or showed an unusually nurturant style. After six months with their mothers, the infants were placed in peer groups. High-reactive infants reared by normal nurturant mothers displayed deficits in exploration and fell to the bottom of peer hierarchies. In contrast, high-reactive infants reared by high-nurturant mothers appeared behaviorally precocious, leaving their mothers to explore earlier and explor-

ing more than all other groups, and rising to high-ranking positions within peer hierarchies. When females in this study became mothers themselves, they displayed a maternal style consistent with that of their foster mothers, regardless of the maternal style of their biological mothers. These findings suggest that the interactions between the infants' genetically inherited temperament and their early experiences of maternal care had long-term effects on the infants' developmental trajectories, even persisting into the subsequent generation.

Developmental Systems and Evolutionary Change

Evolution is the change in organisms (often defined in terms of changes in gene frequencies) and the differentiation of species that occur over geologic time. This process is borne out through the differential survival of individual organisms, known as natural selection. Evolution by natural selection is a relatively simple process that depends on the following four factors: 1) *superfecundity* – within a species, there are more members in each generation than can actually survive; 2) *variation* – there are significant differences in physical and behavioral traits between individuals within a species; 3) *inheritability* – traits are passed on from parents to offspring, making variation heritable; and 4) *selection* – the most adaptive traits, those that best promote survival and reproduction in the local ecology, increase in frequency within the population of a species, while maladaptive traits, which fail to support survival and reproduction, decrease. A good measure of the adaptability of a trait is the degree to which it promotes an organism's *inclusive fitness*, which is the reproductive success of an individual including its own offspring and the offspring of others with whom it shares genes [54]. Among all of the variation within a generation, those individuals that have traits that maximize their inclusive fitness will spread the most copies of their genes to future generations, thus increasing the frequencies of these traits within the population of a species.

Natural selection does not generate variation; it merely acts on the variation that exists between the phenotypes of each generation. There are several mechanisms through which variation may be created. Traditionally, the source of phenotypic variety has been identified as random mutation at the genotypic level. In this description of evolutionary change, a mutation alters the genetic makeup of an individual and this creates a novel phenotype within the population. In contrast, epigenetic theories of evolution recognize the prominent role of gene \times environment \times development interactions in phenotypic creation, and invoke phenotypic plasticity as the primary generator of

novel variation. That is, the creation of novel phenotypes, whether produced by genetic variation or induced via environmental events, provides the material upon which natural selection acts. Along these lines, West-Eberhard [124] proposed that “adaptive evolution – phenotypic improvement due to selection – . . . is a two step process: first the generation of variation by development, then the screening of that variation by selection” (p. 139). West-Eberhard proposed that “new phenotypic subunits begin and evolve as products of developmental plasticity. They originate when an environmental or genetic perturbation causes a shift in genes expression, and they are consolidated under selection for improved regulation and form” (p. 129).

From this perspective, evolution should not be viewed simply as changes in gene frequencies over time, as is the canonical way, but rather as changes in developmental systems. According to Oyama ([96], p. 29), “What is transmitted between generations is not traits, or blueprints, or symbolic representations of traits, but developmental *means* (or *resources*, or *interactants*). These means include genes, the cellular machinery necessary for their functioning, and the larger developmental context, which may include a maternal reproductive system, parental care, or other interaction with conspecifics, as well as relations with other aspects of the animate and inanimate worlds. This context, which is actually a system of partially nested contexts, changes with time, partly as a result of the developmental processes themselves.”

Individuals respond to environmental change by altering their gene expression in ways that diverge from those of the parent generation. Those changes in the developmental system that facilitated by plasticity. If an organism encounters a species-atypical environment and possesses appropriate levels of plasticity to adapt to this environment, its development will diverge in a species-atypical manner, creating a novel phenotype. If this novel phenotype provides the organism with enhanced survival and reproductive fitness relative to a typical phenotype or the phenotype of a less-plastic individual, the species-atypical developmental system will spread within the population. With each environmental change, this process repeats. Thus, developmental plasticity in response to environmental variation facilitates evolutionary change, often preceding genetic mutation. Such plastic responses lead to phenotypic variation within a population and provide the grist upon which natural selection acts [44,45]. Although the underlying genetic contribution of species-atypical behaviors and traits is still unknown, novel experiences may activate previously dormant genes, deactivate genes, alter the onset, offset, or amount of activity of genes, or result in adaptive outcomes for some but not other alleles

of a gene, all actions consistent with the Modern Synthesis.

Traits evolve in response to selection pressures within the environment. Each adaptive trait evolved within a specific *environment of evolutionary adaptedness*, which is the collection of environmental factors that provided pressure for the selection of a particular variation of a trait. Those individuals that have or are able to develop the trait under pressure are more likely to survive than those that do not or cannot.

Selection pressures may exist as passive components of the environment, such as light or gravity, which cannot be changed due to the actions of organisms. Or, selection pressures may arise from the modification of the environment by organisms through a process known as *niche construction* [68,94], defined as “the activities, choices, and metabolic processes of organisms, through which they define, choose, modify, and partly create their own niches” ([65], pp. 132–133). Environmental modification can result from basic life processes of organisms such as consumption and excretion, or through the direct construction of habitats. These habitats may represent an “extended phenotype” of the creatures that build them [31], such as the web of the spider or the beaver’s dam. Through niche construction, organisms are able to create or modify selection pressures within their environments. The niche construction efforts of organisms may constitute the *ecological inheritance* [93] of subsequent generations that live in environments at least partially created by the actions of previous generations. If environmental modification occurs in consistent patterns across generations, niche construction can have significant effects on evolutionary outcomes by creating pressure for genetic change [64,65].

Epigenetic Theories of Evolution

Epigenetic inheritance is the non-genetic transfer of information across generations. Epigenetic theories of evolution have been discussed since the late 1800s, the oldest account named for James Mark Baldwin but proposed contemporaneously by Conway Lloyd Morgan and H. F. Osborn. The *Baldwin effect* occurs when acquired behaviors are transmitted non-genetically via social learning. Baldwin [2] suggested that organisms that survive environmental stress are able to cope because they have sufficient developmental plasticity to allow behavioral adjustment in response to stressors. These individuals reproduce and their offspring learn from them how to survive in the given environment. Eventually the once-acquired behavior comes to be expressed in all members of the population, a process Baldwin termed *organic selection*.

The selective advantage of a plastic developmental system was demonstrated by Waddington in several classic experiments (described in his 1975 collection of essays). For example, in one set of experiments Waddington exposed pupal fruit flies (*Drosophila melanogaster*) to heat shock. Although many individuals did not survive this exposure, some that did survive responded by developing wings with few or no cross veins. Waddington selectively bred those flies without cross veins and exposed their offspring to heat shock. This second generation also responded by developing wings with few or no cross veins. After 14 generations of selecting for this phenotype, some of the flies developed the selected phenotype without being exposed to heat shock. Waddington termed this phenomenon *genetic assimilation*. He was also able to demonstrate genetic assimilation for an adaptive trait by feeding fruit flies a diet high in salt. The surviving flies developed larger anal papillae, which facilitate the excretion of salt, than flies fed a normal diet. Waddington bred the survivors, and after 21 generations produced flies that developed large anal papillae even when grown on a low-salt medium. In both of these cases, a response to an environmental change was facilitated by the organisms’ developmental plasticity and allowed the organisms to adapt and survive in their new environments.

Since Waddington’s experiments, epigenetic inheritance has been demonstrated in a variety of species (see [62,124] for reviews). For example, the asexually reproducing water flea *Daphnia cucullata* will grow a protective helmet if exposed to the larva of predators, and offspring and grandoffspring of helmeted members of this species will also develop the protective trait, even without exposure to a predatory environment [1]. Ho, Tucker, Keeley, and Saunder [58] exposed fruit flies to ether, causing the flies to produce two sets of wings. This novel phenotype appeared in offspring of affected females, but not of affected males, suggesting transgenerational epigenetic inheritance of changes in the chemical machinery of the female gametes. This does not imply a change to the nuclear DNA of any female, and it should be stressed that none of the epigenetic theories presented relies on Lamarckian ideas to explain evolutionary change.

There is also solid evidence of epigenetic inheritance in mammals, dating back to the 1960s [33,103]. For example, in one early study mice from either the C57BL or BALB strains that had been raised by foster parents from the BALB strain performed better on four operant conditioning performance measures than C57BL foster-raised mice, and this effect was maintained into the subsequent generation [103]. These results suggested to Ressler [103] that “a nongenetic system of inheritance based upon trans-

mission of parental influences is potentially available to all mammals” (p. 267).

More recently, a research program by Michael Meaney and his colleagues has provided impressive evidence of epigenetic inheritance in rats, with these effects being mediated by maternal behavior (see [28,84]). In rats, licking/grooming (LG) is highly associated with arched-back nursing (ABN), in which a mother arches her back and splays her legs outward during nursing. Offspring of mothers that engage in high levels of LG-ABN experience less stress than offspring of mothers that engage in low levels of LG-ABN. This effect is mediated by gene expression (see [84]), but is probably not due to genetic inheritance, as demonstrated in a cross-fostering study by Francis and colleagues [38], in which offspring of low LG-ABN mothers were raised by high LG-ABN mothers and vice versa. The stress reactions of the offspring were similar to the behavior of their foster mothers, not their biological mothers. The females of this second generation then served as foster mothers and displayed LG-ABN behavior characteristic of their foster mothers. The same effect was seen in this third generation. The experiment demonstrates that changes in gene expression can be moderated by maternal behavior and that these changes can persist across generations.

Meaney (Fig. 1, p. 1179 in [84]) presented a schema illustrating how stress in an environment could affect maternal anxiety and subsequent maternal care, which in turn influences neural development of the offspring. Neural development affects expression of genes related to chemicals involved in stress responses (e.g., corticotropin-releasing factor) and production of neurotransmitters and hormones such as dopamine and oxytocin, which are involved in emotional expression and regulation. This pattern would affect the behavior of the female animal when she becomes a mother, influencing the neural development and behavior of her offspring. This led Meaney [84] to suggest that “Individual differences in behavioral and neuroendocrine responses to stress in rats are, in part, derived from naturally occurring variations in maternal care. Such effects might serve as a possible mechanism by which selected traits are transmitted from one generation to another” (pp. 1170–1171).

Other research with rhesus macaques has investigated the possibility that abuse of offspring is transmitted from mothers to daughters. In a cross-fostering study, Maestripieri [75] placed daughters of abusive mothers with non-abusive foster mothers and daughters of non-abusive mothers with abusive foster mothers during the first two days after birth. These groups were compared to macaques raised by abusive and non-abusive biological

mothers. Both biological and foster daughters raised by abusive mothers were abused, while daughters of non-abusive mothers were not abused. Subsequent parenting styles of these macaques showed that 9 of the 16 females raised by abusive mothers became abusive toward their own offspring, whereas none of the females raised by non-abusive mothers became abusive. This suggests that transfer of abusive parenting from mother to daughter is mediated by early experience rather than genetic inheritance, but that individual differences in risk or protective factors can also impact whether a female will eventually display an abusive parenting style. Other studies have shown that abusive mothers invest more time in their infants than non-abusive mothers and are consistent in the level of abuse shown toward individual offspring, suggesting that this behavior represents inappropriate means of controlling infants and is pathological [74,76,77]. Early experience of abuse may lead to such pathological behavior because of long-term effects of early abusive experiences on stress responses to provoking situations (see discussion in [76]).

These studies by Maestripieri and colleagues, as well as other studies discussed above, illustrate the significant impact that mothers have on the development of their offspring. In mammals, maternal effects on the phenotypes of offspring begin prenatally and continue throughout the juvenile period, portions of an individual’s life in which the developmental system is most flexible. A mother’s contribution to the development of her offspring can significantly affect her ontogeny and thereby affect the phenotypes that serve as the raw material to be filtered by natural selection. Thus, maternal behavior shapes not only ontogeny, but phylogeny as well [7].

Epigenetic Inheritance, the Enculturation Hypothesis, and the Evolution of Human Social Cognition

Given the emphasis placed on the role of sociality as the impetus for the evolution of human intelligence, some theorists have suggested that one avenue for such changes may have been via epigenetic inheritance, specially via maternal effects [7,10]. As work with rats (e.g., Francis et al., [38]) and monkeys (e.g., [114]) has shown that socioemotional responsivity can be transmitted across generations via maternal effects, might social cognitive abilities similarly be transmitted? Although experimental research of the nature done by Meaney, Suomi, and their colleagues with rats and monkeys cannot be performed with humans to address this question, some experimentation can be done with humans’ closest genetic relatives, chimpanzees (*Pan troglodytes*). According to the *enculturation hypothesis* [22], when chimpanzees are reared much as hu-

man children, specially spoken to and treated as intentional agents – as individuals who are motivated by their own wishes and desires and who understand that the behavior of others is similarly motivated – they develop some social-cognitive abilities in a direction characteristic of human cognitive abilities, particularly the ability to understand others as intentional agents and to imitate others.

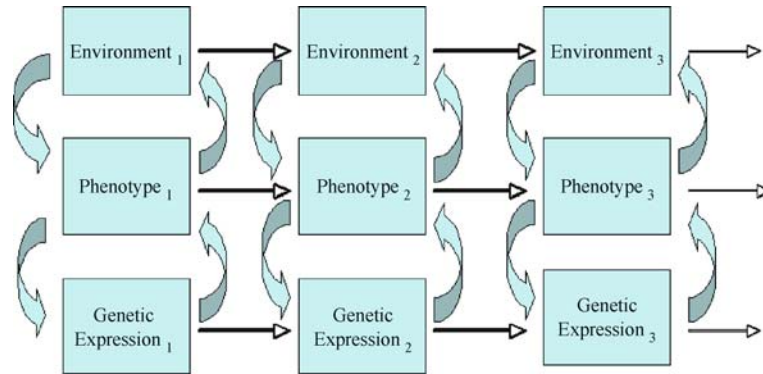
Imitation is an important human ability that facilitates cultural transmission by allowing individuals to understand and assume the goals of others and thereby faithfully replicate the behavior of previous generations [118]. Through imitation, children are able to inherit accumulated cultural knowledge, and a ratchet effect ensues as each generation contributes its own innovations to the culture and passes them on [117]. Much inherited cultural knowledge has significant survival value, such as how to use objects as tools, and shared cultural norms and values create important group bonds. Therefore, intentional understanding and imitation may be considered cognitive adaptations that facilitate the larger human adaptation of culture.

Enculturation effects are best seen in social learning, particularly imitation. Although chimpanzees and the other great apes display impressive social-learning abilities both in the wild (e.g., [125,126]) and in captivity (e.g., [125]), most do not seem to involve *true imitation*, which requires observers to understand the intentions of others, to adopt the another's goals as their own, and to reproduce important aspects of the observed behavior when attempting to achieve the goals [118,120]. Rather, most sophisticated chimpanzee social learning seems to be achieved via *emulation*, where observers attend to the effects on the environment produced by another's actions and reproduce those effects through different means, without understanding the effects as a goal of the observed individual [116,119]. For instance, Horner and Whiten ([59], see also [20,90]) had a human adult demonstrate how to retrieve a treat from within a puzzle box, first performing an irrelevant action that did not contribute to retrieval and then performing a relevant action that facilitated retrieval. Chimpanzees performed the most efficient solution by ignoring the irrelevant action and only replicating the relevant action, while 3- and 4-year-old children performed both the irrelevant and relevant actions. It is unclear why children chose a non-optimal solution, but it could be that a tendency to copy actions overrides emulative capabilities [73,127], perhaps because of the paramount importance of cultural learning.

In some ways, enculturated chimpanzees are more similar to human preschool children than to their nonenculturated conspecifics. For example, in investigating the

imitation of novel behaviors by infants, Gergely, Bekkering, and Király [41] questioned whether one could differentiate true imitation from emulation. In their terms, were infants imitating rationally? To assess this, a model demonstrated a behavior to 14-month-old infants – using her head to press a button to turn on a light. This is a novel action (one would usually use one's hand) and, if babies copied it, would indicate that they were reproducing the exact behaviors the model had witnessed to achieve a goal. Gergely and his colleagues hypothesized that infants may have believed that the model did not use her hands when she could have because using her head may have provided some advantage in turning on the light. In their study, infants watched as a model turned on the light with her head either (a) with her hands free, or (b) with her hands wrapped in a blanket, that is, her hands were occupied and thus were not available to turn on the light. Infants in the “hands free” condition used their head to turn on the light. But the pattern was reversed in the “hands occupied” condition. Now most of the infants (79%) used their hands to turn on the light. That is, when there was a reason why the model did not use her hands (they were wrapped in a blanket), the babies focused on the goal (turn on the light), not the means (use your head), and used their hands as the most efficient way to turn on the light. When no such reason was available, they copied the model's behavior exactly, reflective of imitative learning. Nearly identical results were found in a similarly designed study with enculturated chimpanzees, suggesting that the imitation these animals display is indeed “rational”, reflecting an understanding of the intentions and goals of the model [18].

Other studies have demonstrated that enculturated apes are more apt to display *deferred imitation* (imitating a model after a significant delay) than nonenculturated apes, often comparable to that of human preschool children. Tomasello, Savage-Rumbaugh, and Kruger [121] demonstrated a variety of actions on objects, such as prying open a paint can, and tested the immediate and deferred imitation abilities of both enculturated and mother-reared chimpanzees. The enculturated group replicated actions at significantly higher rates than the mother-reared group and even outperformed 1.5- and 2.5-year-old children on the deferred trials. Subsequent work by Bjorklund, Bering, and their colleagues also found that enculturated chimpanzees displayed deferred imitation ([6,8,13], see [12] for review). In an initial study, apes were allowed to interact with the target objects for a period of several minutes and then observed a human perform actions on the objects, such as placing a plastic nail in a form board and striking it with a plastic hammer. After a delay of 10 minutes, the chimpanzees and orangutans were given the



Development, Evolution, and the Emergence of Novel Behavior, Figure 3
An epigenetic view of inheritance

objects again [6]. These enculturated apes showed greater-than-chance levels of deferred imitation.

The emergence of such human-like behavior in enculturated non-human primates in the domains of social learning indicates that these apes have enough developmental plasticity to respond to human “mothering” in a manner approximating some aspects of human social cognition. This also suggests that our common ancestor had a similar level of plasticity that allowed for phenotypes to develop in response to environments increasing in social complexity and to lay the foundation for the modern human mind. These abilities would have emerged through of an epigenetic process in which the inherited cognitive and behavioral systems of the young animal are sensitive to the ecology in which it is raised [7,12,67]. One important component of that ecology is the infant’s mother, who herself would need the social-cognitive plasticity to modify her behavior in response to environmental pressures. In fact, should a group of mothers experience the same novel environment that induced greater attention to their infants’ intentions, their offspring should be similarly affected, resulting in a cohort of individuals with superior social-learning abilities, that in turn would treat their own offspring as intentional beings, promoting these skills in the population and, as a result, providing new selection pressures.

Conclusion

Although evolution and development both examine change over time, throughout most of the 20th century what happened during the lifetime on an individual was believed to have no consequence on evolution (other than reproducing the next generation). The new field of Evo-Devo and the recognition that genes are always expressed in a developmental context has changed this perspective.

Development matters to evolution, just as an animal’s phylogenetic history matters to its ontogeny. Epigenetic theories of development and evolution realize that activities of the organism and events in the wider environment feed-back upon the genome, so that both development and evolution are best conceived as dynamical systems with bidirectional relations between different levels of organization within and outside of the organism. This suggests that the genocentric view of evolution presented in Fig. 1 requires modification to something more resembling the situation depicted in Fig. 3: Activity of the genome, the form and function of the individual (the phenotype in Fig. 3), and the environment all interact, with the nature of the interaction changing over time, resulting in gene \times environment \times development effects that determine both the course on an individual life and of a species.

Bibliography

1. Agrawal AA, LaForsch C, Tollrian R (1999) Transgenerational induction of defences in animals and plants. *Nature* 401:60–63
2. Baldwin JM (1902) *Development and evolution*. McMillan, New York
3. Barr CS, Newman TK, Shannon C, Parker C, Dvoskin RL, Becker ML et al (2004) Rearing condition and rh5-HTTLPR interact to influence limbic-hypothalamic-pituitary-adrenal axis response to stress in infant macaques. *Biol Psychiatry* 55:733–738
4. Belsky J, Steinberg L, Draper P (1991) Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Dev* 62:647–670
5. Belsky J, Bakermans-Kranenburg MJ, van IJzendoorn MH (2007) For better and worse: Differential susceptibility to environmental influences. *Curr Dir Psychol Sci* 16:300–304
6. Bering JM, Bjorklund DF, Ragan P (2000) Deferred imitation of object-related actions in human-reared juvenile chimpanzees and orangutans. *Dev Psychobiol* 36:218–232
7. Bjorklund DF (2006) Mother knows best: Epigenetic inheritance, maternal effects, and the evolution of human intelligence. *Dev Rev* 26:213–242

8. Bjorklund DF, Bering JM, Ragan P (2000) A two-year longitudinal study of deferred imitation of object manipulation in an enculturated juvenile chimpanzee (*Pan troglodytes*) and orangutan (*Pongo pygmaeus*). *Dev Psychobiol* 37:229–237
9. Bjorklund DF, Ellis BJ, Rosenberg JS (2007) Evolved probabilistic cognitive mechanisms: An evolutionary approach to gene X environment X development interactions. *Adv Child Dev Beh* 35:1–36
10. Bjorklund DF, Grotuss J, Csinady A (2008) Maternal effects, social cognitive development, and the evolution of human intelligence. In: Maestriperi D, Mateo J (eds) *Maternal effects in mammals*. Chicago University Press, Chicago
11. Bjorklund DF, Pellegrini AD (2002) *The origins of human nature: Evolutionary developmental psychology*. American Psychological Association, Washington
12. Bjorklund DF, Rosenberg JS (2005) The role of developmental plasticity in the evolution of human cognition. In: Ellis BJ, Bjorklund DF (eds) *Origins of the social mind: Evolutionary psychology and child development* (pp 45–75). Guilford, New York
13. Bjorklund DF, Yunger JL, Bering JM, Ragan P (2002) The generalization of deferred imitation in enculturated chimpanzees (*Pan troglodytes*). *Anim Cogn* 5:49–58
14. Boyce WT, Ellis BJ (2005) Biological sensitivity to context: I An evolutionary-developmental theory of the origins and functions of stress reactivity. *Dev Psychopathol* 17:271–301
15. Boyce WT, O'Neill-Wagner P, Price CS, Haines M, Suomi SJ (1998) Crowding stress and violent injuries among behaviorally inhibited rhesus macaques. *Health Psychol* 17:285–289
16. Brakefield PM, Gates J, Keys D, Kesbeke F, Wijngaarden PJ, Montelro A, French V, Carroll SB (1996) Development, plasticity and evolution of butterfly eyespot patterns. *Nature* 384:236–242
17. Bull JJ (1980) Sex determination in reptiles. *Q Rev Biol* 55:3–21
18. Buttelman D, Carpenter M, Call J, Tomasello M (2007) Enculturated chimpanzees imitate rationally. *Dev Sci* 10:F31–F38
19. Call J, Carpenter M (2002) Three sources of information in social learning. In: Dautenham K, Nehaniv C (eds) *Imitation in Animals and Artifacts*. MIT Press, Cambridge, pp 211–228
20. Call J, Carpenter M, Tomasello M (2005) Copying results and copying actions in the process of social learning: chimpanzees (*Pan troglodytes*) and human children (*Homo sapiens*). *Anim Cogn* 8:151–163
21. Call J, Tomasello M (1994) The production and comprehension of referential pointing by orangutans. *J Comp Psychol* 108:301–317
22. Call J, Tomasello M (1996) The effects of humans on the cognitive development of apes. In: Russon AE, Bard KA, Parker ST (eds) *Reaching into thought: The minds of the great apes*. Cambridge University Press, New York, pp 371–403
23. Carroll SB (2005) *Endless forms most beautiful: The new science of Evo Devo*. Norton, New York
24. Carroll SB, Grenier J, Weatherbee S (2005) *From DNA to diversity: Molecular genetics and the evolution of animal design*, 2nd edn. Blackwell Science, Medford
25. Caspi A, McClay J, Moffitt TW, Mill J, Martin J, Craig IW et al (2002) Role of genotype in the cycle of violence in maltreated children. *Science* 297:851–854
26. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H et al (2003) Influence of life stress on depression: moderation of a polymorphism in the 5-HTT gene. *Science* 301:386–389
27. Caspi A, Williams B, Kim-Cohen J, Craig IW, Milne BJ, Poulton R, Schalkwyk LC, Taylor A, Helen Werts H, Terrie E Moffitt TE (2007) Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc Natl Acad Sci* 47:18860–18865
28. Champagne FA, Curley JP (2008) The trans-generational influence of maternal care on offspring gene expression and behavior in rodents. In: Maestriperi D, Mateo J (eds) *Maternal effects in mammals*. Chicago University Press, Chicago
29. Champoux M, Bennett A, Shannon C, Higley JD, Lesch KP, Suomi SJ (2002) Serotonin transporter gene polymorphism, differential early rearing, and behavior in rhesus monkey neonates. *Mol Psychiatr* 7:1058–1063
30. Crews D (2003) Sex determination: where environment and genetics meet. *Evol Dev* 5:50–55
31. Dawkins R (1982) *The extended phenotype: The long reach of the gene*. Oxford University Press, Oxford
32. de Beer G (1958) *Embryos and ancestors* (3rd Ed). Clarendon Press, Oxford
33. Denenberg VH, Rosenberg KM (1967) Non-genetic transmission of information. *Nature* 216:549–550
34. Dobzhansky T (1937) *Genetics and the origins of species*. Columbia University Press, New York
35. Ellis BJ (2004) Timing of pubertal maturation in girls: An integrated life history approach. *Psychol Bull* 130:920–958
36. Ellis BJ (2005) Determinants of pubertal timing: An evolutionary developmental approach. In: Ellis BJ, Bjorklund DF (eds) (2005) *Origins of the social mind: Evolutionary psychology and child development*. Guilford Press, New York, pp 164–188
37. Flinn MV (2006) Evolution and ontogeny of stress response to social challenges in the human child. *Dev Rev* 26:138–174
38. Francis DD, Diorio J, Liu D, Meaney MJ (1999) Nongenomic transmission across generations in maternal behavior and stress response in the rat. *Science* 286:1155–1158
39. Garstang W (1922) The theory of recapitulation: A critical re-statement of the biogenetic law. *J Linn Soc Lond Zool* 35:81–101
40. Geary DC, Bjorklund DF (2000) Evolutionary developmental psychology. *Child Dev* 71:57–65
41. Gergely G, Bekkering H, Kiraly I (2002) Rational imitation in preverbal infants. *Nature* 415:755
42. Gilissen R, Bakermans-Kranenburg MJ, van IJendoorn MH, van der Veer R (2008) Parent-child relationship, temperament, and physiological reactions to fear-inducing film clips: Further evidence for differential susceptibility. *J Exp Child Psychol* 99:182195
43. Goodall J (1986) *The chimpanzees of Gombe*. Belknap Press of Harvard University, Cambridge
44. Gottlieb G (1987) The developmental basis of evolutionary change. *J Comp Psychol* 101:262–271
45. Gottlieb G (1992) *Individual development and evolution: The genesis of novel behavior*. Oxford University Press, New York
46. Gottlieb G (1998) Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychol Rev* 105:792–802
47. Gottlieb G (2000) Environmental and behavioral influences on gene activity. *Curr Dir Psychol Sci* 9:93–97
48. Gottlieb G (2002) Developmental-behavioral initiation of evolutionary change. *Psychol Rev* 109:211–218

49. Gottlieb G (2007) Probabilistic epigenesis. *Dev Sci* 10:1–11
50. Gottlieb G, Wahlsten D, Lickliter R (2006) The significance of biology for human development: A developmental psychobiological systems view. In: Damon W, Lerner RM (Gen Eds), *Handbook of Child Psychology* (6th edition), Lerner RM (Vol Ed), vol 1, Theoretical models of human development. Wiley, New York, pp 210–257
51. Gould SJ (1977) *Ontogeny and phylogeny*. Harvard University Press, Cambridge
52. Greene E (1996) Effect of light quality and larval diet on morph induction in the polymorphic caterpillar *Nemoria arizonaria* (Lepidoptera: Geometridae). *Biol J Linn Soc* 58:277–285
53. Haefel GJ, Getchell M., Kuposov R a., Yrigollen CY, DeYoung CG, Klinteberg B, Orelund L, Ruchkin VV, Grigorenko EL (2008) Association between polymorphisms in the dopamine transporter gene and depression: evidence for a gene-environment interaction in a sample of juvenile detainees. *Psychol Sci* 19:62–69
54. Hamilton WD (1964) The genetical theory of social behavior. *J Theor Biol* 7:1–52
55. Harper L (2005) Epigenetic inheritance and the intergenerational transfer of experience. *Psychol Bull* 131:340–360
56. Hattori K (1998) Drivers of intelligence evolution in *Homo*: Sexual behavior, food acquisition and infant neoteny. *Mank Q* 39:127–146
57. Ho M-W (1998) Evolution. In: Greenberg G, Haraway MM (eds) *Comparative psychology: A handbook*. Garland, New York, pp 107–119
58. Ho M-W, Tucker C, Keeley D, Saunder PT (1983) Effects of successive generations of ether treatment on penetrance and expression of the bithorax phenocopy in *Drosophila melanogaster*. *J Exp Zool* 225:357–368
59. Horner V, Whiten A (2005) Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Anim Cogn* 8:164–181
60. Huether G (1998) Stress and the adaptive self-organization of neural connectivity during early childhood. *Int J Dev Neurosci* 16:297–306
61. Jablonka E (2001) The systems of inheritance. In: Oyama S, Griffiths PE, Gray RD (eds) *Cycles of contingency: Developmental systems and evolution* (pp 99–116). MIT Press, Cambridge
62. Jablonka E, Lamb M (1995) *Epigenetic inheritance and evolution: The Lamarckian dimension*. Oxford University Press, New York
63. Juffer F, IJzendoorn M (2005) Behavior problems and mental health referrals of international adoptees: A meta-analysis. *J Am Med Assoc*, 293:2501–2515
64. Laland KN, Odling-Smee FJ, Feldman MW (1999) Evolutionary consequences of niche construction and their implications for ecology. *Proc Natl Acad Sci* 96:10242–10247
65. Laland KN, Odling-Smee FJ, Feldman MW (2000) Niche construction, biological evolution, and cultural change. *Behav Brain Sci* 23:131–175
66. Laland KN, Odling-Smee FJ, Feldman MW (2001) Cultural niche construction and human evolution. *J Evol Biol* 14:22–33
67. Leavens DA, Russell JL, Hopkins WD (2005) Intentionality as measured in the persistence and elaboration of communication by chimpanzees (*Pan troglodytes*). *Child Dev* 76:291–306
68. Lewontin RC (1983) Gene, organism and environment. In: Bendall DS (ed) *Evolution from molecules to men*. Cambridge University Press, Cambridge, pp 273–286
69. Lickliter R (1990) Premature visual stimulation accelerates intersensory functioning in bobwhite quail neonates. *Dev Psychobiol* 23:15–27
70. Lickliter R, Hellewell TB (1992) Contextual determinates of auditory learning in bobwhite quail embryos and hatchlings. *Dev Psychobiol* 25:17–31
71. Lickliter R, Lewkowicz DJ (1995) Intersensory experience and early perceptual development: Attenuated prenatal sensory stimulation affects postnatal auditory and visual responsiveness in bobwhite quail chicks (*Colinus virginianus*). *Dev Psychol* 31:609–618
72. Lickliter R, Schneider SM The role of development in evolutionary change: A view from comparative psychology. *Int J Comp Psychol*, in press
73. Lyons De, Young AG, Keil FC (2007) The hidden structure of overlimitation. *Proc Natl Acad Sci* 104:19751–19756
74. Maestripieri D (1998) Parenting styles of abusive mothers in group-living rhesus macaque monkeys. *Anim Behav* 55:1–11
75. Maestripieri D (2005) Early experience affects the intergenerational transmission of infant abuse in rhesus monkeys. *Proc Natl Acad Sci* 102:9726–9729
76. Maestripieri D, Carroll KA (1998) Child abuse and neglect: usefulness of the animal data. *Psychol Bull* 3:211–223
77. Maestripieri D, Tomaszewski M, Carroll KA (1999) Consistency and change in behavior of rhesus macaque abusive mothers with successive infants. *Dev Psychobiol* 34:29–35
78. Mameli M (2004) Nongenetic selection and nongenetic inheritance. *Br J Philos Sci* 55:35–71
79. Mayr E (1942) *Systematics and the origins of species from the viewpoint of a zoologist*. Columbia University Press, New York
80. Mayr E (1982) *The growth of biological thought: Diversity, evolution, and inheritance*. Belknap Press, Cambridge
81. McGrew WC, Tutin CEG (1978) Evidence for a social custom in wild chimpanzees? *Man* 13:243–251
82. McKinney ML (1998) Cognitive evolution by extending brain development: On recapitulation, progress, and other heresies. In: Langer J, Killen M (eds) *Piaget, evolution, and development*. Erlbaum, Mahwah, pp 9–31
83. McKinney ML (2000) Evolving behavioral complexity by extending development. In: Parker ST, Langer J, McKinney ML (eds) *Biology, brains, and behavior: The evolution of human development*. School of American Research Press, Santa Fe, pp 25–40
84. Meany MJ (2001) Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. *Ann Rev Neurosci* 24:1161–1192
85. Miles L (1990) The cognitive foundations for reference in signing orangutan. In: Parker ST, Gibson KR (eds). *“Language” and intelligence in monkeys and apes*. Cambridge University Press, Cambridge, pp 511–539
86. Miller DB (1998) Epigenesis. In: Breenberg G, Haraway MM (eds) *Comparative psychology: A handbook*. Garland, New York, pp 105–106
87. Moffitt TW, Caspi A, Rutter M (2006) Measured gene-environment interactions in *Psychology: Concepts, research strategies, and implications for research, intervention, and public understanding of genetics*. *Perspect Psychol Sci* 1:5–27

88. Montagu MFA (1962) Time, morphology, and neoteny in the evolution of man. In: Montagu MFA (ed) *Culture and the evolution of man*. Oxford University Press, New York, pp 324–342
89. Montagu A (1989) *Growing young*, 2nd edn. Bergin and Garvey, Grandy
90. Nagell K, Olguin RS, Tomasello M (1993) Processes of social learning in the tool use of chimpanzees (*Pan troglodytes*) and human children (*Homo sapiens*). *J Comp Psychol* 107:174–186
91. Nelson CA, Zeanah CH, Fox NA, Marshall PJ, Smyke AT, Guthrie D (2007) Cognitive recovery in socially deprived young children: The Bucharest Early Intervention Project. *Science* 319:1937–1940
92. O'Connor TG, Rutter M, Beckett C, Keaveny L, Kreppner JM, and the English and Romanian Adoptees Study Team (1999) The effects of global severe privation on cognitive competence: Extension and longitudinal follow-up. *Child Dev* 71:376–390
93. Odling-Smee FJ (1988) Niche constructing phenotypes. In: Plotkin HC (ed) *The role of Behaviour in Evolution*. MIT Press, Cambridge, pp 73–132
94. Odling-Smee FJ, Laland KN, Feldman MW (2003) *Niche construction: The neglected process in evolution*. Princeton University Press, Princeton
95. Oyama S (2000a) *The ontogeny of information: Developmental systems and evolution*, 2nd edn. Duke University Press, Durham
96. Oyama S (2000b) *Evolution's eye: A systems view of biology-culture divide*. Duke University Press, Durham
97. Oyama S, Griffiths PE, Gray RD (eds) (2001) *Cycles of contingency: Developmental systems and evolution*. MIT Press, Cambridge
98. Patterson F (1978) Linguistic capabilities of a lowland gorilla. In: Peng F (ed) *Sign language and language acquisition in man and ape*. Westview Press, Boulder, pp 161–201
99. Ploeger A, van der Maas HLJ, Rajimakers MEJ (2008) Is evolutionary psychology a metatheory for psychology? A discussion of four major issues in psychology from an evolutionary developmental perspective. *Psychol Inq* 19:1–18
100. Povinelli DJ, Nelson K, Boysen S (1992) Comprehension of role reversal in chimpanzees: Evidence of empathy? *Anim Behav* 43:633–640
101. Povinelli DJ, Reaux DJ, Bierschwale DT, Allain AD, Simon BB (1997) Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. *Cogn Dev* 12:423–461
102. Raff RA (1996) *The shape of life: Genes, development, and the evolution of animal form*. Chicago University press, Chicago
103. Ressler RH (1966) Inherited environmental influences on the operant behaviour of mice. *J Comp Phys Psychol* 61:264–267
104. Savage-Rumbaugh ES (1986) *Ape language: From conditioned response to symbol*. Columbia University Press, New York
105. Savage-Rumbaugh ES, McDonald K, Sevcik RA, Hopkins WD, Rubert E (1986) Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*). *J Exp Psychol Gen* 115:211–235
106. Schwartz JH (1999) *Sudden origins: Fossils, genes, and the emergence of species*. Wiley, New York
107. Shea BT (1989) Heterochrony in human evolution: The case for neoteny revisited(ed). *Yearb Phys Anthr* 32:69–101
108. Shea BT (2000) Current issues in the investigation of evolution by heterochrony, with emphasis on the debate over human neoteny. In: Parker ST, Langer J, McKinney ML (eds) *Biology, brains, and behavior: The evolution of human development*. School of American Research Press, Santa Fe, pp 181–213
109. Simpson GG (1944) *Tempo and mode in evolution*. Columbia University Press, New York
110. Skeels HM (1966) Adult status of children with contrasting early life experiences. *Monogr Soc Res Child Dev* 31(3, Serial No. 105)
111. Stright AD, Gallagher KC, Kelly K (2008) Infant temperament moderates relations between maternal parenting in early childhood and children's adjustment in first grade. *Child Dev* 79:186–200
112. Suomi S (1995) Influence of attachment theory on ethological studies of biobehavioral development in nonhuman primates. In: Goldberg S, Muir R, Kerr J (eds) *Attachment theory: Social, developmental and clinical perspectives*. Analytic Press, Hillsdale, pp 185–202
113. Suomi SJ (1999) Attachment in rhesus monkeys. In: Cassidy J, Shaver PR (eds) *Handbook of attachment: Theory, research, and clinical applications*. Guilford Press, New York
114. Suomi SJ (2004) How gene-environment interactions shape biobehavioral development: Lessons from studies with rhesus monkeys. *Res Human Dev* 1:205–222
115. Suomi SJ, Harlow H (1972) Social rehabilitation of isolate-reared monkeys. *Dev Psychol* 6:487–496
116. Tomasello M (1998) Emulation learning and cultural learning. *Behav Brain Sci* 21:703–704
117. Tomasello M (2000) Culture and cognitive development. *Curr Dir Psychol Sci* 9:37–40
118. Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: the origins of cultural cognition. *Behav Brain Sci* 28:675–735
119. Tomasello M, Davis-Dasilva M, Camak L, Bard K (1987) Observational learning of tool-use by young chimpanzees. *Hum Evol* 2:175–183
120. Tomasello M, Kruger AC, Ratner HH (1993) Cultural learning. *Behav Brain Sci* 16:495–552
121. Tomasello M, Savage-Rumbaugh S, Kruger AC (1993) Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child Dev* 64:1688–1705
122. Waddington CH (1975) *The evolution of an evolutionist*. Cornell University Press, Ithaca
123. Weismann A (1892) *Das Keimplasma: Eine Theorie der Vererbung*. Gustav Fischer, Jena
124. West-Eberhard MJ (2003) *Developmental plasticity and evolution*. Oxford University Press, New York
125. Whiten A (2007) Pan African culture: Memes and genes in wild chimpanzees. *Proc Natl Acad Sci* 104:17559–17560
126. Whiten A, Goodall J, McGrew WC, Nishida T, Reynolds V, Sugiyama Y, Tutin CEG, Wrangham RW, Boesch C (1999) Cultures in chimpanzees. *Nature* 399:682–685
127. Whiten A, Cusance DM, Gomez JC, Teixidor P, Bard KA (1996) Imitative learning of artificial fruit processing in children (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). *J Comp Psychol* 110:3–14
128. Yunker JL, Bjorklund DF (2004) An assessment of generalization of imitation in two enculturated orangutans (*Pongo pygmaeus*). *J Comp Psychol* 118:242–246

Diagrammatic Methods in Classical Perturbation Theory

GUIDO GENTILE

Dipartimento di Matematica, Università di Roma Tre,
Roma, Italy

Article Outline

Glossary

Definition of the Subject

Introduction

Examples

Trees and Graphical Representation

Small Divisors

Multiscale Analysis

Resummation

Generalizations

Conclusions and Future Directions

Bibliography

Glossary

Dynamical system Let $W \subseteq \mathbb{R}^N$ be an open set and $f: W \times \mathbb{R} \rightarrow \mathbb{R}^N$ be a smooth function. The ordinary differential equation $\dot{x} = f(x, t)$ on W defines a continuous dynamical system. A discrete dynamical system on W is defined by a map $x \rightarrow x' = F(x)$, with F depending smoothly on x .

Hamiltonian system Let $\mathcal{A} \subseteq \mathbb{R}^d$ be an open set and $\mathcal{H}: \mathcal{A} \times \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ be a smooth function ($\mathcal{A} \times \mathbb{R}^d$ is called the phase space). Consider the system of ordinary differential equations $\dot{q}_k = \partial \mathcal{H}(q, p, t) / \partial p_k$, $\dot{p}_k = -\partial \mathcal{H}(q, p, t) / \partial q_k$, for $k = 1, \dots, d$. The equations are called Hamilton equations, and \mathcal{H} is called a Hamiltonian function. A dynamical system described by Hamilton equations is called a Hamiltonian system.

Integrable system A Hamiltonian system is called integrable if there exists a system of coordinates $(\alpha, A) \in \mathbb{T}^d \times \mathbb{R}^d$, called angle-action variables, such that in these coordinates the motion is $(\alpha, A) \rightarrow (\alpha + \omega(A)t, A)$, for some smooth function $\omega(A)$. Hence in these coordinates the Hamiltonian function \mathcal{H} depends only on the action variables, $\mathcal{H} = \mathcal{H}_0(A)$.

Invariant torus Given a continuous dynamical system we say that the motion occurs on an invariant d -torus if it takes place on a d -dimensional manifold and its position on the manifold is identified through a coordinate in \mathbb{T}^d . In an integrable Hamiltonian system all phase space is filled by invariant tori. In a quasi-integrable

system the KAM theorem states that most of the invariant tori persist under perturbation, in the sense that the relative Lebesgue measure of the fraction of phase space filled by invariant tori tends to 1 as the perturbation tends to disappear. The persisting invariant tori are slight deformations of the unperturbed invariant tori.

Quasi-integrable system A quasi-integrable system is a Hamiltonian system described by a Hamiltonian function of the form $\mathcal{H} = \mathcal{H}_0(A) + \varepsilon f(\alpha, A)$, with (α, A) angle-action variables, ε a small real parameter and f periodic in its arguments α .

Quasi-periodic motion Consider the motion $\alpha \rightarrow \alpha + \omega t$ on \mathbb{T}^2 , with $\omega = (\omega_1, \omega_2)$. If ω_1/ω_2 is rational, the motion is periodic, that is there exists $T > 0$ such that $\omega_1 T = \omega_2 T = 0 \bmod 2\pi$. If ω_1/ω_2 is irrational, the motion never returns to its initial value. On the other hand it densely fills \mathbb{T}^2 , in the sense that it comes arbitrarily close to any point of \mathbb{T}^2 . We say in that case that the motion is quasi-periodic. The definition extends to \mathbb{T}^d , $d > 2$: a linear motion $\alpha \rightarrow \alpha + \omega t$ on \mathbb{T}^d is quasi-periodic if the components of ω are rationally independent, that is if $\omega \cdot v = \omega_1 v_1 + \dots + \omega_d v_d = 0$ for $v \in \mathbb{Z}^d$ if and only if $v = 0$ ($a \cdot b$ is the standard scalar product between the two vectors a, b). More generally we say that a motion on a manifold is quasi-periodic if, in suitable coordinates, it can be described as a linear quasi-periodic motion. The vector ω is usually called the frequency or rotation vector.

Renormalization group By renormalization group one denotes the set of techniques and concepts used to study problems where there are some scale invariance properties. The basic mechanism consists in considering equations depending on some parameters and defining some transformations on the equations, including a suitable rescaling, such that after the transformation the equations can be expressed, up to irrelevant corrections, in the same form as before but with new values for the parameters.

Torus The 1-torus \mathbb{T} is defined as $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, that is the set of real numbers defined modulo 2π (this means that x is identified with y if $x - y$ is a multiple of 2π). So it is the natural domain of an angle. One defines the d -torus \mathbb{T}^d as a product of d 1-tori, that is $\mathbb{T}^d = \mathbb{T} \times \dots \times \mathbb{T}$. For instance one can imagine \mathbb{T}^2 as a square with the opposite sides glued together.

Tree A graph is a collection of points, called nodes, and of lines which connect the nodes. A walk on the graph is a sequence of lines such that any two successive lines in the sequence share a node; a walk is nontrivial if it contains at least one line. A tree is a planar graph with

no closed loops, that is, such that there is no nontrivial walk connecting any node to itself. An oriented tree is a tree with a special node such that all lines of the tree are oriented toward that node. If we add a further oriented line connecting the special node to another point, called the root, we obtain a rooted tree (see Fig. 1 in Sect. “Trees and Graphical Representation”).

Definition of the Subject

Recursive equations naturally arise whenever a dynamical system is considered in the regime of perturbation theory; for an introductory article on perturbation theory see ► [Perturbation Theory](#). A classical example is provided by Celestial Mechanics, where perturbation series, known as Lindstedt series, are widely used; see Gallavotti [21] and ► [Perturbation Theory in Celestial Mechanics](#).

A typical problem in Celestial Mechanics is to study formal solutions of given ordinary differential equations in the form of expansions in a suitable small parameter, the perturbation parameter. In the case of quasi-periodic solutions, the study of the series, in particular of its convergence, is made difficult by the presence of the small divisors – which will be defined later on. Under some non-resonance condition on the frequency vector, one can show that the series are well-defined to any order. The first proof of such a property was given by Poincaré [53], even if the convergence of the series remained an open problem up to the advent of KAM theory – an account can be found in Gallavotti [17] and in Arnold et al. [2]; see also ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#). KAM is an acronym standing for Kolmogorov [47], Arnold [1] and Moser [51], who proved in the middle of last century the persistence of most of invariant tori for quasi-integrable systems.

Kolmogorov and Arnold proofs apply to analytic Hamiltonian systems, while Moser’s approach deals also with the differentiable case; the smoothness condition on the Hamiltonian function was thereafter improved by Pöschel [54]. In the analytic case, the persisting tori turn out to be analytic in the perturbation parameter, as explicitly showed by Moser [52]. In particular, this means that the perturbation series not only are well-defined, but also converge. However, a systematic analysis with diagrammatic techniques started only recently after the pioneering, fundamental works by Eliasson [16] and Gallavotti [18], and were subsequently extended to many other problems with small divisors, including dynamical systems with infinitely many degrees of freedom, such as nonlinear partial differential equations, and non-Hamiltonian systems.

Some of these extensions will be discussed in Sect. “Generalizations”.

From a technical point of view, the diagrammatic techniques used in classical perturbation theory are strongly reminiscent of the Feynman diagrams used in quantum field theory: this was first pointed out by Gallavotti [18]. Also the multiscale analysis used to control the small divisors is typical of renormalization group techniques, which have been successfully used in problems of quantum field theory, statistical mechanics and classical mechanics; see Gallavotti [20] and Gentile & Mastropietro [38] for some reviews.

Note that there exist other renormalization group approaches to the study of dynamical systems, and of KAM-like problems in particular, different from that outlined in this article. By confining ourselves to the framework of problems of KAM-type, we can mention the paper by Bricmont et al. [11], which also stressed the similarity of the technique with quantum field theory, and the so called dynamical renormalization group method – see MacKay [50] – which recently produced rigorous proofs of persistence of quasi-periodic solutions; see for instance Koch [46] and Khanin et al. [45].

Introduction

Consider the ordinary differential equation on \mathbb{R}^d

$$Du = G(u) + \varepsilon F(u), \quad (1)$$

where D is a pseudo-differential operator and G, F are real analytic functions. Assume that (1) admits a solution $u^{(0)}(t)$ for $\varepsilon = 0$, that is $Du^{(0)} = G(u^{(0)})$. The problem we are interested in is to investigate whether there exists a solution of (1) which reduces to $u^{(0)}$ as $\varepsilon \rightarrow 0$. For simplicity assume $G = 0$ in the following.

The first attempt one can try is to look for solutions in the form of power series in ε ,

$$u(t) = \sum_{k=0}^{\infty} \varepsilon^k u^{(k)}(t), \quad (2)$$

which, inserted into (1), when equating the left and right hand sides order by order, gives the list of recursive equations $Du^{(0)} = 0$, $Du^{(1)} = F(u^{(0)})$, $Du^{(2)} = \partial_u F(u^{(0)})u^{(1)}$, and so on. In general to order $k \geq 1$ one has

$$Du^{(k)} = \sum_{s=0}^{k-1} \frac{1}{s!} \partial_u^s F(u^{(0)}) \sum_{\substack{k_1 + \dots + k_s = k-1 \\ k_i \geq 1}} u^{(k_1)} \dots u^{(k_s)}, \quad (3)$$

where $\partial_u^s F$, the s th derivative of F , is a tensor with $s + 1$ indices (s must be contracted with the vectors

$u^{(k_1)}, \dots, u^{(k_s)}$, and the term with $s = 0$ in the sum has to be interpreted as $F(u^{(0)})$ and appears only for $k = 1$.

For instance for $F(u) = u^3$ the first orders give

$$\begin{aligned} Du^{(1)} &= u^{(0)3}, \\ Du^{(2)} &= 3u^{(0)2}u^{(1)}, \\ Du^{(3)} &= 3u^{(0)2}u^{(2)} + 3u^{(0)}u^{(1)2}, \\ Du^{(4)} &= 3u^{(0)2}u^{(3)} + 6u^{(0)}u^{(1)}u^{(2)} + u^{(1)3}, \end{aligned} \quad (4)$$

as is easy to check.

If the operator D can be inverted then the recursions (3) provide an algorithm to compute the functions $u^{(k)}(t)$. In that case we say that (2) defines a *formal power series*: by this we mean that the functions $u^{(k)}(t)$ are well-defined for all $k \geq 0$. Of course, even if this can be obtained, there is still the issue of the convergence of the series that must be dealt with.

Examples

In this section we consider a few paradigmatic examples of dynamical systems which can be described by equations of the form (1).

A Class of Quasi-integrable Hamiltonian Systems

Consider the Hamiltonian system described by the Hamiltonian function

$$\mathcal{H}(\alpha, A) = \frac{1}{2}A^2 + \varepsilon f(\alpha), \quad (5)$$

where $(\alpha, A) \in \mathbb{T}^d \times \mathbb{R}^d$ are angle-action variables, with $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, f is a real analytic function, 2π -periodic in each of its arguments, and $A^2 = A \cdot A$, if (here and henceforth) \cdot denotes the standard scalar product in \mathbb{R}^d , that is $a \cdot b = a_1b_1 + \dots + a_db_d$. Assume also for simplicity that f is a trigonometric polynomial of degree N .

The corresponding Hamilton equations are (we shorten $\partial_x = \partial/\partial_x$)

$$\begin{aligned} \dot{\alpha} &= \partial_A \mathcal{H}(\alpha, A) = A, \\ \dot{A} &= -\partial_\alpha \mathcal{H}(\alpha, A) = -\varepsilon \partial_\alpha f(\alpha), \end{aligned}$$

which can be written as an equation involving only the angle variables:

$$\ddot{\alpha} = -\varepsilon \partial_\alpha f(\alpha), \quad (6)$$

which is of the form (1) with $u = \alpha$, $G = 0$, $F = -\partial_\alpha f$, and $D = d^2/dt^2$.

For $\varepsilon = 0$, $\alpha^{(0)}(t) = \alpha_0 + \omega t$ is a solution of (6) for any choice of $\alpha_0 \in \mathbb{T}^d$ and $\omega \in \mathbb{R}^d$. Take for simplicity $\alpha_0 = 0$: we shall see that this choice makes sense. We

say that for $\varepsilon = 0$ the Hamiltonian function (5) describes a system of d rotators.

We call ω the *frequency vector*, and we say that ω is irrational if its components are rationally independent, that is if $\omega \cdot v = 0$ for $v \in \mathbb{Z}^d$ if and only if $v = 0$. For irrational ω the solution $\alpha^{(0)}(t)$ describes a *quasi-periodic motion* with frequency vector ω , and it densely fills \mathbb{T}^d .

Then (3) becomes

$$\begin{aligned} \ddot{\alpha}^{(k)} &= -[\varepsilon f(\alpha)]^{(k)} \\ &:= -\sum_{s=0}^{k-1} \frac{1}{s!} \partial_\alpha^{s+1} f(\omega t) \sum_{\substack{k_1+\dots+k_s=k-1 \\ k_i \geq 1}} \alpha^{(k_1)} \dots \alpha^{(k_s)}. \end{aligned} \quad (7)$$

We look for a quasi-periodic solution of (6), that is a solution of the form $\alpha(t) = \omega t + h(\omega t)$, with $h(\omega t) = O(\varepsilon)$. We call h the *conjugation function*, as it “conjugates” (that is, maps) the perturbed solution $\alpha(t)$ to the unperturbed solution ωt . In terms of the function h (6) becomes

$$\ddot{h} = -\varepsilon \partial_\alpha f(\omega t + h), \quad (8)$$

where ∂_α denotes derivative with respect to the argument. Then (8) can be more conveniently written in Fourier space, where the operator D acts as a multiplication operator.

If we write

$$h(\omega t) = \sum_{v \in \mathbb{Z}^d} e^{i\omega \cdot vt} h_v, \quad h_v = \sum_{k=1}^{\infty} \varepsilon^k h_v^{(k)}, \quad (9)$$

and insert (9) into (8) we obtain

$$\begin{aligned} (\omega \cdot v)^2 h_v^{(k)} &= [\varepsilon \partial_\alpha f(\alpha)]_v^{(k)} \\ &:= \sum_{s=0}^{k-1} \sum_{\substack{k_1+\dots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{v_0+v_1+\dots+v_s=v \\ v_i \in \mathbb{Z}^d}} \frac{1}{s!} \\ &\quad \cdot (iv_0)^{s+1} f_{v_0} h_{v_1}^{(k_1)} \dots h_{v_s}^{(k_s)}. \end{aligned} \quad (10)$$

These equations are well-defined to all orders provided $[\varepsilon \partial_\alpha f(\alpha)]_v^{(k)} = 0$ for all v such that $\omega \cdot v = 0$. If ω is an irrational vector we need $[\varepsilon \partial_\alpha f(\alpha)]_0^{(k)} = 0$ for the equations to be well-defined. In that case the coefficients $h_0^{(k)}$ are left undetermined, and we can fix them arbitrarily to vanish (which is a convenient choice).

We shall see that under some condition on ω a quasi-periodic solution $\alpha(t)$ exists, and densely fills a d -dimensional manifold. The analysis carried out above for $\alpha_0 = 0$ can be repeated unchanged for all values of $\alpha_0 \in \mathbb{T}^d$: α_0 represents the *initial phase* of the solution, and by varying α_0 we cover all the manifold. Such a manifold can be parametrized in terms of α_0 , so it represents an invariant torus for the perturbed system.

A Simplified Model with No Small Divisors

Consider the same equation as (8) with $D = d^2/dt^2$ replaced by -1 , that is

$$h = \varepsilon \partial_\alpha f(\omega t + h). \quad (11)$$

Of course in this case we no longer have a differential equation; still, we can look again for quasi-periodic solutions $h(\omega t) = O(\varepsilon)$ with frequency vector ω . In such a case in Fourier space we have

$$\begin{aligned} h_v^{(k)} &= [\varepsilon \partial_\alpha f(\alpha)]_v^{(k)} \\ &:= \sum_{s=0}^{k-1} \sum_{\substack{k_1+\dots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{v_0+v_1+\dots+v_s=v \\ v_i \in \mathbb{Z}^n}} \frac{1}{s!} (i v_0)^{s+1} \\ &\quad \cdot f_{v_0} h_{v_1}^{(k_1)} \dots h_{v_s}^{(k_s)}. \end{aligned} \quad (12)$$

For instance if $d = 1$ and $f(\alpha) = \cos \alpha$ the equation, which is known as the *Kepler equation*, can be explicitly solved by the Lagrange inversion theorem [55], and gives

$$h_v^{(k)} = \begin{cases} \frac{-i(-1)^{k+(v-k)/2}}{2^k ((k-v)/2)! ((k+v)/2)!}, & |v| \leq k, \\ & v+k \text{ even}, \\ 0, & \text{otherwise}. \end{cases} \quad (13)$$

We shall show in Sect. “[Small Divisors](#)” that a different derivation can be provided by using the forthcoming diagrammatic techniques.

The Standard Map

Consider the finite difference equation

$$D\alpha = -\varepsilon \sin \alpha, \quad (14)$$

on \mathbb{T} , where now D is defined by

$$D\alpha(\psi) := 2\alpha(\psi) - \alpha(\psi + \omega) - \alpha(\psi - \omega). \quad (15)$$

By writing $\alpha = \psi + h(\psi)$, (14) becomes

$$Dh = -\varepsilon \sin(\psi + h), \quad (16)$$

which is the functional equation that must be solved by the *conjugation function* of the *standard map*

$$\begin{cases} x' = x + y + \varepsilon \sin x, \\ y' = y + \varepsilon \sin x. \end{cases} \quad (17)$$

In other words, by writing $x = \psi + h(\psi)$ and $y = \omega + h(\psi) - h(\psi - \omega)$, with (ψ, ω) solving (17) for $\varepsilon = 0$, that is

$(\psi', \omega') = (\psi + \omega, \omega)$, we obtain a closed-form equation for h , which is exactly (16).

In Fourier space the operator D acts as $D: e^{i v \psi} \rightarrow 4 \sin^2(\omega v/2) e^{i v \psi}$, so that, by expanding h according to (9), we can write (16) as

$$\begin{aligned} h_v^{(k)} &= \frac{1}{4 \sin^2(\omega v/2)} \sum_{s=0}^{k-1} \sum_{\substack{k_1+\dots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{v_0+v_1+\dots+v_s=v \\ v_i \in \mathbb{Z}}} \frac{1}{s!} \\ &\quad \cdot (i v_0)^{s+1} f_{v_0} h_{v_1}^{(k_1)} \dots h_{v_s}^{(k_s)}, \end{aligned} \quad (18)$$

where $v_0 = \pm 1$ and $f_{\pm 1} = 1/2$.

Note that (17) is a discrete dynamical system. However, when passing to Fourier space, (18) acquires the same form as for the continuous dynamical systems previously considered, simply with a different kernel for D . In particular if we replace D with 1 we recover the Kepler equation.

The number ω is called the *rotation number*. We say that ω is irrational if the vector $(2\pi, \omega)$ is irrational according to the previous definition.

Trees and Graphical Representation

Take ω to be irrational. We study the recursive equations

$$\begin{cases} h_v^{(k)} = g(\omega \cdot v) [\varepsilon \partial_\alpha f(\alpha)]_v^{(k)}, & v \neq 0, \\ [\varepsilon \partial_\alpha f(\alpha)]_0^{(k)} = 0, & v = 0, \end{cases} \quad (19)$$

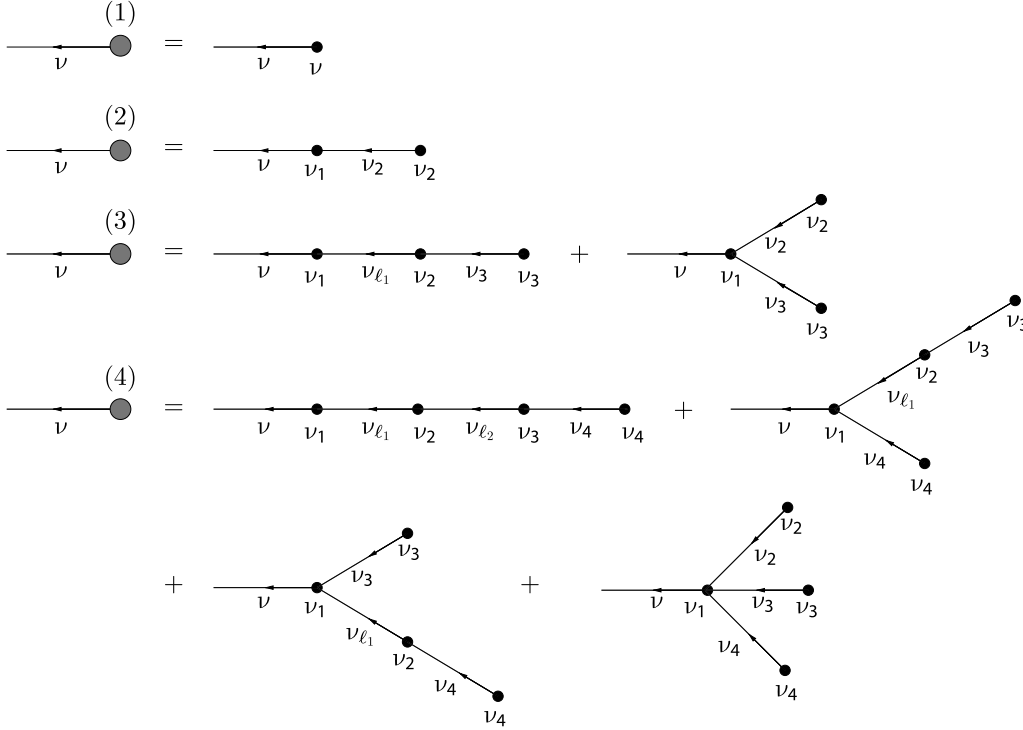
where the form of g depends on the particular model we are investigating. Hence one has either $g(\omega \cdot v) = (\omega \cdot v)^{-2}$ or $g(\omega \cdot v) = 1$ or $g(\omega \cdot v) = (2 \sin(\omega v/2))^{-2}$ according to models described in Sect. “[Examples](#)”.

For $v \neq 0$ we have equations which express the coefficients $h_v^{(k)}$, $v \in \mathbb{Z}^d$, in terms of the coefficients $h_v^{(k')}$, $v \in \mathbb{Z}^d$, with $k' < k$, provided the equations for $v = 0$ are satisfied for all $k \geq 1$. Recursive equations, such as (19), naturally lead to a graphical representation in terms of trees.

Trees

A connected graph \mathcal{G} is a collection of points (nodes) and lines connecting all of them. Denote with $N(\mathcal{G})$ and $L(\mathcal{G})$ the set of nodes and the set of lines, respectively. A path between two nodes is the minimal subset of $L(\mathcal{G})$ connecting the two nodes. A graph is planar if it can be drawn in a plane without graph lines crossing.

A *tree* is a planar graph \mathcal{G} containing no closed loops. Consider a tree \mathcal{G} with a single special node v_0 : this introduces a natural partial ordering on the set of lines and



Diagrammatic Methods in Classical Perturbation Theory, Figure 4
Trees of lower orders

the second tree, and so on. Moreover one has to sum over all possible choices of the labels ν , $\nu \in N(\theta)$, which sum up to ν .

Given any tree $\theta \in \mathcal{T}_{k,\nu}$ we associate with each node $\nu \in N(\theta)$ a *node factor* F_ν and with each line $\ell \in L(\theta)$ a *propagator* g_ℓ , by setting

$$F_\nu := \frac{1}{s_\nu!} (i\nu_\nu)^{s_\nu+1} f_{\nu_\nu}, \quad g_\ell := g(\omega \cdot \nu_\ell), \quad (22)$$

and define the *value* of the tree θ as

$$\text{Val}(\theta) := \left(\prod_{\nu \in N(\theta)} F_\nu \right) \left(\prod_{\ell \in L(\theta)} g_\ell \right). \quad (23)$$

The propagators g_ℓ are scalars, whereas each F_ν is a tensor with $s_\nu + 1$ indices, which can be associated with the $s_\nu + 1$ lines entering or exiting ν . In (23) the indices of the tensors F_ν must be contracted: this means that if a node ν is connected to a node ν' by a line ℓ then the indices of F_ν and $F_{\nu'}$ associated with ℓ are equal to each other, and eventually one has to sum over all the indices except that associated with the root line.

For instance the value of the tree in Fig. 4 contributing to $h_\nu^{(2)}$ is given by

$$\text{Val}(\theta) = (i\nu_1)^2 f_{\nu_1} (i\nu_2) f_{\nu_2} g(\omega \cdot \nu) g(\omega \cdot \nu_2),$$

with $\nu_1 + \nu_2 = \nu$, while the value of the last tree in Fig. 4 contributing to $h_\nu^{(4)}$ is given by

$$\begin{aligned} \text{Val}(\theta) = & \frac{(i\nu_1)^4}{3!} f_{\nu_1} (i\nu_2) f_{\nu_2} (i\nu_3) f_{\nu_3} (i\nu_4) f_{\nu_4} \\ & \cdot g(\omega \cdot \nu) g(\omega \cdot \nu_2) g(\omega \cdot \nu_3) g(\omega \cdot \nu_4), \end{aligned}$$

with $\nu_1 + \nu_2 + \nu_3 + \nu_4 = \nu$.

It is straightforward to prove that one can write

$$h_\nu^{(k)} = \sum_{\theta \in \mathcal{T}_{k,\nu}} \text{Val}(\theta), \quad \nu \neq 0, \quad k \geq 1. \quad (24)$$

This follows from the fact that the recursive equations (19) can be graphically represented through Fig. 3: one iterates the graphical representation of Fig. 3 until only graph elements of order $k = 1$ appear, and if θ is of order 1 (cf. Fig. 4) then $\text{Val}(\theta) = (i\nu) f_\nu g(\omega \cdot \nu)$.

Each line $\ell \in L(\theta)$ can be seen as the root line of the tree consisting of all nodes and lines preceding ℓ . The choice $h_0^{(k)} = 0$ for all $k \geq 1$ implies that no line can have zero momentum: in other words we have $\nu_\ell \neq 0$ for all $\ell \in L(\theta)$.

Therefore in order to prove that (9) with $h_\nu^{(k)}$ given by (24) solves formally, that is order by order, the Eqs. (19),

we have only to check that $[\varepsilon \partial_\alpha f(\omega t + h(\omega t))]_0^{(k)} = 0$ for all $k \geq 1$.

If we define $g_\ell = 1$ for $v_\ell = 0$, then also the second relation in (19) can be graphically represented as in Fig. 3 by setting $v = 0$ and requiring $h_0^{(k)} = 0$, which yields that the sum of the values of all trees on the right hand side must vanish. Note that this is not an equation to solve, but just an identity that has to be checked to hold at all orders.

For instance for $k = 2$ (the case $k = 1$ is trivial) the identity $[\varepsilon \partial_\alpha f(\omega t + h(\omega t))]_0^{(2)} = 0$ reads (cf. the second line in Fig. 4)

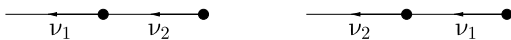
$$\sum_{v_1+v_2=0} (iv_1)^2 f_{v_1}(iv_2) f_{v_2} g(\omega \cdot v_2) = 0,$$

which is found to be satisfied because the propagators are even in their arguments.

Such a cancellation can be graphically interpreted as follows. Consider the tree with mode labels v_1 and v_2 , with $v_1 + v_2 = 0$: its value is $(iv_1)^2 f_{v_1}(iv_2) f_{v_2} g(\omega \cdot v_2)$. One can detach the root line from the node with mode label v_1 and attach it to the node with mode label v_2 , and reverse the arrow of the other line so that it points toward the new root line. In this way we obtain a new tree (cf. Fig. 5): the value of the new tree is $(iv_1) f_{v_1}(iv_2)^2 f_{v_2} g(\omega \cdot v_1)$, where $g(\omega \cdot v_1) = g(-\omega \cdot v_2) = g(\omega \cdot v_2)$, so that the values of the two trees contain a common factor $(iv_1) f_{v_1}(iv_2) f_{v_2} g(\omega \cdot v_2)$ times an extra factor which is (iv_1) for the first tree and (iv_2) for the second tree. Hence the sum of the two values gives zero.

The cancellation mechanism described above can be generalized to all orders. Given a tree θ one considers all trees which can be obtained by detaching the root line and attaching to the other nodes of the tree, and by reversing the arrows of the lines (when needed) to make them point toward the root line. Then one sums together the values of all the trees so obtained: such values contain a common factor times a factor iv_v , if v is the node which the root line exits (the only nontrivial part of the proof is to check that the combinatorial factors match each other: we refer to Gentile & Mastropietro [37] for details). Hence the sum gives zero, as the sum of all the mode labels vanishes.

For instance for $k = 3$ the cancellation operates by considering the three trees in Fig. 5: such trees can be considered to be obtained from each other by shifting the root line and consistently reversing the arrows of the lines.



Diagrammatic Methods in Classical Perturbation Theory, Figure 5

Trees to be considered together to prove that $[\varepsilon \partial f(\alpha)]_0^2 = 0$

In such a case the combinatorial factors of the node factors are different, because in the second tree the node factor associated with the node with mode label v_2 contains a factor $1/2$: on the other hand if $v_1 \neq v_3$ there are two nonequivalent trees with that shape (with the labels v_1 and v_3 exchanged between themselves), whereas if $v_1 = v_3$ there is only one such tree, but then the first and third trees are equivalent, so that only one of them must be counted. Thus, by using that $v_1 + v_2 + v_3 = 0$ – which implies $g(\omega \cdot (v_2 + v_3)) = g(-\omega \cdot v_1)$ and $g(\omega \cdot (v_1 + v_2)) = g(-\omega \cdot v_3)$ – in all cases we find that the sum of the values of the trees gives a common factor $(iv_1) f_{v_1}(iv_2)^2 f_{v_2}(iv_3) f_{v_3} g(\omega \cdot v_3) g(\omega \cdot v_1)$ times a factor 1 or $1/2$ times $i(v_1 + v_2 + v_3)$, and hence vanishes: once more the property that g is even is crucial.

Small Divisors

We want to study the convergence properties of the series

$$h(\omega t) = \sum_{v \in \mathbb{Z}^d} e^{i\omega \cdot vt} h_v, \quad h_v = \sum_{k=1}^{\infty} \varepsilon^k h_v^{(k)}, \quad (25)$$

which has been shown to be well-defined as a formal power series for the models considered in Sect. “Examples”.

Recall that the number of unlabeled trees of order k is bounded by 2^{2k} . To sum over the labels we can confine ourselves to the mode labels, as the momenta are uniquely determined by the mode labels. If f is a trigonometric polynomial of degree N , that is $f_v = 0$ for all v such that $|v| := |v_1| + \dots + |v_d| > N$, we have that $h_v^{(k)} = 0$ for all $|v| > kN$ (which can be easily proved by induction), and moreover we can bound the sum over the mode labels of any tree of order k by $(2N + 1)^{dk}$. Finally we can bound

$$\prod_{v \in N(\theta)} |v_v|^{s_{v+1}} \leq \prod_{v \in N(\theta)} N^{s_{v+1}} \leq N^{2k}, \quad (26)$$

because of (20).

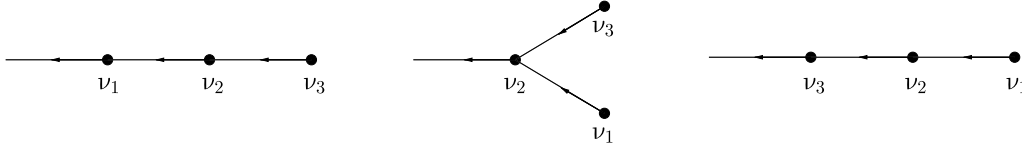
For the model (11), where $g_\ell = 1$ in (22), we can bound

$$\begin{aligned} |h_v^{(k)}| &\leq \sum_{v \in \mathbb{Z}^d} |h_v^{(k)}| \leq 2^{2k} (2N + 1)^{dk} N^{2k} \Phi^k, \\ \Phi &= \max_{|v| \leq N} |f_v|, \end{aligned} \quad (27)$$

which shows that the series (25) converges for ε small enough, more precisely for $|\varepsilon| < \varepsilon_0$, with

$$\varepsilon_0 := C_0 (4N^2 \Phi (2N + 1)^d)^{-1}, \quad (28)$$

where $C_0 = 1$. Hence the function $h(\omega t)$ in that case is analytic in ε . For $d = 1$ and $f(\alpha) = \cos \alpha$, we can easily



Diagrammatic Methods in Classical Perturbation Theory, Figure 6
Trees to be considered together to prove that $[\varepsilon \partial f(\alpha)]_0^3 = 0$

provide an exact expression for the coefficients $h_v^{(k)}$: all the computational difficulties reduce to a combinatorial check, which can be found in Gentile & van Erp [42], and the formula (13) is recovered.

However for the models where $g_\ell \neq 1$, the situation is much more involved: the propagators can be arbitrarily close to zero for v large enough. This is the so-called *small divisor problem*. The series (25) is formally well-defined, assuming only an irrationality condition on ω . But to prove the convergence of the series, we need a stronger condition. For instance one can require the *standard Diophantine condition*

$$|\omega \cdot v| > \frac{\gamma}{|v|^\tau} \quad \forall v \neq 0, \quad (29)$$

for suitable positive constants γ and τ . For fixed $\tau > d-1$, the sets of vectors which satisfy (29) for some constant $\gamma > 0$ has full Lebesgue measure in \mathbb{R}^d [17]. We can also impose a weaker condition, known as the *Bryuno condition*, which can be expressed by requiring

$$\mathcal{B}(\omega) := \sum_{k=0}^{\infty} \frac{1}{2^k} \log \frac{1}{\min_{0 < |v| \leq 2^k} |\omega \cdot v|} < \infty. \quad (30)$$

We call *Diophantine vectors* and *Bryuno vectors* the vectors satisfying (29) and (30), respectively. Note that any Diophantine vector satisfies (30).

If we assume only analyticity on f – that is, we remove the assumption that f be a trigonometric polynomial, – then we need a Diophantine condition, such as (29) or (30), also to show that the series (25) are well-defined as formal power series: the condition is needed, as one can easily check, in order to sum over the mode labels.

In the case of the standard map $|\omega \cdot v|$ must be replaced with $\min_{p \in \mathbb{Z}} |\omega v - p|$, and the function $\mathcal{B}(\omega)$ can be expressed in terms of the *best approximants* of the number ω [15].

For the models with small divisors considered in Sect. “Examples” convergence of the series (25) can be proved for ω satisfying (29) or (30). But this requires more work, and, in particular, a detailed discussion of the product of propagators in (24). In the following we shall consider the case of Diophantine vectors, by referring to the

bibliography for the Bryuno vectors (cf. Sect. “Generalizations”).

Multiscale Analysis

Consider explicitly the case $g(\omega \cdot v) = (\omega \cdot v)^{-2}$ in (19) and ω satisfying (29). We can introduce a label characterizing the size of the propagator: we say that $v \in \mathbb{Z}^d$ is on *scale*

$$\begin{cases} n \geq 1, & \text{if } 2^{-n}\gamma \leq |\omega \cdot v| < 2^{-(n-1)}\gamma, \\ n = 0, & \text{if } \gamma \leq |\omega \cdot v|, \end{cases} \quad (31)$$

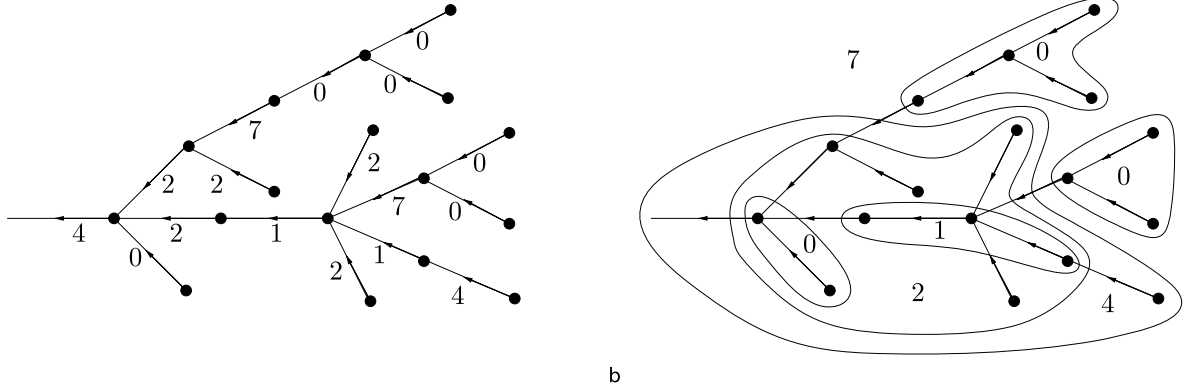
where γ is the constant appearing in (29), and we say that a line ℓ has a scale label $n_\ell = n$ if v_ℓ is on scale n . If $\mathfrak{N}_n(\theta)$ denotes the number of lines $\ell \in L(\theta)$ with scale $n_\ell = n$, then we can bound in (23)

$$\left| \prod_{\ell \in L(\theta)} g_\ell \right| \leq \gamma^{-2k} \prod_{n=0}^{\infty} 2^{2n\mathfrak{N}_n(\theta)}, \quad (32)$$

so that the problem is reduced to bounding $\mathfrak{N}_n(\theta)$.

The product of propagators gives problems when the small divisors “accumulate”. To make more precise the idea of accumulation we introduce the notion of cluster. Once all lines of a tree θ have been given their scale labels, for any $n \geq 0$ we can identify the maximal connected sets of lines with scale not larger than n . If at least one among such lines has scale equal to n we say that the set is a *cluster* on scale n . For instance consider the tree in Fig. 1, and assign the mode labels to the nodes: this uniquely fixes the momenta, and hence the scale labels, of the lines. Suppose that we have found the scales as in Fig. 7a. Then a cluster decomposition as in Fig. 7b follows. Given a cluster T call $L(T)$ the set of lines of θ contained in T , and denote by $N(T)$ the set of nodes connected by such lines. We define $k_T = |N(T)|$ the order of the cluster T .

Any cluster has either one or no exiting line, and can have an arbitrary number of entering lines. We call *self-energy clusters* the clusters which have one exiting line and only one entering line and are such that both lines have the same momentum. This means that if T is a self-energy



Diagrammatic Methods in Classical Perturbation Theory, Figure 7
Example of clusters with their scales

cluster and ℓ_1 and ℓ_2 are the lines entering and exiting T , respectively, then $v_{\ell_1} = v_{\ell_2}$, so that

$$\sum_{v \in N(T)} v_v = 0. \quad (33)$$

By definition of scale, both lines ℓ_1 and ℓ_2 have the same scale, say n , and that, by definition of cluster, one has $n_\ell < n$ for all $\ell \in L(T)$.

We define the *value* of the self-energy cluster T whose entering line has momentum v as the matrix

$$\mathcal{V}_T(\omega \cdot v) := \left(\prod_{v \in N(T)} F_v \right) \left(\prod_{\ell \in L(T)} g_\ell \right), \quad (34)$$

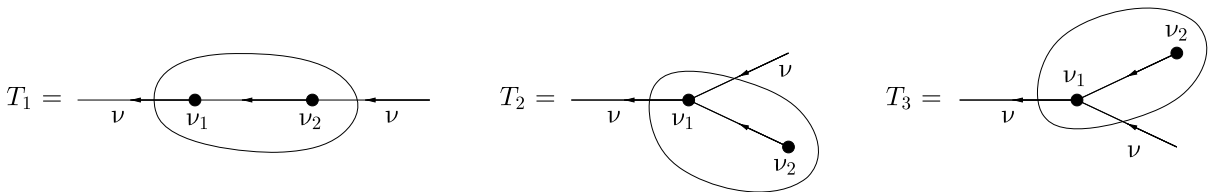
where all the indices of the node factors must be contracted except those associated with the line ℓ_1 entering T and with the line ℓ_2 exiting T .

We can extend the notion of self-energy clusters also to a single node, by saying that v is a self-energy cluster if $s_v = 1$ and the line entering v has the same momentum as the exiting line. In that case (34) has to be interpreted as $\mathcal{V}_T(\omega \cdot v) = F_v$; in particular it is independent of $\omega \cdot v$.

The simplest self-energy cluster one can think of consists of only one node v , but then (33) implies $v_v = 0$,

so that $F_v = 0$, see (22), hence the corresponding value is zero. Thus the simplest non-trivial self-energy clusters contain at least two nodes, and are represented by the clusters T_1 , T_2 and T_3 of Fig. 8. In all cases one has $v_1 + v_2 = 0$. By using the definition (34) one has $\mathcal{V}_{T_1}(x) = (iv_1)^2 f_{v_1}(iv_2)^2 f_{v_2} g(\omega \cdot v_2 + x)$ and $\mathcal{V}_{T_2}(x) = \mathcal{V}_{T_3}(x) = (iv_1)^3 f_{v_1}(iv_2) f_{v_2} g(\omega \cdot v_2)/2$, with $x = \omega \cdot v$. Hence for $x = 0$ the sum of the three values gives zero. It is not difficult to see that also the first derivatives of the values of all self-energy clusters of order 2 sum up to zero, that is the sum of the values of all possible self-energy clusters of order $k = 2$ gives zero up to order x^2 . (The only *caveat* is that, if we want to derive $\mathcal{V}_T(x)$, the sharp multiscale decomposition in (31) can be a little annoying, hence it can be more convenient to replace it with a smooth decomposition through C^∞ compact support functions; we refer to the literature for details). This property generalizes to all orders k , and the underlying cancellation mechanism is essentially the same that ensures the validity of the second relation in (19).

The reason why it is important to introduce the self-energy clusters is that if we could neglect them then the product of small divisors would be controlled. Indeed, let us denote by $\mathfrak{N}_n(\theta)$ the number of lines on scale n which do exit a self-energy cluster, and set $\mathfrak{N}_n^*(\theta) = \mathfrak{N}_n(\theta) -$



Diagrammatic Methods in Classical Perturbation Theory, Figure 8
Self-energy clusters of order $k = 2$

$\mathfrak{R}_n(\theta)$. Then an important result, known as the *Siegel–Bryuno lemma*, is that

$$\mathfrak{R}_n^*(\theta) \leq c 2^{-n/\tau} k, \quad (35)$$

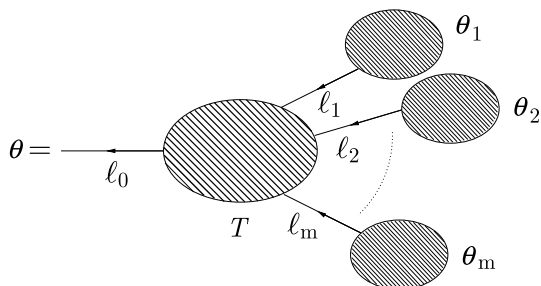
for some constant c , where k is the order of θ and τ is the Diophantine exponent in (29).

The bound (35) follows from the fact that if $\mathfrak{R}_n^*(\theta) \neq 0$ then $\mathfrak{R}_n^*(\theta) \leq E(n, k) := 2Nk2^{-(n-2)/\tau} - 1$, which can be proved by induction on k as follows. Given a tree θ let ℓ_0 be its root line, let ℓ_1, \dots, ℓ_m , $m \geq 0$, be the lines on scales $\geq n$ which are the closest to ℓ_0 , and let $\theta_1, \dots, \theta_m$ the trees with root lines ℓ_1, \dots, ℓ_m , respectively (cf. Fig. 9 – note that by construction all lines ℓ in the subgraph T have scales $n_\ell < n$, so that if $n_{\ell_0} \geq n$ then T is necessarily a cluster). If either ℓ_0 is not on scale n or it is on scale n but exits a self-energy cluster then $\mathfrak{R}_n^*(\theta) = \mathfrak{R}_n^*(\theta_1) + \dots + \mathfrak{R}_n^*(\theta_m)$ and the bound $\mathfrak{R}_n^*(\theta) \leq E(n, k)$ follows by the inductive hypothesis. If ℓ_0 does not exit a self-energy cluster and $n_{\ell_0} = n$ then $\mathfrak{R}_n^*(\theta) = 1 + \mathfrak{R}_n^*(\theta_1) + \dots + \mathfrak{R}_n^*(\theta_m)$, and the lines ℓ_1, \dots, ℓ_m enter a cluster T with $k_T = k - (k_1 + \dots + k_m)$, where k_1, \dots, k_m are the orders of $\theta_1, \dots, \theta_m$, respectively. If $m \geq 2$ the bound $\mathfrak{R}_n^*(\theta) \leq E(n, k)$ follows once more by the inductive hypothesis. If $m = 0$ then $\mathfrak{R}_n^*(\theta) = 1$; on the other hand for ℓ_0 to be on scale $n_{\ell_0} = n$ one must have $|\omega \cdot v_{\ell_0}| < 2^{-n+1}\gamma$ (see (31)), which, by the Diophantine condition (29), implies $Nk \geq |v_{\ell_0}| > 2^{(n-1)/\tau}$, hence $E(n, k) > 1$. If $m = 1$ call v_1 and v_2 the momenta of the lines ℓ_0 and ℓ_1 , respectively. By construction T cannot be a self-energy cluster, hence $v_1 \neq v_2$, so that, by the Diophantine condition (29),

$$2^{-n+2}\gamma \geq |\omega \cdot v_1| + |\omega \cdot v_2| \geq |\omega \cdot (v_1 - v_2)| > \frac{\gamma}{|v_1 - v_2|^\tau}, \quad (36)$$

because $n_{\ell_0} = n$ and $n_{\ell_1} \geq n$. Thus, one has

$$Nk_T \geq \sum_{v \in N(T)} |v_v| \geq |v_1 - v_2| > 2^{(n-2)/\tau}, \quad (37)$$



Diagrammatic Methods in Classical Perturbation Theory, Figure 9

Construction for the proof of the Siegel–Bryuno lemma

hence T must contain “many nodes”. In particular, one finds also in this case $\mathfrak{R}_n^*(\theta) = 1 + \mathfrak{R}_n^*(\theta_1) \leq 1 + E(n, k_1) \leq 1 + E(n, k) - E(n, k_T) \leq E(n, k)$, where we have used that $E(n, k_T) \geq 1$ by (37).

The argument above shows that small divisors can accumulate only by allowing self-energy clusters. That accumulation really occurs is shown by the example in Fig. 10, where a tree θ of order k containing a chain of p self-energy clusters is depicted. Assume for simplicity that $k/3$ is an integer: then if $p = k/3$ the subtree θ_1 with root line ℓ is of order $k/3$. If the line ℓ entering the rightmost self-energy cluster T_p has momentum v , also the lines exiting the p self-energy clusters have the same momentum v . Suppose that $|v| \approx Nk/3$ and $|\omega \cdot v| \approx \gamma/|v|^\tau$ (this is certainly possible for some v). Then the value of the tree θ grows like $a_1^k(k!)^{a_2}$, for some constants a_1 and a_2 : a bound of this kind prevents the convergence of the perturbation series (25).

If no self-energy clusters could occur (so that $\mathfrak{R}_n(\theta) = 0$) the Siegel–Bryuno lemma would allow us to bound in (32)

$$\prod_{n=0}^{\infty} 2^{2n\mathfrak{R}_n(\theta)} = \prod_{n=0}^{\infty} 2^{2n\mathfrak{R}_n^*(\theta)} \leq \exp\left(C_1 k \sum_{n=0}^{\infty} n 2^{-n/\tau}\right) \leq C_2^k, \quad (38)$$

for suitable constants C_1 and C_2 . In that case convergence of the series for $|\varepsilon| < \varepsilon_0$ would follow, with ε_0 defined as in (26) with $C_0 = \gamma^2/C_2$. However, there are self-energy clusters and they produce factorials, as the example in Fig. 10 shows, so that we have to deal with them.

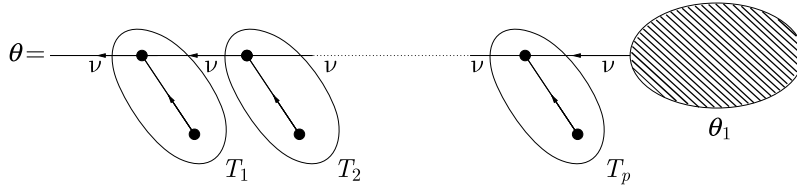
Resummation

Let us come back to the Eq. (10). If we expand $g(\omega \cdot v)[\varepsilon \partial_\alpha f(\alpha)]_v^{(k)}$ in trees according to the diagrammatic rules described in Sect. “Trees and Graphical Representation”, we can distinguish between contributions in which the root line exits a self-energy cluster T , that we can write as

$$g(\omega \cdot v) \sum_{T: k_T < k} \mathcal{V}_T(\omega \cdot v) h_v^{(k-k_T)}, \quad (39)$$

and all the other contributions, that we denote by $[\varepsilon \partial_\alpha f(\alpha)]_v^{(k)*}$. We can shift the contributions (39) to the left hand side of (10) and divide by $g(\omega \cdot v)$, so to obtain

$$D(\omega \cdot v) h_v^{(k)} - \sum_{T: k_T < k} \mathcal{V}_T(\omega \cdot v) h_v^{(k-k_T)} = [\varepsilon \partial_\alpha f(\omega t + h)]_v^{(k)*}. \quad (40)$$



Diagrammatic Methods in Classical Perturbation Theory, Figure 10

Example of accumulation of small divisors because of the self-energy clusters

where $D(\omega \cdot v) = 1/g(\omega \cdot v) = (\omega \cdot v)^2$. By summing over k and setting

$$M(\omega \cdot v; \varepsilon) = \sum_{k=1}^{\infty} \varepsilon^k \sum_{T: k_T=k} \mathcal{V}_T(\omega \cdot v), \quad (41)$$

then (40) gives

$$\begin{aligned} \mathcal{D}(\omega \cdot v) h_v &= [\varepsilon \partial_\alpha f(\omega t + h)]_v^*, \\ \mathcal{D}(\omega \cdot v) &:= D(\omega \cdot v) - M(\omega \cdot v; \varepsilon). \end{aligned} \quad (42)$$

The motivation for proceeding in this way is that, at the price of changing the operator D into \mathcal{D} , hence of changing the propagators, lines exiting self-energy clusters no longer appear. Therefore, in the tree expansion of the right hand side of the equation, we have eliminated the self-energy clusters, that is the source of the problem of accumulation of small divisors.

Unfortunately the procedure described above has a problem: $M(\omega \cdot v; \varepsilon)$ itself is a sum of self-energy clusters, which can still contain some other self-energy clusters on lower scales. So finding a good bound for $M(\omega \cdot v; \varepsilon)$ could have the same problems as for the values of the trees.

To deal with such a difficulty we modify the prescription by proceeding recursively, in the following sense. Let us start from the momenta v which are on scale $n = 0$. Since there are no self-energy clusters with exiting line on scale $n = 0$, for such v one has $M(\omega \cdot v; \varepsilon) = 0$. Next, we consider the momenta v which are on scale $n = 1$, and we can write (42), where now all self-energy clusters T whose values contribute to $M(\omega \cdot v; \varepsilon)$ cannot contain any self-energy clusters, because the lines $\ell \in T$ are on scale $n_\ell = 0$. Then, we consider the momenta v which are on scale $n = 2$: again all the self-energy clusters contributing to $M(\omega \cdot v; \varepsilon)$ do not contain any self-energy clusters, because the lines on scale $n = 0, 1$ cannot exit self-energy clusters by the construction of the previous step, and so on. The conclusion is that we have obtained a different expansion for $h(\omega t)$, that we call a *resummed series*,

$$h(\omega t) = \sum_{v \in \mathbb{Z}^d} e^{i\omega \cdot vt} h_v, \quad h_v = \sum_{k=1}^{\infty} \varepsilon^k h_v^{[k]}(\varepsilon), \quad (43)$$

where the self-energy clusters do not appear any more in the tree expansion and the propagators must be defined recursively: the propagator g_ℓ of a line ℓ on scale $n_\ell = n$ and momentum $v_\ell = v$ is the matrix

$$g_\ell := g^{[n]}(\omega \cdot v; \varepsilon) = (D(\omega \cdot v) - \mathcal{M}^{[n]}(\omega \cdot v; \varepsilon))^{-1}, \quad (44)$$

with

$$\mathcal{M}^{[n]}(\omega \cdot v; \varepsilon) := \sum_{T: n_T < n} \varepsilon^{k_T} \mathcal{V}_T(\omega \cdot v), \quad (45)$$

where the value $\mathcal{V}_T(\omega \cdot v)$ is written in accord with (34), with all the lines $\ell' \in L(T)$ on scales $n_{\ell'} < n$ and the corresponding propagators $g_{\ell'}$ expressed in terms of matrices $\mathcal{M}^{[n_{\ell'}]}(\omega \cdot v_{\ell'}; \varepsilon)$ as in (44). By construction, the new propagators depend on ε , so that the coefficients $h_v^{[k]}(\varepsilon)$ depend explicitly on ε : hence (43) is not a power series expansion.

The coefficients $h_v^{[k]}(\varepsilon)$ still admit a tree expansion

$$\begin{aligned} h_v^{[k]}(\varepsilon) &= \sum_{\theta \in \mathcal{T}_{k,v}^R} \text{Val}(\theta), \quad v \neq 0, k \geq 1, \\ \text{Val}(\theta) &:= \left(\prod_{v \in N(\theta)} F_v \right) \left(\prod_{\ell \in L(\theta)} g^{[n_\ell]}(\omega \cdot v_\ell; \varepsilon) \right), \end{aligned} \quad (46)$$

which replaces (24). In particular $\mathcal{T}_{k,v}^R$ is defined as the set of *renormalized trees* of order k and momentum v , where “renormalized” means that the trees do not contain any self-energy clusters. Now the propagators are matrices: the contractions of the indices in (46) yields that if ℓ connects v to v' the indices of F_v and $F_{v'}$ associated with ℓ must be equal to the column and row indices of g_ℓ , respectively.

Since for any tree $\theta \in \mathcal{T}_{k,v}^R$ one has $\mathfrak{N}_n(\theta) = \mathfrak{N}_n^*(\theta)$, we can bound the product of propagators according to (36) provided the propagators on scale n can still be bounded proportionally to 2^{2n} . This is certainly not obvious, because of the extra term $\mathcal{M}^{[n]}(\omega \cdot v; \varepsilon)$ appearing in (44).

It is a remarkable cancellation that $\mathcal{M}^{[n]}(x; \varepsilon)$ vanishes up to second order, that is $\mathcal{M}^{[n]}(x; \varepsilon) = O(x^2)$, so that, by

taking into account also that $\mathcal{V}_T(x) \neq 0$ requires $k_T \geq 2$, we can write, for some constant C ,

$$\mathcal{M}^{[n]}(x; \varepsilon) = \varepsilon^2 x^2 \overline{\mathcal{M}}^{[n]}(x; \varepsilon), \quad \left\| \overline{\mathcal{M}}^{[n]}(x; \varepsilon) \right\| \leq C, \quad (47)$$

where $\|\cdot\|$ denotes – say – the uniform norm. The cancellation leading to (47) can be proved as discussed in Sect. “Multiscale Analysis” – where it has been explicitly proved for $k = 2$ in the absence of resummation. Now the propagators are matrices, but one can prove (by induction on n) that $\mathcal{M}^{[n]}(x; \varepsilon) = (\mathcal{M}^{[n]}(-x; \varepsilon))^T = (\mathcal{M}^{[n]}(x; \varepsilon))^\dagger$, with T and \dagger denoting transposition and adjointness, respectively, and this is enough to see that the same cancellation mechanism applies [23].

Thus (47) implies that

$$\|g^{[n]}(x; \varepsilon)\| = \left\| \left(x^2 - \varepsilon^2 x^2 \overline{\mathcal{M}}^{[n]}(x; \varepsilon) \right)^{-1} \right\| \leq \frac{2}{x^2}, \quad (48)$$

and the same argument as used in Sect. “Small Divisors” implies that the series in (43) for h_v converges for $|\varepsilon| < \varepsilon_0$, where ε_0 is defined as in (26) with $C_0 = 2\gamma^2/C_2$. Moreover (47) also implies that $h(\omega t)$ is analytic in ε (notwithstanding that the expansion is not a power expansion), so that we can say a posteriori that the original power series (25) also converges.

Generalizations

The case where the function $G(u)$ in (1) does not vanish is notationally more involved, and it is explicitly discussed in Gentile [28].

Extensions of the results described above to Hamiltonian functions more general than (5) require only some notational complications, and can be found in Gentile & Mastropietro [37] for anisochronous systems and in Bartuccelli & Gentile [3] for isochronous systems. The case (5) with f a real analytic function was first discussed by using trees in Chierchia & Falcolini [12].

Lower-Dimensional Tori

A tree formalism for hyperbolic lower-dimensional tori was introduced and studied in Gallavotti [19], Gentile [26] and Gallavotti, Gentile & Mastropietro [25], for systems consisting of a set of rotators interacting with a pendulum. In that case also the stable and unstable manifolds (*whiskers*) of the tori were studied with diagrammatic techniques.

For Hamiltonian functions of the form (5) solutions of the form (43) describe quasi-periodic motions with

frequency vector ω on a d -dimensional manifold. If we fix t , say $t = 0$, and keep α_0 as a parameter, we obtain a parametrization of the manifold in terms of $\alpha_0 \in \mathbb{T}^d$, hence the manifold is an invariant torus. We say that the torus is a *maximal torus*.

We can study the problem of persistence of *lower-dimensional tori* by considering unperturbed solutions $\alpha(t) = \alpha_0 + \omega t$, where the components of ω are rationally dependent. For instance we can imagine that there exist s linearly independent vectors $\hat{v}_1, \dots, \hat{v}_s \in \mathbb{Z}^d$ such that $\omega \cdot \hat{v}_k = 0$ for $k = 1, \dots, s$. In that case, we say that the unperturbed torus is a *resonant* torus of order s . We can imagine performing a linear change of variables which transforms the frequency vector into a new vector, that we still denote with ω , such that $\omega = (\bar{\omega}, 0)$, where $\bar{\omega} \in \mathbb{R}^r$ and $0 \in \mathbb{R}^s$, with $r + s = d$, and $\bar{\omega}$ is irrational. This naturally suggests that we write $\alpha = (\bar{\alpha}, \beta)$, with $\bar{\alpha} \in \mathbb{T}^r$ and $\beta \in \mathbb{T}^s$. More generally, here and henceforth in this subsection for any vector $v \in \mathbb{R}^d$ we denote by \bar{v} the vector in \mathbb{R}^r whose components are the first r components of v . For instance for the initial phase α_0 we write $\alpha_0 = (\bar{\alpha}_0, \beta_0)$.

In general the resonant torus is destroyed by the perturbation, and only some lower-dimensional tori persist under perturbation. We assume on $\bar{\omega}$ one of the Diophantine conditions of Sect. “Small Divisors” in \mathbb{R}^r , for instance $|\bar{\omega} \cdot \bar{v}| > \gamma/|\bar{v}|^\tau$ for all $\bar{v} \in \mathbb{Z}^r$, $\bar{v} \neq 0$.

To prove the existence of a maximal torus for the system with Hamiltonian function (5) we needed no condition on the perturbation f . On the contrary to prove existence of lower-dimensional tori, we need some *non-degeneracy condition*: by defining

$$f_0(\beta) = \int_{\mathbb{T}^r} \frac{d\bar{\alpha}}{(2\pi)^r} f(\bar{\alpha}, \beta), \quad (49)$$

if $\partial_\beta f_0(\beta_*) = 0$ for some β_* then we assume that the matrix $\partial_\beta^2 f_0(\beta_*)$ is positive definite (more generally one could assume it to be non-singular, that is $\det \partial_\beta^2 f_0(\beta_*) \neq 0$).

The formal analysis can be carried out as in Sect. “Trees and Graphical Representation”, with the only difference being that now the compatibility conditions $[\varepsilon \partial_\alpha f(\alpha)]_v^{(k)} = 0$ have to be imposed for all v such that $\bar{v} = 0$, because $\omega \cdot v = \bar{\omega} \cdot \bar{v}$. It turns out to be an identity only for the first r components (this is trivial for $k = 1$, whereas it requires some work for $k > 1$). For the last s components, for $k = 1$ it reads $\partial_\beta f_0(\beta_0) = 0$, hence it fixes β_0 to be a stationary point for $f_0(\beta)$, while for higher values of k it fixes the corrections of higher order of these values (to do this we need the non-degeneracy condition). Thus, we are free to choose only $\bar{\alpha}_0$ as a free parameter, since the last s components of α_0 have to be fixed.

Clusters and self-energy clusters are defined as in Sect. “Multiscale Analysis”. Note that only the first r components $\tilde{\nu}$ of the momenta ν intervene in the definition of the scales – again because $\omega \cdot \nu = \tilde{\omega} \cdot \tilde{\nu}$. In particular, in the definition of self-energy clusters, in (33) we must replace ν_ν with $\tilde{\nu}_\nu$. Thus, already to first order the value of a self-energy cluster can be non-zero: for $k_T = 1$, that is for T containing only a node ν with mode label $(\tilde{\nu}_\nu, \tilde{\nu}_\nu) = (0, \tilde{\nu}_\nu)$, the matrix $\mathcal{V}_T(x; \varepsilon)$ is of the form

$$\mathcal{V}_T(x) = \begin{pmatrix} 0 & 0 \\ 0 & b_{\tilde{\nu}_\nu} \end{pmatrix}, \quad (50)$$

$$(b_{\tilde{\nu}_\nu})_{i,j} = e^{i\tilde{\nu}_\nu \cdot \beta_0} (i\tilde{\nu}_{\nu,i})(i\tilde{\nu}_{\nu,j}) f_{(0,\tilde{\nu}_\nu)},$$

with $i, j = r+1, \dots, d$. If we sum over $\tilde{\nu}_\nu \in \mathbb{Z}^s$ and multiply times ε , we obtain

$$M_0 := \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon B \end{pmatrix}, \quad (51)$$

$$B_{i,j} = \sum_{\tilde{\nu} \in \mathbb{Z}^s} e^{i\tilde{\nu} \cdot \beta_0} (i\tilde{\nu}_i)(i\tilde{\nu}_j) f_{(0,\tilde{\nu})} = \partial_{\beta_i} \partial_{\beta_j} f_0(\beta_0).$$

The $s \times s$ block B is non-zero: in fact, the non-degeneracy condition yields that it is invertible.

To higher orders one finds that the matrix $\mathcal{M}^{[n]}(x; \varepsilon)$, with $x = \omega \cdot \nu = \tilde{\omega} \cdot \tilde{\nu}$, is a self-adjoint matrix and $\mathcal{M}^{[n]}(x; \varepsilon) = (\mathcal{M}^{[n]}(-x; \varepsilon))^T$, as in the case of maximal tori. Moreover the corresponding eigenvalues $\lambda_i^{[n]}(x; \varepsilon)$ satisfy $\lambda_i^{[n]}(x; \varepsilon) = O(\varepsilon^2 x^2)$ for $i = 1, \dots, r$ and $\lambda_i^{[n]}(x; \varepsilon) = O(\varepsilon)$ for $i = r+1, \dots, d$; this property is not trivial because of the off-diagonal blocks (which in general do not vanish at orders $k \geq 2$), and to prove it one has to use the self-adjointness of the matrix $\mathcal{M}^{[n]}(x; \varepsilon)$. More precisely one has $\lambda_i^{[n]}(x; \varepsilon) = \varepsilon a_i + O(\varepsilon^2)$ for $i > r$, where a_{r+1}, \dots, a_d are the s eigenvalues of the matrix B in (51). From this point on the discussion proceeds in a very different way according to the sign of ε (recall that we are assuming that $a_i > 0$ for all $i > r$).

For $\varepsilon < 0$ one has $\lambda_i^{[n]}(x; \varepsilon) = a_i \varepsilon + O(\varepsilon^2) < 0$ for $i > r$, so that we can bound the last s eigenvalues of $x^2 - \mathcal{M}^{[n]}(x; \varepsilon)$ with x^2 , and the first r with $x^2/2$ by the same argument as in Sect. “Resummation”. Hence we obtain easily the convergence of the series (43); of course, analyticity at the origin is prevented because of the condition $\varepsilon < 0$. We say in that case that the lower-dimensional tori are *hyperbolic*. We refer to Gallavotti & Gentile [22] and Gallavotti et al. [23] for details.

The case of *elliptic* lower-dimensional tori – that is $\varepsilon > 0$ when $a_i > 0$ for all $i > r$ – is more difficult. Essentially the idea is as follows (we only sketch the strat-

egy: the details can be found in Gentile & Gallavotti [36]). One has to define the scales recursively, by using a variant, first introduced in Gentile [27], of the resummation technique described in Sect. “Resummation”. We say that ν is on scale 0 if $|\tilde{\omega} \cdot \tilde{\nu}| \geq \gamma$ and on scale $[\geq 1]$ otherwise: for ν on scale 0 we write (42) with $M = M_0$, as given in (51). This defines the propagators of the lines ℓ on scale $n = 0$ as

$$g_\ell = g^{[0]}(\omega \cdot \nu_\ell) = ((\tilde{\omega} \cdot \tilde{\nu}_\ell)^2 - M_0)^{-1}. \quad (52)$$

Denote by λ_i the eigenvalues of M_0 : given ν on scale $[\geq 1]$ we say that ν is on scale 1 if $2^{-1}\gamma \leq \min_{i=1,\dots,d} \sqrt{|(\tilde{\omega} \cdot \tilde{\nu})^2 - \lambda_i|}$, and on scale $[\geq 2]$ if $\min_{i=1,\dots,d} \sqrt{|(\tilde{\omega} \cdot \tilde{\nu})^2 - \lambda_i|} < 2^{-1}\gamma$. For ν on scale 1 we write (42) with M replaced by $\mathcal{M}^{[0]}(\omega \cdot \tilde{\nu}; \varepsilon)$, which is given by M_0 plus the sum of the values of all self-energy clusters T on scale $n_T = 0$. Then the propagators of the lines ℓ on scale $n_\ell = 1$ is defined as

$$g_\ell = g^{[1]}(\omega \cdot \nu_\ell) = ((\tilde{\omega} \cdot \tilde{\nu}_\ell)^2 - \mathcal{M}^{[0]}(\tilde{\omega} \cdot \tilde{\nu}_\ell; \varepsilon))^{-1}. \quad (53)$$

Call $\lambda_i^{[n]}(x; \varepsilon)$ the eigenvalues of $\mathcal{M}^{[n]}(x; \varepsilon)$: given ν on scale $[\geq 2]$ we say that ν is on scale 2 if $2^{-2}\gamma \leq \min_{i=1,\dots,d} \sqrt{|(\tilde{\omega} \cdot \tilde{\nu})^2 - \lambda_i^{[0]}(\tilde{\omega} \cdot \tilde{\nu}; \varepsilon)|}$, and on scale $[\geq 3]$ if $\min_{i=1,\dots,d} \sqrt{|(\tilde{\omega} \cdot \tilde{\nu})^2 - \lambda_i^{[0]}(\tilde{\omega} \cdot \tilde{\nu}; \varepsilon)|} < 2^{-2}\gamma$. For ν on scale 2 we write (42) with M replaced by $\mathcal{M}^{[1]}(\tilde{\omega} \cdot \tilde{\nu}; \varepsilon)$, which is given by $\mathcal{M}^{[0]}(\tilde{\omega} \cdot \tilde{\nu}; \varepsilon)$ plus the sum of the values of all self-energy clusters T on scale $n_T = 1$. Thus, the propagators of the lines ℓ on scale $n_\ell = 2$ will be defined as

$$g_\ell = g^{[2]}(\omega \cdot \nu_\ell) = ((\tilde{\omega} \cdot \tilde{\nu}_\ell)^2 - \mathcal{M}^{[1]}(\tilde{\omega} \cdot \tilde{\nu}_\ell; \varepsilon))^{-1}, \quad (54)$$

and so on. The propagators are self-adjoint matrices, hence their norms can be bounded through the corresponding eigenvalues. In order to proceed as in Sects. “Multiscale Analysis” and “Resummation” we need some Diophantine conditions on these eigenvalues. We can assume for some $\tau' > \tau$

$$\left| |\tilde{\omega} \cdot \tilde{\nu}| - \sqrt{|\lambda_i^{[n]}(\tilde{\omega} \cdot \tilde{\nu}; \varepsilon)|} \right| > \frac{\gamma}{|\tilde{\nu}|^{\tau'}}, \quad \forall \tilde{\nu} \neq 0, \quad (55)$$

for all $i = 1, \dots, d$ and $n \geq 0$. These are known as the *first Melnikov conditions*.

Unfortunately, things do not proceed so plainly. In order to prove a bound like (35), possibly with a different τ' replacing τ , we need to compare the propagators of the

lines entering and exiting clusters T which are not self-energy clusters. This requires replacing (36) with

$$2^{-n+2}\gamma \geq \left| \bar{\omega} \cdot (\bar{v}_1 - \bar{v}_2) \pm \sqrt{|\lambda_i^{[n]}(\bar{\omega} \cdot \bar{v}_1; \varepsilon)|} \pm \sqrt{|\lambda_j^{[n]}(\bar{\omega} \cdot \bar{v}_2; \varepsilon)|} \right| > \frac{\gamma}{|\bar{v}_1 - \bar{v}_2|^{\tau'}}, \quad (56)$$

for all $i, j = 1, \dots, d$ and choices of the signs \pm , and hence introduces further Diophantine conditions, known as the *second Melnikov conditions*.

The conditions in (56) turn out to be too many, because for all $n \geq 0$ and all $\bar{v} \in \mathbb{Z}^r$ such that $\bar{v} = \bar{v}_1 - \bar{v}_2$ there are infinitely many conditions to be considered, one per pair (\bar{v}_1, \bar{v}_2) . However we can impose both the conditions (55) and (56) not for the eigenvalues $\lambda_i^{[n]}(\bar{\omega} \cdot \bar{v}; \varepsilon)$, but for some quantities $\underline{\lambda}_i^{[n]}(\varepsilon)$ independent of \bar{v} and then use the smoothness of the eigenvalues in x to control $(\bar{\omega} \cdot \bar{v})^2 - \lambda_i^{[n]}(\bar{\omega} \cdot \bar{v}; \varepsilon)$ in terms of $(\bar{\omega} \cdot \bar{v})^2 - \underline{\lambda}_i^{[n]}(\varepsilon)$. Eventually, beside the Diophantine condition on $\bar{\omega}$, we have to impose the Melnikov conditions

$$\left| |\bar{\omega} \cdot \bar{v}| - \sqrt{|\underline{\lambda}_i^{[n]}(\varepsilon)|} \right| > \frac{\gamma}{|\bar{v}|^{\tau'}}, \quad (57)$$

$$\left| |\bar{\omega} \cdot \bar{v}| \pm \sqrt{|\underline{\lambda}_i^{[n]}(\varepsilon)|} \pm \sqrt{|\underline{\lambda}_j^{[n]}(\varepsilon)|} \right| > \frac{\gamma}{|\bar{v}|^{\tau'}},$$

for all $\bar{v} \neq 0$ and all $n \geq 0$. Each condition in (57) leads us to eliminate a small interval of values of ε . For the values of ε which are left define $h(\omega t)$ according to (43) and (46), with the new definition of the propagators. If ε is small enough, say $|\varepsilon| < \varepsilon_0$, then the series (43) converges. Denote by $\mathfrak{E} \subset [0, \varepsilon_0]$ the set of values of ε for which the conditions (57) are satisfied. One can prove that \mathfrak{E} is a *Cantor set*, that is a perfect, nowhere dense set. Moreover \mathfrak{E} has large relative Lebesgue measure, in the sense that

$$\lim_{\varepsilon \rightarrow 0} \frac{\text{meas}(\mathfrak{E} \cap [0, \varepsilon])}{\varepsilon} = 1, \quad (58)$$

provided τ' in (57) is large enough with respect to τ . The property (58) yields that, notwithstanding that we are eliminating infinitely many intervals, the measure of the union of all these intervals is small.

If $a_i < 0$ for all $i > r$ we reason in the same way, simply exchanging the role of positive and negative ε . On the contrary if $a_i = 0$ for some $i > r$, the problem becomes much more difficult. For instance if $s = 1$ and $a_{r+1} = 0$, then in general perturbation theory in ε is not possible, not even at a formal level. However, under some conditions,

one can still construct fractional series in ε , and prove that the series can be resummed [24].

Other Ordinary Differential Equations

The formalism described above extends to other models, such as skew-product systems [29] and systems with strong damping in the presence of a quasi-periodic forcing term [32].

As an example of skew-product system one can consider the linear differential equation $\dot{x} = (\lambda A + \varepsilon f(\omega t))x$ on $\text{SL}(2, \mathbb{R})$, where $\lambda \in \mathbb{R}$, ε is a small real parameter, $\omega \in \mathbb{R}^n$ is an irrational vector, and $A, f \in \mathfrak{sl}(2, \mathbb{R})$, with A is a constant matrix and f an analytic function periodic in its arguments. Trees for skew-products were considered by Iserles and Nørsett [44], but they used expansions in time, hence not suited for the study of global properties, such as quasi-periodicity.

Quasi-periodically forced one-dimensional systems with strong damping are described by the ordinary differential equations $\dot{x} + \gamma x + g(x) = f(\omega t)$, where $x \in \mathbb{R}$, $\varepsilon = 1/\gamma$ is a small real parameter, $\omega \in \mathbb{R}^n$ is irrational, and f, g are analytic functions (g is the “force”), with f periodic in its arguments.

We refer to the bibliography for details and results on the existence of quasi-periodic solutions.

Bryuno Vectors

The diagrammatic methods can be used to prove the any unperturbed maximal torus with frequency vector which is a Bryuno vector persists under perturbation for ε small enough [31]. One could speculate whether the Bryuno condition (30) is optimal. In general the problem is open. However, in the case of the standard map – see Sect. “The Standard Map”, – one can prove [6,15] that, by considering the radius of convergence ε_0 of the perturbation series as a function of ω , say $\varepsilon_0 = \rho_0(\omega)$, then there exists a universal constant C such that

$$|\log \rho_0(\omega) + 2B(\omega)| \leq C. \quad (59)$$

In particular this yields that the invariant curve with rotation number ω persists under perturbation if and only if ω is a Bryuno number. The proof of (59) requires the study of a more refined cancellation than that discussed in Sect. “Resummation”. We refer to Berretti & Gentile [6] and Gentile [28] for details.

Extensions to Bryuno vectors for lower-dimensional tori can also be found in Gentile [31]. For the models considered in Sect. “Other Ordinary Differential Equations” we refer to Gentile [29] and Gentile et al. [33].

Partial Differential Equations

Existence of quasi-periodic solutions in systems described by one-dimensional nonlinear partial differential equations (finite-dimensional tori in infinite-dimensional systems) was first studied by Kuksin [48], Craig and Wayne [14] and Bourgain [9]. In these systems, even the case of periodic solutions yields small divisors, and hence requires a multiscale analysis. The study of persistence of periodic solutions for nonlinear Schrödinger equations and nonlinear wave equations, with the techniques discussed here, can be found in Gentile & Mastropietro [39], Gentile et al. [40], and Gentile & Procesi [41].

The models are still described by (1), with $G(u) = 0$, but now D is given by $D = \partial_t^2 - \Delta + \mu$ in the case of the wave equation and by $D = i\partial_t - \Delta + \mu$ in the case of the Schrödinger equation, where Δ is the Laplacian and $\mu \in \mathbb{R}$. In dimension 1, one has $\Delta = \partial_x^2$. If we look for periodic solutions with frequency ω it can be convenient to pass to Fourier space, where the operator D acts as

$$D: e^{i\omega nt + imx} \rightarrow (-\omega^2 n^2 + m^2 + \mu) e^{i\omega nt + imx}, \quad (60)$$

for the wave equation; a similar expression holds for the Schrödinger equation. Therefore the kernel of D can be arbitrarily close to zero for n and m large enough.

Then one can consider, say, (1) for $x \in [0, \pi]$ and $F(u) = u^3$, with Dirichlet boundary conditions $u(0) = u(\pi) = 0$, and study the existence of periodic solutions with frequency ω close to some of the unperturbed frequencies. We refer to the cited bibliography for results and proofs.

Conclusions and Future Directions

The diagrammatic techniques described above have been applied also in cases where no small divisors appear; cf. Berretti & Gentile [4] and Gentile et al. [34]. Of course, such problems are much easier from a technical point of view, and can be considered as propaedeutic examples to become familiar with the tree formalism. Also the study of lower-dimensional tori becomes easy for $r = 1$ (periodic solutions): in that case one has $|\tilde{\omega} \cdot \tilde{v}| \geq |\tilde{\omega}|$ for all $\tilde{v} \neq 0$, so that the product of the propagators is bounded by $|\tilde{\omega}|^{-2k}$, and one can proceed as in Sect. “Small Divisors” to obtain analyticity of the solutions.

In the case of hyperbolic lower-dimensional tori, if ω is a two-dimensional Diophantine vector of constant type (that is, with $\tau = 1$) the conjugation function h can be proved to be Borel summable [13]. Analogous considerations hold for the one-dimensional systems in the presence

of friction and of a quasiperiodic forcing term described in Sect. “Other Ordinary Differential Equations”; in that case one has Borel summability also for one-dimensional ω , that is for periodic forcing [33]. It would be interesting to investigate whether Borel summability could be obtained for higher values of τ .

Recently existence of finite-dimensional tori in the nonlinear Schrödinger equation in higher dimensions was proved by Bourgain [10]. It would be interesting to investigate how far the diagrammatic techniques extend to deal with such higher dimensional generalizations. The main problem is that (the analogues of) the second Melnikov conditions in (57) cannot be imposed.

In certain cases the tree formalism was extended to non-analytic systems, such as some quasi-integrable systems of the form (5) with f in a class of C^p functions for some finite p [7,8]. However, up to exceptional cases, the method described here seems to be intrinsically suited in cases in which the vector fields are analytic. The reason is that in order to exploit the expansion (3), we need that F be infinitely many times differentiable and we need a bound on the derivatives. It is a remarkable property that the perturbation series can be given a meaning also in cases where the solutions are not analytic in ε .

An advantage of the diagrammatic method is that it allows rather detailed information about the solutions, hence it could be more convenient than other techniques to study problems where the underlying structure is not known or too poor to exploit general abstract arguments.

Another advantage is the following. If one is interested not only in proving the existence of the solutions, but also in explicitly constructing them with any prefixed precision, this requires performing analytical or numerical computations with arbitrarily high accuracy. Then high perturbation orders have to be reached, and the easiest and most direct way to proceed is just through perturbation theory: so the approach illustrated here allows a unified treatment for both theoretical investigations and computational ones.

The resummation technique described in Sect. “Resummation” can also be used for computational purposes. With respect to the naive power series expansion it can reduce the computation time required to approximate the solution within a prefixed precision. It can also provide accurate information on the analyticity properties of the solution. For instance, for the Kepler equation, Levi-Civita at the beginning of the last century described a resummation rule (see [49]), which gives immediately the radius of convergence of the perturbation series. Of course, in the case of small divisor problems, everything becomes much more complicated.

Bibliography

Primary Literature

- Arnold VI (1963) Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian (Russian). *Uspehi Mat Nauk* 18(5):13–40
- Arnold VI, Kozlov VV, Neishtadt AI (1988) *Dynamical systems III*, Encyclopaedia of Mathematical Sciences, vol 3. Springer, Berlin
- Bartuccelli MV, Gentile G (2002) Lindstedt series for perturbations of isochronous systems: a review of the general theory. *Rev Math Phys* 14(2):121–171
- Berretti A, Gentile G (1999) Scaling properties for the radius of convergence of a Lindstedt series: the standard map. *J Math Pures Appl* 78(2):159–176
- Berretti A, Gentile G (2000) Scaling properties for the radius of convergence of a Lindstedt series: generalized standard maps. *J Math Pures Appl* 79(7):691–713
- Berretti A, Gentile G (2001) Bryuno function and the standard map. *Comm Math Phys* 220(3):623–656
- Bonetto F, Gallavotti G, Gentile G, Mastropietro V (1998) Quasi linear flows on tori: regularity of their linearization. *Comm Math Phys* 192(3):707–736
- Bonetto F, Gallavotti G, Gentile G, Mastropietro V (1998) Lindstedt series, ultraviolet divergences and Moser's theorem. *Ann Scuola Norm Sup Pisa Cl Sci* 26(3):545–593
- Bourgain J (1998) Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equations. *Ann of Math* 148(2):363–439
- Bourgain J (2005) Green's function estimates for lattice Schrödinger operators and applications. *Annals of Mathematics Studies* 158. Princeton University Press, Princeton
- Bricmont J, Gawędzki K, Kupiainen A (1999) KAM theorem and quantum field theory. *Comm Math Phys* 201(3):699–727
- Chierchia L, Falcolini C (1994) A direct proof of a theorem by Kolmogorov in Hamiltonian systems. *Ann Scuola Norm Sup Pisa Cl Sci* 21(4):541–593
- Costin O, Gallavotti G, Gentile G, Giuliani A (2007) Borel summability and Lindstedt series. *Comm Math Phys* 269(1):175–193
- Craig W, Wayne CE (1993) Newton's method and periodic solutions of nonlinear wave equations. *Comm Pure Appl Math* 46(11):1409–1498
- Davie AM (1994) The critical function for the semistandard map. *Nonlinearity* 7(1):219–229
- Eliasson LH (1996) Absolutely convergent series expansions for quasi periodic motions. *Math Phys Electron J* 2(4):1–33
- Gallavotti G (1983) *The elements of mechanics*. Texts and Monographs in Physics. Springer, New York
- Gallavotti G (1994) Twistless KAM tori. *Comm Math Phys* 164(1):145–156
- Gallavotti G (1994) Twistless KAM tori, quasi flat homoclinic intersections, and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems. A review. *Rev Math Phys* 6(3):343–411
- Gallavotti G (2001) Renormalization group in statistical mechanics and mechanics: gauge symmetries and vanishing beta functions, Renormalization group theory in the new millennium III. *Phys Rep* 352(4–6):251–272
- Gallavotti G (2006) Classical mechanics. In: Françoise JP, Naber GL, Tsun TS (eds) *Encyclopedia of Mathematical Physics*, vol 1. Elsevier, Oxford
- Gallavotti G, Gentile G (2002) Hyperbolic low-dimensional invariant tori and summations of divergent series. *Comm Math Phys* 227(3):421–460
- Gallavotti G, Bonetto F, Gentile G (2004) *Aspects of ergodic, qualitative and statistical theory of motion*. Texts and Monographs in Physics. Springer, Berlin
- Gallavotti G, Gentile G, Giuliani A (2006) Fractional Lindstedt series. *J Math Phys* 47(1):1–33
- Gallavotti G, Gentile G, Mastropietro V (1999) A field theory approach to Lindstedt series for hyperbolic tori in three time scales problems. *J Math Phys* 40(12):6430–6472
- Gentile G (1995) Whiskered tori with prefixed frequencies and Lyapunov spectrum. *Dyn Stab Syst* 10(3):269–308
- Gentile G (2003) Quasi-periodic solutions for two-level systems. *Comm Math Phys* 242(1–2):221–250
- Gentile G (2006) Brjuno numbers and dynamical systems. *Frontiers in number theory, physics, and geometry*. Springer, Berlin
- Gentile G (2006) Diagrammatic techniques in perturbation theory. In: Françoise JP, Naber GL, Tsun TS (eds) *Encyclopedia of Mathematical Physics*, vol 2. Elsevier, Oxford
- Gentile G (2006) Resummation of perturbation series and reducibility for Bryuno skew-product flows. *J Stat Phys* 125(2):321–361
- Gentile G (2007) Degenerate lower-dimensional tori under the Bryuno condition. *Ergod Theory Dyn Syst* 27(2):427–457
- Gentile G, Bartuccelli MV, Deane JHB (2005) Summation of divergent series and Borel summability for strongly dissipative differential equations with periodic or quasiperiodic forcing terms. *J Math Phys* 46(6):1–21
- Gentile G, Bartuccelli MV, Deane JHB (2006) Quasiperiodic attractors, Borel summability and the Bryuno condition for strongly dissipative systems. *J Math Phys* 47(7):1–10
- Gentile G, Bartuccelli MV, Deane JHB (2007) Bifurcation curves of subharmonic solutions and Melnikov theory under degeneracies. *Rev Math Phys* 19(3):307–348
- Gentile G, Cortez DA, Barata JCA (2005) Stability for quasiperiodically perturbed Hill's equations. *Comm Math Phys* 260(2):403–443
- Gentile G, Gallavotti G (2005) Degenerate elliptic tori. *Comm Math Phys* 257(2):319–362
- Gentile G, Mastropietro V (1996) Methods for the analysis of the Lindstedt series for KAM tori and renormalizability in classical mechanics. A review with some applications. *Rev Math Phys* 8(3):393–444
- Gentile G, Mastropietro V (2001) Renormalization group for one-dimensional fermions. A review on mathematical results. Renormalization group theory in the new millennium III. *Phys Rep* 352(4–6):273–437
- Gentile G, Mastropietro V (2004) Construction of periodic solutions of nonlinear wave equations with Dirichlet boundary conditions by the Lindstedt series method. *J Math Pures Appl* 83(8):1019–1065
- Gentile G, Mastropietro V, Procesi M (2005) Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions. *Comm Math Phys* 256(2):437–490
- Gentile G, Procesi M (2006) Conservation of resonant periodic solutions for the one-dimensional nonlinear Schrödinger equation. *Comm Math Phys* 262(3):533–553

42. Gentile G, van Erp TS (2005) Breakdown of Lindstedt expansion for chaotic maps. *J Math Phys* 46(10):1–20
43. Harary F (1969) *Graph theory*. Addison-Wesley, Reading
44. Iserles A, Nørsett SP (1999) On the solution of linear differential equations in Lie groups. *Royal Soc Lond Philos Trans Ser A Math Phys Eng Sci* 357(1754):983–1019
45. Khanin K, Lopes Dias J, Marklof J (2007) Multidimensional continued fractions, dynamical renormalization and KAM theory. *Comm Math Phys* 270(1):197–231
46. Koch H (1999) A renormalization group for Hamiltonians, with applications to KAM tori. *Ergod Theory Dyn Syst* 19(2):475–521
47. Kolmogorov AN (1954) On conservation of conditionally periodic motions for a small change in Hamilton's function (Russian). *Dokl Akad Nauk SSSR* 98:527–530
48. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. *Lecture Notes in Mathematics*, vol 1556. Springer, Berlin
49. Levi-Civita T (1954) *Opere matematiche*. Memorie e note. Zanichelli, Bologna
50. MacKay RS (1993) Renormalisation in area-preserving maps. *Advanced Series in Nonlinear Dynamics*, 6. World Scientific Publishing, River Edge
51. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. *Nachr Akad Wiss Göttingen Math-Phys Kl.* II 1962:1–20
52. Moser J (1967) Convergent series expansions for quasi-periodic motions. *Math Ann* 169:136–176
53. Poincaré H (1892–1899) *Les méthodes nouvelles de la mécanique céleste*. Gauthier-Villars, Paris
54. Pöschel J (1982) Integrability of Hamiltonian systems on Cantor sets. *Comm Pure Appl Math* 35(5):653–696
55. Wintner A (1941) *The analytic foundations of celestial mechanics*. Princeton University Press, Princeton

Books and Reviews

- Berretti A, Gentile G (2001) Renormalization group and field theoretic techniques for the analysis of the Lindstedt series. *Regul Chaotic Dyn* 6(4):389–420
- Gentile G (1999) *Diagrammatic techniques in perturbations theory, and applications*. Symmetry and perturbation theory. World Science, River Edge

Differential Games

MARC QUINCAMPOIX

Université de Bretagne Occidentale, Brest, France

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Qualitative and Quantitative Differential Games](#)

[Existence of a Value for Zero Sum Differential Games](#)

[Nonantagonist Differential Games](#)

[Miscellaneous](#)

Future Directions

Bibliography

Glossary

Dynamics This is the law which governs the evolution of the system: for differential games it is a differential equation.

Strategies This is the way a player chooses his control as a function of the state of the system and of the action of his opponents.

Information This is the set of parameters known by the player in order to build his strategy.

Definition of the Subject

Differential games is a mathematical theory which is concerned with problems of conflicts modeled as game problems in which the state of the players depends on time in a continuous way. The positions of the players are the solution to differential equations. Differential games can be described from two different points of view, depending mainly on the field of application. Firstly, they can be considered as games where time is continuous. This aspect is often considered for applications in economics or management sciences. Secondly, they also can be viewed as control problems with several controllers having different objectives. In this way, differential games are a part of control theory with conflicts between the players. The second aspect often concerns classical applications of control theories: engineering sciences.

The importance of the subject was emphasized by J. von Neuman in 1946 in his pioneer book “Theory of Games and Economic Behaviour” [35]: *We repeat most emphatically that our theory is thoroughly static. A dynamic theory would unquestionably be more complete and therefore preferable*. But at that time, static aspects of game theory were mostly treated. The true birth of the domain was pursuit differential games (motivated by military applications during the “Cold War”) developed in the 1950s concurrently by R. Isaacs in the USA and L. Pontryagin in the Soviet Union. Now differential games have a wide range of applications from economics to engineering sciences but also more recently in biology and behavioral ecology. The present article is focused on two-player zero sum and antagonist differential games.

Introduction

The best-known example of a differential game is the pursuit game “Lion and Man” where in an infinite plane a Lion wants to catch a Man. This example is elementary enough

for describing clearly the problematic and we will use this example as an illustration throughout the article. The position of the Man at time t is $y(t)$ while the Lion's position is $z(t)$. At any time the Man can choose his velocity $y'(t)$ in any direction but with a maximum intensity, while the Lion can choose his velocity $z'(t)$ with an intensity less than or equal to L (here $L > M$). The objective of the Lion is to catch the Man as soon as possible, the aim of the Man is to escape the Lion as long as possible. In this very intuitive game, we wish to introduce the important elements needed for defining a differential game: *the dynamics, the actions of the players, the objectives, the rules of the game (the strategies)*. The *dynamics* is then

$$y'(t) = u(t), \quad u(t) \in U, \quad z'(t) = v(t) \in V; \quad t \geq 0 \quad (1)$$

where $u(t)$ is the *action* chosen by the first player (the Man), $u(t)$ belongs to a set $U = \{u \mid \|u\| \leq M\}$, similarly $v(t)$ belongs to the set $V = \{v \mid \|v\| \leq L\}$. The *objectives* of the two players are of antagonist nature: The Lion wants to minimize the time before capture while the Man wants to maximize it. The rule of the game is the way the two players choose their actions: here we can assume that they choose their instantaneous velocity as a function of the current positions

$$u(t) = \bar{u}(x(t), y(t)), \quad v(t) = \bar{v}(x(t), y(t)). \quad (2)$$

Clearly the time of capture is T in a function of the strategies \bar{u} and \bar{v} ; the Lion wants to minimize this function, the Man wants to maximize it. Solving the game means finding the *optimal value* namely the time of capture which is minimum over all strategies \bar{v} and maximum over all strategies \bar{u} . In the elementary example of "Lion and Man," one can easily check that

$$\begin{aligned} \bar{v}(y, z) &= L \frac{y - z}{\|y - z\|}, \\ \bar{u}(y, z) &= -M \frac{y - z}{\|y - z\|}, \\ T &= \frac{\|y(0) - z(0)\|}{L - M}, \end{aligned}$$

which is a rather intuitive solution: the Lion runs in a direction towards the Man at his maximal possible speed L , while the Man runs in the opposite direction with his maximal speed M . We leave this illustrative example which has allowed us to underline the concepts of differential games.

The present article is restricted to the two player case. A state variable x is governed by a differential equation

$$x'(t) = f(x(t), u(t), v(t)). \quad (3)$$

The players act on the state variable x by choosing two controls u and v that are functions of the time variable t taking their values in spaces U and V . (Note that for the Lion and the Man $x = (y, z)$ and each player controls a part of the state: the game is said to be *separated*). For making the notations clearer, the first player which chooses u is called Ursula and the second player which chooses v is called Victor.

A classification of differential games according to the level of conflict between the players is very relevant. The case where the two players have opposite objectives concerns *antagonist games* or *zero-sum games*, and in the case of a nonopposite objective these are *cooperative* or *nonantagonist* differential games.

The objectives of the game can be quantitative – players want to minimize or maximize a *payoff* (a function depending on the initial state and their action, for instance the time of capture in the Lion and Man case) – or qualitative for instance one player wants the state to stay in a subset of the state space while the other player wants the state to reach a given target. In Isaacs' terminology, these games are *games of degree* or *games of kind*. The definition of the rules of the games (or strategies) is one of the most difficult topics in differential games. This problem does not exist neither static games nor for elementary pursuit games like the Lion and Man game. To illustrate this fact, suppose that we have to solve the game with a strategy of *feedback form*, i. e. when controls depend on the current state (as in Eq. (2)), then we have to solve the differential equation

$$x'(t) = f(x(t), \bar{u}(x(t)), \bar{v}(x(t)))$$

which has neither uniqueness nor existence of solutions, even if Eq. (3) enjoys good properties of existence and uniqueness of solutions. This point is crucial and is of pure mathematical nature. A way to overcome this difficulty could be to impose regularity properties on the feedback strategies. In the – young – history of differential games, this was the "time of paradoxes" in the 1960s and 1970s, where specific regularity of feedback was adequate for a class of problem but led to "paradoxes" (for quantitative games nonexistence of the value, for qualitative games existence of an initial position starting from which players are not able either to win or to lose) in slightly modified problems. This was solved in the 1970s by enlarging the class of feedback strategies allowing the choice of the control to depend not only on the current state space but also on passed values of the control of the opponents (this is the so-called nonanticipative strategies introduced by Varaiya, Roxin, Elliot, and Kalton [16,31,33,34], another class of strategies was introduced by Krasovski [21]).

For differential games, another important feature is to express the fact that the two players act simultaneously on the system. One way to translate this fact in a mathematical way is to prove that the order of the actions of the player do not modify the final result. Section “Qualitative and Quantitative Differential Games” is concerned with this question. For instance, in the case of quantitative games like the Lion and Man, by interchanging the order of operations “minimum” and “maximum” we must obtain the same result. This is the problem of existence of the value which is of first importance for differential games. We discuss this question in Sect. “Existence of a Value for Zero Sum Differential Games”. The fifth section is devoted to cooperative quantitative games. Some other aspects or applications of differential games are evoked in Sect. “Miscellaneous”. In the last section, we give short descriptions of present domains of research such as problems of information or impulsive differential games.

Qualitative and Quantitative Differential Games

Throughout this article, we will make suppositions such that as soon as the initial position is fixed and the controls u and v are given, we have the existence of a unique associated trajectory defined for every time (this could be easily obtained with assumptions on the function f). We also denote X the state space to which the state variable x belongs.

A nonanticipative strategy for the first player associates to any action of the second player (to any control v) a control u of the first player such that *at any time* v depends only on past values of v . It could be shown that every regular feedback strategy is such a nonanticipative strategy. So a nonanticipative strategy of the first player is a map α from \mathcal{V} (the set of measurable control $v[0, +\infty] \mapsto V$) to \mathcal{U} (the set of measurable controls u), which has furthermore the nonanticipative property. Similarly, one can define a strategy of the second player as a nonanticipative function β from \mathcal{U} to \mathcal{V} .

We will now present a rather general description of a target game viewed as a game of kind (qualitative game) or as a game of degree. For this, we consider a given set in the state space – called the target – that Victor wants the state of the system to reach while Ursula wants to avoid the target. For the Lion and Man pursuit game, the target can be considered as the set of (y, z) such that $y = z$.

Qualitative Target Games

The problems consist of finding the victory domains, i. e. the set of initial positions such that one player can win. This leads to the precise following definition.

Victor Victory domain W_V is the set of initial position x_0 (not in C) such that he can find a strategy β such that there exists a time T such that whatever is the control u chosen by Ursula, the target is reached before the time T (cf. [1]).

In fact, to make this definition mathematically correct we have to add a small number ε and to say that the trajectory reaches a ε neighborhood of C :

$$W_V := \{x_0 \notin C, \exists \beta, \exists T, \varepsilon > 0, \text{ such that} \\ \forall u \in \mathcal{U}, \exists \tau > 0, \text{ dist}(x[x_0, u, \beta(u)](\tau), C) \leq \varepsilon\}.$$

In the above formula dist means the distance and $t \mapsto x[x_0, u, \beta(u)](t)$ denotes the trajectory associated with the control u and the strategy β .

In a parallel way, the victory domain W_U of Ursula is the set of initial positions x_0 for which Ursula can find a strategy α such that whatever is the control played by Victor, the associate trajectory never reaches the target C .

$$W_U := \{x_0 \notin C, \exists \alpha, \text{ such that} \\ \forall v \in \mathcal{V}, x[x_0, \alpha(v), v](t) \notin C, \forall t \geq 0\}.$$

In Fig. 1, the reader can see a figure corresponding to a specific differential game.

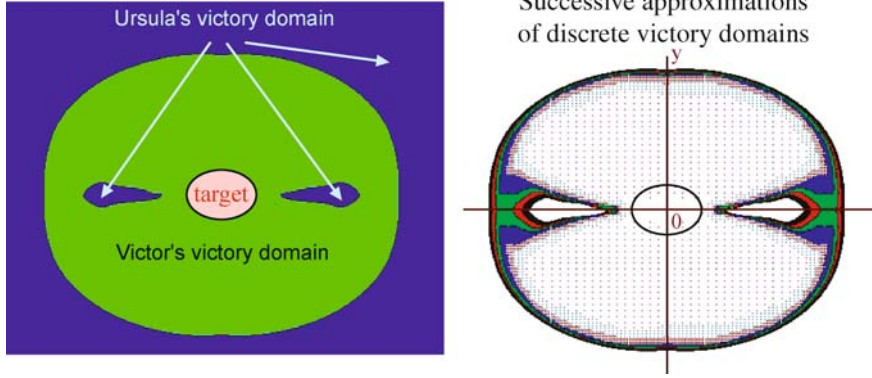
Before going further, it could be surprising to the reader that the game is not “symmetric” due to the presence of $\varepsilon > 0$. This is a mathematical problem that exceeds the scope of the present paper, for a better understanding the reader can imagine that $\varepsilon = 0$ and see [10] for deeper analysis.

The main problem of such a qualitative game is the following alternative problem. Roughly speaking, if one player does not win, the other player must win. Of course this appears to be the minimal requirement so that the modelization of the problem is correctly formulated. Nevertheless, the alternative is not an easy problem; for instance, it is not obvious that the intuitive notion of feedback strategies is not suitable for giving a positive answer to this question. The alternative can be expressed by the fact that the sets CW_U and W_V form a partition of the whole space X :

$$X = C \cup W_U \cap W_V, \quad \text{and} \quad W_U \cap W_V = \emptyset.$$

The above alternative is valid under several technical conditions (for instance C is an open set) we will not describe and under the following crucial condition – called *Isaacs’ condition*

$$\min_{v \in \mathcal{V}} \max_{u \in \mathcal{U}} \langle f(x, u, v), p \rangle = \max_{u \in \mathcal{U}} \min_{v \in \mathcal{V}} \langle f(x, u, v), p \rangle \quad (4)$$



Differential Games, Figure 1

On the left: Ursula and Victor's victory domains. On the right, successive approximations of Victor's victory domain. $(\begin{smallmatrix} x'(t) \\ y'(t) \end{smallmatrix}) = w_p \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \frac{w_p}{0.8} \begin{pmatrix} -y(t) \\ x(t) \end{pmatrix} v(t) + w_e \min(2\sqrt{x^2 + y^2}, 1) \begin{pmatrix} \cos(u(t)) \\ \sin(u(t)) \end{pmatrix}$ where $u \in [-\pi, \pi]$ and $v \in [-1, 1]$. Numerical values for the pursuer and evader's velocities are $w_p = 1$ and $w_e = 1.1$.

for any direction p . This could be understood as the existence of a saddle point for the *static game* with payoff $\langle f(x, u, v), p \rangle$ ($\langle \cdot, \cdot \rangle$ is the scalar product). The expression of Eq. (4) above is called the *Hamiltonian* of the game; it is a function $(x, p) \mapsto H(x, p)$.

The alternative theorem was first obtained by Krassovskii and Subbotin [21] for a slightly different notion of strategy, and by Cardaliaguet [10] for a nonanticipative strategy. It is worth pointing out that even for the elementary Lion and Man game played in a circular “arena” (instead of the whole plane) the problem is not obvious [19]. For more complex forms of “arenas” and for general differential games with restricted space domain the alternative problem was only solved in 2001 [14].

Once the alternative property is known, the second interesting step is to describe the victory domains. In several examples, Isaacs discovered that the boundaries of the domains satisfy a geometric equation called *the Isaacs equations*

$$H(x, \nu_x) = 0 \quad \text{for any } x \text{ on the boundary of the victory domain} \quad (5)$$

where ν_x denotes the normal of the victory domain. Hypersurfaces satisfying such an equation are *semipermeable barriers*: one player can prevent the other from crossing the barrier. From this property it is possible to obtain much information on the winning strategies. But unfortunately the victory domains are seldom regular enough to have a normal. Very often they have “corners” even for very elementary games. So there are two different ways to treat this question. A first approach initiated by Isaacs himself, and developed by Breakwell, Bernhard, and

Melykian consists of a precise and fine geometrical analysis of semipermeable barriers [6,8,24]. When this approach is possible, this gives very precise information on the behavior of the players. Unfortunately, this is hardly possible in high-dimension games or/and when the victory domain is not smooth enough. A second approach consists of proving that the boundary of the victory domain satisfies (5) in a suitable generalized sense [11,28] and to use this information to approach numerically the victory domain (cf. [13] for a detailed exposition of this method).

Quantitative Target Games

Here the goal of the player is of quantitative nature: Victor wants to minimize the time to reach the target C while Ursula wants to maximize it. For describing the game, we associate to any trajectory of the dynamics $t \in [0, +\infty) \mapsto x(t)$ the first time $x(\cdot)$ reaches C :

$$\vartheta_C(x(\cdot)) = \inf\{t \geq 0 \mid x(t) \in C\},$$

with, by convention, $\vartheta_C(x(\cdot)) = +\infty$ if $x(\cdot)$ does not reach C .

The modelization problem of such a quantitative game is to express the fact that both players act *simultaneously* on the system. So roughly speaking, we must check that if Ursula chooses her strategy before the second player, the result is the same as that in the case where Victor chooses his strategy first. This means that the following *value functions* of the game do coincide.

$$\begin{aligned} \vartheta^b(x_0) &= \inf_{\beta} \sup_{u \in \mathcal{U}} \vartheta_C(x[x_0, u, \beta(u)]) \quad (\text{lower value}), \\ \vartheta^\#(x_0) &= (\sup_{\alpha} \inf_{v \in \mathcal{V}} \vartheta_C(x[x_0, \alpha(v), v])) \quad (\text{upper value}). \end{aligned}$$

When $\vartheta^b = \vartheta^\#$, one says that the game has a value. Because the question of existence of value in a differential game is an essential question in game theory, Sect. “[Existence of a Value for Zero Sum Differential Games](#)” is devoted to the exposition of this feature.

It is worth pointing out, that in an opposite sense to many static games, this quantitative differential game is not in a *normal form*, namely a player does not play a strategy against a strategy of his opponents. This is another main difference with static games. In fact, it is sometimes possible to write the game in normal form allowing the nonanticipative strategies to have a small delay [15]. The normal form will be used in Sect. “[Nonantagonist Differential Games](#)” for nonzero sum games.

The *Dynamic programming principle* says, roughly speaking, that the game maintains the same structure if the time changes. Indeed, if starting from time 0 the players play the games until time t , then at time t both players are facing a differential game of the same nature as the initial game. This can be expressed by the dynamic programming equation

$$\begin{cases} \vartheta^b(x_0) = \inf_{\beta} \sup_{u \in \mathcal{U}} \{t + \vartheta^b(x[x_0, u, \beta(u)](t))\} \\ \vartheta^\#(x_0) = \sup_{\alpha} \inf_{v \in \mathcal{V}} \{t + \vartheta^\#(x[x_0, \alpha(v), v](t))\} \end{cases} \quad (6)$$

which is available as soon as the trajectories do not reach the target.

Existence of a Value for Zero Sum Differential Games

Using the dynamic programming principle, it is possible to deduce an infinitesimal characterization of the value functions (formally we subtract the two sides of one line of (6), divide by t and we let t tend to 0^+). Under the Isaacs condition (4), when the value functions are differentiable, it is not difficult to show that the values $\vartheta^\#$ and ϑ^b are solutions to a partial differential equation: the following is the *Hamilton–Jacobi–Isaacs equation*

$$H(x, D\vartheta(x)) = 1 \text{ in } \mathbb{R}^N \setminus \mathcal{C}. \quad (7)$$

This fact was noticed at the very beginning of the history of differential games. Furthermore, if the Hamilton–Jacobi equation has a continuously differentiable solution, then this solution is the value function and the game can be solved using feedback strategies (for instance the Lion and Man game). This is the famous *Isaacs verification theorem*. This is a completely satisfactory solvation of the game when the value is smooth. Unfortunately, early on it was noticed that the values are not differentiable and so the previous approach is not possible.

Before going further, it is worth pointing out that this is not only a mathematical question of regularity of the value functions. Indeed the smoothness of the value function is closely related to the number of optimal strategies in the game. Most differential games, even very elementary ones have not smooth values. The reader can convince himself by considering the Lion and Man game in a circular arena with furthermore a round pillar in the center of the arena (which of course the players cannot cross).

One of the more substantial advancements in the theory of the differential game in the 1980s was due to the use of viscosity solution theory [3]. In fact, even when the values $\vartheta^\#$ and ϑ^b are not smooth, they are both solutions in a *viscosity sense* to the Hamilton–Jacobi–Isaacs equation (7). Moreover, with very weak assumptions (Lipschitzian continuity of the data) the partial differential Eq. (7) has a unique solution. Consequently the values coincide. This existence of the value result was due to Evans and Souganidis [17] and can be viewed as a generalization of Isaacs verification theorem. We refer the reader to [3] for viscosity solutions. In several cases, for instance in the state constraint games case, the value is neither smooth nor Lipschitz (even not continuous), so it is necessary to extend the Evans and Souganidis scheme. The main idea is to reduce the quantitative game to a qualitative game in a higher dimensional space [7,14]; this is the object of the *Viability theory* approach to differential games. We refer the reader to [23], pp. 3–37, for a survey of this method. Another advantage of this *reduction to qualitative* method lies in the fact that numerical analysis tools of the qualitative approach [13] can be used to build algorithms computing the value.

At this point, it is important to note that the scheme for obtaining the existence of the value is valid for a wide class of payoff. This fact can be explained by considering a game on a finite time interval $[0, T]$ such that at any initial condition $x(t_0) = x_0$ and any pair of controls u and v is associated the cost

$$C(t_0, x_0, u, v) := \int_{t_0}^T L(x(s), u(s), v(s)) ds + g(x(T)) \quad (8)$$

where L and g are given and $x(\cdot)$ is the solution to (3) on $[t_0, T]$ with initial condition $x(t_0) = x_0$. The payoff (the cost) is then the sum of an integral cost and of a terminal cost. Victor wants to minimize this cost while Ursula wants to maximize it. This leads to the following values

$$\begin{aligned} V^b(t_0, x_0) &:= \inf_{\beta} \sup_{u \in \mathcal{U}} C(t_0, x_0, u, \beta(u)), \\ V^\#(t_0, x_0) &:= \sup_{\alpha} \inf_{v \in \mathcal{V}} C(t_0, x_0, \alpha(v), v) \end{aligned}$$

which are both the solution to the following Hamilton–Jacobi–Isaacs equation

$$\begin{cases} \frac{\partial V}{\partial t}(t, x) + H(t, x, \frac{\partial V}{\partial x}(t, x)) = 0 \\ V(T, x) = g(x) \end{cases} \quad \begin{matrix} \text{for all } (t, x) \in [0, T] \times \mathbb{R}^N \\ \text{for all } x \in \mathbb{R}^N. \end{matrix} \quad (9)$$

Here the Isaacs condition takes the following form

$$\begin{aligned} \min_{v \in V} \max_{u \in U} \{L(x, u, v) + \langle f(x, u, v), p \rangle\} \\ = \max_{u \in U} \min_{v \in V} \{L(x, u, v) + \langle f(x, u, v), p \rangle\} \end{aligned}$$

and the Hamiltonian $H(x, p)$ is equal to the above expression (we refer to [7,27] in the discontinuous or constrained case).

Nonantagonist Differential Games

This section is devoted to a two player game on $[0, T]$ with different objectives: Ursula wants to maximize a cost

$$C_1(t_0, x_0, u, v) := \int_{t_0}^T L_1(x(s), u(s), v(s))ds + g_1(x(T)),$$

while Victor wants to maximize a payoff

$$C_2(t_0, x_0, u, v) := \int_{t_0}^T L_2(x(s), u(s), v(s))ds + g_2(x(T)).$$

Observe that if $C_2 + C_1 = 0$ (i. e. $L_1 + L_2 = 0$ and $g_1 + g_2 = 0$), the game reduces to the game studied in Sect. “[Existence of a Value for Zero Sum Differential Games](#)”. Here we consider the general case where $C_1 + C_2$ is non-necessarily equal to 0. In a nonantagonist game, it is important to play a strategy against a strategy, so we introduced the concept of nonanticipative strategies with delay on the time interval $[0, T]$. A nonanticipative strategy with delay for the first player associates to any action of the second player (to any control v) a control u of the first player such that there exists a delay $r > 0$ such that at any time t if two controls v_1 and v_2 coincide on $[0, t]$ then the associated controls $u_1 = \alpha(v_1)$ and $u_2 = \alpha(v_2)$ do coincide on $[0, t + r]$. The nonanticipative strategy with delay β for the second player is defined in a similar way. Clearly any nonanticipative strategy with delay is nonanticipative. Moreover, it is possible to prove that for an initial condition (t_0, x_0) for any pair (α, β) of nonanticipative strategies with delays there exists a *unique pair* of control (u, v) satisfying

$$\alpha(v) = u, \quad \beta(u) = v.$$

Hence it is possible to associate a trajectory to (α, β) which is the trajectory associated to (u, v) and we denote

$C_1(t_0, x_0, \alpha, \beta)$ and $C_1(t_0, x_0, \alpha, \beta)$ the associated costs. This enables us to write the game in a normal form.

The antagonist differential game problem consists of finding *Nash Equilibria* defined as follows. Fix (t_0, x_0) an initial condition, a pair of real numbers (e_1, e_2) is a Nash equilibria payoff if and only if there exists a pair of nonanticipative strategies with delays $(\tilde{\alpha}, \tilde{\beta})$ such that

$$e_1 = C_1(t_0, x_0, \tilde{\alpha}, \tilde{\beta}), \quad e_2 = C_2(t_0, x_0, \tilde{\alpha}, \tilde{\beta})$$

and such that for any other pair of nonanticipative strategies with delay (α, β) the following inequalities hold true

$$C_1(t_0, x_0, \tilde{\alpha}, \tilde{\beta}) \geq C_1(t_0, x_0, \alpha, \beta),$$

$$C_2(t_0, x_0, \tilde{\alpha}, \tilde{\beta}) \geq C_2(t_0, x_0, \alpha, \beta).$$

In a completely rigorous mathematical sense the above definition must be understood with a small $\varepsilon > 0$ error (the correct statement is: For all $\varepsilon > 0$, there exists $(\tilde{\alpha}, \tilde{\beta})$ such that

$$|e_i - C_i(t_0, x_0, \tilde{\alpha}, \tilde{\beta})| \leq \varepsilon, \quad i = 1, 2$$

and for any other pair of strategies (α, β) we have

$$C_1(t_0, x_0, \tilde{\alpha}, \tilde{\beta}) \geq C_1(t_0, x_0, \alpha, \beta) - \varepsilon,$$

$$C_2(t_0, x_0, \tilde{\alpha}, \tilde{\beta}) \geq C_2(t_0, x_0, \alpha, \beta) - \varepsilon$$

cf. [12]). Nash equilibria were studied in detail in [20] for another concept of strategy. Now the remaining part of this section concerns first the characterization of the Nash equilibria through zero-sum games and second the existence of Nash equilibria. For doing this, we introduce the following value function associated with two auxiliary zero-sum games where one player is the minimizer while its opponent is the maximizer.

$$V_1(t_0, x_0) = \inf_{\beta} \sup_{u \in \mathcal{U}(t_0)} C_1(t_0, x_0, u, \beta(u)),$$

$$V_2(t_0, x_0) = \inf_{\alpha} \sup_{v \in \mathcal{V}(t_0)} C_2(t_0, x_0, \alpha(v), v).$$

Recall that under the Isaacs condition, the value of the above auxiliary games does exist. Then the characterization of Nash payoff can be expressed: the pair (e_1, e_2) is a Nash equilibrium payoff if and only if there exist two controls (u, v) such that

$$e_1 = C(t_0, x_0, u, v), \quad e_2 = C(t_0, x_0, u, v)$$

and for all time $t \in [t_0, T]$

$$e_1 \geq V_1(t, x[t_0, x_0, u, v](t)),$$

$$e_2 \geq V_2(t, x[t_0, x_0, u, v](t)).$$

Observe that this is not an intuitive result (once again a correct statement is, a $\varepsilon > 0$ is needed to have a rigorous

statement: For any ϵ there exists a pair (u, v) satisfying the above inequality up to an error ϵ . From this result, it is possible – but not easy – to obtain the existence of a Nash equilibrium pair under Isaacs condition.

Observe that at a first glance, it is very surprising to obtain the existence of the Nash equilibria assuming Isaacs condition which means the existence of a saddle point in static games. One could assume instead the existence of Nash equilibria of static games. Unfortunately, except for very specific games like some linear quadratic differential games, this method fails. Another way to understand this lies in the fact that Nash equilibria are very unstable (cf. pp. 57–67 in [23]). We refer the reader to the bibliography for noncooperative games in specific cases and applications.

Miscellaneous

This part is devoted to the description of Stochastic differential games (games with randomness in the dynamics) and Worst Case Design where games are used as a tool to prevent the worst case caused by nature or uncertainty.

Stochastic Differential Games

The dynamics is given by a stochastic differential equation

$$dX_t = f(X_t, u_t, v_t)dt + \sigma(X_t, u_t, v_t)dW_t, \quad (10)$$

where W is a Brownian motion on a given probability space. Victor and Ursula choose the controls u and v which are adapted processes (with respect to the filtration generated by the Brownian motion). The notion of nonanticipative strategies is similar to the determinist case but the strategies are nonanticipative in time *almost surely*. The rigorous description of the game is rather technical due to the need of stochastic analysis. The reader can refer to the nice article [30] for a precise formulation. We want just to stress some important aspects of stochastic zero-sum games. Of course the notions of qualitative and quantitative games are relevant. The qualitative aspect is not yet well-studied. Oppositely the quantitative aspect is now well studied. Consider for instance a payoff of the form

$$C(t_0, x_0, u, v) \\ := E \left[\int_{t_0}^T L(x(s), u(s), v(s))ds + g(x(T)) \right],$$

where E is the expectation. The existence of the value was obtained in [18], by adopting the same scheme as described in Subsect. “Quantitative Target Games”: The values are the unique viscosity solution of a – second order – Hamilton–Jacobi–Isaacs equation which has a unique so-

lution. The nonantagonist case could be solved similarly to the zero-sum case up to hard technicalities due to stochastic analysis [9].

Worst Case Design

This concerns a zero-sum quantitative game where a controller wants to drive a system against another action on the system (uncertainty, nature) considered as a second player. The controller wants to find the most *robust* strategy which prevents the worst case of nature. Clearly the question of existence of a value is not relevant in this case. Only one value is interesting in this case.

The linear quadratic case was extensively studied in the book [4]. In the fully nonlinear case, the Hamilton–Jacobi–Isaacs equation becomes infinite dimensional. An alternative approach can be made by setting Hamilton–Jacobi equations on the space of closed sets [29].

Future Directions

Among several future directions of investigations, let us mention first the case of impulsive differential games where the player can at any time either follow a continuous dynamic of the form (3) or make a jump. These games combine the effects and difficulties of static and differential games. Only very preliminary results are available (pp. 223–249 in [23]) for pursuit games when both players cannot “jump” simultaneously. Another interesting class of pursuit impulsive game, is the case where each player can choose at every time point between two dynamics. Once again, this research domain is still wide open.

Another important and very active field of research is devoted to information phenomena. This is well understood in the case of (discrete) repeated games (cf. [2]) with the help of mixed strategies but still open for differential games. In more concrete terms: Consider the two-player quantitative game of Sect. “Qualitative and Quantitative Differential Games”, with dynamics (3) and cost (8). We consider that the initial position belongs to a set of $I \times J$ possible initial positions

$$x_0^{i,j}, \quad i = 1, \dots, I, \quad j = 1 \dots J.$$

At the initial time the game is played in two steps, first the integer i is chosen according to a probability p (belonging to the set $\Delta(I)$ of probabilities on $\{1, 2, \dots, I\}$) while j is chosen according to a probability $q \in \Delta(J)$; second the index i is communicated to Ursula only and the index j is communicated to the player j . As previously stated Victor wants to minimize the cost C and Ursula want to maximize it. Each player has a *private information* (the knowledge

of i respectively j) which cannot allow him/her to compute the current position of the game. Each player can only observe his player behavior (choice of controls) trying to deduce from this observation his missing information. The problem consists of checking the existence of a value for suitably defined mixed strategies.

Very recently, this specific game was solved, but the more general question of taking into account the information phenomena in differential games has been little studied and a wide field of research is just starting to rise in this direction.

Bibliography

Primary Literature

- Aubin JP (1991) Viability Theory. Birkhäuser, Basel
- Aumann RJ, Maschler M (1995) Repeated games with incomplete information. MIT Press, Cambridge
- Bardi M, Capuzzo Dolcetta I (1997) Optimal control and viscosity solutions of Hamilton–Jacobi–Bellman equations. Systems and Control: Foundations and Applications, vol xvii. Birkhäuser, Boston, pp 570
- Basar T, Bernhard P (1995) H^∞ -optimal control and related minimax design problems. A dynamic game approach, 2nd edn. Systems and Control: Foundations and Applications. Birkhäuser, Basel
- Berkovitz LD (1994) A theory of differential games. In: Basar T et al (eds) Advances in dynamic games and applications, Birkhäuser, Basel; Ann Int Soc Dyn Games 1:3–22
- Bernhard P (1988) Differential games in Systems and control Encyclopedia. In: Singh MG (ed) Theory Technology Application. Pergamon Press, Oxford
- Bettiol P, Cardaliaguet P, Quincampoix M (2006) Zero-sum state constrained differential games: Existence of value for Bolza problem. Int J Game Theor 34(4):495–527
- Breakwell JV (1977) Zero-sum differential games with terminal payoff. In: Hagedorn P, Knobloch HW, Olsder GH (eds) Differential Game and Applications. Lecture Notes in Control and Information Sciences, vol 3. Springer, Berlin
- Buckdahn R, Cardaliaguet P, Rainer C (2004) Nash equilibrium payoffs for nonzero-sum stochastic differential games. SIAM J Control Optim 43(2):624–642
- Cardaliaguet P (1996) A differential game with two players and one target. SIAM J Control Optim 34(4):1441–1460
- Cardaliaguet P (1997) Nonsmooth semi-permeable barriers, Isaacs equation, and application to a differential game with one target and two players. Appl Math Optim 36:125–146
- Cardaliaguet P, Plaskacz S (2003) Existence and uniqueness of a Nash equilibrium feedback for a simple non zero-sum differential game. Int J Game Theor 32(4):561–593
- Cardaliaguet P, Quincampoix M, Saint-Pierre P (1999) Numerical methods for optimal control and numerical games. In: Bardi M, Parthasarathy T, Raghavan TES (eds) Annals of International Society of Dynamical Games. Birkhäuser, Basel, pp 177–249
- Cardaliaguet P, Quincampoix M, Saint-Pierre P (2001) Pursuit differential games with state constraints. SIAM J Control Optim 39(5):1615–1632
- Cardaliaguet P, Quincampoix M (2008) Determinist Differential games under probability knowledge of initial condition. Int Game Theor Rev 6(1):1–16
- Elliot N, Kalton N (1972) The existence of value in differential games. Mem Am Math Soc 126:67
- Evans LC, Souganidis PE (1984) Differential games and representation formulas for solutions of Hamilton–Jacobi equations. Indiana Univ Math J 282:487–502
- Fleming W, Souganidis P (1989) On the existence of value functions of two-player, zero-sum stochastic differential games. Indiana Univ Math J 38(2):293–314
- Flynn J (1973) Lion and Man: The Boundary Constraints. SIAM J Control 11:397
- Kleimenov AF (1993) Nonantagonistic positional differential games. Nauka, Ekaterinburg (in Russian)
- Krasovskii NN, Subbotin AI (1988) Game-Theoretical Control Problems. Springer, New York
- Isaacs R (1965) Differential Games. Wiley, New York
- Jorgensen S, Quincampoix M, Vincent T (2007) Advances in Dynamic Games Theory. Annals of International Society of Dynamical Games. Birkhäuser, Basel
- Melikyan AA (1998) Generalized characteristics of first order PDEs. Applications in optimal control and differential games. Birkhäuser, Boston
- Petrosjan LA (2004) Cooperation in games with incomplete information. Nonlinear analysis and convex analysis. Yokohama Publ, pp 469–479
- Pontryagin N (1968) Linear Differential Games, I, II. Sov Math Doklady 8(3–4):769–771; 910–913
- Plaskacz S, Quincampoix M (2000) Discontinuous Mayer control problem under state-constraints. Topol Method Nonlinear Analysis 15:91–100
- Quincampoix M (1992) Differential inclusions and target problems. SIAM J Control Optim 30(2):324–335
- Quincampoix M, Veliov V (2005) Optimal control of uncertain systems with incomplete information for the disturbance. SIAM J Control Optim 43(4):1373–1399
- Rainer C (2007) On two different approaches to nonzero-sum stochastic differential games. Appl Math Optim 56:131–144
- Roxin E (1969) The axiomatic approach in differential games. J Optim Theor Appl 3:153–163
- Subbotin AI (1995) Generalized solutions of first-order PDEs, The dynamical optimization perspective. Translated from the Russian. Systems and Control: Foundations and Applications. Birkhäuser, Boston
- Varaiya P (1967) The existence of solution to a differential game. SIAM J Control Optim 5:153–162
- Varaiya P, Lin J (1967) Existence of saddle points in differential game. SIAM J Control Optim 7(1):141–157
- Von Neumann J, Morgenstern O (1946) Theory of Games and Economic Behaviour. Princeton University Press, Princeton

Books and Reviews

- Bardi M, Raghavan TES, Parthasarath T (1999) Stochastic and differential games. Theory and numerical methods, Dedicated to Prof. Subbotin AI. Annals of the International Society of Dynamical Games, vol 4. Birkhäuser, Boston
- Basar T, Olsder GJ (1999) Dynamic noncooperative game theory, 2nd edn. Classics in Applied Mathematics, vol 23. SIAM, Society for Industrial and Applied Mathematics, Philadelphia

- Blaquière A, Gérard F, Leitman G (1969) Quantitative and Qualitative Games. Academic Press, New York
- Dockner E, Jrgensen S, Van Long N, Sorger G (2000) Differential games in economics and management science. Cambridge University Press, New York
- Hajek O (1975) Pursuit games. Academic Press, New York
- Patsko VS, Turova VL (2000) Numerical study of differential games with the homicidal chauffeur dynamics. Russian Academy of Sciences, Institute of Mathematics and Mechanics, Ekaterinburg
- Petrosyan LA (1993) Differential games of pursuit. Series on Optimization, vol 2. World Scientific, Singapore

Diffusion of Innovations, System Dynamics Analysis of the

PETER M. MILLING¹, FRANK H. MAIER²

¹ Industrieseminar der Universität Mannheim, Mannheim University, Mannheim, Germany

² International University in Germany, Bruchsal, Germany

Article Outline

Glossary

Definition of the Subject

Introduction

Principle Structures to Model the Diffusion of Innovations

Representing Managerial Decision Making in Innovation Diffusion Models

Managerial Implications

Future Directions

Bibliography

Glossary

Adopters The cumulated number of persons who have bought a product over time.

Diffusion The spread of a new product, process or concept in the market. The process of bringing innovation into wide use.

Invention The process of bringing new technology into being.

Innovator A customer with general interest in innovations making his buying decision independent of others.

Innovation The process of bringing new technology into use.

Installed base Installed base is defined as the amount of users in a network system.

Imitator An imitator buy a new product because he observed or communicated with customers who have already bought the product. The buying decision of imitators is influenced by the adoption of other customers.

Network effects A product is characterized by a network effect, if the utility of that product is a function of the installed base. The utility increases with the installed base.

Definition of the Subject

The article describes how system dynamics-based models can contribute to the understanding and improved management of the diffusion of innovations. It emphasizes the importance of an integrated feedback-oriented view of the different stages of innovation processes. The aim is to generate insight in the complexity and the dynamics of innovation processes. Based on the classical Bass model of innovation diffusion, the system dynamics perspective is introduced. In a systematic approach several structures to model the complexity and dynamics of managerial decision-making in the context of the diffusion of innovation are described and analyzed. Aspects covered consider market structure, network externalities, dynamic pricing, manufacturing related decisions and the link between research and development and the diffusion of a new product in the market place. The article concludes with managerial implications.

Introduction

Continuous activities to renew a company's range of products are crucial for the survival in a competitive environment. However, to improve the competitive position or the competitive advantage, ongoing innovation activity through the development, test, and introduction of new products is necessary. At least since the 1970s, it could be observed that new and technically more complex and sophisticated products have to be developed in a shorter time span. Resources have to be allocated to research and development (R&D) projects that are expected to be economically successful. New products have to be introduced to global markets with severe competition. Decisions about the adequate time to market and appropriate pricing, advertising, and quality strategies have to be made.

The complexity and difficulties to manage innovation activities partly derive from the comprehensiveness of the innovation processes. To be competitive, companies have to be successful in all stages of the innovation process, i. e., the process of invention, innovation, and diffusion. This becomes obvious when new product failure rates and innovation costs are analyzed. Figure 1 illustrates the cas-

cading process of innovation activity and the related innovation costs.

For one successful new product in the market place, 64 promising ideas must be channeled through the process of invention and innovation. The cost at each stage of the invention and innovation process increases from a \$1000 to \$5 million per attempt. Not only is failure more expensive in later stages – which requires an effective management to reduce the failure rates – successful new products have to earn all necessary resources for the whole process. This requires the following: (1) to manage R&D projects and processes effectively and efficiently – including thorough and educated assessment of the economic potential of a new product – to reduce failure rates in later stages and (2) to increase management attention in the final stages since failures in late stages of the process are much more expensive.

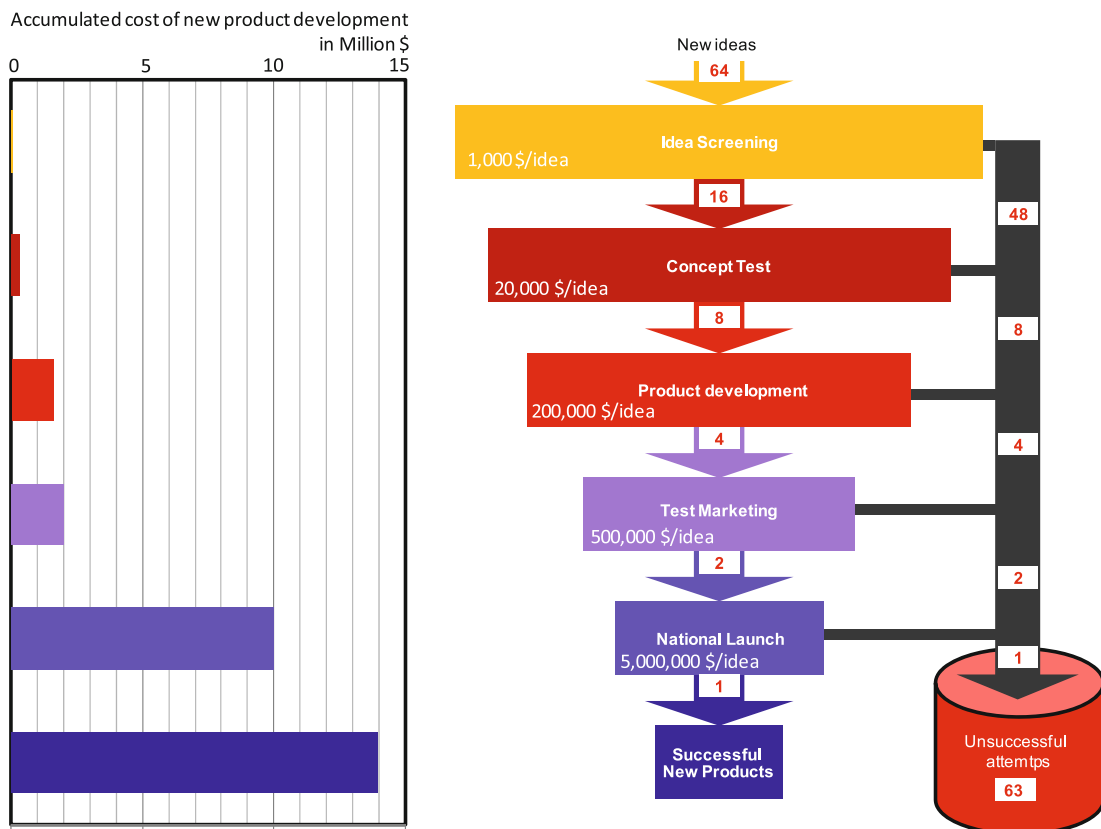
Models of innovation diffusion can support the complex and highly dynamic tasks. The article will briefly examine how system dynamics-based analysis of innovation diffusion can contribute to the understanding of the struc-

tures and forces driving the processes of innovation and diffusion. It will show how system dynamics models can support the decision-making and how they can help to reduce failures in the later stages of innovation activities.

Principle Structures to Model the Diffusion of Innovations

Traditional Innovation Diffusion Models from a System Dynamics Perspective

In literature discusses plenty of models about the diffusion of innovations. Many models are based on Frank M. Bass' model of innovation diffusion. In this model, product purchases result from two distinct forms of buying behavior, i. e., innovative purchases and imitative purchases. According to the original Bass model, innovative purchases of a period can be calculated as a fraction of the remaining market potential ($N - X_{t-1}$) with N being the market potential and $X_{t-1} = \sum_{\tau=0}^{t-1} S_{\tau}$ representing the accumulation of all past purchases of the product S_{τ} until period $t - 1$.



Diffusion of Innovations, System Dynamics Analysis of the, Figure 1
Outcome of activities along the process of invention and innovation

According to this, innovative purchases S_t^{inno} can be calculated as

$$S_t^{\text{inno}} = \alpha \cdot \left(N - \sum_{\tau=0}^{t-1} S_{\tau} \right) \quad (1)$$

where α represents the coefficient of innovation. In the original model, this coefficient is a constant essentially representing the fraction of innovators of the remaining market potential at any point of time. Imitative purchases, however, are influenced by the number of purchases in the past. Potential adopters of an innovation make their purchasing decision depending on the spread of the product in the market place. The more customers have adopted the product in the past, the higher is the social pressure to purchase the product as well. Imitative demand of a period S_t^{imit} hence can be calculated as

$$S_t^{\text{imit}} = \beta \cdot \frac{\sum_{\tau=0}^{t-1} S_{\tau}}{N} \cdot \left(N - \sum_{\tau=0}^{t-1} S_{\tau} \right) \quad (2)$$

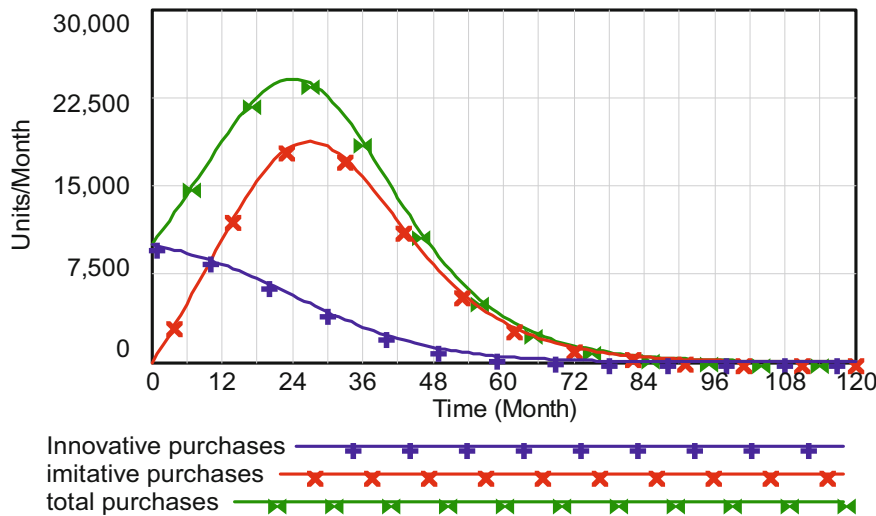
with β representing the coefficient of imitation – a probability that a purchase takes place by someone who observed the use of a product. Together, the total purchases in a period S_t^{total} equal $S_t^{\text{inno}} + S_t^{\text{imit}}$ and hence are calculated as

$$S_t^{\text{total}} = S_t^{\text{inno}} + S_t^{\text{imit}} = \alpha \cdot \left(N - \sum_{\tau=0}^{t-1} S_{\tau} \right) + \beta \cdot \frac{\sum_{\tau=0}^{t-1} S_{\tau}}{N} \cdot \left(N - \sum_{\tau=0}^{t-1} S_{\tau} \right) \quad (3)$$

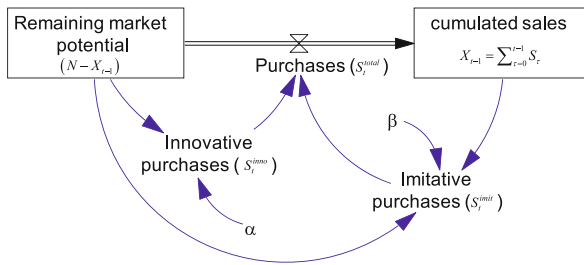
Innovative and imitative purchases together create the typical product life cycle behavior of the diffusion of an innovation in the market place as shown in Fig. 2.

The model above is a simple mathematical representation of the product life cycle concept, a key framework in business management. It describes the time pattern a product follows through subsequent stages of introduction, growth, maturity, and decline. Because of its mathematical simplicity and its ability to represent the diffusion of an innovation, the Bass model has been used for parameter estimation and therefore serves as a base for projections of future sales. Although the concept is a powerful heuristic, many models generating this typical behavior do not consider e. g., actual economic environment, competition, capital investment, cost and price effects. Innovation diffusion models, which do not comprise the relevant decision variables, exhibit a significant lack of policy content. They do not explain how structure conditions behavior. They cannot indicate how actions of a firm can promote but also impede innovation diffusion. For an improved understanding of innovation dynamics generated by feedback structures that include managerial decision variables or economic conditions, the system dynamics approach is highly suitable.

Equations (1) to (3) can easily be transformed into the system dynamics terminology. $(N - X_{t-1})$ represents the stock of the remaining market potential at any point in time and X_{t-1} represents the accumulation of all product purchases over time. The sales of a period S_t^{total} are the flows connecting these two stocks as shown in the Fig. 3.



Diffusion of Innovations, System Dynamics Analysis of the, Figure 2
Product life cycle behavior generated by the Bass model

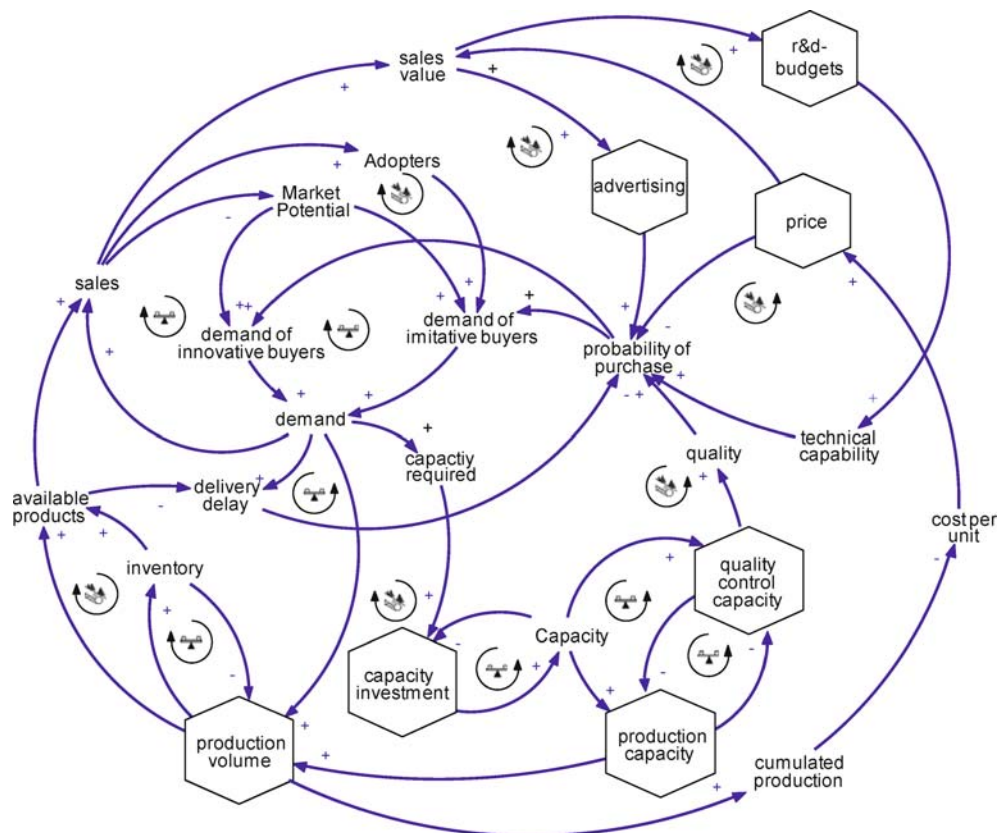


Diffusion of Innovations, System Dynamics Analysis of the, Figure 3
Stock-flow view of the Bass model

The coefficients α and β represent the probability of a purchase taking place; they are constants in the original Bass model and independent of any decisions or changes over time. For this reason, the model has been criticized and subsequent models have been developed that make the coefficients depending on variables like price or advertising budget. Most of the extensions, however, include no feedback between the diffusion process and these decision

variables. This is a severe shortcoming since in the market place, diffusion processes are strongly influenced by feedback. What in classical innovation diffusions models typically is referred to as word-of-mouth processes is nothing else than a reinforcing feedback process. Adopters of an innovation – represented by the cumulated sales X_{t-1} – communicate with potential customers ($N - X_{t-1}$) and – by providing information about the product – influence their behavior. However, feedback in innovation diffusion goes beyond the pure word-of-mouth processes. It also involves the decision processes of a company and the outcome generated by the purchasing decision of the customers like the sales volume generated.

Figure 4 describes as a causal loop diagram the diversity of potential influences of corporate decision variables (marked with hexagons) on demand of the products by making the probability of a purchase – the coefficients α and β – depending on decision variables. It also shows how corporate decisions are interconnected through several feedback structures and influence the diffusion of a new product in the market place. Although be-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 4
Feedback structures driving innovation processes

ing far from a comprehensive structure of potential feedback, the figure gives an impression of the complex dynamic nature of innovation diffusion processes.

Decision variables like pricing or advertising directly influence the purchase probability of a potential customer. The higher the advertising budgets and the lower the price, the higher will be demand for the products of a company. Furthermore, there are indirect and/or delayed effects on the speed of the spread of a new product in the market. Actual sales of a product may be limited by insufficient production and inventory levels which increases delivery delays (perceived or actual) and therefore reduce demand. Growing demand, however, motivates the company to expand its capacity and to increase the volume of production. This leads to higher cumulated production and through experience curve effects to decreasing costs per unit, lower prices, and further increased demand. Other influences might reflect that a certain percentage of total available production capacity has to be allocated to ensure the quality of the output – either by final inspection or during the production process. Quality control then will improve product quality, which directly affects demand.

Models developed in this manner can serve as simulators to analyze the consequences of strategies and to improve understanding. They can show e. g., how pricing and investment strategies depend on each other and quantify the impact of intensified quality control on production and sales. They are suitable tools to investigate the effects resulting from the impact of a particular management problem on the dynamic complexity of innovation diffusion. Creating an understanding of the processes and interactions is the main purpose of system dynamics-based innovation diffusion models. Subsequently, a base structure of a system dynamics-based model will be described.

Base Structure of a System Dynamics-Based Model of Innovation Diffusion

First, a model will be discussed that maps the diffusion of an innovation in a monopolistic situation or can serve as an industry level model. Secondly, competition between potential and existing companies is introduced. Thirdly, substitution between successive product generations is considered. Each step adds complexity to the model. This approach allows for a better understanding of the forces driving the spread of a new product in the market.

In the following, the coarse structure of a model generating the life cycle in the market of a new product is presented and analyzed in its dynamic implications in Sect. “[Representing Managerial Decision Making in Innovation Diffusion Models](#)”. Figure 5 gives an aggregated

view the main model structure. It also introduces – in contrast to the mathematical terms known from the Bass model, variable names, which are informative and consistent with the use in system dynamics models.

The diffusion of a new product is generated by the behavior of the before mentioned two different types of buyers: innovators and imitators. If the potential customers (PC) – i. e., the remaining market potential of a product – decide to purchase, either as innovators or as imitators, they become adopters (*ADOP*). The variables *PC* and *ADOP* and their associated transfer rates are the basic variables of the core diffusion process. The untapped market (*UM*) covers latent demand that can be activated by appropriate actions and leads to an increase in the number of potential customers and therefore increases the remaining market potential. Besides the growth resulting from the influx from the untapped market, a decline in market volume can be caused by the loss of potential customers to competitors. This lost demand (*LD*) turned to competing products that are more attractive, e. g., products of a higher level of technological sophistication, quality or lower price.

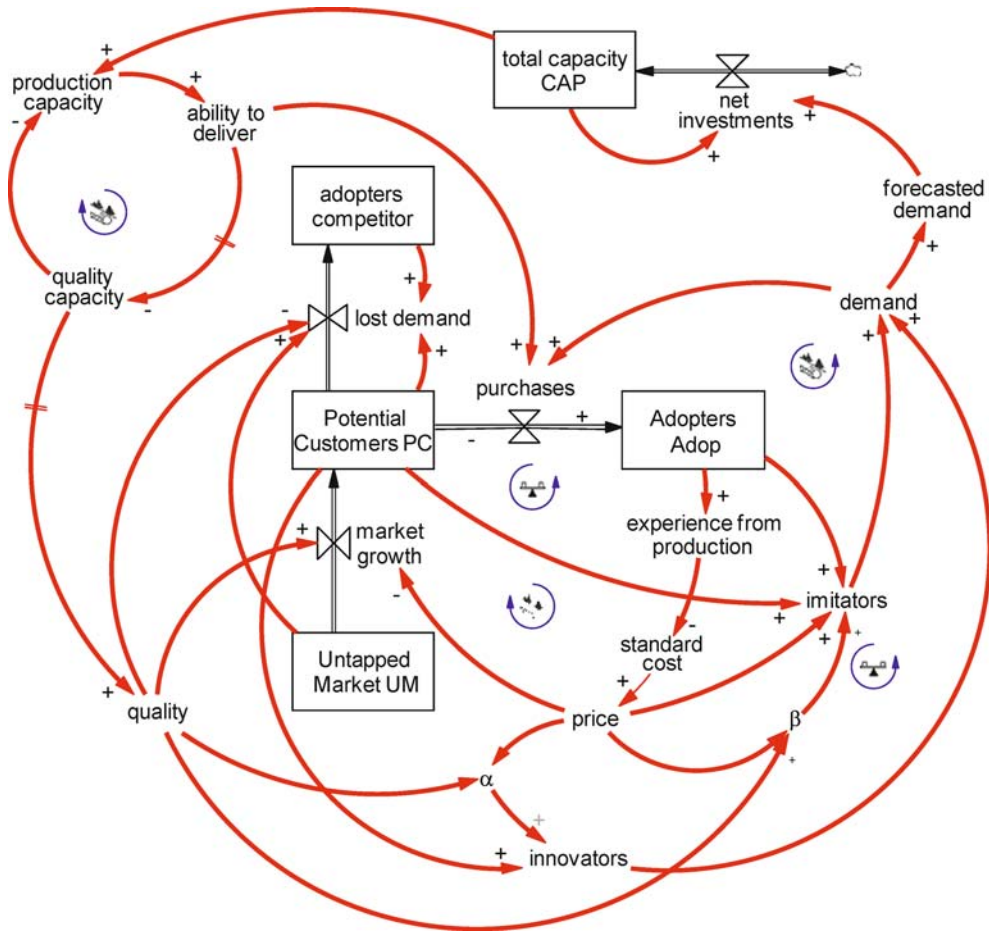
The differentiation into the two buying categories “innovators” and “imitators” refers to the Bass model of innovation diffusion as described in Subsect. “[Traditional Innovation Diffusion Models from a system dynamics Perspective](#)”. The distinction is made because these two types of buyers react differently to prices charged, product quality offered, advertisements or the market penetration already achieved. The term “innovator” refers to customers who make their purchasing decision without being influenced by buyers who already purchased the product, the adopters. In the beginning of an innovation diffusion process, innovators take up the new product because they are interested in innovations. The number of innovators is a function of the potential customers. Mathematically, the purchasing decision of innovators D^{Inno} is defined by a coefficient of innovation α times the number of potential customers *PC*.

$$D_{(t)}^{\text{Inno}} = \alpha_{(t)} \cdot PC_{(t)} \quad (4)$$

with:

$D_{(t)}^{\text{Inno}}$	Demand from innovators
$\alpha_{(t)}$	Coefficient of innovation
$PC_{(t)}$	Potential customers .

The purchasing decision of “imitators” is calculated differently. Imitators buy a new product because they observe or communicate with customers who have already adopted the product. They imitate the observed buying



Diffusion of Innovations, System Dynamics Analysis of the, Figure 5
Coarse structure of the innovation diffusion model

behavior. Innovators initiate new product growth, but the diffusion gains momentum from the word-of-mouth process between potential customers and the increasing level of adopters. The driving force behind the imitation process is communication – either personal communication between an adopter and someone who still does not own the product or just observation of someone already owning and using the product. Although, the Bass model describes how the imitators' purchases can be calculated – as shown in Eq. (2) – the equation can also be derived from a combinatorial analysis of the number of possible contacts between the adopters and the potential customers. If N is the total number of people in a population consisting of potential customers PC and adopters $ADOP$, the amount of possible combinations C_N^k is

$$C_N^k = \binom{N}{k} = \frac{N!}{k!(N-k)!} \quad (5)$$

Here we are only interested in paired combinations ($k = 2$) between the elements in N

$$\begin{aligned} C_N^2 &= \binom{N}{2} = \frac{N!}{2!(N-2)!} \\ &= \frac{N(N-1)}{2!} = \frac{1}{2}(N^2 - N) \end{aligned} \quad (6)$$

Since N represents the sum of elements in PC and in $ADOP$, ($N = PC + ADOP$), the number of combinations between potential customers and adopters is

$$\begin{aligned} &= \frac{1}{2} [(PC + ADOP)^2 - (PC + ADOP)] \\ &= \frac{1}{2} [PC^2 + 2 \cdot PC \cdot ADOP + ADOP^2 - PC - ADOP] \end{aligned} \quad (7)$$

and after regrouping and collecting terms, we get

$$= \frac{1}{2} \left(\underbrace{2 \cdot PC \cdot ADOP}_{\text{Communication between PC and ADOP}} + \underbrace{PC^2 - PC}_{\text{Communication within PC}} + \underbrace{ADOP^2 - ADOP}_{\text{Communication within ADOP}} \right). \quad (8)$$

Internal communications, both within *PC* and *ADOP*, generate no incentive to purchase the new product and are neglected; the process of creating imitative buying decisions in Eq. (9) is, therefore, reduced to the first term in Eq. (8), the information exchange between potential customers and adopters.

$$D_{(t)}^{\text{imit}} = \beta^* \cdot PC_{(t)} \cdot ADOP_{(t)} \quad (9)$$

with:

$D_{(t)}^{\text{imit}}$	Demand from imitators
$\beta_{(t)}^*$	Coefficient of imitation = $\frac{\beta_{(t)}}{N}$
$ADOP_{(t)}$	Adopters
N	Initial market potential.

The coefficient of imitation $\beta_{(t)}^*$ represents the original coefficient of innovation β from the Bass model divided by the initial market potential N . β can be interpreted as the probability that the possible contacts between members in *PC* and *ADOP* have been established, relevant information has been exchanged, and a purchasing decision is made.

The sum of the demand of innovators and imitators in each period, $D_{(t)}$, establishes the basic equation for the spread of a new product in the market. Together with the state variables of potential customers and adopters the flows of buyers (innovators and imitators) constitute the core model of innovation diffusion, which generates the typical s-shaped pattern of an adoption process over time.

$$D_{(t)} = D_{(t)}^{\text{inno}} + D_{(t)}^{\text{imit}} \quad (10)$$

$$= \alpha_{(t)} \cdot PC_{(t)} + \beta_{(t)}^* \cdot PC_{(t)} \cdot ADOP_{(t)}.$$

Although Eqs. (3) and (10) are based on different interpretations and explanations, they are structurally identical since *PC* equals $(N - X_{t-1})$ and *ADOP* equals $X_{t-1} = \sum_{\tau=0}^{t-1} S_{\tau}$. The only difference is that the coefficients of innovation and imitation, in the context of the model based on (10) are now a variable – rather than a constant – depending on corporate decision variables like price or quality. Furthermore, corporate decisions are not just set as predefined time paths; they are endogenously calculated and depend on the outcome of the diffusions process itself. Model simulations of this extended

innovation diffusion model will be discussed in Sect. “Representing Managerial Decision Making in Innovation Diffusion Models”.

Extending the Base Structure to Include Competition

In the model described above, competition is not modeled explicitly. The model only assumes a potential loss in demand, if price, quality or ability to deliver are not within the customers’ expectations. The internal corporate structures of competition are not explicitly represented. To generate diffusion patterns that are influenced by corporate decisions and the resulting dynamic interactions of the different competitors in a market, a more sophisticated way to incorporate competition is needed. Therefore, a subscript i ($i = 1, 2, \dots, k$) representing a particular company is introduced as a convenient and efficient way to model the different competitors. In a competitive innovation diffusion model the calculation of innovative and imitative demand of a company has to be modified. Equation (4) that determines the innovative demand in a monopolistic situation becomes Eq. (11) – in the following discussion, the time subscript (t) is omitted for simplicity. The coefficient of innovation α has to be divided by the number of competitors N to ensure that each company will have the same share of innovative demand as long as there is no differentiation among the competitors’ products through, e.g., through pricing or advertising. The subscript i in the coefficient of innovation is necessary because it considers that the decisions of an individual company regarding product differentiation influences its proportion of innovative buyers. A third modification is necessary, because the number of competitors may vary over time. Therefore, the term φ_i represents a factor to model different dates of market entry. It takes the value 1 if a company i is present at the market, otherwise it is 0. Hence, the demand of company i is 0, as long as it is not present at the market and $\sum_{i=1}^k \varphi_i$ represents the actual number of competitors. The variable potential customers *PC* has no subscript because all companies in the market compete for a common group of potential customers, whereas innovative demand has to be calculated for each company.

$$D_i^{\text{inno}} = \frac{\alpha_i}{NC} \cdot PC \cdot \varphi_i \quad (11)$$

with:

α_i	coefficient of innovation for company i
NC	number of active competitors = $\sum_{i=1}^k \varphi_i$
φ_i	factor of market presence company i

i subscript representing the companies
 $i = (1, 2, \dots, k)$.

The buying decisions of imitators are influenced by observation of, or communication with the adopters ($ADOP$). In a competitive environment two alternative approaches can be used to calculate imitative demand. These different approaches are a result of different interpretations of the object of the communication processes. In the first interpretation, the ‘product related communication’, the adopters of a particular company’s product communicate information about the product they have purchased e.g., an electronic device like a MP3 player of a particular company. In this case, the calculation of imitative demand has to consider the number of potential contacts between the potential customers PC and the adopters of the products of company i ($ADOP_i$) as shown in Eq. (12).

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot ADOP_i \cdot PC \cdot \varphi_i \quad (12)$$

with:

β_i coefficient of imitation for company i .

The second interpretation about the object of communication is the ‘product form-related communication’. Here, the adopters communicate information about a product form, for example, DVD players in general and not about an MP3 player of a particular company. The equation to calculate imitative demand for the model of product form

related communication is shown in Eq. (13). The sum of adopters for each company i ($\sum_{i=1}^k ADOP_i$) represents the total number of adopters in the market. The product of the total adopters and the potential customers then represents the total number of potential contacts in the market. Imitative demand of a company i depends on the share of total adopters $\frac{ADOP_i}{\sum_{i=1}^k ADOP_i}$ this company holds.

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot \frac{ADOP_i}{\sum_{i=1}^k ADOP} \cdot PC \cdot \sum_{i=1}^k ADOP_i \cdot \phi_i \quad (13)$$

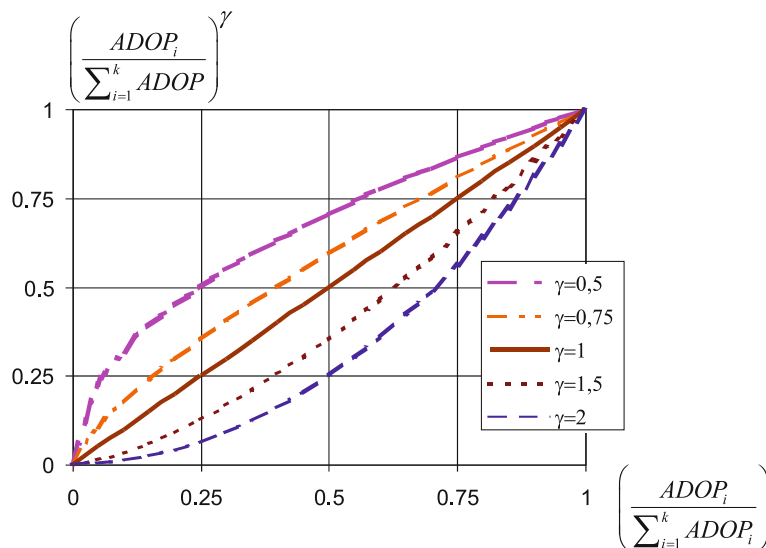
If the term that represents a company’s share of the total adopters of a market $\frac{ADOP_i}{\sum_{i=1}^k ADOP_i}$ is raised to the power of γ as in Eq. (14), weaker ($0 < \gamma < 1$) or stronger ($\gamma > 1$) influences of a company’s share of total adopters on demand can be represented explicitly. For $\gamma = 1$, Eq. (14) is identical to Eq. (13).

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot \left(\frac{ADOP_i}{\sum_{i=1}^k ADOP} \right)^\gamma \cdot PC \cdot \sum_{i=1}^k ADOP_i \cdot \phi_i \quad (14)$$

with:

γ factor representing customers’ resistance to “Me-too”-pressure.

Figure 6 shows the effects of a company’s share of the total adopters for different γ . For a given share of total adopters



Diffusion of Innovations, System Dynamics Analysis of the, Figure 6
 Effects of a company’s share of adopters for different γ

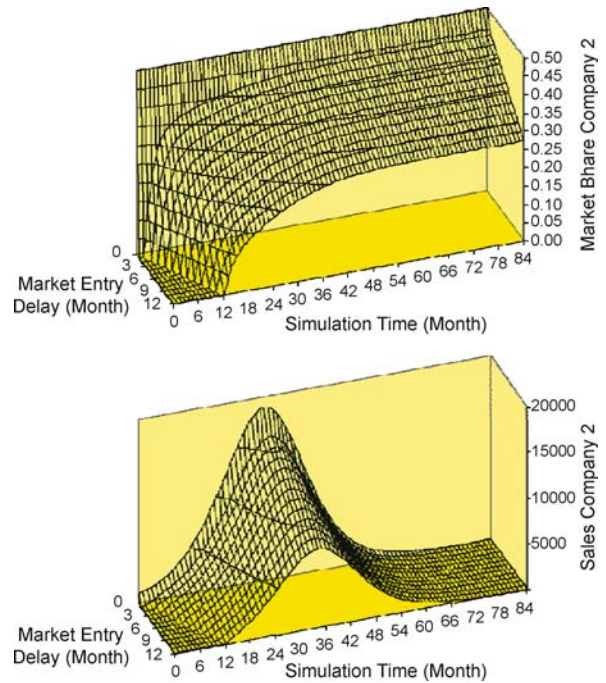
this means: the higher γ , the lower is the value of the term

$$\left(\frac{ADOP_i}{\sum_{i=1}^k ADOP_i} \right)^\gamma$$

and the stronger is the importance of a high share of total adopters. The parameter γ can be interpreted as a measure of the importance of customer loyalty or as resistance to “me-too” pressure.

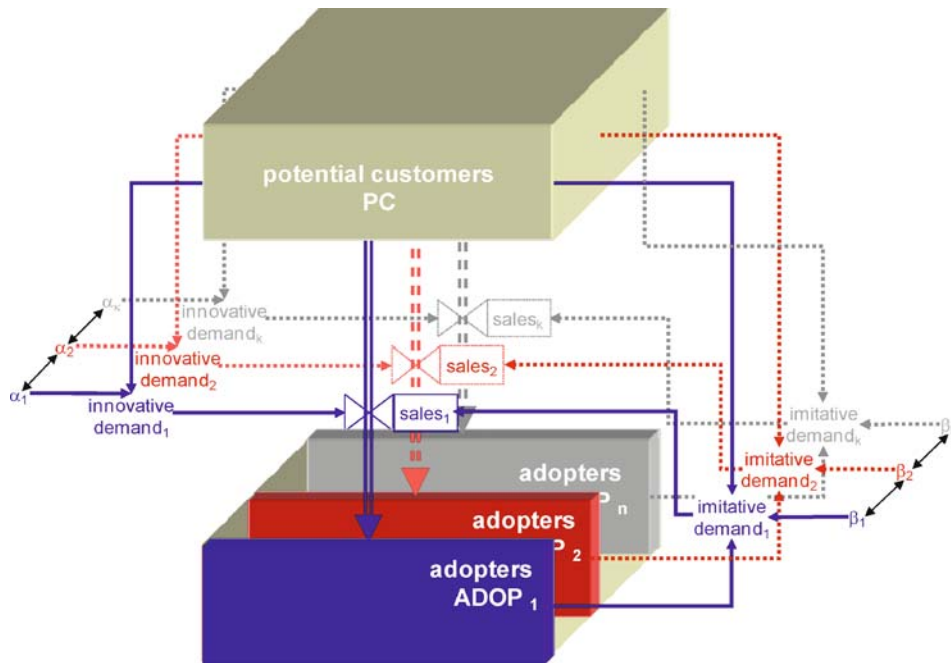
Figure 7 illustrates the coarse structure of an oligopolistic innovation diffusion model as described by Eqs. (11) and (14). The hexahedron at the top represents the stock of potential customers PC for the whole market. The blocks with the different shading represent for each company i the level of adopters, i. e., the cumulated sales of the company. The total number of adopters of the product form corresponds to the addition of these blocks.

Since the sales are calculated separately for each company i there are n outflows from the stock of potential customers to the adopters. Again, sales comprise innovative and imitative demand, which are influenced by the coefficient of innovation α_i and imitation β_i . Both coefficients are influenced by managerial decisions of each company i like pricing, advertising, quality, market entry timing, etc. and measure the relative influence of the decisions compared to the competitor's decisions. Therefore, the values



Diffusion of Innovations, System Dynamics Analysis of the, Figure 8

Follower's market share and sales for different market entry times



Diffusion of Innovations, System Dynamics Analysis of the, Figure 7
Coarse structure of an oligopolistic innovation diffusion model

α_i and β_i not only depend on the decisions of company i , they also depend on the competitor's decisions. Both variables are crucial for the speed and the maximum volume of demand for the products of a company i .

Figure 8 shows the results of simulations based on Eq. (11) for innovative demand and Eq. (14) for imitative demand with the effects of a market entry delay of the second company – the influences of other decision variables are switched off. Several model simulations have been made assuming a market entry delay of company 2 between 0 and 12 months.

The plots in Fig. 8 show the development of market share and sales of the second company over time. Since there is no further product differentiation, both competitors have the same market share when they enter the market at the same time. With each month delay of the second company the market share that can be achieved at the end of the simulation decreases. A three months delay reduces the finally achieved market share to 40%; a 12-month delay even causes a decrease in market share down to approximately 25%. Accordingly, the maximum sales volume decreases significantly with each month delay in market entry time.

Representing Network Externalities

In the following, we will investigate the diffusion of a specific type of goods in order to show the importance of understanding the diffusion of goods with network effects (based on [22]). The trend towards an information society has stressed the relevance of goods satisfying information and communication needs. Many products of this market segment such as electronic mail contain attributes that necessitate a specific examination, since the diffusion of goods showing network effects differs from that of conventional products. The main difference between conventional products and products with network effects is that the utility of the latter cannot be regarded as a constant value. With regard to these products, utility is an endogenous variable which results in a specific diffusion behavior. Two effects are responsible for this particular behavior: the bandwagon effect and the penguin effect. A refined system dynamics model supports a better understanding of this special diffusion process.

The fact that the utility is not constant can be reasoned by a concept commonly referred to as “network effect”. A product is characterized by a network effect, if the utility of that product is a function of the installed base, which is defined as the amount of users in a network system. The utility increases with the installed base. This leads to a virtual interdependency of the users inside the network. The

interdependency is based on the bandwagon effect and the penguin effect. Starting from the fundamental diffusion model of Bass the characteristics of network effects are integrated into the model in order to simulate the diffusion behavior.

Many definitions for network effects (sometimes also called “positive demand externalities”) can be found in the literature. Basically, it can be stated that network externalities exist if the utility of a product for a customer depends on the number of other customers who have also bought and use this product. These network externalities can be indirect or direct, whereby we concentrate on the latter. A typical example for a product with direct network effects is the telephone. The utility that a telephone system can generate increases with the amount of feasible communication connections. Other examples are e-mail, fax, instant messaging, etc. each of which satisfies communication needs. But none of these products can generate any utility for a user on its own. In order to create utility the existence of other users adopting this product is required. Accordingly, the product's utility changes dynamically with the number of users, i. e., the installed base.

The installed base B_t determines the utility of products influenced by network externalities. In terms of direct network externalities the utility is based on B_t exclusively since utility can only be generated by interconnections within the underlying network solely. Accordingly, the utility of a product with direct network externalities is a function of the number of feasible connections I_t . The number of connections is determined by the number of users on the one hand and the technological restriction of the network on the other hand, whereby the latter one represents the number of users being able to communicate via the network simultaneously (for instance, classical telephone system $r = 2$, telephone conferencing $r \leq 2$). Thus, U_t can be calculated by the formula:

$$U_t = U_t(I_t) = \sum_{k=2}^n \binom{B_t}{r} = \sum_{k=2}^n \frac{B_t!}{r! (B_t - r)!}, \quad \text{whereby } r, B_t > 0. \quad (15)$$

Since the achievable utility of a product with direct network externalities depends exclusively on the network size the adoption process depends on the decision of potential users influencing the diffusion process significantly. This leads to two different effects: Firstly, the utility for each user grows exponentially with an increasing amount of actual users according to the formula. This implies that the more people are attracted the more are part of the network leading to an exponential growth of the diffusion process which is referred to as the “bandwagon effect”: The higher

the number of people the more are decoyed as well resulting in a reinforcing process. Although this effect occurs with conventional products as well, in case of products with direct network externalities, it is much stronger since the exponentially growing utility has a greater impact on the diffusion process.

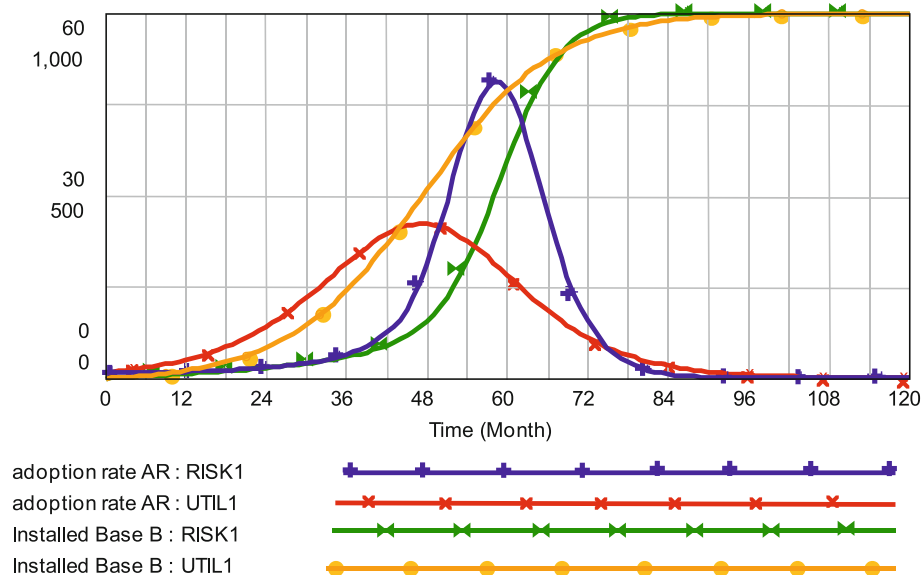
Secondly, utility must be created by the utilization of the users first in order to establish a new communication network which determines the diffusion process significantly since products influenced by direct network externalities cannot generate an original utility by itself. Accordingly, early adopters of a network are confronted with the risk that they cannot derive sufficient benefit from the network so that they must rely on potential users to follow entering the network in the future. Therefore, the adoption decision of a potential user depends on the future decision of other potential users. All in all, this leads to a hesitating behavior of all potential adopters resulting in an imaginary network barrier, which is based on the risk of backing the wrong horse, which is also known as the “penguin effect”.

Finally, another important aspect must be considered when analyzing the diffusion process of products with network externalities. In terms of conventional products the decision to buy a product is the final element of the decision process. Contrary to that, concerning products with network externalities the adoption process is not finished with the decision to enter a network since the subsequent utilization of the product is important for the diffusion

process. If the expected utility of the communication product cannot be achieved users may stop using this product leading to a smaller installed base and a lower network utility which is important for other users that may stop their utilization as well and for the adoption process of potential users.

In the following, the basic structure of the underlying model will be described. Analogously to the model presented in the preceding paragraphs there exists a group of potential users (in this model, we only focus on the core diffusion process without considering competitors or latent demand in order to keep the complexity of the model low.) If these potential users decide to adopt the communication product, they become part of the installed base B . The adoption process is illustrated by the adoption rate AR , which is primarily influenced by the variable *word of mouth*. In order to consider the average utility per user – as it is necessary for analyzing products with network externalities – the imitation coefficient β has been endogenized contrary to the classical Bass model. Therefore, the variable β is influenced by the “word-of-mouth” effect which depends on the average utility per user. If actual utility is bigger than the desired utility all individuals in contact with users adopt and buy the product. If it is smaller, however, only a fraction adopts. The size of this fraction depends on the distance between actual and desired utility.

Figure 9 depicts two simulation runs showing the system behavior of the diffusion process of conventional



Diffusion of Innovations, System Dynamics Analysis of the, Figure 9
Comparison of diffusion behavior

products and products influenced by direct network externalities. Graph UTIL1 represents the *adoption rate*, i. e., the amount of buyers of a conventional product per period. The graph UTIL1 shows the behavior of the variable *installed base B* which is the accumulation of *adoption rate* (note that the graphs have a different scale). The graphs RISK1 show the system behavior for products influenced by direct network externalities, i. e., the *adoption rate AR* and the corresponding *installed base B*.

A comparison of both simulation runs shows that diffusion needs longer to take off in terms of products influenced by direct network externalities, but showing a steeper proceeding of *Installed Base B* in later periods. This behavior can be verified comparing the adoption rates of the two runs: although adoption starts later with an endogenously generated adoption fraction, it nevertheless has a higher amplitude. This behavior can be interpreted as the penguin effect and the bandwagon effect.

Finally, it has to be taken into account that some users of the installed base might quit to use the product since they are disappointed from its utility. Accordingly, it is an important issue to find ways in order to raise the patience of users to stay within the network. That gives potential users the chance to follow into the network which will increase the utility for the user as well.

From the simulation analysis the following conclusions can be drawn. The importance of the installed base for a success diffusion process is shown. Without a sufficient amount of users it is not possible to generate a utility on a satisfying level which prevents potential users to enter the network or even making users leave the network. Accordingly, ways must be found to increase the utility that a network creates for a user in order to reach the critical mass. This can be done in several ways of which some will be discussed briefly. One possible way is to increase the installed base by compatibility to other networks. Furthermore, the risk to back the wrong horse can be mitigated by product pre-announcements in order to lower the imaginary network barrier by making potential users familiar with the product. Another possibility is to increase the group of relevant users, i. e., to enlarge the average group size within the network, since not all users are equally important for a potential user. Furthermore, the technological potential can be improved by introducing multilateral interconnections between the members of a network.

Representing Managerial Decision Making in Innovation Diffusion Models

Subsequently the basic structures of innovation diffusion processes described above will be extended and simulated

to demonstrate the impact of managerial decision-making on the diffusion of innovations. The model used for the simulations serves as a simulator to determine how individual strategies can accelerate or hamper market penetration and profit performance. The models are not designed to predict the basic market success or failure of innovations. Although, they are rather comprehensive, several assumptions apply here as for all models. E.g., in all model runs, the basic market acceptance of the innovation is assumed. Furthermore, the simulations assume for the moment that no competition exists.

Dynamic Pricing Without Direct Competition

In a first step the basic model from Subsect. “[Base Structure of a System Dynamics-Based Model of Innovation Diffusion](#)” is extended to generate dynamic cost behavior as suggested in Fig. 5. Standard costs are the basis for the calculation of prices – an important decision variable. Experience curve effects are modeled based on cumulated production in order to map the long-term behavior of standard cost. The actual costs of a product in a certain period are derived from the standard cost modified for variations resulting from capacity utilization.

The concept of experience curve effects suggests a direct relationship between cumulated production $X_{(t)}$ and average standard cost per unit $c_{(t)}^s$, adjusted for inflation; where c^s defines standard unit cost at the planned level of production. Every doubling of $X_{(t)}$ is associated with a cost reduction in real terms by a constant percentage according to:

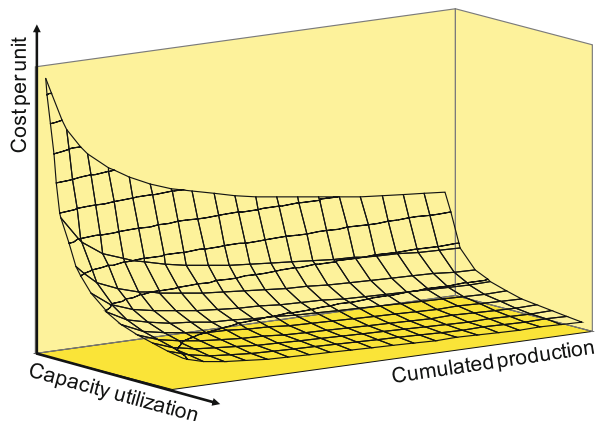
$$c_{(t)}^s = c^n \left(\frac{X_{(t)}}{n} \right)^{-\delta} \quad (16)$$

where c^n stands for the cost of unit n ($n \subseteq X$) and δ represents a constant depending on the experience rate. For many businesses experience rates of 10% to 20% have been observed and ample empirical evidence for this relationship is available.

The costs of a product in each period of time $C_{(t)}$ are a function of cumulated production $X_{(t)}$ and capacity utilization determined by the production volume of a period $x_{(t)}$ as defined in Eq. (17). Figure 10 shows the behavior of the dynamic cost function

$$C_{(t)} = \Phi(X_{(t)}, x_{(t)}) \quad (17)$$

Furthermore, the model comprises elements of (i) market development, (ii) product pricing and its impact on the profits from producing and selling the products, i. e., the operating results, and (iii) resource allocation, e. g., capital investment, production volume, and quality control.



Diffusion of Innovations, System Dynamics Analysis of the, Figure 10
Dynamic cost function

Pricing and quality affects the coefficients of innovation α and imitation β from Eq. (10). Figure 11 shows the run of a model version including market development.

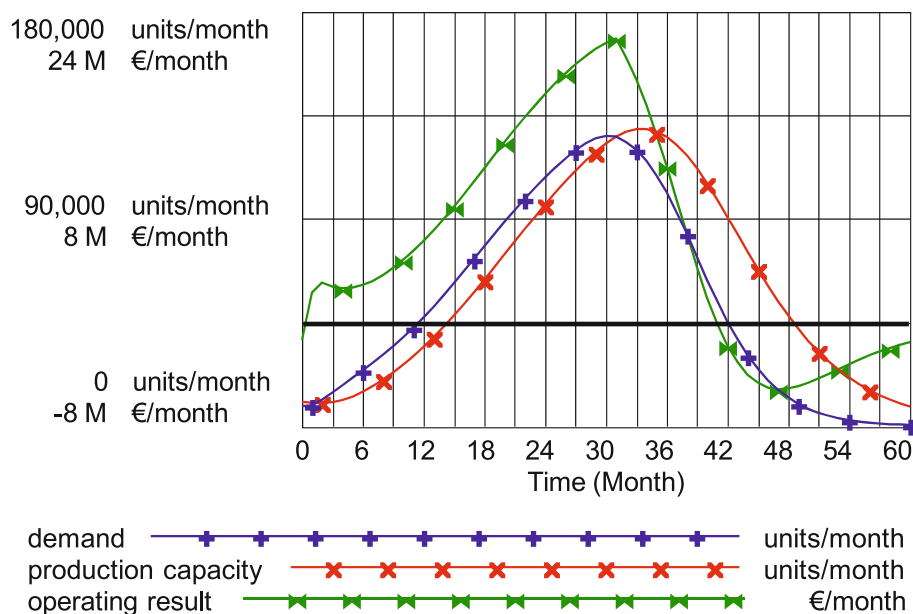
The time behavior of production, demand, and operating results duplicate usual characteristics of the life cycle of a successful innovation. After the product launch, additional customers can be gained from an untapped market as diffusion and thereby product awareness proceeds and prices decline. The maximum of demand from imitators – the quantitatively most important fraction of demand – is

reached when the amount of possible communications between potential customers and adopters reaches its maximum. The decreasing level of potential customers and the depletion of the untapped market cause the decline towards the end of the simulation. The behavior also shows that demand rises much faster than the company can increase its production capacity. The behavior of Fig. 11 will serve as reference mode for further analysis.

Pricing strategies and decisions are additional important elements, which require an extension of the model. The problem of the “right price” for a new product is essential but still unsolved in the area of innovation management. Difficulties to recommend the optimal pricing policy derive in particular from the dynamics in demand interrelations, cost development, potential competition, and the risk of substitution through more advanced products. Regardless of this complex framework, several attempts in management science try to derive and to apply optimal pricing policies. However, they are faced with difficulties, both mathematical and practical. Their results are too complicated to support actual pricing decisions. Therefore simulation studies found more frequently their way into management science.

The extended model includes four predefined pricing policies to investigate their impact on market development on operating results:

Myopic profit maximization assuming perfect information about cost and demand. The optimal price p^{opt} is de-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 11
Reference mode of the basic innovation diffusion model

rived from elasticity of demand ε_t and per unit standard cost $c_{(t)}^s$ considering the impact of short term capacity utilization:

$$p_{(t)}^{\text{opt}} = c_{(t)}^s \cdot \frac{\varepsilon_t}{\varepsilon_t - 1} . \quad (18)$$

Skimming price strategy aims at serving innovative customers with high reservation prices and then subsequently reduces prices. The model applies a simple decision rule modifying $p_{(t)}^{\text{opt}}$ through an exponential function that raises the price during the first periods after market introduction:

$$p_{(t)}^{\text{skim}} = p_{(t)}^{\text{opt}} \cdot \left(1 + a \cdot e^{\frac{-t}{T}}\right) . \quad (19)$$

Full cost coverage, i. e., standard cost per unit plus a profit margin π to assure prices above cost level even during the

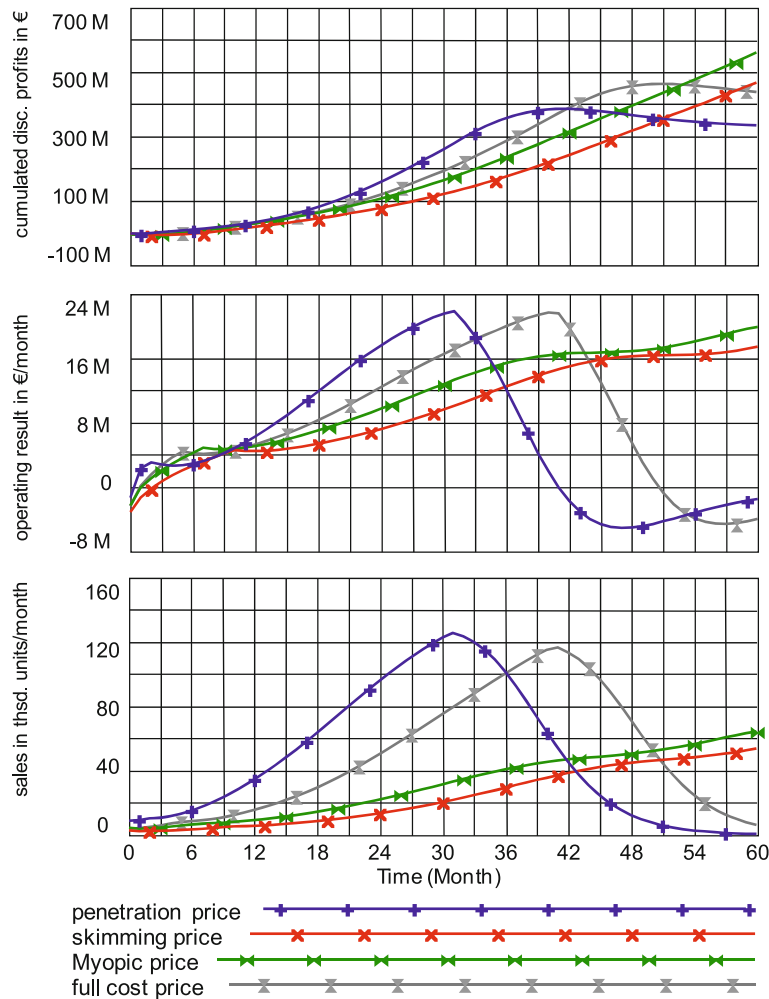
early stages of the life cycle:

$$p_{(t)}^{\text{fcc}} = c_{(t)}^s \cdot \pi . \quad (20)$$

Penetration pricing aims at rapidly reaching high production volumes to benefit from the experience curve and to increase the number of adopters. It uses a similar policy as for the skimming price, but instead of a surcharge it decreases prices early after market introduction:

$$p_{(t)}^{\text{pen}} = c_{(t)}^s \cdot \pi \cdot \left(1 - a \cdot e^{\frac{-t}{T}}\right) . \quad (21)$$

The simulation runs shown in Fig. 12 give an overview of the development of profits, cumulated profits, and sales for the four pricing strategies discussed above. The model assumes the following: (1) there is an inflow from the un-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 12
Comparison of the outcome of pricing strategies

tapped market, which depends on the dynamic development of prices; (2) there is no risk of competition; (3) repeat purchases do not occur. Taking profits into account, Fig. 12 indicates that – over the time horizon observed –, the classic pricing rule of profit optimization leads to superior results from a financial point of view. However, if judged by the market development, the strategy of penetration prices is the appropriate strategy. This strategy allows rapid penetration of the market by setting relatively low prices, especially in the early stages of the life cycle. The combined price and diffusion effects stimulate demand and reduce the risk of losing potential customers to upcoming substitution products.

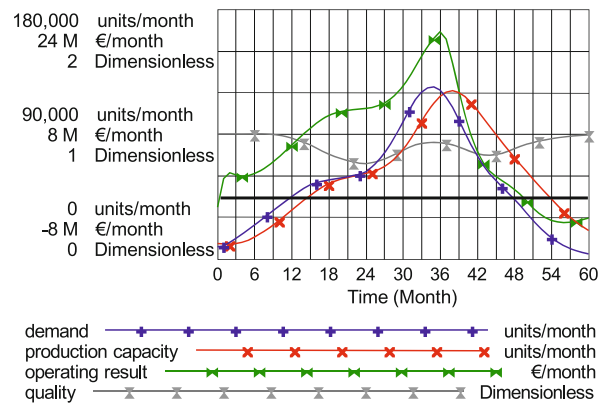
Figure 12 also indicates a disadvantage of the penetration strategy. Since the market is already completely satisfied after period 54, there is only little time to develop and introduce a new product in the market successfully. The slower market growth of the skimming and optimum price strategy leaves more time for the development of a new product, but the attractive profit situation and the slow development also increase the risk that competitors might enter the market. In a dynamic demand situation where prices influence market growth, where substitution or competition can occur, and where delivery delays eventually accelerate the decision of potential buyers to turn to other products, a strategy of rapid market penetration seems to be the most promising one. It will, therefore, be the basis for the following simulation runs investigating manufacturing's role in innovation management.

Linking Manufacturing-Related Decision Variables

The role of manufacturing is important for the successful management of innovations. Manufacturing has to provide sufficient capacity to produce the goods sold. The investments to adjust capacity influence a company's ability to meet demand and deliver on time. It is assumed that the necessary financial resources for the investments are available. The aggregated capacity provided by the company includes both, machinery equipment and production personnel. Since the manufacturing function also has to ensure the quality of the output through dedicating a portion of its total available capacity to quality control, the capacity resources can be used to either manufacture the products or to assure the desired level of quality. Capacity allocation to improve quality takes away capacity for production. This additional feedback structure – as indicated in Fig. 5 – maps the allocation of resources for quality control to the achieved ability to meet product demand. If manufacturing capacity does not meet demand, a temporary reduction of capacity for quality assurance seems a plausible

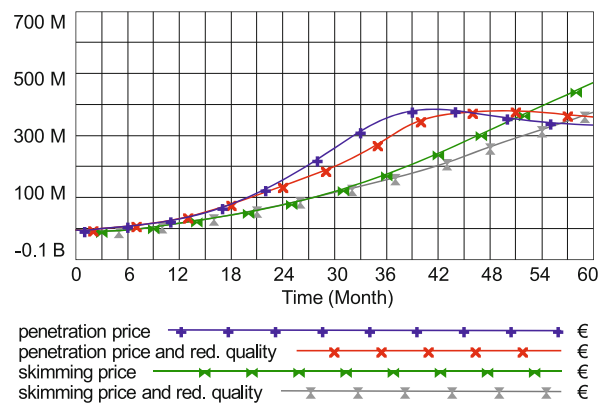
strategy. Quality control resources than are allocated to manufacturing rather than testing whether quality standards are met. In this scenario, it would be expected that total cost remain unchanged and the additional manufacturing capacity gained through the reallocation can be used to provide more products to the customers, increase sales, and improve the overall results.

Figure 13 shows the simulation assuming the same scenario as in the base mode together with penetration prices and reduced quality resources if demand exceeds production capacity. It also shows a quality index plotted as an additional variable. Quality is defined to be 1, if the actual quality capacity equals a standard value of quality resources necessary. It is assumed that 10% of total production capacity is necessary to assure 100% quality. For values above the 10%-level, quality is better; for values be-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 13

Reduced quality control



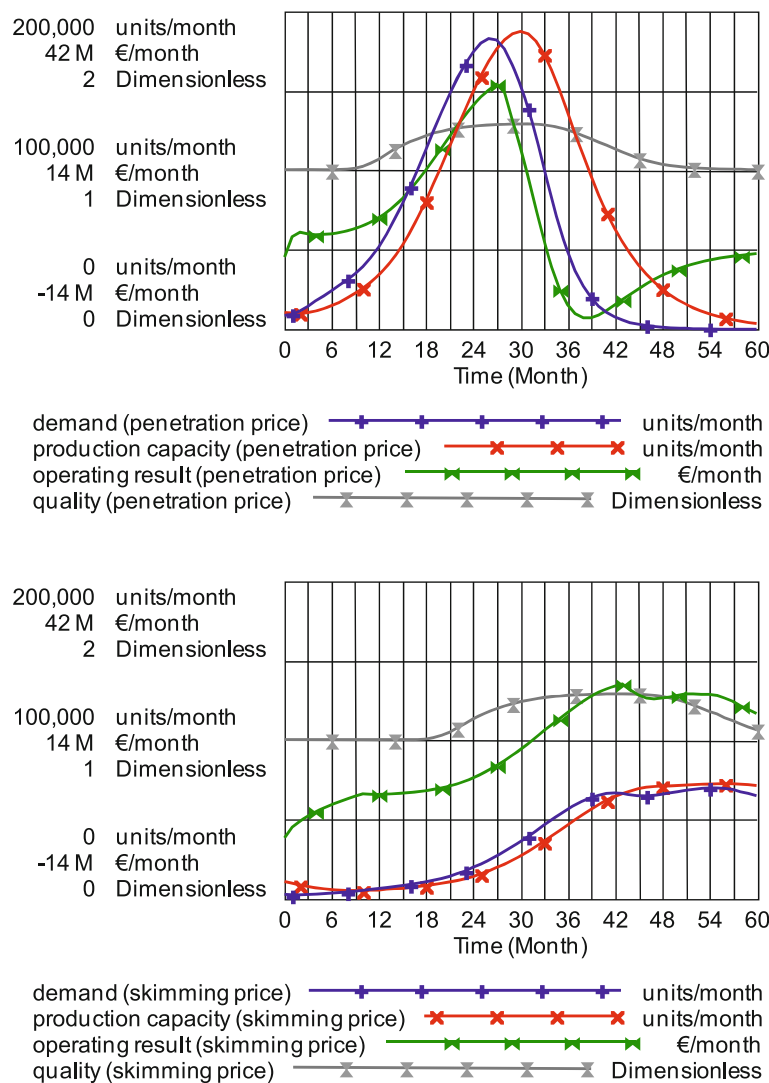
Diffusion of Innovations, System Dynamics Analysis of the, Figure 14

Cumulated discounted profits – penetration vs. skimming pricing in combination with quality control strategies

low, it is poorer. The simulation indicates that the policy of reduced quality resources successfully decreases the discrepancy between demand and production as seen in the reference mode of Fig. 11. This results from the increased proportion of capacity used for production and an additional effect caused by lower product quality, which then decreases demand. Although the maximum sales are nearly the same in the simulation of reduced quality control strategy, the peak demand occurs around 5 months later. Instead of gaining higher sales only the shape of the life cycle changed. However, operating results had improved, in particular the sharp decline of profits in the base mode of the simulation could be slowed down and

losses could be avoided. The reduced quality control strategy caused a slower capacity build-up and therefore, when product sales declined capacity adjustment was easier to achieve. From the financial point of view the strategies of penetration prices and reduced quality control fit quiet well.

The results are different if a strategy of quality reduction is used in combination with a strategy of skimming pricing. Figure 14 compares the outcome of cumulated discounted profits for the strategy of reduced quality and penetration prices or skimming prices with the development of the reference mode – the simulations without quality adjustment. The behavior indicates that in the case



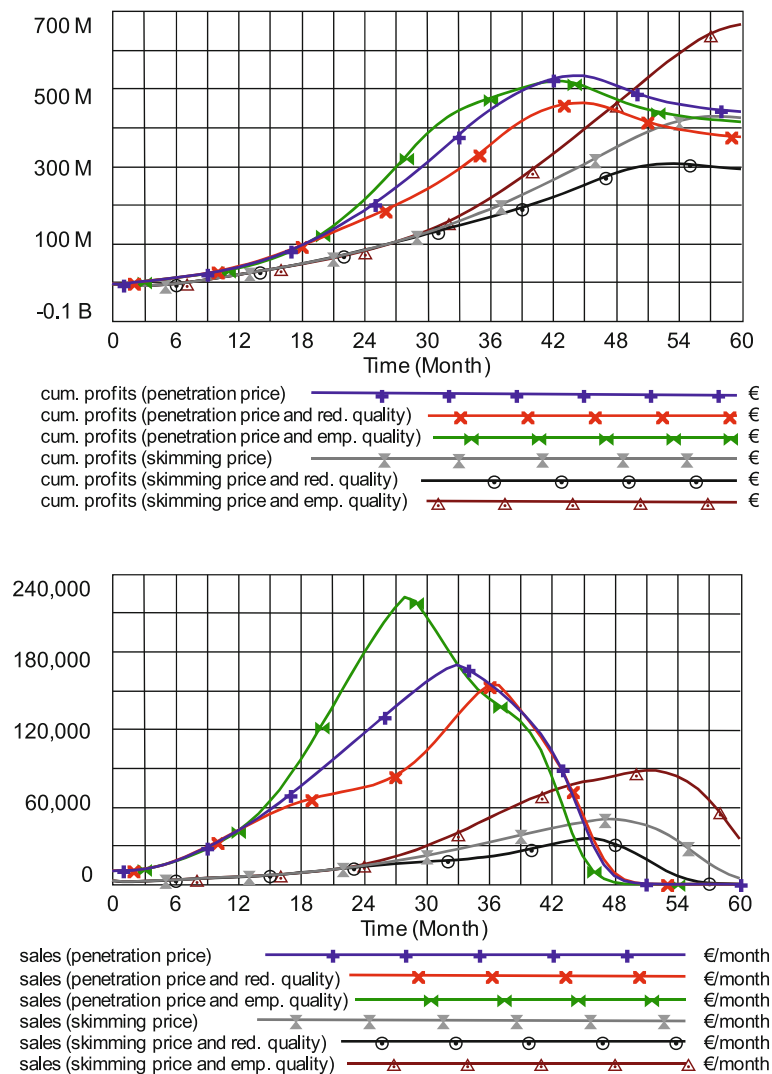
Diffusion of Innovations, System Dynamics Analysis of the, Figure 15
Emphasized quality in all innovation stages

of skimming prices, quality reductions slow down the development of the market and cumulated profits significantly.

The simulation results raise the question whether emphasizing quality when demand is higher than capacity would be a more appropriate way to react. As the upper part of Fig. 15 points out, the strategy of emphasized quality leads to an accelerated product life cycle in the case of the penetration pricing strategy. Tremendous capacity build-up is necessary after the introduction of the new product. As demand declines, a plenty of capacity is idle, causing significant losses during the downswing of the product life cycle.

Emphasizing quality turns out to be more effective in the case of skimming prices. The additional demand gained from quality improvements also accelerates the product life cycle, but at a much slower rate and leads to improved cumulated profits. Emphasizing quality in combination with skimming or optimum prices leads to improved cumulated profits, compared to both, the simulation without quality reaction and the quality reduction run.

The simulations show the importance of a detailed judgment of strategic choices. Strategies must be consistent with each other and with the real world structures mapped by the model. The simulations above assume a sit-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 16
Behavior of the base model including simple competitive structures

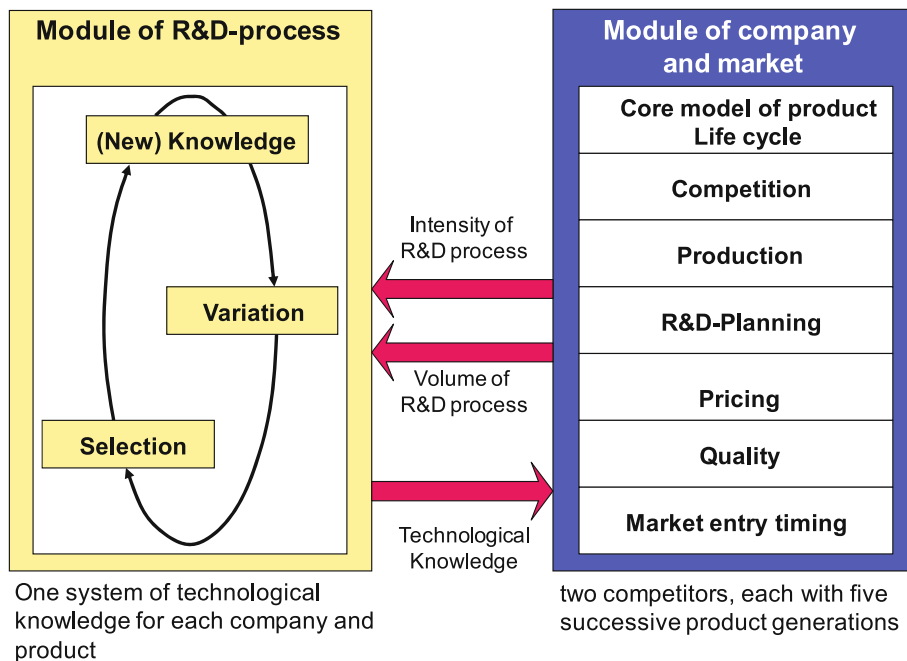
uation without existing or potential competition. In such an environment there is no interest paid for fast market penetration. Hence, a penetration pricing strategy is the most unfavorable alternative. However, this changes if structural elements are added to the model that incorporate competition – even as in the simple structure from Fig. 5, which considers the loss of demand to a competitor. Lost demand therefore is represented as a process equivalent to the imitative demand from Eq. (9). The calculation of lost demand starts in period 15 through an initial switching of a potential customer to the competitor. This switch starts a process that drives demand for the competitors' products and is influenced through the quality the company offers. If the company provides poor quality, more potential customers and market potential from the untapped market will directly move to the competitor. The accumulation of lost demand corresponds to the number of adopters the competitors gained over time. Simulations with these additional structures give some additional insights (Fig. 16).

Penetration pricing leads again to the fastest market development. In the competitive surrounding, however, emphasizing quality accelerates the market development and leads to better performance than quality reductions. This is in contrast to the simulations without competition shown in Fig. 13 to Fig. 15. Skimming prices in combi-

nation with reduced quality control shows the poorest financial and market performance. A strategy of reduced quality control causes in the competitive environment the demand to increase at a slower rate than in the base run, where no quality adjustments were made when demand exceeded capacity. In both cases, the skimming and the penetration price scenario, quality reductions lead to the poorest performance.

Linking R&D and New Product Development

The models discussed above are able to generate under different conditions the typical diffusion patterns of new products in the market place. However, these models do not consider the stage of new product development. New products have to be developed before they can be introduced into the market. A costly, lengthy, and risky period of R&D has to be passed successfully. The diverging trends of shortening product life cycles and increasing R&D costs show the importance of an integrated view of all innovation stages. In the remainder, a comprehensive model comprising both, the process of R&D and an oligopolistic innovation diffusion with subsequent product generations is used to investigate the interrelations between the stages of innovation processes. The integration of both modules is shown in Fig. 17.



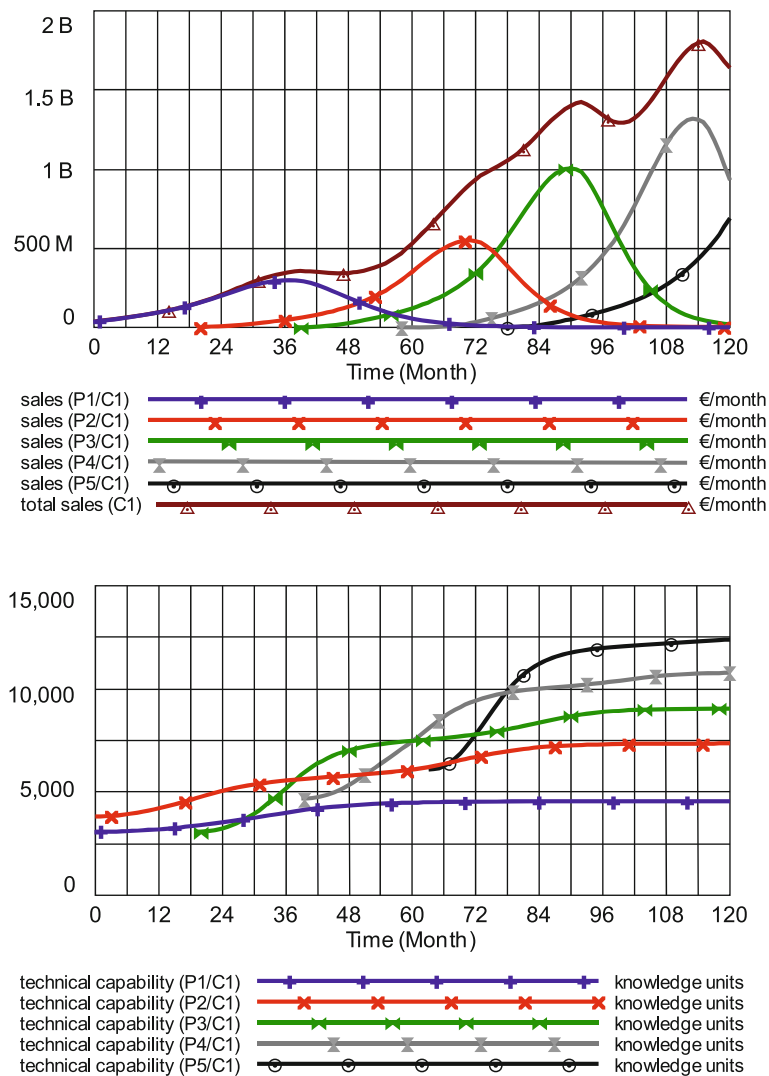
Diffusion of Innovations, System Dynamics Analysis of the, Figure 17
Linking R&D-processes with corporate and market structures

The volume and the intensity of the research and development activities feed the R&D-process. The number of research personnel determines the volume. Since R&D personnel requires resources like laboratory equipment, material for experiments etc., the intensity of R&D depends on the budget available for each person working in the R&D sector. This information is calculated in a more comprehensive model in the sector of R&D planning, which also includes policies about resource allocation within the research and development stages, i.e., mainly the question of how much to spend on which new product development project.

Depending on the volume and the intensity of R&D, the technological knowledge of each product generation

for each company evolves over time. The module of the R&D-process feeds back the current state of the technological knowledge for each company and product generation.

The basic assumptions of the model are as follows. The model maps the structures of two competitors. Both competitors can introduce up to five successive product generations. The initial values of the model ensure that all competitors start from the same point. All firms have already introduced the first product generation and share the market equally. The resources generated by the first product are used to develop subsequent product generations. In the base run each company follows the same set of strategies. Therefore, except for minor differences resulting from the



Diffusion of Innovations, System Dynamics Analysis of the, Figure 18
Exemplary behavior of the integrated innovation model

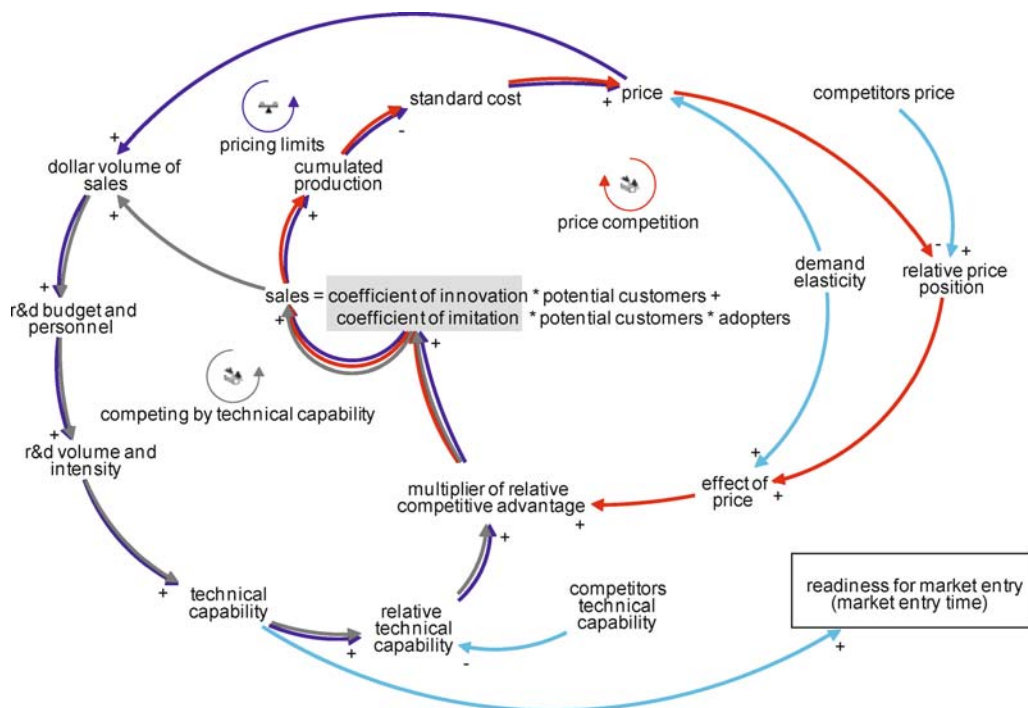
stochastic nature of the R&D-process, they show the same behavior over time. Figure 18 provides a simulation run of the model with all modules and sectors coupled.

The curves show for a single company the development of the sales of the products and the total sales. They emphasize the importance of a steady flow of new and improved products. Without on-time replacement of older products, the total sales of the products will flatten or deteriorate like in the simulation around periods 44, 92, and 116. The model also generates the typical s-shaped curves of technological development (lower part of Fig. 18). Each product generation has a higher technological potential and the knowledge developed for the preceding product generations partly can be used by the successive product generations. For this reason the subsequent product generations start at a level different from zero.

In a dynamic environment such as the computer industry, where investments in R&D and manufacturing equipment are high, the product life cycles are short, and time-to-market as well as time-to-volume are essential variables, it is important to understand the dynamic consequences of decisions and strategies in the different areas. Figure 19 describes some of the important feedback loops linking the process of invention to the processes of innovation and diffusion.

Central element in the figure is the calculation of the sales of a company according to Eqs. (11) and (14). The coefficients of innovation and imitation are influenced by the multiplier of relative competitive advantage, which depends on the relative technical capability and the price advantage of a company. The technical capability of the products is influenced by the strength of its R&D-processes and the total amount of R&D expenditures. Empirical studies in Germany have shown that measures like sales volume, profits or R&D budgets of earlier periods are quite common as a basis for R&D budgeting. However, using historic sales volume as a basis to determine R&D budgets invokes the positive feedback loop “competing by technical capability”. With an increasing number of products sold and growing value of sales the budget and the number of personnel for R&D grow. This leads to an improved competitive position, if the technical capabilities of a product increases. The higher the sales volume, the better is the resulting competitive position. This produces increasing coefficients of innovation and imitation and leads to higher sales. This budgeting strategy is implemented in the model for the next simulation runs.

The second loop “price competition” links pricing strategies to sales volume. The actual price of a product is influenced by three factors. The first factor, standard



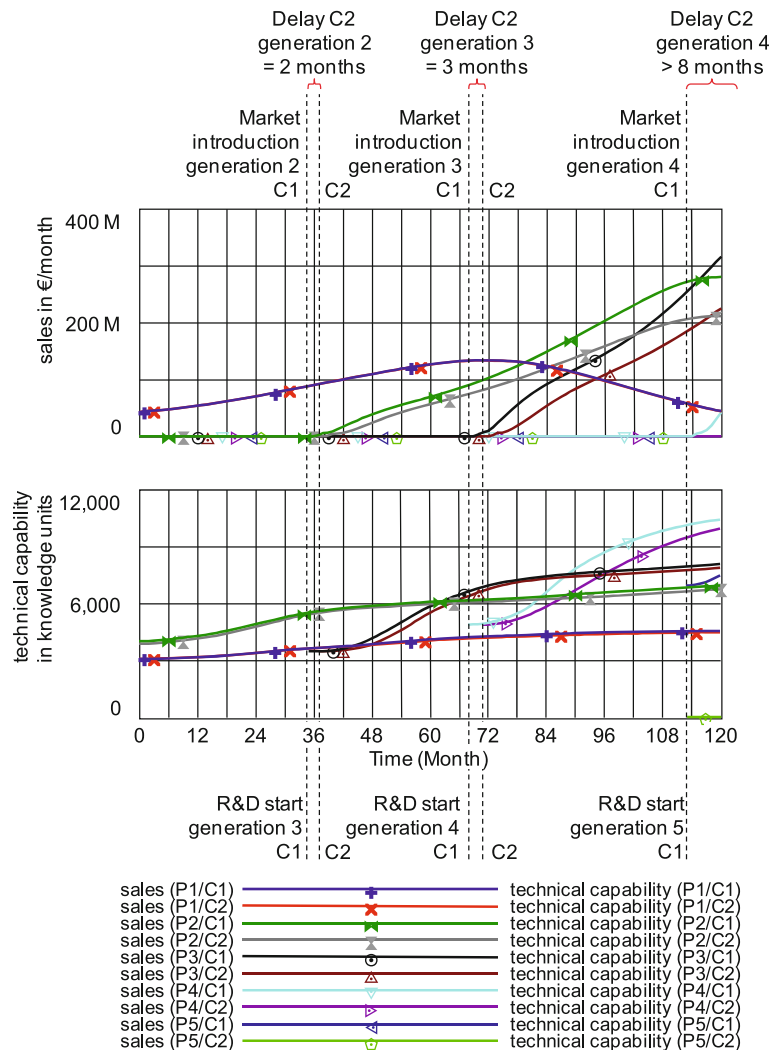
Diffusion of Innovations, System Dynamics Analysis of the, Figure 19
Feedback structure influencing the diffusion process

costs, is endogenous. As cumulated production increases, the experience gained from manufacturing causes declining standard costs. The second and third elements influencing the calculation of prices are exogenous: parameters which define the pricing strategy and demand elasticity. Caused by increasing cumulated production, standard costs fall over the life cycle and prices are also declining. Lower prices affect the relative price and improve the effect of price on the coefficients of innovation and imitation, which leads to increased sales and higher cumulated production.

The loop “pricing limits” reduces the effects of the reinforcing loops described above to some extent. The standard cost and price reductions induce – *ceteris paribus* –

a decrease in the sales volume and set off all the consequences on the R&D-process, the technical know-how, the market entry time and sales shown in the first feedback loop – but in the opposite direction. Additionally, since standard cost cannot be reduced endlessly this feedback loop will show a goal seeking behavior.

With equivalent initial situations and the same set of strategies, both companies behave in an identical way for all product generations. If one company has a competitive advantage, the reinforcing feedback loops suggest that this company will achieve a dominating position. In the simulation shown in Fig. 20, both competitors have the same competitive position for the first product generation. But the first company will be able to enter the market 2 months

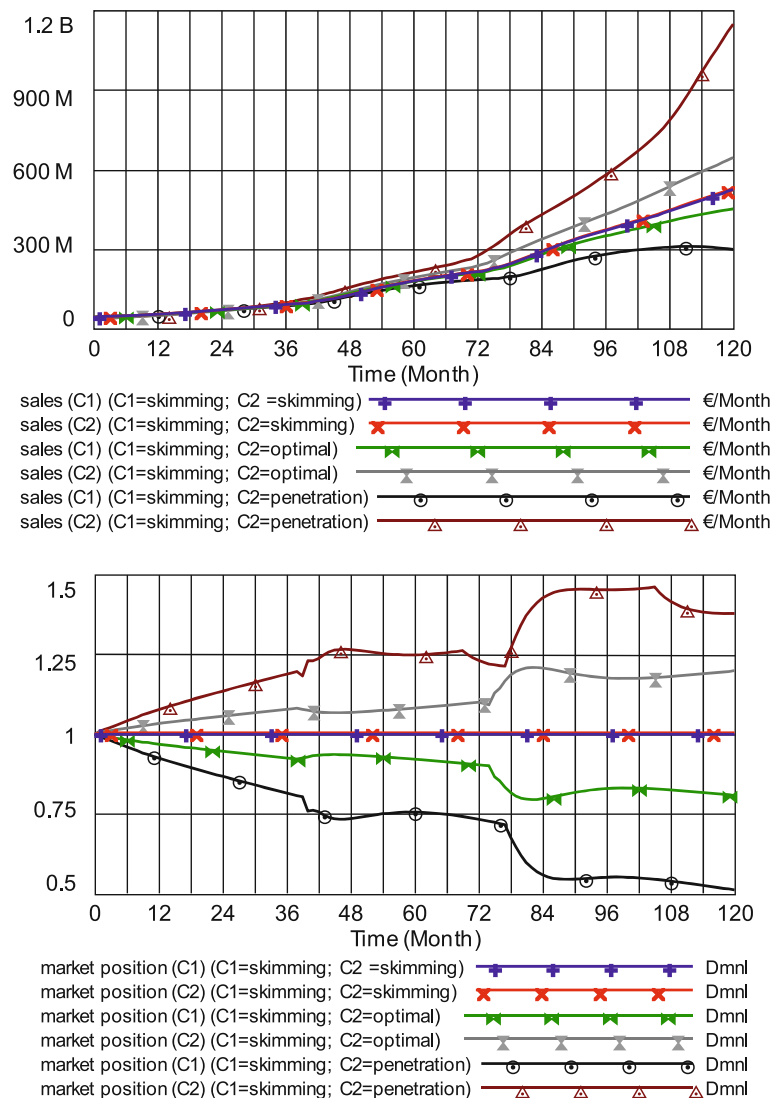


Diffusion of Innovations, System Dynamics Analysis of the, Figure 20
Reinforcing effects of initial competitive advantage

earlier than the competitor, because the initial outcome of the R&D process is slightly better than the second company's second product generation. Both competitors follow a strategy of skimming prices and demand elasticity has the value -2 .

The initial gain in the outcome of the R&D-process initiates a process of sustained and continuing competitive advantage for the first company. It will improve continuously, since the positive feedback loop "competing by technical capability" dominates. The first company's advantage in the market introduction leads to an increasing readiness for market entry. It is able to launch the third

product generation 3 months earlier than the follower and will introduce the fourth product generation in period 112. The follower is not able to introduce its fourth generation during the time horizon of the simulation, i.e., the pioneers advantage has extended to more than 8 months. The first company's competitive advantage is a result of the slightly higher initialization of the knowledge system and the dominance of the positive feedback loops, which causes shortened time-to-market and higher sales volume over all successive product life cycles. Additionally, the technical capabilities of both competitors' product generations show the same reinforcing effect. The difference



Diffusion of Innovations, System Dynamics Analysis of the, Figure 21
Sales volume and market position for different pricing strategies

between the technical capability of both competitors increases in favor of company 1 until they approach the boundaries of the technology.

Although literature discusses a variety of models to find optimal pricing strategies, these models usually only consider the market stage of a new product and neglect the interactions with the development stage of a new product. Pricing decisions not only drive the diffusion of an innovation, but they also have a strong impact on the resources available for research and development. Since the comprehensive innovation model links the stages of developing and introducing a new product, the following simulations will show the impact of pricing strategies on performance in a competitive environment. In the analysis shown the first company uses the strategy of skimming price for all product generations. The second company alternatively uses a skimming price strategy in the first model run, myopic profit maximization strategy in the second run, and the strategy of penetration prices in the third run. The initial conditions are identical, except the price strategy settings. Sales volume, market position, and cumulated dis-

counted profits are used to judge the advantages of the alternative pricing strategies. Market entry time is used as a measure of time-to-market.

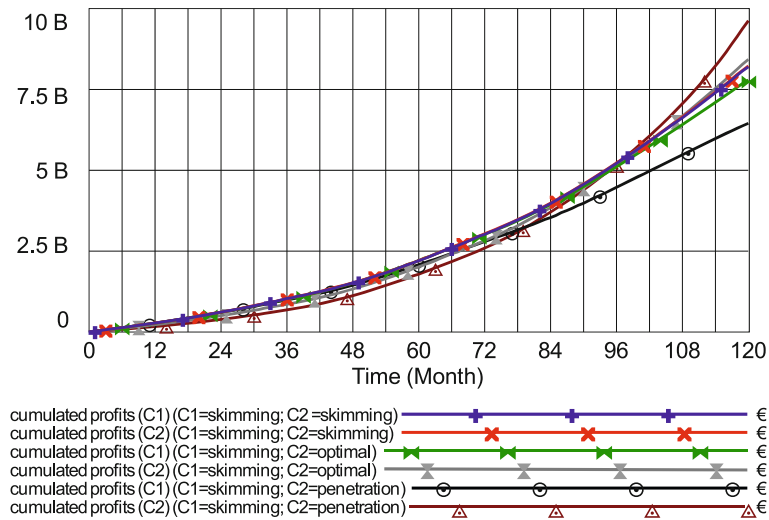
The logic behind the skimming price strategy is to sell new products with high profit margins in the beginning of a life cycle to receive high returns on investment, achieve short pay off periods, and high resources for the R&D-process. However, in a dynamic competitive setting the strategy of myopic profit maximization and penetration prices achieve better results (Fig. 21). Company 1 which uses a skimming price strategy achieves the lowest sales volume. Myopic profit maximization prices and penetration prices of the second competitor causes the sales to increase stronger through the combined price and diffusion effect.

The results are confirmed if the variable market position – an aggregate of the market share a company has for its different products – is taken into account. For values greater than 1 the market position is better than the one of the competitor. Using the penetration strategy, company 2 can improve its market share, achieve higher sales vol-

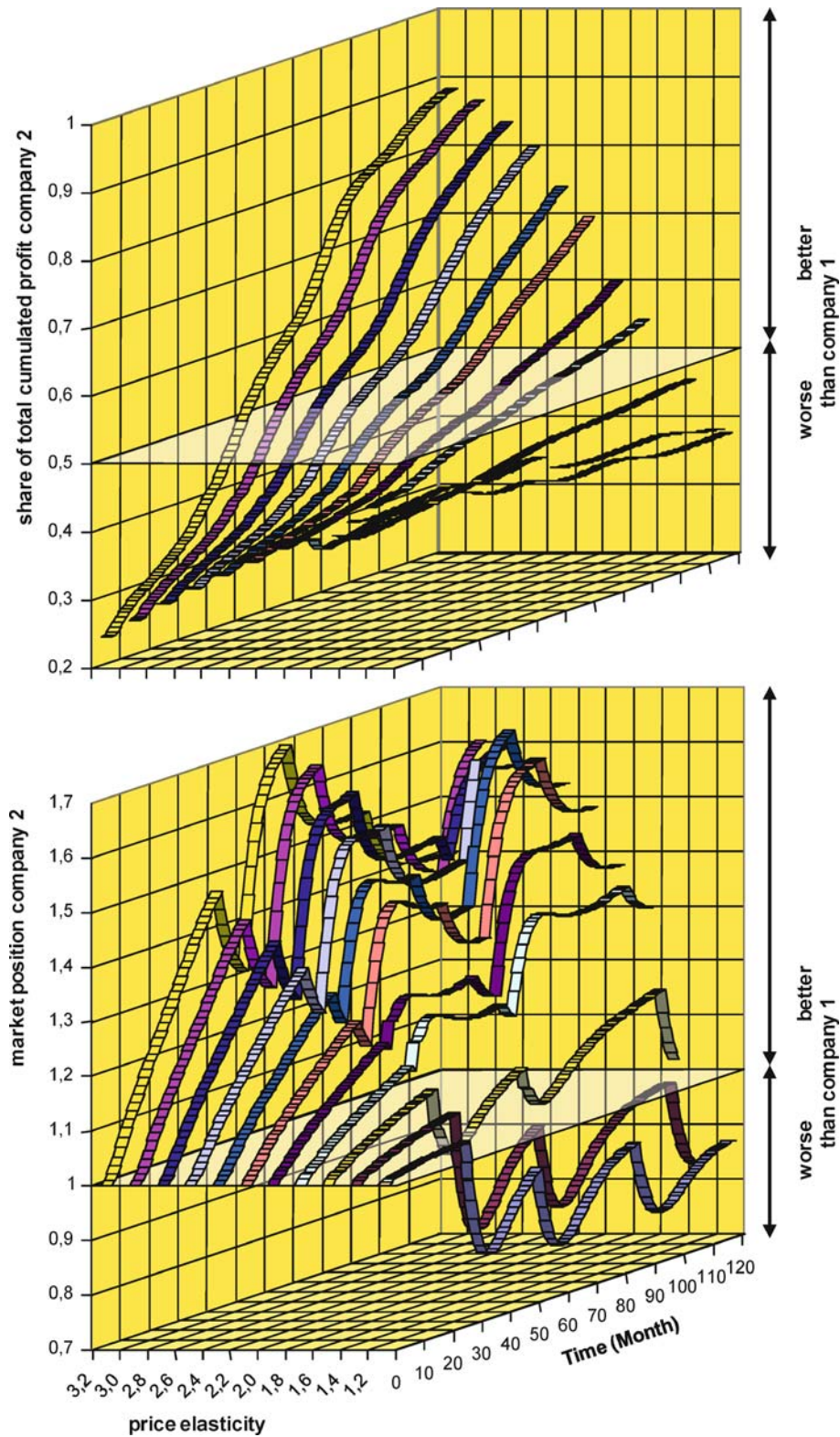
Diffusion of Innovations, System Dynamics Analysis of the, Table 1
Consequences of pricing strategies on market entry time

	Product generation 2			Product generation 3			Product generation 4		
Pricing strategy C2*	C1	C2	Delay C1 to C2	C1	C2	Delay C1 to C2	C1	C2	Delay C1 to C2
Skimming prices	38	38	0	71	71	0	n.i.	n.i.	–
Profit maximization	36	35	1	71	69	2	n.i.	118	> 2
Penetration prices	37	35	2	74	66	8	n.i.	102	> 8

*C1 uses skimming prices in all simulations; **n.i. = product was not introduced



Diffusion of Innovations, System Dynamics Analysis of the, Figure 22
Time path of cumulated profits



Diffusion of Innovations, System Dynamics Analysis of the, Figure 23
Impact of demand elasticity on performance measures

ume and therefore has more resources available for R&D. This enables it to launch new products earlier than company 1. As shown in Table 1, the advantage of time-to-market increases from product generation to product generation.

The improvement in time-to-market for the first company's second product generation results from the slightly higher sales volume compared to the use of skimming pricing strategies for both competitors. The second company achieves the strongest improvements in time-to-market if it uses a penetration pricing strategy.

In terms of cumulative profits (Fig. 22) one would expect that skimming prices should generate the highest cumulated profits, however, this is not true. Penetration prices generate the highest results followed by skimming prices. The strategy of myopic profit maximization shows the least favorable outcome.

The simulations so far assumed a price response function with a constant price elasticity ε of -2 . Since price elasticity influences both, the demand for a product as well as the price level (cf. Fig. 19), the influence of price elasticities have to be investigated before recommendations can be made. Assuming that company 1 uses a strategy of skimming prices and the second competitor follows a strategy of penetration pricing, Fig. 23 shows the time path of cumulated discounted profits and market position for ε between -3.2 and -1.2 .

Due to the different profit margins – resulting from myopic profit maximization being the basis for price calculation – the use of the absolute value of the cumulated profits is not appropriate. Therefore, the second company's share of the total cumulated profits is used for evaluation purposes. The measure is calculated as

$$\left(\frac{\text{cum.profits}_2}{\sum_{i=1}^2 \text{cum.profits}_i} \right).$$

The first graph in Fig. 23 shows that the initial disadvantage of the second company rises with increasing demand elasticity. However, its chance of gaining an advantage increases as well. In the case of lower demand elasticities ($\varepsilon > -1.7$) firm 2 cannot make up the initial disadvantage during the whole simulation. For demand elasticities ($\varepsilon > -1.4$) the cumulated profits ratio even deteriorates. Considering the market position the picture is similar. For demand elasticities $\varepsilon > -1.6$ the penetrations strategy leads to a loss in the market position in the long run. The improvements resulting from the introduction of the successive product generations are only temporary.

Managerial Implications

The simulations above lead to the insight that general recommendations for strategies are not feasible in such complex and dynamic environments. The specific structures like competitive situation, demand elasticity, or strategies followed by the competitors have to be taken into account. Recommendations only can be given in the context of the specific situation. Furthermore, the evaluation of strategies depends on the objectives of a company. If a firm wants to enhance its sales volume or the market share, the strategy of penetration pricing is the superior one. Viewing cumulative profits and the readiness for market entry as prime objectives, the strategy of skimming prices is the best. However, these recommendations hold only for high demand elasticities. Furthermore, the model does not consider price reactions of competitors. The evaluation of improved strategic behavior would become even more difficult. The outcome and the choice of a particular strategy depend on many factors that influence the diffusion process. The dynamics and the complexity of the structures make it almost unfeasible to find optimal solutions. Improvements of the system behavior gained through a better understanding, even if they are incremental, are steps into the right direction.

Future Directions

The series of models presented here are designed in a modular fashion. They offer the flexibility to be adapted to different types of innovations, to different structures, initial conditions and situations. The models provide the opportunity to investigate courses of action in the setting of a management laboratory. They allow one to investigate different strategies and to learn in a virtual reality. They emphasize the process of learning in developing a strategy rather than the final result. To support learning processes, the models could be combined with an easy-to-use interface and serve as a management flight simulator which allows one to gain experience and understanding from playing.

Although the models cover a variety of different aspects in the management of innovations, they still can be extended. Besides more detailed mapping of corporate structures behind managerial decision processes the structures representing the diffusion process can be extended in various ways. Although some research already discusses the problems of mapping the substitution among successive product generations, this area deserves further attention. In particular in high-tech industries with short product life cycles the interrelations between successive product generations strongly influence the overall success of

a company. Furthermore, the diffusion structures could be extended to include cross-buying and up-buying behavior of customers and by that link models of innovation diffusion to the field of customer equity marketing.

Bibliography

Primary Literature

- Bass FM (1969) A New Product Growth Model for Consumer Durables. *Manag Sci* 15:215–227
- Bental B, Spiegel M (1995) Network Competition, Product Quality, and Market Coverage in the presence of network externalities. *J Ind Econ* 43(2):197–208
- Boston Consulting Group (1972) Perspectives on Experience. Boston Consulting Group Inc., Boston
- Brockhoff K (1987) Budgetierungsstrategien für Forschung und Entwicklung. *Z Betriebswirtschaft* 75:846–869
- Brynjolfsson E, Kemerer CF (1996) Network Externalities in Microcomputer Software: An Econometric Analysis of the Spreadsheet Market. *Manag Sci* 42(12):1627–1647
- Church J, Gandal N (1993) Complementary network externalities and technological adoption. *Int J Ind Organ* 11:239–260
- Farrell J, Saloner G (1986) Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation. *Am Econ Review* 76(5):940–955
- Forrester JW (1961) *Industrial Dynamics*. MIT Press, Cambridge
- Jeuland AP, Dolan RJ (1982) An Aspect of New Product Planning: Dynamic Pricing. In: Zoltners AA (ed) *TIMS Studies in the Management Sciences* 18. North Holland, Amsterdam, pp 1–21
- Katz ML, Shapiro C (1985) Network Externalities, Competition, and Compatibility. *Am Econ Review* 75(3):424–440
- Kotler P (1994) *Marketing Management*. Prentice-Hall, Englewood Cliffs
- Leibenstein H (1950) Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. *Q J Econ* 64(2):183–207
- Mahajan V, Muller E (1979) Innovation Diffusion and New Product Growth Models in Marketing. *J Mark* 43:55–68
- Maier FH (1998) New Product Diffusion Models in Innovation Management – A System Dynamics Perspective. *Syst Dyn Review* 14:285–308
- Milling P (1996) Modeling Innovation Processes for Decision Support and Management Simulation. *Syst Dyn Review* 12(3):221–234
- Milling P, Maier F (1996) *Invention, Innovation und Diffusion*. Duncker & Humboldt, Berlin
- Milling P, Maier F (2004) R&D, Technological Innovations and Diffusion. In: Barlas Y (ed) *System Dynamics: Systemic Feedback Modeling for Policy Analysis*. In: *Encyclopedia of Life Support Systems (EOLSS)*, Developed under the Auspices of the UNESCO. EOLSS-Publishers, Oxford, p 39; <http://www.eolss.net>
- Schmalen H (1989) Das Bass-Modell zur Diffusionsforschung – Darstellung, Kritik und Modifikation. *Schmalenbachs Z betriebswirtschaftliche Forsch (zfbf)* 41:210–225
- Senge PM (1994) *Microworlds and Learning Laboratories*. In: Senge PM et al (eds) *The Fifth Discipline Fieldbook*. Doubleday, New York, pp 529–531
- Sterman JD (1992) Teaching Takes Off – Flight Simulators for Management Education. In: *OR/MS Today*, October 1992. Lionheart, Marietta, pp 40–44
- Sterman JD (2000) *Business Dynamics*. Irwin McGraw-Hill, Boston
- Thun JH, Größler A, Milling P (2000) The Diffusion of Goods Considering Network Externalities – A System Dynamics-Based Approach. In: Davidsen P, Ford DN, Mashayekhi AN (eds) *Sustainability in the Third Millennium*. Systems Dynamics Society, Albany, pp 204:1–204:14
- Xie J, Sirbu M (1995) Price Competition and Compatibility in the Presence of Positive Demand Externalities. *Manag Sci* 41(5):909–926

Books and Reviews

- Abernathy JW, Utterback JM (1988) Patterns of Industrial Innovation. In: Burgelman RA, Maidique MA (eds) *Strategic Management of Technology and Innovation*. Homewood, Irwin, pp 141–148
- Bailey NTJ (1957) *The Mathematical Theory of Epidemics*. Griffin, London
- Bass FM (1980) The Relationship between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durable Technological Innovations. *J Bus* 53:50–67
- Bower JL, Hout TM (1988) Fast-Cycle Capability for Competitive Power. *Harvard Bus Review* 66(6):110–118
- Bye P, Chanaron J (1995) Technology Trajectories and Strategies. *Int J Technol Manag* 10(1):45–66
- Clarke DG, Dolan RJ (1984) A Simulation Analysis of Alternative Pricing Strategies for Dynamic Environments. *J Bus* 57:179–200
- Dumaine B (1989) How Managers can Succeed Through Speed. *Fortune* 4 February 13:30–35
- Easingwood C, Mahajan V, Muller E (1983) A Non-Uniform Influence Innovation Diffusion Model of New Product Acceptance. *Mark Sci* 2:273–295
- Fisher JC, Pry RH (1971) A Simple Substitution Model of Technological Change. *Technol Forecast Soc Chang* 3:75–88
- Ford DN, Sterman JD (1998) Dynamic Modeling of Product Development Processes. *Syst Dyn Review* 14:31–68
- Forrester JW (1968) *Industrial Dynamics – After the First Decade*. *Manag Sci* 14:389–415
- Forrester JW (1981) Innovation and Economic Change. *Futures* 13(4):323–331
- Georgescu-Roegen N (1971) *The Entropy Law and the Economic Process*. Harvard University Press, Cambridge
- Graham AK, Senge PM (1980) A Long Wave Hypothesis of Innovation. *Technol Forecast Soc Chang* 17:283–311
- Homer JB (1983) A Dynamic Model for Analyzing the Emergence of New Medical Technologies. Ph.D. Thesis. MIT Sloan School of Management
- Homer JB (1987) A Diffusion Model with Application to Evolving Medical Technologies. *Technol Forecast Soc Chang* 31(3):197–218
- Kern W, Schröder HH (1977) *Forschung und Entwicklung in der Unternehmung*. Rowohlt Taschenbuch Verlag, Reinbek
- Linstone HA, Sahal D (eds) (1976) *Technological Substitution*. Forecast Techniques Appl. Elsevier, New York
- Maier FH (1992) R&D Strategies and the Diffusion of Innovations. In: Vennix JAM (ed) *Proceedings of the 1992 International System Dynamics Conference*. System Dynamics Society, Utrecht, pp 395–404

- Maier FH (1995) Die Integration wissens- und modellbasierter Konzepte zur Entscheidungsunterstützung im Innovationsmanagement. Duncker & Humboldt, Berlin
- Maier FH (1995) Innovation Diffusion Models for Decision Support in Strategic Management. In: Shimada T, Saeed K (eds) System Dynamics '95. vol II. System Dynamics Society, Tokyo, pp 656–665
- Maier FH (1996) Substitution among Successive Product Generations – An Almost Neglected Problem in Innovation Diffusion Models. In: Richardson GP, Sterman JD (eds) System Dynamics '96. System Dynamics Society, Boston, pp 345–348
- Mansfield EJ, Rapoport J, Schnee S, Wagner S, Hamburger M (1981) Research and Innovation in the Modern Corporation: Conclusions. In: Rothberg RR (ed) Corporate Strategy and Product Innovation. Norton, New York, pp 416–427
- Meieran ES (1996) Kostensenkung in der Chip-Fertigung. Siemens Z Special FuE Frühjahr:6–10
- Milling P (1986) Diffusionstheorie und Innovationsmanagement. In: Zahn E (ed) Technologie- und Innovationsmanagement. Duncker & Humboldt, Berlin, pp 49–70
- Milling PM (1986) Decision Support for Marketing New Products. In: Aracil J, Machuca JAD, Karsky M (eds) System Dynamics: On the Move. The System Dynamics Society, Sevilla, Spain, pp 787–793
- Milling PM (1987) Manufacturing's Role in Innovation Diffusion and Technological Innovation. In: Proceedings of the 1987 International Conference of the System Dynamics Society. The System Dynamics Society, Shanghai, China, pp 372–382
- Milling PM (1989) Production Policies for High Technology Firms. In: Murray-Smith D, Stephenson J, Zobel RN (eds) Proceedings of the 3rd European Simulation Congress. Society for Computer Simulation International, Edinburgh, pp 233–238
- Milling PM (1991) An Integrative View of R&D and Innovation Processes. In: Mosekilde E (ed) Modelling and Simulation. Simulation Councils, San Diego, pp 509–514
- Milling PM (1991) Quality Management in a Dynamic Environment. In: Geyer F (ed) The Cybernetics of Complex Systems – Self-organization, Evolution, and Social Change. InterSystems Publications, Salinas, pp 125–136
- Milling PM, Maier FH (1993) Dynamic Consequences of Pricing Strategies for Research and Development and the Diffusion of Innovations. In: Zepeda E, Machuca JAD (eds) The Role of Strategic Modeling in International Competitiveness – System Dynamics '93. The System Dynamics Society, Cancun, Mexico, pp 358–367
- Milling PM, Maier FH (1993) The Impact of Pricing Strategies on Innovation Diffusion and R&D Performance. Syst Dyn Int J Policy Model 6:27–35
- Norton JA, Bass FM (1987) A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. Manag Sci 33:1069–1086
- Norton JA, Bass FM (1992) Evolution of Technological Generations: The Law of Capture. Sloan Manag Review 33:66–77
- Paich M, Sterman JD (1993) Boom, Bust, and Failure to Learn in Experimental Markets. Manag Sci 39:1439–1458
- Pearl R (1924) Studies in Human Biology. Williams and Wilkins, Baltimore
- Robinson B, Lakhani C (1975) Dynamic Price Models for New Product Planning. Manag Sci 21:1113–1122
- Roberts EB (1964) The Dynamics of Research and Development. Harper and Row Publishers, New York Evanston London
- Roberts EB (1978) Research and Development Policy Making. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 283–292
- Roberts EB (1978) A Simple Model of R&D Project Dynamics. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 293–314
- Rogers EM (1983) Diffusion of Innovation, 3rd edn. The Free Press, New York
- Schumpeter JA (1961) Konjunkturzyklen – Eine theoretisch, historische und statistische Analyse des kapitalistischen Prozesses, Erster Band. Vandenhoeck & Ruprecht, Göttingen
- Steele LW (1989) Managing Technology. McGraw-Hill, New York
- Sterman JD (1989) Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Environment. Manag Sci 35:321–339
- Sterman JD (1994) Learning in and about Complex Systems. Syst Dyn Review 10:291–330
- Weil HB, Bergan TB, Roberts EB (1978) The Dynamics of R&D Strategy. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 325–340

Discovery Systems

PETRA POVALEJ, MATEJA VERLIC, GREGOR STIGLIC
Faculty of Electrical Engineering and Computer Science,
University of Maribor, Maribor, Slovenia

Article Outline

Glossary

Definition of the Subject

Introduction

Knowledge Discovery and Data Mining Process

Application Domain Understanding

Data Understanding

Data Preparation and Identification of DM Technology

Applying Data Mining

Interpretation and Evaluation of Results

Utilization of Results

Knowledge Discovery Frameworks and Tools

Conclusions and Future Directions

Bibliography

Glossary

Accuracy (rate) Used for evaluating quality of induced model.

Average class accuracy One of the simplest metrics for estimating the quality of a model. Classification accuracy is calculated for each class of the target variable and then the average of all accuracies per class is calculated.

Aggregation Process of combining two or more objects into single one. Typical statistical aggregation functions for quantitative attributes are sum and average.

Attribute A property or characteristic of data object, which may vary in time and also from object to object. Attributes have usually assigned values or symbols for the purpose of analysis. Other frequently used names for an attribute are variable and feature.

Binarization Transformation of continuous or discrete attributes into binary attributes. Binary attributes have only two possible values.

Classification Classification of data objects is a process of assigning classes or class labels to data objects. It is a type of predictive modeling and it is used for predicting discrete target variable.

Classification accuracy See description under *Accuracy (rate)*.

Classifier A model based on data used for classification.

Confusion matrix A matrix of results from testing model versus predicted class values. It is very useful visual tool for understanding results of testing a classification model.

Data cleaning Step of KDDM usually involving detection and correction of data quality problems, removal of noise, defining outliers, and dealing with missing values.

Data mining A technology combining traditional data analysis methods and sophisticated algorithms for automatically processing large volumes of data and finding and extracting novel, useful and usually hidden patterns.

Data object A record of attribute values about an object or person. Other common names for data object are data record, case, point, sample, observation or entity.

Data preprocessing/preparation A phase of KDDM related to the preparation and transformation of data for data mining. It comprises of several techniques for selecting relevant data and attributes and creating or changing the attributes.

Data set A collection of similar or related data objects. Data objects are usually collected for a particular study.

Dimensionality reduction Reduction in the number of attributes. It is used for eliminating irrelevant features and noise by creating new attributes as a combination of the old attributes. Feature subset selection or feature selection is other type of dimensionality reduction, where dimensionality is reduced by selecting and using only a subset of old attributes.

Discretization A transformation of continuous numerical attributes into categorical or discrete attributes.

Ensemble Also known as committee or multiple classifier system is a group of classifiers. Ensemble approaches exploit the classification abilities of multiple classifiers. The integration of classifiers usually enhances the performance of final classification.

Feature A feature (variable) is a synonym for attribute. It is frequently used in data-mining domain. See *Attribute*.

Feature extraction A process of creating new features from the original raw data. It is highly domain-specific.

Knowledge discovery and data mining (KDDM) The “umbrella” term for the overall process of knowledge discovery.

Knowledge discovery (KD) Nontrivial process of mapping low-level data into other more meaningful forms that are easier to understand, like patterns, rules, summaries or even graphs.

Noise Result of erroneous measurements. It can involve distortion of values or addition of unauthentic data objects. Unlike outlier, noise is not legitimate data.

Outlier An anomalous object or atypical value of an attribute. Outliers can be legitimate data objects or values. Detecting outliers is especially important in fraud detection or network intrusion detection.

Pattern In KDDM defined as a high-level description of a subset of data and can be in many forms, e. g. statistical or predictive models of data, relationships among parts of data sets, association rules, clusters, graphs, summaries, or classification rules, tree structures, linear equations etc.

Precision Fraction of positive samples correctly classified as positive among all samples classified as positive.

Recall See *Sensitivity*.

Regression Regression is a type of predictive modeling used for predicting continuous target variable.

Sampling A process of selecting a subset of data or sample, for the data analysis. Basic sampling techniques are simple random sampling with or without replacement, stratified sampling and adaptive or progressive sampling.

Sensitivity (recall) Proportion of samples correctly classified as positive (true positives) of all positive samples tested. If sensitivity is 1, all positive samples have been identified as positive.

Specificity Proportion of samples classified as negative of all negative samples tested.

Definition of the Subject

By definition, to *discover* is to see, get knowledge of, learn of, find or find out; gain sight or knowledge of something

previously unseen or unknown [18], therefore a discovery system can be defined as a system that supports the process of finding new knowledge. Results of a simple query for *discovery system* on the World Wide Web returns different types of discovery systems: from knowledge discovery systems in databases, internet-based knowledge discovery, service discovery systems and resource discovery systems to more specific, like for example drug discovery systems [10], gene discovery systems [43], discovery system for personality profiling [48], and developmental discovery systems [17] among others. As illustrated variety of discovery systems can be found in many different research areas, but we will focus on knowledge discovery and knowledge discovery systems from the computer science perspective. Inconsistent definitions of terms knowledge discovery (abbreviated, KD), knowledge discovery in databases (abbreviated, KDD) and data mining (abbreviated, DM) frequently confuse potential a novice in the field of knowledge discovery.

DM is a technology combining traditional data analysis methods and sophisticated algorithms for automatically processing large volumes of data and finding and extracting novel, useful and usually hidden patterns. DM is also associated with other names, such as knowledge extraction, information harvesting, data archaeology, data pattern processing, and knowledge discovery in databases. DM and KDD are often used as synonyms, although DM is an integral part of KDD.

KD is a process of mapping low-level data into other more meaningful forms that are easier to understand [21] and knowledge discovery systems are systems that implement knowledge discovery process. Also, KD can be defined as a process that seeks new knowledge about specific problem or application domain.

KDD is a KD process applied to databases [30]. In the most popular definition by Fayyad et al. [21] KDD is defined as a non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in data. Note that although databases are considered as a primary source of data in KDD, the definition itself is not limited to one type of data sources only.

The umbrella terms *knowledge discovery* and *data mining* (abbreviated, KDDM) was proposed as the most appropriate name for the overall process of KD [30] instead of *knowledge discovery process* (abbreviated, KDP) and it will be used from now on to describe the overall process of knowledge discovery from different sources of data including methods for storing and accessing data, scalability of algorithms for large data sets, interpretation and visualization of results, and modeling of human-machine interaction [21].

The importance of KD systems lies in discovering the knowledge that may otherwise remain hidden and to integrate extracted knowledge into decision support systems used by scientist and businesses. The need for automated or at least semi-automated data analysis arose when computers revolutionized our way of living and making business. The digital information era brought advancements and benefits but also a very serious pitfall - data overload or data flood. Waste amounts of data, currently measured in terabytes, are collected daily, much more than human analysts can examine in foreseeable time. Still, the importance of human involvement in KD is not to be neglected. KD can only assist people with finding hidden patterns and relationships in data; it cannot tell whether discovered patterns are valuable for the organization.

Introduction

Knowledge discovery is not a new approach to analyzing data. People soon discovered that data need to be analyzed to make sense. The term *knowledge discovery in databases* was introduced in 1989, when first workshop on KDD was organized [35]. Soon after that, in 1991, Frawley [23] published a definition of KDD, which was later revised by Fayyad et al. [21]. The revised definition of KDD became very popular in KD community.

Classical approach to data analysis involves one or more analysts who need to be closely acquainted with the data. The analysts act as an interface between the data and the end-users of gained knowledge. Although this approach is perhaps appropriate for small amounts of data or with data with small number of variables or attributes, it has many drawbacks if larger amount of data or large number of variables are involved. Unfortunately, these drawbacks cannot be afforded in the modern high-paced business world. First, analyzing large amounts of data manually is too slow to effectively apply the results of the analysis. Second, this approach is too expensive because experts need to be very familiar with the data (their expertise is valuable) and third, the level of subjectivity of results is too high. Hence, the urge for at least semi-automated data analysis became even stronger.

Several computational approaches to data analysis have been proposed as an alternative to the classical approach. They are based on the idea that while on one hand computers contribute to data overload, they can on the other hand assist humans in extracting useful knowledge by finding meaningful patterns from growing volumes of data. For example, statistical, artificial intelligence and machine learning and even cognitive approaches can all deal in the same problem domain but from different perspec-

tives and in different ways using different models, methods, and techniques.

In contrast to single-disciplinary approaches, knowledge discovery in databases is an interdisciplinary approach. It evolved from several research fields and it incorporates many techniques and findings from statistics, artificial intelligence, machine learning, pattern recognition, databases, and data visualization among others. All those disciplines are combined in KDD with one common goal: extracting high-level knowledge from low-level data.

Before the beginning of the 21st century Piatetsky-Shapiro [36] had already identified three generations in KD systems. First-generation KD systems (early-1990s) were intended for expert users and provided only single data mining technique for data analysis per system. They had very weak support for the overall KDDM process and had only limited commercial success. The main research was focused on the development of new data mining algorithms. Second-generation KD systems (mid-1990s), called suites, were collections of multiple data analysis methods with the support for data cleaning, preprocessing and visualization. They were certainly a step forward, but the idea of KDDM process model was still not implemented. Third-generation KD systems (late-1990s) introduced different approach by addressing specific business problems like fraud detection. They provided interface to hide complexity of underlying data mining techniques and introduced first KDDM process models. Now, when we are approaching the second decade of the 21st century, it is possible to identify new generation of KD systems. This new generation of KD systems aims at integration and interoperability of modern KDDM models through use of popular industrial standards like eXtensible Markup Language (abbreviated, XML) and Predictive Model Markup Language (abbreviated, PMML) or some other approaches [30].

After the first workshop on KDD the idea of KDDM model evolved simultaneously with the development of KD systems. In 1996 Fayyad et al. [22] laid the foundation for KDDM process model by proposing the basic structure of the model. After the publication of their book the development of KDDM models followed two different streams: data-centric and human-centric. Data-centric (or data-driven) models focused on the iterative and interactive nature of the data analysis task, while the human-centric models focused on a series of knowledge-intensive tasks with complex interactions between human and the database [30]. Furthermore, the models can be broadly divided into academic, usually not considering industrial aspects of KDDM projects, and industrial, which address specific industrial issues. All process models con-

sist of multiple sequential steps with loops and interactions; they differ in the number and the scope of proposed specific steps. Also, all models emphasize the iterative nature of the model.

Although many models have been developed, only five of them left significant impact and were applied in at least several real KDDM projects [30]. The nine step academic data-centric KDDM model by Fayyad et al. [22] in 1996 was the first reported model. In the same year the CRISP-DM (Cross-Industry Standard Process for DM) model was proposed by a consortium of four companies: SPSS, NCR, Daimler Chrysler and OHRA, but its mature version was released in 2000. The second model, this time from industrial area, was suggested in 1998 by Cabena et al. [9] and it consists of five steps. Third model was an eight-step academic model by Anand and Buchner [1], developed at about the same time as the second model. In 2000 the fourth model, already mentioned CRISP-DM model consisting of six steps, was officially released [44]. This model is still strongly supported by the industry, CRISP-DM Special Interest Group and the European Commission. The last, fifth model consisting of six steps, was proposed by Cios et al. [11]. This model adopted CRISP-DM model to the needs of academic research community [30]. The study by Kurgan and Musilek [30] offers detailed description and an exhaustive comparison of the major five existing KDDM models. They also introduced a generic model, which tries to capture main steps and tasks of the mentioned KDDM models. In following sections and subsections this generic model was used in order to avoid favoring any particular KDDM model. Using KDDM models for knowledge discovery is essential to ensure that useful knowledge is discovered; blind application of data mining methods, notoriously known as data dredging or data fishing can easily lead to the discovery of potentially interesting but meaningless patterns.

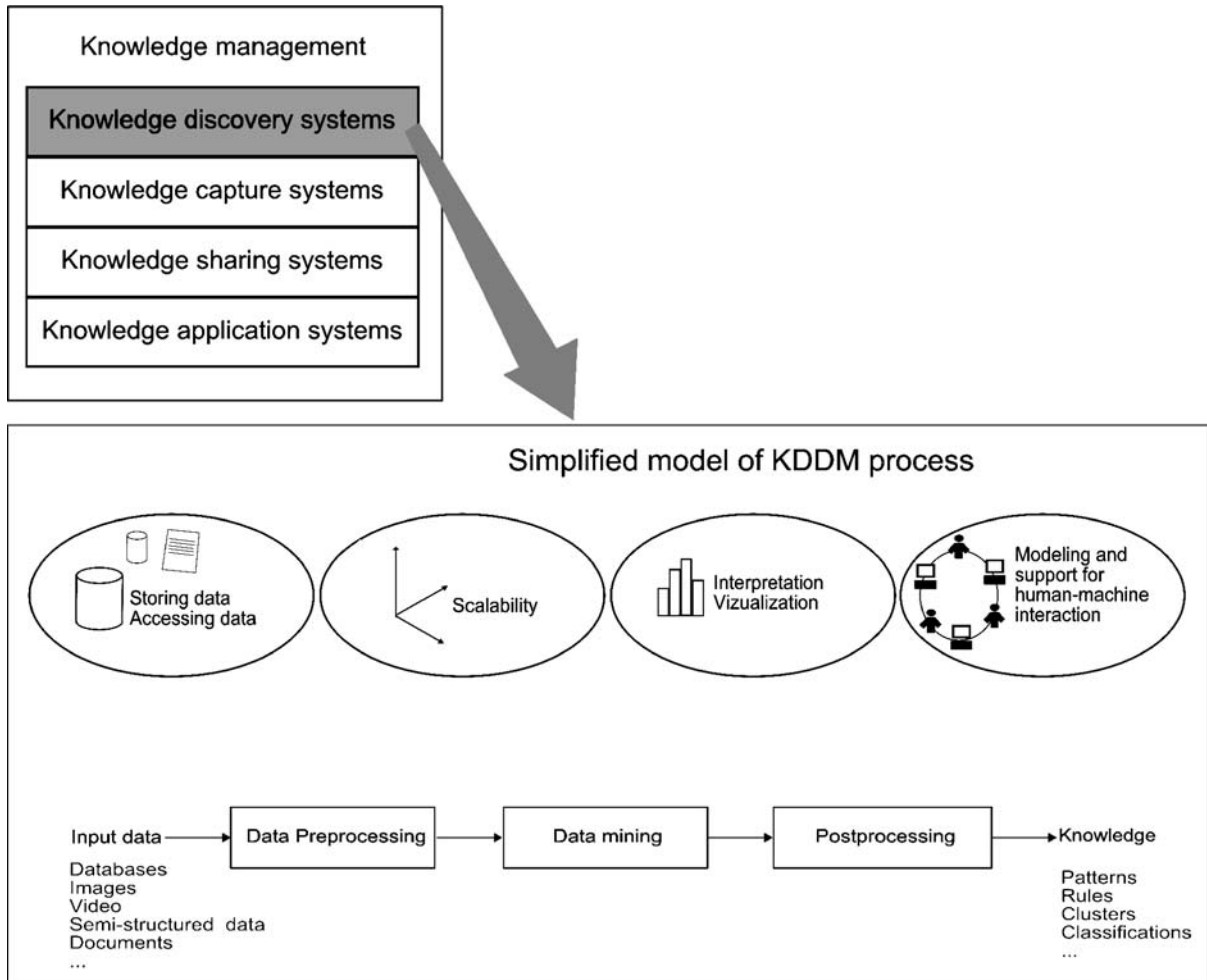
Other pioneers, who were not explicitly mentioned above but also contributed significantly to the development of contemporary KDDM, are P. Smyth [21,26] and H. Manilla [26,32].

In the following sections knowledge discovery process will be described in more details. Each step of the process will be briefly explained to show how a particular step contributes to the results of the entire KDDM process. Detailed description of the KDDM process will be followed by a section on KDDM frameworks, which will include a review of contemporary frameworks. In next section several applications of KDDM will be mentioned so that we can recognize the real value of applying KD to the real world problems. In the last section some guidelines for the future development of KDDM will be suggested.

Knowledge Discovery and Data Mining Process

The simplest definition of KDDM is an overall process of transforming raw data into knowledge. As already mentioned in the introductory section, KDDM process includes, among others, methods for this transformation, which is the central part of KDDM process. Fayyad et al. [21] defined this transformation as KDD and described it as ‘the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data’. Before we describe KDDM in more details, it is important to understand the meaning of basic terms in this definition. **Data** is a set of facts and a **pattern** is an expression (high-level description) that describes a subset of the data or a model applicable to this subset. Patterns, as mentioned in definition, can take many forms;

for example, they can be statistical or predictive models of data, relationships among parts of data sets, classification rules, association rules, summaries, linear equations, clusters, graphs, tree structures, or recurrent patterns in time series. In this context a pattern is considered as **knowledge** if it exceeds a certain level of **interestingness**, usually set by user or even data itself. The notion of interestingness was introduced by Fayyad et al. [21] and it represents an overall measure of pattern’s value, novelty, usefulness, and understandability. Because functions and thresholds for measuring interestingness are chosen by user, this definition of knowledge is purely user oriented and domain specific. As we may intuitively know, knowledge is much more than just patterns or relationships in data. For instance, researchers from the field of knowledge management define knowledge as a mix of experience, values and



Discovery Systems, Figure 1

Relation between knowledge management, knowledge discovery systems, and data mining

insights that provides some sort of framework for evaluating new experiences and information. It is embedded not only in documents but also in organizational routines, processes, practices and norms [14]. This broader definition of knowledge is certainly better than the definition of knowledge as specific patterns, but unfortunately it is not very useful in the context of KDDM.

Figure 1 shows the relationship between knowledge management, knowledge discovery systems, which implement KDDM process, and data mining. KDDM process is essential part of an even broader process named knowledge management. While knowledge management deals with aspects of gathering, organizing, sharing and analyzing intellectual capital – knowledge [4], KDDM focuses on the overall process of knowledge discovery from data, including methods for storing and accessing data, scalability of algorithms for large data sets, interpretation and visualization of results, and modeling of human-machine interaction. Simplified model of KDDM process clearly illustrates the central role of DM in KDDM process.

KDDM is a process of several steps which transforms inputs in form of raw data into outputs – applicable knowledge. Different KDDM models define steps involved in this process, but the number and the scope of particular step vary from model to model. We decided to describe the process of KDDM using the generalized model of KDDM process proposed by Kurgan and Musilek [30]. Figure 2 shows the steps of generalized model around the schematic diagram of general sequential structure of KDDM process [12] in the center. The sequential structure includes

many feedback loops between steps, because the process of revision can show that some of the steps need to be repeated.

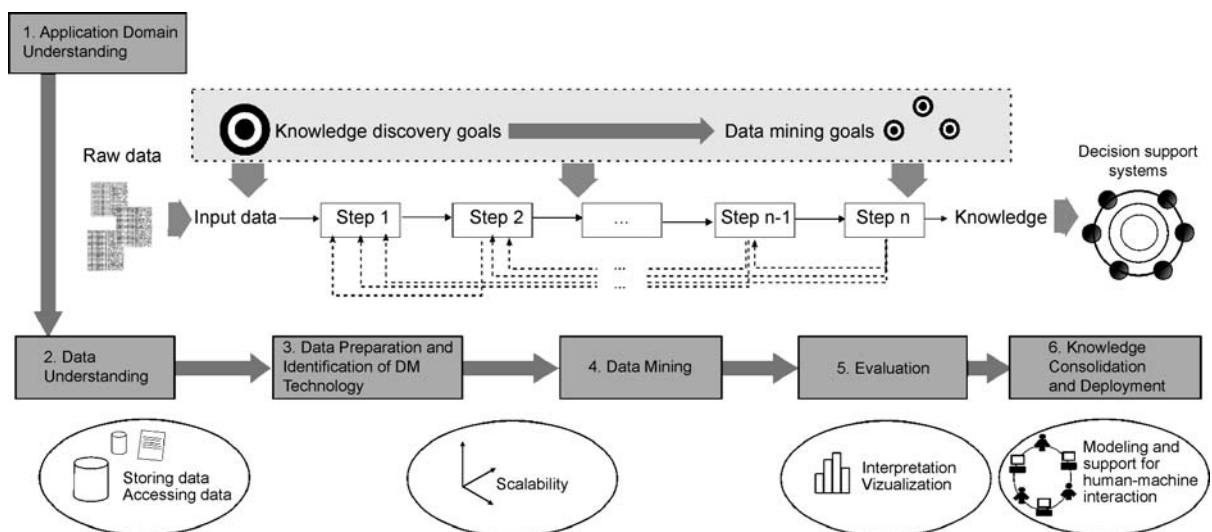
The generalized KDDM model has six steps:

- (1) Application domain understanding,
- (2) Data understanding,
- (3) Data preparation and identification of DM technology,
- (4) Data mining,
- (5) Evaluation and
- (6) Knowledge consolidation and deployment.

Each step will be described in more details in the following sections.

Application Domain Understanding

Before analyzing data it is recommended to define the goals or objectives of the analysis. First, we have to understand the domain objectives that are important for the end-user, for example a customer. We have to carefully consider constraints, assumptions and any other factors that may affect our search for new knowledge. Second, some project management efforts are needed to identify human and technical resources, decompose KDDM project into tasks, estimate the duration of each task, define inputs, outputs, and dependencies to successfully apply data mining techniques. Each task should also be associated with particular KD goal. Third, at this point business goals or KD goals need to be translated into more spe-



Discovery Systems, Figure 2
Generalized model of KDDM process

age age of the patients, it is senseless to calculate average ID, because the purpose of using IDs is to distinguish data objects. Also, it can happen that an attribute is measured in such a way that measurements do not capture all the properties of the attribute. Thus, knowing the type of an attribute is important. It tells us which properties of the measured values are consistent with the properties of the attribute, so we can avoid senseless actions.

According to the properties of numbers that correspond to the properties of attribute we can define four types of attributes: nominal, ordinal, interval, and ratio. **Nominal** attributes are described with symbols or numbers which distinguish one object from another. Nominal attributes from the example in Table 1 are identification number, gender and name. **Ordinal** attributes provide information to order the objects. Examples of ordinal attributes are level of tiredness and blood sugar. For **interval** attributes, the difference between values are important and unit of measurement exists. An example of interval attribute is body temperature in the Celsius scale. For **ratio** attributes, both differences and ratios are important. Examples of ratio attributes are age, height and weight.

Nominal and ordinal attributes are **categorical** or **qualitative** attributes, while interval and ratio attributes are **quantitative** or **numeric** attributes. Qualitative attributes take on values that are names or labels and do not have properties of numbers. Even if their values are numeric, they should be treated more like symbols. Quantitative attributes are represented by numbers and have most of the properties of numbers.

If an attribute can have any value between two specified values, then it is **continuous**; but if it can take only a limited or finite number of values it is **discrete**. Example of continuous attributes is weight, while gender is an example of discrete attribute.

Discrete attributes are often represented using integer values. **Binary attributes** are special case of discrete attributes. They can have only two values like yes/no, true/false, male/female or 1/0 and are usually represented with Boolean variables or as integer variables. Gender, level of tiredness and blood sugar from Table 1 are ex-

table is a record and attributes are person identification number, and body temperature

Person ID	Name	Age	Gender	Height (cm)	Weight (kg)	Level of tiredness	Blood sugar	Body temp. (°C)
-----------	------	-----	--------	-------------	-------------	--------------------	-------------	-----------------

[illegible]

amples of discrete attributes. Continuous attributes have values which are real numbers and are typically represented as floating point variables. Height, weight and body temperature from Table 1 are examples of continuous attributes.

Different types of data sets exist. Most common types are data sets with record data (e.g. transaction data, data matrix, and sparse data matrix), graph-based data (e.g. data with relationships, data as graphs) and ordered data (e.g. sequential data, time series data, and spatial data). Data sets may even display some special characteristics. For example, when we are working with time series data, we should consider temporal autocorrelation; or when we are dealing with spatial data, we should not ignore spatial autocorrelation. Furthermore, some data sets might even include explicit relationships within data (e.g. linked documents in the WWW). Many data sets share three general characteristics that significantly affect data mining techniques: dimensionality, sparsity, and resolution.

The **dimensionality** of a data set is the number of attributes that are used to describe data. Sometimes it is possible to come across the term **curse of dimensionality**. This term is used for difficulties that usually occur during analysis of high-dimensional data and therefore dimensionality reduction is an important part of data pre-processing phase.

In some data sets only few attributes of an object have non-zero values. In this case we can talk about **sparsity**. Attributes, for which only non-zero values are important, are called **asymmetric attributes**. Sparsity in data set can be considered as an advantage because we don't have to store, manipulate or analyze non-important (zero) values. Some data mining algorithms work well only for sparse data sets.

Data is often obtained at different levels of resolution. The properties of data at different resolution levels are different. For example, we measure weight of babies in grams, while the weight of adults is measured in kilograms. If the resolution is too fine, a pattern might not be visible or too much noise is present; on the other hand, if resolution is too coarse, the pattern may disappear.

In this phase we also have to deal with several data-related issues: the types of data in our selected data sets, the quality of data, how to improve data quality or how to modify data for the needs of data mining methods. Data is never perfect: values may be missing, duplicate objects may exist, inconsistencies may occur due to human error, hardware and software limitations, or flaws in the data collection process. Detection and correction of data quality problems is often called **data cleaning**. When dealing with data quality we have to consider measure-

ment and data collection issues and also issues related to applications. We will briefly describe some of most frequently occurring data errors: noise, outliers and missing values.

We can encounter a variety of problems related to **measurement error**, when recorded value differs from 'true' value to a certain extent, and **data collection error**, which refers to omitting data objects or attribute values or including data object in inappropriate way. Both errors can be systematic or random. Good news is that for some types of data errors well-developed techniques for detecting and correcting these errors exist.

Noise is an example of measurement error. Often it occurs randomly and it may involve distortion of values or addition of unauthentic objects. Usually, the term noise is used in connection with temporal or spatial data, but is not uncommon in other types of data as well. The elimination of noise is difficult, therefore **robust** (insensitivity for noise) data mining algorithms are desired. Precision, bias and accuracy are also important factors for measuring the quality of data.

Outliers are data objects that are different from most of the other data objects or values of an attribute that are atypical for this attribute. Outliers are considered anomalous objects or values. It is important to distinguish between noise and outliers. Unlike noise, outliers can be legitimate data objects or values and may sometimes be interesting, especially in fraud detection or network intrusion detection.

Missing values are not uncommon in data sets. A data object can have one or more missing attribute values; sometimes the values were not collected or attributes are not applicable to all objects or attributes are optional. Usually, for the sake of simplicity, all attributes (fields) are stored, but during the data analysis all missing values should be included.

The problem of **duplicate data** should also be taken into an account in data cleaning. Data duplication can occur due to inconsistent values and these values must be resolved. However, duplication can also occur due to accidentally combining data objects, that are similar but not the same. For example, two or even more persons can have identical names.

Data Preparation and Identification of DM Technology

Data pre-processing phase is related to preparation and transformation of data into more suitable form for data mining. This particular step is by far the most time-consuming part of the KDDM process [12].

Data pre-processing strategies and techniques can be roughly divided into two groups – strategies and techniques for:

- (1) Selecting relevant data object and attributes, and
- (2) Creating or changing the attributes.

We will familiarize with terms aggregation, sampling, dimensionality reduction, feature subset selection, feature creation, discretization, binarization, and variable transformation. In data pre-processing the term feature or variable is frequently used as a synonym for attribute.

The idea behind **aggregation** is to combine two or more objects into a single object. Typically, quantitative attributes are aggregated by simple statistic functions like a sum or an average. Qualitative attributes are more difficult to deal with. They can be either omitted or summarized in some way. If aggregation is used, then resulting smaller data sets require less computer resources (memory and processing time) and thus more expensive data mining techniques can be used. Aggregation can also serve as a high-level view of the data. However, aggregation can also lead to potential loss of interesting details.

Sampling is an approach for selecting a subset of data objects for the data analysis. It has been long used in statistics and it can be also very useful for data mining. In statistics sampling is used because obtaining the entire set of data of interest is too expensive or even impossible; in data mining sampling is used because processing of all the data would be too expensive or time consuming. Sampling can, similar to aggregation, reduce the size of data set and thus more expensive data mining algorithms can be used. Sampling is effective only if the sample is **representative**; in other words, a sample should have very similar or even identical characteristics as the original (entire) set of data or **population**. Different sampling approaches have been developed. One of the basic sampling techniques is **simple random sampling**, which has two variations:

- (1) Sampling without replacement and
- (2) Sampling with replacement.

In random sampling all items in the population have equal opportunity to be selected. In sampling without replacement an item that was selected is removed from the population. In sampling with replacement selected item is not removed from the population, so the same item can be selected more than once. Because simple random sampling is not suitable for less frequent types of objects (rare cases), different sampling method is needed for such data sets. **Stratified sampling** takes into account differing frequencies of objects types. It starts with predefined groups of objects. In one version equal number of objects is drawn

from each group, regardless of potentially different sizes of groups. In another version, the number of objects drawn from each group is proportional to the size of corresponding group. After selecting sampling technique, we have to decide on the **sample size**. Sample size is very important, because if we choose large sample sizes the probability of having a representative sample is higher, but we lose advantage of sampling. On the other hand, if we use a smaller sample size we may miss some patterns or even detect erroneous patterns. Because defining proper sample size can be difficult, we can sometimes use **adaptive** or **progressive sampling**. In this kind of approach, sample size is progressively increased until sufficient size is obtained, but we need a way to determine if the sample size is large enough.

Data sets can have a large number of attributes or features, even tens of thousands attributes. Because many data mining algorithms are more suitable for data sets with smaller number of attributes, reduction in the number of attributes or **dimensionality reduction** is desired. Dimensionality reduction has many benefits, a key one is that data mining algorithms work better with lower dimensionality of data. This is partly due to elimination of irrelevant features and noise reduction. Dimensionality reduction can also lead to more understandable model with fewer attributes, data can be more easily visualized and data mining algorithms require less computer resources. Dimensionality reduction techniques usually reduce the dimensionality by creating new attributes as a combination of the old attributes. Dimensionality of data sets can be also reduced by selecting and using only a subset of old attributes. This reduction is known as **feature subset selection** or **feature selection**. Redundant features duplicate much or all information contained in one or more features while **irrelevant features** contain little or no information that could be used in data mining. Redundant and irrelevant features can even reduce classification accuracy and the quality of clustering and eliminating them can only be beneficial. Some of features can be eliminated immediately, but otherwise selecting the best subset of features requires a systematic approach. Three standard approaches to feature selection exist: embedded (feature selection occurs naturally as a part of data mining), filter (features are selected before data mining), and wrapper (learning algorithm itself is used as part of the evaluation function for selecting features). Sometimes we want to create new set of attributes from original ones, especially if new features capture important information more effectively. Creating new features from the original raw data is known as **feature extraction**, but unfortunately this approach is highly domain-specific. For some cases data need to be mapped to a new space to reveal interesting and important features,

for example, using a Fourier transformation in time series analysis. In situation, when the features in the original data set have the necessary information but they are not in the form suitable for data mining, constructing one or more features from the original ones can be more useful. The most common approach to **feature construction** is using domain expertise.

Several data mining algorithms, especially in classification or association analysis, require data in the form of categorical attributes or binary attributes. Normally attributes in data sets are not only categorical or binary and transformation of attributes is needed. Transforming continuous attributes into categorical attributes is **discretization** and transformation of continuous or discrete attributes into binary attributes is **binarization**. There are many different approaches to discretization and binarization, but they all try to produce the best result for data mining algorithm we intend to use for the data analysis. Another type of transformation is **variable transformation** that is applied to all values of an attribute or variable. Two important types of variable transformations are simple functional transformations and normalization. For **simple functional transformations** a simple mathematical function is applied individually to each value. For example, if x is a variable such simple functions are x^k , $\log x$, e^x , \sqrt{x} , $1/x$, $\sin x$, or $|x|$. Note that variable transformations should be applied with caution, because they change the nature of data (e.g. changing magnitudes, order, or negative values). Standardization or normalization of a variable is concerned with transforming entire set of values in such a way, that they share particular property. This transformation is often necessary if are combining different attributes (variables) in some way and we want to omit the dominance of large values.

Some researchers recommend choosing DM technology before data preparation, others say that methods should be selected we have prepared, but in practice these two things are intermingled. Data need to be prepared at least in such a way that we can decide which DM methodology is most appropriate for selected DM task. Among core DM tasks are predictive modeling, anomaly detection, association analysis and cluster analysis.

With **predictive modeling** we build a model for the target, or dependent variable as a function of other independent variables. Well-known types of predictive modeling are **classification**, which is used for predicting discrete target variable, and **regression**, which is used for predicting continuous target variables. The goal of predictive modeling is to minimize the error between predicted and true values of the target variable [47]. **Anomaly detection** is identification of data samples, known as **anoma-**

lies or **outliers** that are significantly different from the rest of data. The main goal of this task is to identify the real outliers not the noise. **Association analysis** is used to discover strongly associated features and patterns that occur in data because of these associations. Discovered patterns are usually in the form of implication rules or feature subsets. The main goal of this analysis is efficient extraction of the most interesting patterns, for example, which items are frequently bought together by customers. **Cluster analysis** is a task of finding groups or **clusters** of data samples. Data samples in a cluster are more similar to each other than samples of other clusters.

After identifying DM tasks, we can choose one or more of appropriate algorithms for selected DM task.

Applying Data Mining

When applying DM process we can generally pursue two different types of objectives: either to verify some user predefined hypotheses (**verification**) or find new knowledge upon which a user can base decisions using (**knowledge discovery**). Both verification and knowledge discovery are done by building a model of a real world based on collected data in a specific application domain. The result of model building is a description of patterns and relationships in data, which can be used for verification of predefined knowledge (hypothesis), discovery of new knowledge for either **prediction**, where the system autonomously finds patterns for predicting future behavior of some entities (class attributes), or **description**, where the system finds patterns for a presentation in a human-understandable form. The boundary between prediction and description is blurred – some prediction models can also be understandable to some degree.

After selecting specific goals of DM process and the type of prediction, which is the most important for mining the data from the data set, a DM model type can be chosen. Several traditional statistical models exist, for example discriminant analysis, general linear models and logistic regression, but there are also some models from the fields of machine learning and pattern recognition, such as neural net to perform regression, a decision tree or a decision rules for the classification.

Furthermore, several algorithms are available for building selected model. For example, we may build a neural net using backpropagation or radial basis functions; for inducing a decision tree we can choose among C4.5 [39], CART (classification and regression trees) [7], CHAID (CHI-squared automatic interaction detector) [29], QUEST (quick, unbiased, efficient, statistical tree) [31] and many other algorithms. Algorithms tend

to differ primarily in the goodness-of-fit criterion used to evaluate model fit or to find the good fit in a process of searching.

In predictive models, the values (classes) we are predicting with a model are called **response/dependent/target variables** and the values used as an input to build a model are called **prediction/independent variables**. The process of building predictive models requires a well-defined training and validation protocol in order to ensure the most accurate and robust predictions. This kind of protocol is called **supervised learning**. The idea of supervised learning is to *train* a model on a portion of data from the data set (**a training set**), then test and validate it on remainder of the data (**test set**). For each observation in a data set values predicted by the model are compared with actual (known) values of the response variable. For contrast, some of the descriptive techniques like clustering can build a model without known values of the response variable. These types of techniques are sometimes referred to as **unsupervised learning**.

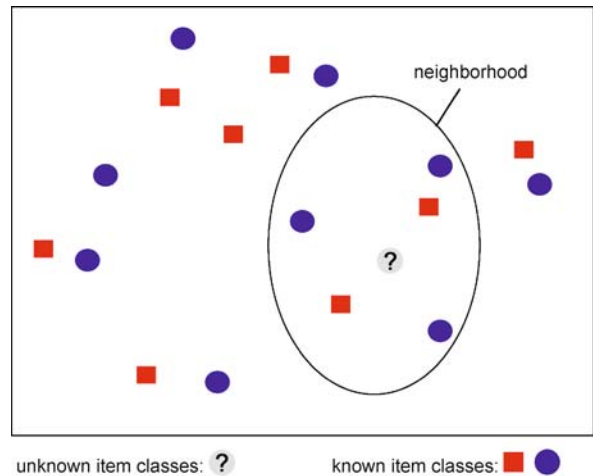
It is important to remember that no model or algorithm should be used exclusively. Model building is an iterative process. The nature of the problem and the data itself affect the choice of models and algorithms. There is no best model or algorithm in general. Selection of the most appropriate model type and algorithm can be crucial for the success of DM process. Therefore for a given problem different models should be explored to find the one producing the most useful solution.

In the following sections, some of the types of models and algorithms used to mine data will be briefly examined. In general, most of the model types can be viewed as an extension or a hybrid of some primary types of models. Advanced descriptions of models can be found in other sections of this chapter.

K-Nearest Neighbor (k-NN)

K-nearest neighbor (k-NN) is a classification technique that defines how to classify an observed item on the basis of the majority class of k items in the neighborhood (Fig. 3).

The idea of k-NN is based on measuring a distance between attributes in the database. If the attributes are numerical the calculation of distances is easy, but categorical attributes require special handling (e.g. what is the distance between blue and gray?). The next task is a selection of the neighborhood of pre-classified cases to be used for classification of a new case and possibly the influence of neighbors considering the distance from the new case. For example, nearest neighbors can have a higher in-



Discovery Systems, Figure 3

k-Nearest neighbor. ? has unknown class; ● and ■ are known classes of the neighbors. In most simple algorithm the class of unknown item ? would be ● since 3 items in the neighborhood have class ● and 2 items have class ■

fluence (weight) on classification of a new case than the neighbors that are farther away.

Since k-NN models are computationally demanding they are mostly useful when only a few prediction variables are included. Different data types of variables can be included into the model under only one basic requirement: the metric for measuring distance has to be defined. The advantage of k-NN models is their understandability.

Memory-Based Reasoning (MBR)

MBR is one of the memory based technologies that belongs to the family of nearest-neighbor-like models. It is used on databases with large number of cases where the data is kept in a system's memory in order to speed up computations. As in k-NN model, MBR algorithms search for the most similar problem from the database of pre-classified cases and apply its class to a new case. There is no learning phase and no adaptive or learning parameters to manipulate.

A basic advantage of MBR algorithms is their effectiveness (on a proper database) that is comparable (and sometimes even better) to neural networks. The disadvantages of MBR algorithms are: a lack of generalization (a large number of good examples are needed), no data compression, large computational demands and consequently large memory requirements that all make them unsuitable for on-line learning.

It has been often used for object recognition and classification of free text samples taken from very large databases.

Rule Induction

Rule induction algorithms derive a set of independent rules from the database of pre-defined cases to classify a new case. Decision rules can be in a simple form, like:

IF $condition_j$ **THEN** $CLASS = class_i$.

Example of a simple rule is: *IF body temperature is higher than 38 degrees THEN patient is ill*. More complex rules can include conjunction (logical AND) and/or disjunction (logical OR), like:

IF $condition_a$ **AND** $condition_b$ **OR** $condition_c$
THEN $CLASS = class_i$.

An example of a more complex rule is: *IF body temperature is higher than 37.5 degrees AND diarrhea is true THEN patient is ill*. More complex dependent rules can form a decision tree.

Rules can also include ELSE statement, like:

IF $condition_j$ **THEN** $CLASS = class_i$
ELSE $CLASS = class_k$.

An example of IF-THAN-ELSE decision rule is: *IF body temperature is higher than 37.5 degrees AND diarrhea is true THEN patient is ill ELSE patient is healthy*.

Disadvantages of rule induction that have to be considered are:

- (1) A set of induced rules does not necessarily cover all possible situations (it is possible that the system will not be able to classify a new case),
- (2) Two or more induced rules may conflict in their predictions. A usual solution to the second disadvantage is to assign a confidence to each rule and for classification of a new case use the most confident one or alternatively to use weighted voting of rules according to their confidence.

An important advantage of decision rule induction is certainly their simplicity and understandability.

Decision Trees

Decision trees are a way of representing a series of rules that lead to classification of known classes. For example, we might wish to diagnose young children with asthma on the basis of gathered data about family history, blood examinations and other indicators of asthma in adults. Fig-

ure 4 shows a simple decision tree that solves this problem while illustrating basic components of a decision tree: internal nodes, branches and external nodes (also called leaves).

Decision trees are induced on the basis of iterative splitting of data from the training set into discrete groups, where the goal is to minimize the 'distance' among groups at each split. Therefore one of the distinctions among decision tree algorithms is the splitting criteria. The most commonly used splitting criteria are: information gain and information gain ratio used by Quinlan [38] in ID3, C4.5 and C5.0 algorithms; chi square goodness of fit [46] used in CHAID (SPSS answer tree) [29]; gini index used by Breiman in CART algorithm [7]; j-measure or cross-entropy introduced by Smyth and Goodman [45], etc.

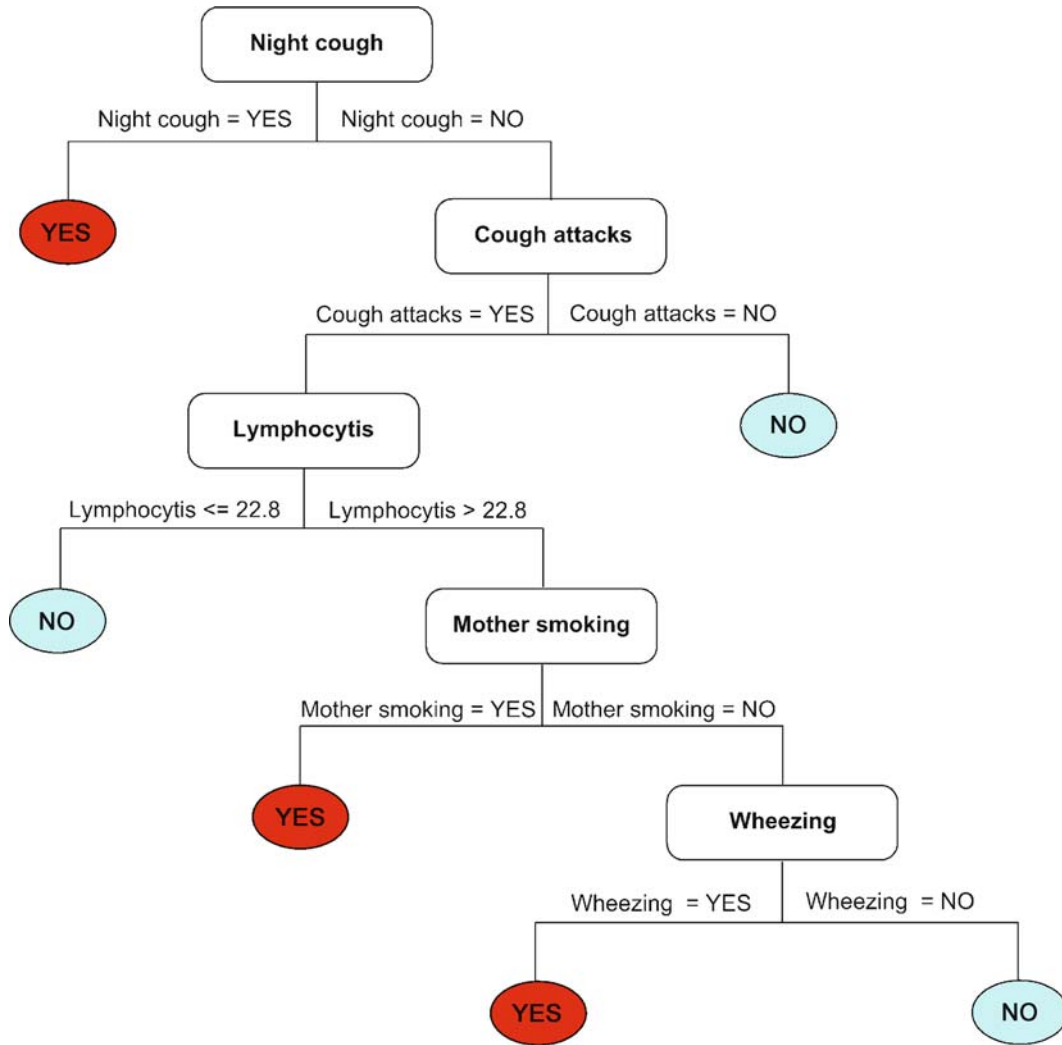
Decision trees models are extensively used in data mining for prediction purposes such as decision support and knowledge extraction in various fields. Decision trees provide very important advantage compared to many of other data mining models - the possibility of explaining the decisions in a way understandable by humans. A main disadvantage of classic decision tree induction is sensitivity to noise (missing or corrupted data).

Decision trees are able to process both numerical and categorical variables. However, on the basis of the type of prediction variable we distinguish between **classification trees** used to predict categorical variables, and **regression trees** for continuous prediction variable.

Neural Networks

A neural network is a powerful model modeled after the human brain. As the human brain consists of millions of neurons interconnected by synapses, neural networks are formed from large number of artificial neurons (nodes) connected to each other to form a network. It can be used to capture and represent complex relationships between prediction variables (input) and target variables (output) or to find patterns in data. Neural networks may be applied to classification problems (where the output is categorical variable) as well to regression (where the output variable is continuous).

A neural network (Fig. 5) starts with an **input layer**, where each node (neuron) corresponds to a predictor variable. Each input node is connected to every node in a **hidden layer**. The nodes in a hidden layer can be connected to nodes in another hidden layer or to an **output layer**. The output layer consists of one or more response variables. Each connection has a **weight** assigned which reflects strength of a connection between the connected neurons. Like in human brain the strength of a connection can



Discovery Systems, Figure 4
A simple decision tree

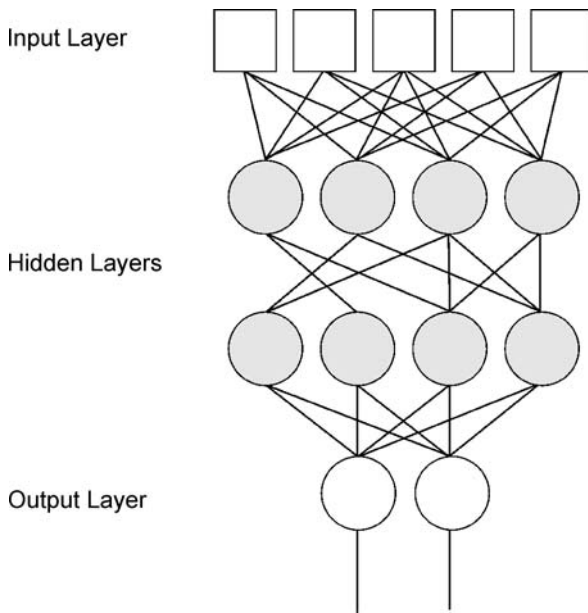
change in response to a presented stimulation or an obtained output, which enables the network to learn. After the input layer, each node (in the first hidden layer) gets a set of inputs which are then multiplied with connection weights and summed together to form a linear combination of values of nodes from the previous layer. An activation function is applied to the linear combination of values and the output is passed to the node(s) in the next layer.

In the beginning of model building (learning a network) the connection weights are unknown parameters which have to be learned on the basis of training data using one of the training algorithms. One of the most known training algorithms is the backpropagation algorithm [42,51], however newer methods include ge-

netic algorithms [41], quasi-Newton [15], conjugate gradient [49], and many others.

The mayor advantages of neural networks are: good results on large databases, low sensitivity to noise and data distribution (on large databases or when proper mechanisms for preventing overfitting are used), applicability to multivariate non-linear problems, easy to implement to run on parallel computers, applicability to several problem domains.

Neural networks also have some disadvantages that have to be mentioned. The complexity of neural network model makes it hard to interpret the results. They belong to a group of so called black box models which means that the relationships among prediction and target vari-



Discovery Systems, Figure 5

A neural network with two hidden layers

ables cannot be represented in humanly understandable (symbolic) form. Therefore they are often not applicable to problems that demand an explanation of induced models.

Second disadvantage of neural networks is their proneness to overfitting due to large number of adjustable parameters which can adopt to every data if the size of a neural network is large enough. In order to reduce overfitting

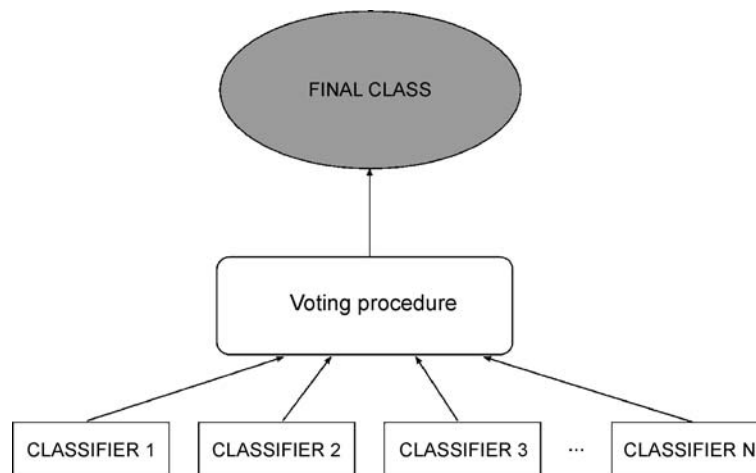
some mechanisms such as weight decay and/or cross validation have to be used.

The next disadvantage is an extensive amount of computational time needed for training a network (especially on large databases). However, as stated before they can be run on massively parallel computers with each node simultaneously doing its own calculations.

Neural networks have been successfully applied to a broad spectrum of data mining applications, such as: character recognition for optical character recognition (OCR) applications, target recognition used in military applications, machine diagnosis, medical diagnosis for image data (MRIs, X-rays), voice recognition, intelligent searching for internet search engines, fraud detection, quality control, etc.

Ensemble Approaches

In recent years there has been a growing interest in the area of combining classifiers into ensembles also known as **committees** or **multiple classifier systems** (Fig. 6). The intuitive concept of ensemble approaches is that no single classifier can claim to be uniformly superior to any other, and that the integration of several single approaches will enhance the performance of final classification [53]. Hence, using the classification capabilities of multiple classifiers, where each classifier may make different and perhaps complementary errors, tend to yield an improved performance over single classifiers. Some ensemble approaches are actively trying to perturb some aspects of the training set, such as training samples, attributes or classes, in order to ensure classifier diversity. One of the most



Discovery Systems, Figure 6

Combining classifiers into an ensemble

popular perturbation approaches are bootstrap aggregation (bagging) and boosting.

Bagging was first introduced by Breimen [6] in 1996 manipulates the training samples and forms replicate training sets. The final classification is based on a majority vote. Boosting, introduced by Freund and Schapire in 1996 [24] combines classifiers with weighted voting and is more complex since the distribution of training samples in training set is adaptively changed according to the performance of sequentially constructed classifiers.

In general ensemble approaches can be divided into three groups:

- (1) Ensemble approaches that combine different independent classifiers (such as: Bayesian voting, majority voting, etc.),
- (2) Ensemble learning approaches which construct a set of classifiers on the basis of one base classifier with perturbation of a training set (such as: bagging, boosting, windowing, etc.) and
- (3) A combination of (1) and (2).

A detailed empirical study is presented in [19].

Hybrid Approaches

The hybrid approaches rest on the assumption that only in the synergetic combination of single models can unleash their full power [28]. Each of the single model types has its advantages, but also inherent limitations and disadvantages, which must be taken into account when using the particular model type. Therefore the logical step is to combine different models (classic approaches) to overcome the disadvantages and limitations of a model.

According to the way of combining different models into a hybrid, four basic types of hybrids can be differentiated:

Sequential hybrid the inputs in the hybrid are fed as inputs in model 1. The outputs from model 1 represent inputs for model 2 and the outputs from model 2 are also outputs of the sequential hybrid (Fig. 7a). Both models in the hybrid are autonomous and can be also used independently.

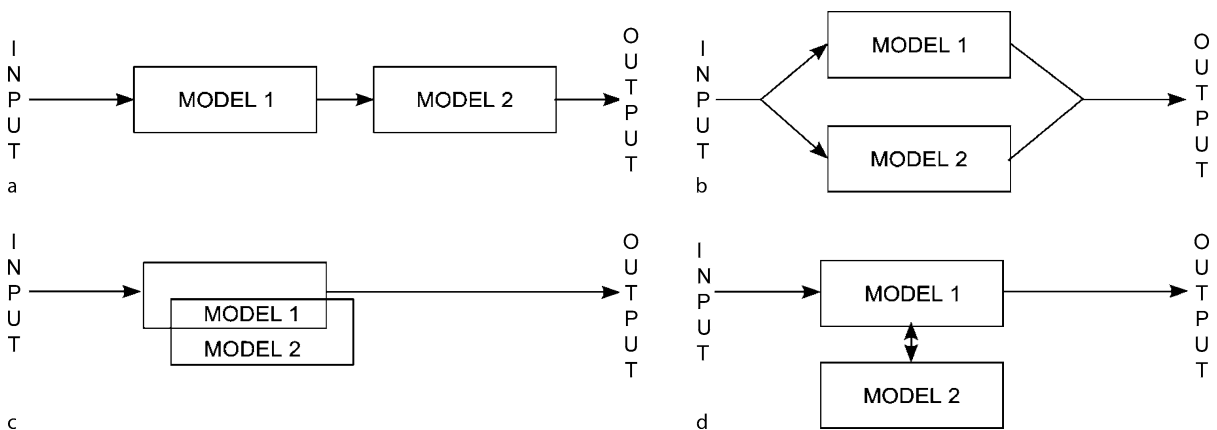
External hybrid model 1 represents the main method which forwards the data to model 2. Model 2 processes the data and sends the results back to model 1 (Fig. 7b). Model 2 totally depends on model 1 and its outputs cannot be used as hybrid outputs.

Embedded hybrid has the most powerful connection between both models which are so strong interlaced, that none of them can function independently (Fig. 7c).

Parallel hybrid models in the parallel hybrid are totally independent. They both operate on the same inputs, but produce separate outputs. The outputs of the hybrid are determined within the arbitrary mechanism, embedded in the hybrid (Fig. 7d).

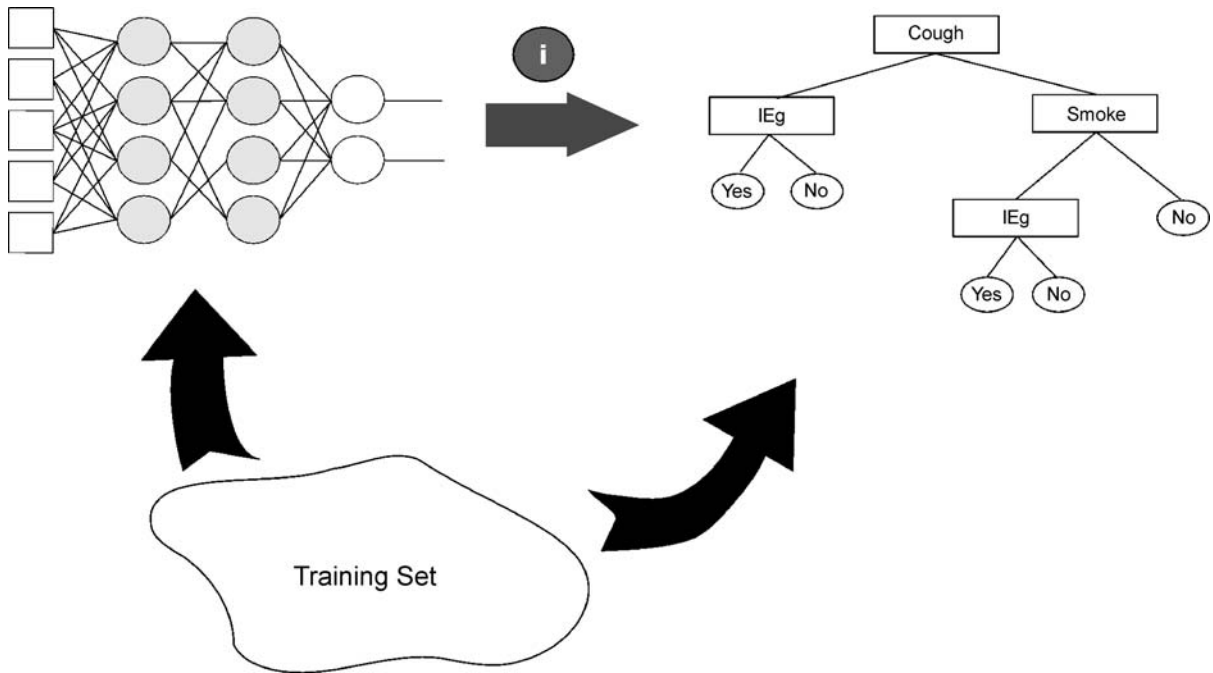
These four types of hybrids differ in type and strength of connection between the embedded methods. Such hybrids can also be building blocks for making more complex hybrids.

For example, in 1999 Boz [8] presented a method for converting a trained backpropagation neural network into a decision tree. In this approach the neural network was used to determine the list of attributes that influence the outcomes of the network the most. That list of attributes



Discovery Systems, Figure 7

Basic types of hybrids: a sequential hybrid, b external hybrid, c embedded hybrid and d parallel hybrid



Discovery Systems, Figure 8

External hybrid of a decision tree and neural network

was then used in the process of the decision tree induction instead of the splitting criteria [Fig. 8].

Interpretation and Evaluation of Results

Evaluation of an induced model is a very important step of KDDM process. Testing of an induced model is crucial for evaluation. For that purpose, an initial data set is usually randomly split into two data sets: a training set (70% of randomly chosen samples) and a test set (30% of samples). When a training set is large it can be divided into two sets: a training set and a validation set. A validation set is used to assess how well the model fits the data, to adjust the model, or to select the best model among those that have been validated.

Beside random selection of training and testing samples from the database, the distribution of classes in training set is also very important for effective learning. In most cases it is best to preserve the natural distribution of classes (as it appears in initial data set) in the training set. However, in real world problems (especially from the medical field) we often come across the data sets with very unequal distribution of classes (for example: less than 10% of samples belong to a specific class). In cases like that the use of natural distribution is not appropriate simply because there are not enough training samples to effectively learn

a model of basic concepts. Equalization of class distribution is therefore a better solution [50].

The quality of the induced model is most commonly evaluated with the **accuracy rate**. The accuracy rate is calculated separately on the training set and the test set. Training accuracy indicates the extent of concepts learned, whereas the accuracy of correctly classified samples from test set provides us more information about the predictive abilities of a model and its applicability to practice. Accuracy is calculated as follows:

$$\text{accuracy} = \frac{\text{num of correctly classified samples}}{\text{num of all samples}}.$$

Unfortunately, the accuracy by itself is not necessarily the right metric for selecting the best model. As mentioned before, the distribution of target variable classes in real world data sets is often very biased. In those cases we can obtain a model with high accuracy; however the accuracy of classifying the samples of non-dominant class can be catastrophically low. Therefore use of other metrics that expose more information of the types of errors and the costs associated with them are a necessity.

Average class accuracy is one of the simple metrics that gives more information about model's quality. The classification accuracy is calculated separately for each class of a target variable and the average across all classes

Discovery Systems, Table 2

A sample confusion matrix for three classes

Predicted	Actual		
	Class 1	Class 2	Class 3
Class 1	30	2	0
Class 2	10	25	3
Class 3	5	7	43

is calculated:

$$\text{average class accuracy} = \frac{\sum_c \text{accuracy}_c}{\text{num of classes}},$$

where c is one of the possible classes and accuracy_c is:

$$\text{accuracy}_c = \frac{\text{num of correctly classified objects in class } c}{\text{num of all objects in class } c}.$$

When accuracy is equal or similar to average class accuracy, one can speculate that the model does not have problems with classification of samples belonging to a specific class. Knowing the accuracies for each class of target variable gives us additional information about problems in the database or model.

For classification problems a **confusion matrix** [37] is a very useful tool for understanding the results of testing a model. A confusion matrix shows the results of actual versus predicted class values. Table 2 presents a sample of a confusion matrix where the columns show actual classes and the rows show predicted classes. Therefore the diagonal shows all correctly classified samples (98 samples out of 125). Class accuracies can also be calculated. For example, 30 samples out of 45 belonging to Class 1 were correctly classified (class accuracy for Class 1 is 67%); 10 samples were misclassified as Class 2 and 5 samples were misclassified as Class 3).

The confusion matrix for binary classification can be interpreted with two commonly used matrices: **specificity** and **sensitivity**. They are most often used to evaluate a performance of medical test. Sensitivity is the proportion of samples that were classified as positive of all the positive samples tested. In the case of classifying people with illness it can be seen as *the probability that the test is positive given that the patient is sick*. Consequently, the higher the sensitivity, the fewer sick people are undetected.

Specificity is the proportion of samples that were classified as negative of all the negative samples tested. In the case of classifying people with illness it can be seen as *the probability that the test is negative given that the patient is not sick*. Consequently, the higher the specificity, the fewer healthy people are classified as sick.

Discovery Systems, Table 3

A binary confusion matrix

Predicted	Actual	
	Positive	Negative
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

From the definitions above we can conclude the following: *Sensitivity and specificity are inversely related (higher sensitivity leads to lower specificity).*

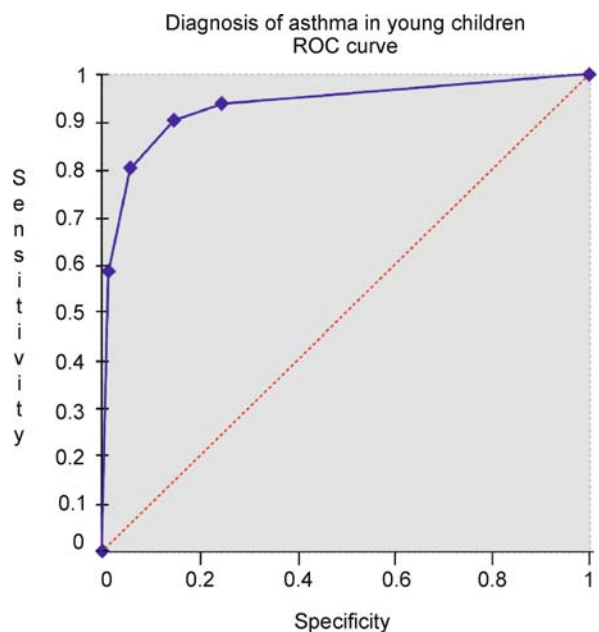
Sensitivity and specificity are calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{FP + TN}$$

where the notations are explained in Table 3.

The relationship between specificity and sensitivity can be visualized with the **ROC graph** [5].

ROC (receiver operating characteristics) graph [5] is a two dimensional plot with the false positive rate (1-specificity) on the x-axis and the true positive rate (sensitivity) on y-axis (Fig. 9). The best possible prediction model would yield a point (0,1) in the upper left corner of ROC graph, which represents 100% sensitivity (all true positive samples correctly classified) and 100% specificity (no false positive samples classified). On the contrary, the point (1,0) in the lower right corner of ROC graph represents the worst possible prediction model with 0% sensitivity and

Discovery Systems, Figure 9
ROC graph

0% specificity. A model with completely random classification would give a point along diagonal line from the left bottom to the top right corner ($y = x$). The diagonal line divides the ROC space into two areas: points above diagonal line represent good models and points below diagonal line represent bad models. A more precise way of characterizing this is simply to calculate the **area under the ROC curve** [3]. The closer the area is to 0.5, the worse is the model, and the closer it is to 1.0, the better is the model.

All presented evaluation techniques express the quality of induced model in dependence of data gathered in the data set. However, no matter how well the model performs on data sets used for testing, there is no 100% guarantee that the model will perform well in the real world. One of the reasons can be found in the quality of data used to build the model (a lot of noise, missing data, implicit assumptions about choosing prediction variables, etc.). When the training data fails to match the real world, the model can learn incorrect concepts and will therefore perform very poor in the practice. For that purpose it is very important to test the model in the real world. For example: if the model is used for diagnosing asthma in young children, it should be tested on a small set of symptomatic children.

An evaluation of a model by a **domain expert** (usually also an end-user) is also very important before presenting the model to the real world. In that case symbolic models (such as: decision rules, decision trees, etc.) have essential advantage over the so called black-box models (such as: neural networks, support vector machines, etc.) – the possibility to interpret the relationships among prediction and target variables.

Data mining extracts already known and expected information from the database, but on the other hand, useful relationships between variables that are non-intuitive are the jewels that data mining hopes to find. In order to enable a domain expert to evaluate the model and possibly to extract some new knowledge, it is very important to choose a proper visualization of a model. This requires understanding end-user's needs. Visualizing a model allows a user to discuss and explain the logic behind the model with colleagues, customers, and other users. It also builds the user's trust in the results. Symbolic models are relatively easy to visualize, however black-box models can pose more problems but novel solutions are starting to appear.

Utilization of Results

Once a data mining model is built and evaluated it can be used in one of the following ways: (1) analyzing a model

and extracting knowledge or (2) applying a model in practice for a decision support.

In (1) the analyst (usually a domain expert) studies the model and the results in order to extract some new knowledge about the relationships among variables (for example: comparison of different clusters, interpretation of induced rules, etc).

In (2) the model is applied into practice and used on new data in order to perform some classification or regression tasks. Prediction models can be used for decision support independently. For instance, a model for diagnosing asthma in young children can support the physician in making a diagnosis. However, usually models are used in some business process and therefore they are most commonly incorporated into business applications. For instance, a model for predicting a risky loan applicant can be integrated into the loan application.

Some prediction models turn out to be very time and resource consuming in practical use (for example: monitoring credit card transactions, fraud detections, etc). In those cases parallel systems can provide an effective solution.

Since all practical knowledge and experiences of domain experts are often not captured in prediction models built in data mining process, they may be used in combination with a model and as such applied into practice. Usually a combination of expert knowledge and knowledge discovered in data mining process can be very effective.

Since we are all aware that over time all systems evolve (all things are changing – nothing is static) it is very important to continuously monitor the performance of a model used in practice. When a significant reduction in performance of a model is detected, the model has to be retrained or sometimes even completely rebuilt.

Knowledge Discovery Frameworks and Tools

As already stated, knowledge discovery is a process of searching through large amounts of data and picking out relevant information. It is very often used in science but has become a very important process in the business and other areas of everyday life, too. Due to a huge demand for knowledge discovery and data mining tools, there are a lot of different tools and frameworks that are becoming available over the Internet. Most of them are coming from academic circles and can be obtained on open source or general public license basis. In this chapter we mostly focus on knowledge discovery and data mining frameworks and tools that are freely available and are used by a wide range of users from different fields.

The following tools that cover the biggest share of KD tools users will be presented:

- Weka/Pentaho
- YALE/Rapid-i
- Orange
- Tanagra.

Weka/Pentaho

Originally created in 1993, the Weka project [52] was established by the University of Waikato as a platform for research and testing of advanced machine learning techniques. Since then, Weka has developed a large and loyal following in both academic and industry circles. It is so popular because of easy tailoring and extending of powerful analytical techniques with advanced visualization and industry-specific algorithms. In 2006 it was acquired by Pentaho [20], which is an even larger suite of tools that provides a full spectrum of open source business intelligence (BI) capabilities including reporting, analysis, dashboards, data mining, data integration, and a BI platform that have made it the world's most popular open source BI suite. Pentaho is also the primary sponsor and owner of popular open source projects including Mondrian, JFreeReport, Kettle, and Weka. By this acquisition, Weka got even wider range of users and established itself in the BI knowledge discovery circles.

YALE/RapidMiner

Similar to Weka, YALE [33] was also transformed into a more commercial, but still freely available set of tools for machine learning based tasks that can be used in knowledge discovery process. Both have their roots in academic circles and are based on most recent research papers. YALE was renamed into RapidMiner after the acquisition of YALE by Rapid-I in May 2007.

YALE/RapidMiner is a rapid prototyping system for knowledge discovery and data mining. With more than 400 data mining operators it is one of the most comprehensive open-source data mining tools. It is widely used by a large number of organizations covering a wide range of different branches.

As the main focus of the YALE authors has been flexibility, it is reasonable to expect most advantage for YALE in this area. But flexibility in creating experiments, reusing previous work and preprocessing data to work on, were main advantages only in the early versions of YALE compared to Weka. However, visualization of results is still better supported in YALE.

Orange

Orange [16] is component-based data mining software. It includes a range of preprocessing, modeling and data exploration techniques. It is based on C++ components, that are accessed either directly, through Python scripts, or through GUI objects called Orange Widgets.

Orange core objects and Python modules support various data mining tasks that span from data preprocessing to modeling and evaluation. Among other it supports techniques for:

- Data input, providing the support for various popular data formats,
- Data manipulation and preprocessing, like sampling, filtering, scaling, discretization, construction of new attributes, and alike,
- Methods for development of classification models, including classification trees, naïve Bayesian classifier, instance-based approaches, logistic regression and support vector machines,
- Regression methods, including linear regression, regression trees, and instance-based approaches,
- Various wrappers, like those for calibration of probability predictions of classification models,
- Ensemble approaches, like boosting and bagging,
- Various state-of-the-art constructive induction methods, including function decomposition,
- Association rules and data clustering methods,
- Evaluation methods, different hold-out schemes and range of scoring methods for prediction models including classification accuracy, AUC, Brier score, and alike. Various hypothesis testing approaches are also supported,
- Methods to export predictive models to PMML.

Orange is based on C++, which could be the main advantage when fast knowledge discovery tools are needed in comparison to before mentioned Java based frameworks.

Tanagra

Tanagra [40] is a free data mining and knowledge discovery software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. In comparison to other presented tools, Tanagra contains more statistical functions than other frameworks, but is aimed mainly to academic circles. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present standards of the software development in this domain (especially in the design of its graphical user

interface and the way in which it is used), and allowing one to analyze either real or synthetic data.

Another approach to data analysis is R [13], which is heavily used for statistical analysis of data and also features a number of machine learning algorithms. Being very efficient with handling data, it lacks a graphical user interface for building projects and uses a rather nonintuitive syntax one has to get used to. Using the extensions that are constantly becoming available one can expect that R will feature most of the knowledge discovery functionality in the future.

Conclusions and Future Directions

The aggressive rate of data that is stored on modern computer systems has far outpaced our ability to process and analyze this data. This way data is often deposited to merely rest in peace, never to be accessed again. Novel applications of knowledge discovery systems in supporting scientific discoveries, business exploitation or complex decision making, have awakened the growing commercial interest in knowledge discovery and data mining techniques. That has stimulated new interest in automatic knowledge extraction from examples stored in large databases.

Knowledge discovery can be broadly defined as automated discovery of novel and useful information from databases. Unfortunately there is no method or formal model that would be able to define the optimal knowledge discovery solution that should be applied for a certain problem. This problem will become even more important in the future as a growing amount of collected data will not allow time costs that are caused by empirical evaluation of all possible knowledge discovery techniques to find the most appropriate one. The future directions and possible solutions for this problem were defined in [34] where different meta-learning techniques along with a knowledge-driven framework for a KD system that contains a limited number of data mining techniques are presented.

Another interesting direction for future discoveries is application of KD techniques in the field of structured knowledge or textual documents that can be transformed into formally structured form. This field of research is heading toward advanced techniques from recent text mining research like sentiment analysis and tries to combine them with classical data mining techniques. This will only be possible when we will be able to transform unstructured text into machine readable form that would formalize parts of text and extract the underlying messages in the text.

Despite the rapid growth, KD systems are still evolving and will represent an important part of the future intelli-

gent systems in various fields. It is obvious that KD systems are also becoming an important part of the large business systems that are fortunately available through different forms of open source initiatives that can help in their sustainable development for the future.

Bibliography

Primary Literature

1. Anand S, Buchner A (1998) Decision support using data mining. Financial Time Management, London
2. Baeck T (1996) Evolutionary algorithms in theory and practice. Oxford University Press, New York
3. Barley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30(7):1145–1159
4. Becerra-Fernandez I, Gonzalez A, Sabherwal R (2004) Knowledge management: Challenges, solutions, and technologies. Prentice Hall, Upper Saddle River
5. Beck JR, Shultz E (1986) The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 110:13–20
6. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
7. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Belmont
8. Boz O (2000) Converting a trained neural network to a decision tree dectext – decision tree extractor. Ph D thesis, Computer Science and Engineering, Lehigh University. <http://citeseer.ist.psu.edu/boz00converting.html>. Accessed 12 Nov 2007
9. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A (1998) Discovering data mining: From concepts to implementation. Prentice Hall, Upper Saddle River
10. Caspase Drug Discovery Systems. drug discovery system. http://www.biomol.com/Online_Catalog/Online_Catalog/Products/36/?categoryid=420. Accessed 6 Nov 2007
11. Cios K, Teresinska A, Konieczna S, Potocka J, Sharma S (2000) Diagnosing myocardial perfusion from PECT bull's-eye maps – a knowledge discovery approach. IEEE Eng Med Biol Mag, Special Issue Med Data Mining Knowl Discov 19(4):17–25
12. Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA (2007) Data mining. A knowledge discovery approach. Springer, New York
13. Dalggaard P (2002) Introductory statistics with R. Springer, New York
14. Davenport TH, Prusak L (1997) Information ecology: Mastering the information and knowledge environment. Oxford University Press, New York
15. Dennis JE Jr, Schnabel RB (1989) A view of unconstrained optimization. In: Nemhauser GL, Runnooy Kan AHG, Todd MJ (eds) Handbook in operations research and management science, vol 1 Optimization. Elsevier, Amsterdam
16. Demsar J, Zupan B (2004) Orange: From experimental machine learning to interactive data mining. White Paper. Faculty of Computer and Information Science, University of Ljubljana. <http://www.ailab.si/orange>
17. Developmental Discovery System (TM). Developmental discovery system. <http://www.gotofocus.com/>. Accessed 6 Nov 2007

18. Dictionary.com Unabridged (v 1.1). discover. <http://dictionary.reference.com/browse/discover>. Accessed 5 Nov 2007
19. Dietterich TG (2000) Ensemble methods in machine learning. In: First International Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science. Springer, New York, pp 1–15
20. Dixon J (2005) Pentaho Open Source Business Intelligence Platform Technical White Paper. Pentaho Corporation, Orlando. http://sourceforge.net/project/showfiles.php?group_id=140317
21. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases (a survey). *AI Mag* 17(3):37–54
22. Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) (1996) *Advances in knowledge discovery and data mining*. AAAI Press, Menlo Park
23. Frawley W, Piatetsky-Shapiro G, Matheus C (1991) Knowledge discovery in databases: An overview. In: Piatetsky-Shapiro G, Frawley W (eds) *Knowledge Discovery in Databases*. AAAI/MIT Press, pp 1–27, Menlo Park
24. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: *Proceedings Thirteenth International Conference on Machine Learning*. Morgan Kaufman, San Francisco, pp 148–156
25. Goldberg DE (1989) *Genetic algorithms in search, optimization, and machine learning*. Addison, Reading
26. Hand D, Mannila H, Smyth P (eds) (2001) *Principles of data mining*. MIT Press, Cambridge
27. Holland JH (1975) *Adaptation in natural and artificial systems*. MIT Press, Cambridge
28. Iglesias CJ (1996) The role of hybrid systems in intelligent data management: The case of fuzzy/neural hybrids. *Control Eng Pract* 4(6):839–845
29. Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. *Appl Stat* 29:119–127
30. Kurgan L, Musilek P (2006) A survey of Knowledge Discovery and Data Mining process models. *Knowl Eng Rev* 21(1):1–24
31. Loh W, Shih Y (1997) Split selection methods for classification trees. *Stat Sinica* 7:815–840
32. Mannila H (2000) Theoretical frameworks of data mining. *SIGKDD Explor* 1:30–32
33. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. In: *Proc of the 12th ACM SIGKDD. International Conference on Knowledge Discovery and Data Mining*, Philadelphia, pp 1–6
34. Pechenizkiy M, Tsymbal A, Puuronen S (2005) Meta-knowledge management in multistrategy process-oriented knowledge discovery systems. Technical Report, Dublin, Trinity College Dublin, Department of Computer Science, TCD-CS-2005–30, p 12
35. Piatetsky-Shapiro G (1991) Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. *AI Mag* 11(5): 68–70
36. Piatetsky-Shapiro G (1999) The data mining industry coming to age. *IEEE Intel Syst* 14(6):32–33
37. Provost F, Fawcett T, Kohavi R (1998) The case against accuracy estimation for comparing classifiers. In: *Proceedings of the Fifteenth International Conference on Machine Learning*, (ICML-98), San Francisco
38. Quinlan JR (1986) Induction of decision trees. In: *Machine Learning*, vol 1. Kluwer, Hingham
39. Quinlan R (1993) *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco
40. Rakotomalala R (2005) TANAGRA: Un logiciel gratuit pour l'enseignement et la recherche. In: *Proc of the 5th Journées d'Extraction et Gestion des Connaissances* 2:697–702
41. Reeves CR (ed) (1993) *Modern heuristic techniques for combinatorial problems*. Wiley, New York
42. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323:533–536
43. Sano M, Katoa Y, Taira K (2005) Functional gene-discovery systems based on libraries of hammerhead and hairpin ribozymes and short hairpin RNAs. *Mol Biosyst* 1:27–35
44. Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. *J Data Wareh* 15(4):13–19
45. Smyth P, Goodman RM (1991) Rule induction using information theory. In: Piatetsky-Schapiro G, Frawley WJ (eds) *Knowledge Discovery in Databases*. AAAI Press, pp 159–176, Menlo Park
46. Snedecor GW, Cochran WG (1989) *Statistical methods*, 8th edn. Iowa State University Press, Ames
47. Tan P, Steinbach M, Kumar V (2005) *Introduction to data mining*. Addison, Boston
48. The Discovery System. discovery system for personality profiling. <http://www.insights.com/core/English/TheDiscoverySystem/default.shtm>. Accessed 6 Nov 2007
49. Towsey M, Alpsan D, Sztrihai L (1995) Training a neural network with conjugate gradient methods. *IEEE Proc Neural Netw* 1:373–378
50. Weiss GM, Provost F (2001) The effect of class distribution on classifier learning. Technical Report ML-TR 43, Department of Computer Science, Rutgers University
51. Werbos PJ (1994) *The roots of backpropagation*. Wiley, New York
52. Witten IH, Frank E (2005) *Data mining: Practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, San Francisco
53. Wolpert D, Macready W (1997) No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1(1):67–82

Books and Reviews

- Berthold M, Hand DJ (2003) *Intelligent data analysis: An introduction*, 2nd edn. Springer, New York
- Lin TY, Ohsuga S, Liau CJ, Hu X, Tsumoto S (eds) (2005) *Foundations of data mining and knowledge discovery. Studies in Computational Intelligence*, vol 6. Springer, New York

Discrete Control Systems

TAHEYOUNG LEE¹, MELVIN LEOK²,
HARRIS MCCLAMROCH¹

¹ Department of Aerospace Engineering,
University of Michigan, Ann Arbor, USA

² Department of Mathematics, Purdue University,
West Lafayette, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Discrete Lagrangian and Hamiltonian Mechanics
 Optimal Control of Discrete Lagrangian
 and Hamiltonian Systems
 Controlled Lagrangian Method
 for Discrete Lagrangian Systems
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Discrete variational mechanics A formulation of mechanics in discrete-time that is based on a discrete analogue of *Hamilton's principle*, which states that the system takes a trajectory for which the action integral is stationary.

Geometric integrator A numerical method for obtaining numerical solutions of differential equations that preserves geometric properties of the continuous flow, such as symplecticity, momentum preservation, and the structure of the configuration space.

Lie group A differentiable manifold with a group structure where the composition is differentiable. The corresponding *Lie algebra* is the tangent space to the Lie group based at the identity element.

Symplectic A map is said to be symplectic if given any initial volume in phase space, the sum of the signed projected volumes onto each position-momentum subspace is invariant under the map. One consequence of symplecticity is that the map is volume-preserving as well.

Definition of the Subject

Discrete control systems, as considered here, refer to the control theory of discrete-time Lagrangian or Hamiltonian systems. These discrete-time models are based on a discrete variational principle, and are part of the broader field of geometric integration. Geometric integrators are numerical integration methods that preserve geometric properties of continuous systems, such as conservation of the symplectic form, momentum, and energy. They also guarantee that the discrete flow remains on the manifold on which the continuous system evolves, an important property in the case of rigid-body dynamics.

In nonlinear control, one typically relies on differential geometric and dynamical systems techniques to prove

properties such as stability, controllability, and optimality. More generally, the geometric structure of such systems plays a critical role in the nonlinear analysis of the corresponding control problems. Despite the critical role of geometry and mechanics in the analysis of nonlinear control systems, nonlinear control algorithms have typically been implemented using numerical schemes that ignore the underlying geometry.

The field of discrete control systems aims to address this deficiency by restricting the approximation to the choice of a discrete-time model, and developing an associated control theory that does not introduce any additional approximation. In particular, this involves the construction of a control theory for discrete-time models based on geometric integrators that yields numerical implementations of nonlinear and geometric control algorithms that preserve the crucial underlying geometric structure.

Introduction

The dynamics of Lagrangian and Hamiltonian systems have unique geometric properties; the Hamiltonian flow is symplectic, the total energy is conserved in the absence of non-conservative forces, and the momentum maps associated with symmetries of the system are preserved. Many interesting dynamics evolve on a non-Euclidean space. For example, the configuration space of a spherical pendulum is the two-sphere, and the configuration space of rigid body attitude dynamics has a Lie group structure, namely the special orthogonal group. These geometric features determine the qualitative behavior of the system, and serve as a basis for theoretical study.

Geometric numerical integrators are numerical integration algorithms that preserve structures of the continuous dynamics such as invariants, symplecticity, and the configuration manifold (see [14]). The exact geometric properties of the discrete flow not only generate improved qualitative behavior, but also provide accurate and efficient numerical techniques. In this article, we view a geometric integrator as an intrinsically discrete dynamical system, instead of concentrating on the numerical approximation of a continuous trajectory.

Numerical integration methods that preserve the symplecticity of a Hamiltonian system have been studied (see [28,36]). Coefficients of a Runge–Kutta method are carefully chosen to satisfy a symplecticity criterion and order conditions to obtain a symplectic Runge–Kutta method. However, it can be difficult to construct such integrators, and it is not guaranteed that other invariants of the system, such as a momentum map, are preserved. Alternatively, variational integrators are constructed by dis-

cretizing Hamilton's principle, rather than discretizing the continuous Euler–Lagrange equation (see [31,34]). The resulting integrators have the desirable property that they are symplectic and momentum preserving, and they exhibit good energy behavior for exponentially long times (see [2]). Lie group methods are numerical integrators that preserve the Lie group structure of the configuration space (see [18]). Recently, these two approaches have been unified to obtain Lie group variational integrators that preserve the geometric properties of the dynamics as well as the Lie group structure of the configuration space without the use of local charts, reprojection, or constraints (see [26,29,32]).

Optimal control problems involve finding a control input such that a certain optimality objective is achieved under prescribed constraints. An optimal control problem that minimizes a performance index is described by a set of differential equations, which can be derived using Pontryagin's maximum principle. Discrete optimal control problems involve finding a control input for a discrete dynamic system such that an optimality objective is achieved with prescribed constraints. Optimality conditions are derived from the discrete equations of motion, described by a set of discrete equations. This approach is in contrast to traditional techniques where a discretization appears at the last stage to solve the optimality condition numerically. Discrete mechanics and optimal control approaches determine optimal control inputs and trajectories more accurately with less computational load (see [19]). Combined with an indirect optimization technique, they are substantially more efficient (see [17,23,25]).

The geometric approach to mechanics can provide a theoretical basis for innovative control methodologies in geometric control theory. For example, these techniques allow the attitude of satellites to be controlled using changes in its shape, as opposed to chemical propulsion. While the geometric structure of mechanical systems plays a critical role in the construction of geometric control algorithms, these algorithms have typically been implemented using numerical schemes that ignore the underlying geometry. By applying geometric control algorithms to discrete mechanics that preserve geometric properties, we obtain an exact numerical implementation of the geometric control theory. In particular, the method of controlled Lagrangian systems is based on the idea of adopting a feedback control to realize a modification of either the potential energy or the kinetic energy, referred to as potential shaping or kinetic shaping, respectively. These ideas are applied to construct a real-time digital feedback controller that stabilizes the inverted equilibrium of the cart-pendulum (see [10,11]).

In this article, we will survey discrete Lagrangian and Hamiltonian mechanics, and their applications to optimal control and feedback control theory.

Discrete Lagrangian and Hamiltonian Mechanics

Mechanics studies the dynamics of physical bodies acting under forces and potential fields. In Lagrangian mechanics, the trajectory of the object is derived by finding the path that extremizes the integral of a Lagrangian over time, called the action integral. In many classical problems, the Lagrangian is chosen as the difference between kinetic energy and potential energy. The Legendre transformation provides an alternative description of mechanical systems, referred to as Hamiltonian mechanics.

Discrete Lagrangian and Hamiltonian mechanics has been developed by reformulating the theorems and the procedures of Lagrangian and Hamiltonian mechanics in a discrete time setting (see, for example, [31]). Therefore, discrete mechanics has a parallel structure with the mechanics described in continuous time, as summarized in Fig. 1 for Lagrangian mechanics. In this section, we describe discrete Lagrangian mechanics in more detail, and we derive discrete Euler–Lagrange equations for several mechanical systems.

Consider a mechanical system on a configuration space Q , which is the space of possible positions. The Lagrangian depends on the position and velocity, which are elements of the tangent bundle to Q , denoted by TQ . Let $L: TQ \rightarrow \mathbb{R}$ be the Lagrangian of the system. The discrete Lagrangian, $L_d: Q \times Q \rightarrow \mathbb{R}$ is an approximation to the exact discrete Lagrangian,

$$L_d^{\text{exact}}(q_0, q_1) = \int_0^h L(q_{01}(t), \dot{q}_{01}(t)) dt, \quad (1)$$

where $q_{01}(0) = q_0$, $q_{01}(h) = q_1$, and $q_{01}(t)$ satisfies the Euler–Lagrange equation in the time interval $(0, h)$. A discrete action sum $\mathcal{G}_d: Q^{N+1} \rightarrow \mathbb{R}$, analogous to the action integral, is given by

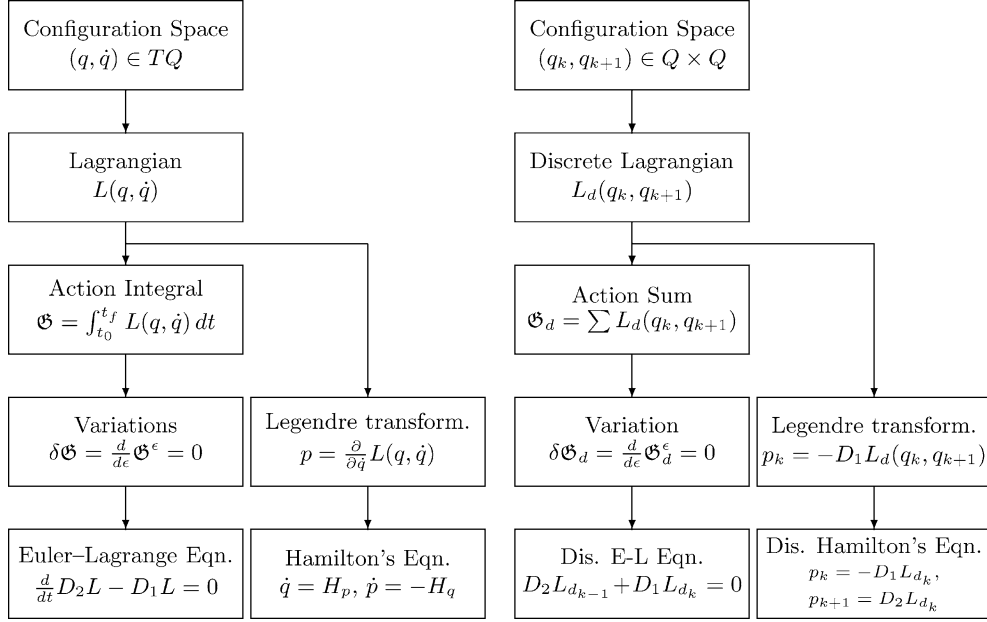
$$\mathcal{G}_d(q_0, q_1, \dots, q_N) = \sum_{k=0}^{N-1} L_d(q_k, q_{k+1}). \quad (2)$$

The discrete Hamilton's principle states that

$$\delta \mathcal{G}_d = 0$$

for any δq_k , which yields the *discrete Euler–Lagrange (DEL)* equation,

$$D_2 L_d(q_{k-1}, q_k) + D_1 L_d(q_k, q_{k+1}) = 0. \quad (3)$$



Discrete Control Systems, Figure 1

Procedures to derive continuous and discrete equations of motion

This yields a discrete Lagrangian flow map $(q_{k-1}, q_k) \mapsto (q_k, q_{k+1})$. The discrete Legendre transformation, which from a pair of positions (q_0, q_1) gives a position-momentum pair $(q_0, p_0) = (q_0, -D_1 L_d(q_0, q_1))$ provides a discrete Hamiltonian flow map in terms of momenta.

The discrete equations of motion, referred to as variational integrators, inherit the geometric properties of the continuous system. Many interesting Lagrangian and Hamiltonian systems, such as rigid bodies evolve on a Lie group. Lie group variational integrators preserve the nonlinear structure of the Lie group configurations as well as geometric properties of the continuous dynamics (see [29,32]). The basic idea for all Lie group methods is to express the update map for the group elements in terms of the group operation,

$$g_1 = g_0 f_0, \quad (4)$$

where $g_0, g_1 \in G$ are configuration variables in a Lie group G , and $f_0 \in G$ is the discrete update represented by a right group operation on g_0 . Since the group element is updated by a group operation, the group structure is preserved automatically without need of parametrizations, constraints, or re-projection. In the Lie group variational integrator, the expression for the flow map is obtained from the discrete variational principle on a Lie group, the same procedure presented in Fig. 1. But, the infinitesimal variation of a Lie group element must be carefully ex-

pressed to respect the structure of the Lie group. For example, it can be expressed in terms of the exponential map as

$$\delta g = \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} g \exp \epsilon \eta = g \eta,$$

for a Lie algebra element $\eta \in \mathfrak{g}$. This approach has been applied to the rotation group $SO(3)$ and to the special Euclidean group $SE(3)$ for dynamics of rigid bodies (see [24,26,27]). Generalizations to arbitrary Lie groups gives the generalized *discrete Euler-Poincaré (DEP)* equation,

$$T_e^* L_{f_0} \cdot D_2 L_d(g_0, f_0) - \text{Ad}_{f_0}^* \cdot (T_e^* L_{f_1} \cdot D_2 L_d(g_1, f_1)) + T_e^* L_{g_1} \cdot D_1 L_d(g_1, f_1) = 0, \quad (5)$$

for a discrete Lagrangian on a Lie group, $L_d: G \times G \rightarrow \mathbb{R}$. Here $L_f: G \rightarrow G$ denotes the left translation map given by $L_f g = fg$ for $f, g \in G$, $T_g L_f: T_g G \rightarrow T_{fg} G$ is the tangential map for the left translation, and $\text{Ad}_g: \mathfrak{g} \rightarrow \mathfrak{g}$ is the adjoint map. A dual map is denoted by a superscript $*$ (see [30] for detailed definitions).

We illustrate the properties of discrete mechanics using several mechanical systems, namely a mass-spring system, a planar pendulum, a spherical pendulum, and a rigid body.

Example 1 (Mass-spring System) Consider a mass-spring system, defined by a rigid body that moves along a straight frictionless slot, and is attached to a linear spring.

Continuous equation of motion: The configuration space is $Q = \mathbb{R}$, and the Lagrangian $L: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is given by

$$L(q, \dot{q}) = \frac{1}{2} m \dot{q}^2 - \frac{1}{2} \kappa q^2, \quad (6)$$

where $q \in \mathbb{R}$ is the displacement of the body measured from the point where the spring exerts no force. The mass of the body and the spring constant are denoted by $m, \kappa \in \mathbb{R}$, respectively. The Euler–Lagrange equation yields the continuous equation of motion.

$$m\ddot{q} + \kappa q = 0. \quad (7)$$

Discrete equation of motion: Let $h > 0$ be a discrete time step, and a subscript k denotes the k th discrete variable at $t = kh$. The discrete Lagrangian $L_d: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is an approximation of the integral of the continuous Lagrangian (6) along the solution of (7) over a time step. Here, we choose the following discrete Lagrangian.

$$\begin{aligned} L_d(q_k, q_{k+1}) &= hL\left(\frac{q_k + q_{k+1}}{2}, \frac{q_{k+1} - q_k}{h}\right) \\ &= \frac{1}{2h} m (q_{k+1} - q_k)^2 - \frac{h\kappa}{8} (q_k + q_{k+1})^2. \end{aligned} \quad (8)$$

Direct application of the discrete Euler–Lagrange equation to this discrete Lagrangian yields the discrete equations of motion. We develop the discrete equation of motion using the discrete Hamilton’s principle to illustrate the principles more explicitly. Let $\mathcal{G}_d: \mathbb{R}^{N+1} \rightarrow \mathbb{R}$ be the discrete action sum defined as $\mathcal{G}_d = \sum_{k=0}^{N-1} L_d(q_k, q_{k+1})$, which approximates the action integral. The infinitesimal variation of the action sum can be written as

$$\begin{aligned} \delta \mathcal{G}_d &= \sum_{k=0}^{N-1} \delta q_{k+1} \left\{ \frac{m}{h} (q_{k+1} - q_k) - \frac{h\kappa}{4} (q_k + q_{k+1}) \right\} \\ &\quad + \delta q_k \left\{ -\frac{m}{h} (q_{k+1} - q_k) - \frac{h\kappa}{4} (q_k + q_{k+1}) \right\}. \end{aligned}$$

Since $\delta q_0 = \delta q_N = 0$, the summation index can be rewritten as

$$\begin{aligned} \delta \mathcal{G}_d &= \sum_{k=1}^{N-1} \delta q_k \left\{ -\frac{m}{h} (q_{k+1} - 2q_k + q_{k-1}) \right. \\ &\quad \left. - \frac{h\kappa}{4} (q_{k+1} + 2q_k + q_{k-1}) \right\}. \end{aligned}$$

From discrete Hamilton’s principle, $\delta \mathcal{G}_d = 0$ for any δq_k . Thus, the discrete equation of motion is given by

$$\begin{aligned} \frac{m}{h} (q_{k+1} - 2q_k + q_{k-1}) \\ + \frac{h\kappa}{4} (q_{k+1} + 2q_k + q_{k-1}) = 0. \end{aligned} \quad (9)$$

For a given (q_{k-1}, q_k) , we solve the above equation to obtain q_{k+1} . This yields a discrete flow map $(q_{k-1}, q_k) \mapsto (q_k, q_{k+1})$, and this process is repeated. The discrete Legendre transformation provides the discrete equation of motion in terms of the velocity as

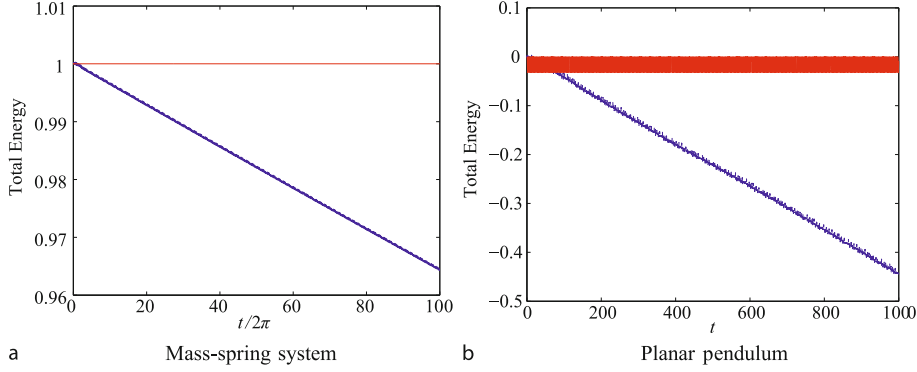
$$\left(1 + \frac{h^2 \kappa}{4m}\right) q_{k+1} = h\dot{q}_k + \left(1 - \frac{h^2 \kappa}{4m}\right) q_k, \quad (10)$$

$$\dot{q}_{k+1} = \dot{q}_k - \frac{h\kappa}{2m} q_k - \frac{h\kappa}{2m} q_{k+1}. \quad (11)$$

For a given (q_k, \dot{q}_k) , we compute q_{k+1} and \dot{q}_{k+1} by (10) and (11), respectively. This yields a discrete flow map $(q_k, \dot{q}_k) \mapsto (q_{k+1}, \dot{q}_{k+1})$. It can be shown that this variational integrator has second-order accuracy, which follows from the fact that the discrete action sum is a second-order approximation of the action integral.

Numerical example: We compare computational properties of the discrete equations of motion given by (10) and (11) with a 4(5)th order variable step size Runge–Kutta method. We choose $m = 1$ kg, $\kappa = 1$ kg/s² so that the natural frequency is 1 rad/s. The initial conditions are $q_0 = \sqrt{2}$ m, $\dot{q}_0 = 0$, and the total energy is $E = 1$ Nm. The simulation time is 200π sec, and the step-size $h = 0.035$ of the discrete equations of motion is chosen such that the CPU times are the same for both methods. Figure 2a shows the computed total energy. The variational integrator preserves the total energy well. The mean variation is 2.7327×10^{-13} Nm. But, there is a notable dissipation of the computed total energy for the Runge–Kutta method

Example 2 (Planar Pendulum) A planar pendulum is a mass particle connected to a frictionless, one degree-of-freedom pivot by a rigid massless link under a uniform gravitational potential. The configuration space is the one-sphere $\mathbb{S}^1 = \{q \in \mathbb{R}^2 \mid \|q\| = 1\}$. While it is common to parametrize the one-sphere by an angle, we develop parameter-free equations of motion in the special orthogonal group $\text{SO}(2)$, which is a group of 2×2 orthogonal matrices with determinant 1, i.e. $\text{SO}(2) = \{R \in \mathbb{R}^{2 \times 2} \mid R^T R = I_{2 \times 2}, \det[R] = 1\}$. $\text{SO}(2)$ is diffeomorphic to the one-sphere. It is also possible to develop global equations of motion on the one-sphere directly, as shown



Discrete Control Systems, Figure 2

Computed total energy (RK45: blue, dotted, VI: red, solid)

in the next example, but here we focus on the special orthogonal group in order to illustrate the key steps to develop a Lie group variational integrator.

We first exploit the basic structures of the Lie group $SO(2)$. Define a hat map $\hat{\cdot}$, which maps a scalar Ω to a 2×2 skew-symmetric matrix $\hat{\Omega}$ as

$$\hat{\Omega} = \begin{bmatrix} 0 & -\Omega \\ \Omega & 0 \end{bmatrix}.$$

The set of 2×2 skew-symmetric matrices forms the Lie algebra $\mathfrak{so}(2)$. Using the hat map, we identify $\mathfrak{so}(2)$ with \mathbb{R} . An inner product on $\mathfrak{so}(2)$ can be induced from the inner product on \mathbb{R} as $\langle \hat{\Omega}_1, \hat{\Omega}_2 \rangle = 1/2 \text{tr}[\hat{\Omega}_1^T \hat{\Omega}_2] = \Omega_1 \Omega_2$ for any $\Omega_1, \Omega_2 \in \mathbb{R}$. The matrix exponential is a local diffeomorphism from $\mathfrak{so}(2)$ to $SO(2)$ given by

$$\exp \hat{\Omega} = \begin{bmatrix} \cos \Omega & -\sin \Omega \\ \sin \Omega & \cos \Omega \end{bmatrix}.$$

The kinematics equation for $R \in SO(2)$ can be written in terms of a Lie algebra element as

$$\dot{R} = R \hat{\Omega}. \quad (12)$$

Continuous equations of motion: The Lagrangian for a planar pendulum $L: SO(2) \times \mathfrak{so}(2) \rightarrow \mathbb{R}$ can be written as

$$\begin{aligned} L(R, \hat{\Omega}) &= \frac{1}{2} m l^2 \Omega^2 + m g l e_2^T R e_2 \\ &= \frac{1}{2} m l^2 \langle \hat{\Omega}, \hat{\Omega} \rangle + m g l e_2^T R e_2, \end{aligned} \quad (13)$$

where the constant $g \in \mathbb{R}$ is the gravitational acceleration. The mass and the length of the pendulum are denoted by $m, l \in \mathbb{R}$, respectively. The second expression is used to define a discrete Lagrangian later. We choose the bases

of the inertial frame and the body-fixed frame such that the unit vector along the gravity direction in the inertial frame, and the unit vector along the pendulum axis in the body-fixed frame are represented by the same vector $e_2 = [0; 1] \in \mathbb{R}^2$. Thus, for example, the hanging attitude is represented by $R = I_{2 \times 2}$. Here, the rotation matrix $R \in SO(2)$ represents the linear transformation from a representation of a vector in the body-fixed frame to the inertial frame.

Since the special orthogonal group is not a linear vector space, the expression for the variation should be carefully chosen. The infinitesimal variation of a rotation matrix $R \in SO(2)$ can be written in terms of its Lie algebra element as

$$\delta R = \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} R \exp \epsilon \hat{\eta} = R \hat{\eta} \exp \epsilon \hat{\eta} \Big|_{\epsilon=0} = R \hat{\eta}, \quad (14)$$

where $\eta \in \mathbb{R}$ so that $\hat{\eta} \in \mathfrak{so}(2)$. The infinitesimal variation of the angular velocity is induced from this expression and (12) as

$$\begin{aligned} \delta \hat{\Omega} &= \delta R^T \dot{R} + R^T \delta \dot{R} \\ &= (R \hat{\eta})^T \dot{R} + R^T (\dot{R} \hat{\eta} + R \dot{\hat{\eta}}) \\ &= -\hat{\eta} \hat{\Omega} + \hat{\Omega} \hat{\eta} + \dot{\hat{\eta}} = \dot{\hat{\eta}}, \end{aligned} \quad (15)$$

where we used the equality of mixed partials to compute $\delta \dot{R}$ as $d/dt(\delta R)$.

Define the action integral to be $\mathcal{G} = \int_0^T L(R, \hat{\Omega}) dt$. The infinitesimal variation of the action integral is obtained by using (14) and (15). Hamilton's principle yields the following continuous equations of motion.

$$\hat{\Omega} + \frac{g}{l} e_2^T R e_1 = 0, \quad (16)$$

$$\dot{R} = R \hat{\Omega}. \quad (17)$$

If we parametrize the rotation matrix as

$$R = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix},$$

these equations are equivalent to

$$\ddot{\theta} + \frac{g}{l} \sin \theta = 0. \quad (18)$$

Discrete equations of motion: We develop a Lie group variational integrator on $\text{SO}(2)$. Similar to (4), define $F_k \in \text{SO}(2)$ such that

$$R_{k+1} = R_k F_k. \quad (19)$$

Thus, $F_k = R_k^T R_{k+1}$ represents the relative update between two integration steps. If we find $F_k \in \text{SO}(2)$, the orthogonal structure is preserved through (19) since multiplication of orthogonal matrices is also orthogonal. This is a key idea of Lie group variational integrators.

Define the discrete Lagrangian $L_d: \text{SO}(2) \times \text{SO}(2) \rightarrow \mathbb{R}$ to be

$$\begin{aligned} L_d(R_k, F_k) &= \frac{1}{2h} m l^2 \langle F_k - I_{2 \times 2}, F_k - I_{2 \times 2} \rangle \\ &\quad + \frac{h}{2} m g l e_2^T R_k e_2 + \frac{h}{2} m g l e_2^T R_{k+1} e_2 \\ &= \frac{1}{2h} m l^2 \text{tr}[I_{2 \times 2} - F_k] + \frac{h}{2} m g l e_2^T R_k e_2 \\ &\quad + \frac{h}{2} m g l e_2^T R_{k+1} e_2, \end{aligned} \quad (20)$$

which is obtained by an approximation $h\hat{\Omega}_k \simeq R_k^T(R_{k+1} - R_k) = F_k - I_{2 \times 2}$, applied to the continuous Lagrangian given by (13).

As for the continuous time case, expressions for the infinitesimal variations should be carefully chosen. The infinitesimal variation of a rotation matrix is the same as (14), namely

$$\delta R_k = R_k \hat{\eta}_k, \quad (21)$$

for $\eta_k \in \mathbb{R}$, and the constrained variation of F_k is obtained from (19) as

$$\begin{aligned} \delta F_k &= \delta R_k^T R_{k+1} + R_k^T \delta R_{k+1} \\ &= -\hat{\eta}_k F_k + F_k \hat{\eta}_{k+1} = F_k (\hat{\eta}_{k+1} - F_k^T \hat{\eta}_k F_k) \\ &= F_k (\hat{\eta}_{k+1} - \hat{\eta}_k), \end{aligned} \quad (22)$$

where we use the fact that $F \hat{\eta} F^T = \hat{\eta}$ for any $F \in \text{SO}(2)$ and $\hat{\eta} \in \mathfrak{so}(2)$.

Define an action sum $\mathcal{G}_d: \text{SO}(2)^{N+1} \rightarrow \mathbb{R}$ as $\mathcal{G}_d = \sum_{k=0}^{N-1} L_d(R_k, F_k)$. Using (21) and (22), the variation of the action sum is written as

$$\begin{aligned} \delta \mathcal{G}_d &= \sum_{k=1}^{N-1} \left\langle \frac{1}{2h} m l^2 (F_{k-1} - F_{k-1}^T) \right. \\ &\quad \left. - \frac{1}{2h} (F_k - F_k^T) - h m g l \widehat{e_2^T R_k e_1}, \hat{\eta}_k \right\rangle. \end{aligned}$$

From the discrete Hamilton's principle, $\delta \mathcal{G}_d = 0$ for any $\hat{\eta}_k$. Thus, we obtain the Lie group variational integrator on $\text{SO}(2)$ as

$$(F_k - F_k^T) - (F_{k+1} - F_{k+1}^T) - \frac{2h^2 g}{l} \widehat{e_2^T R_{k+1} e_1} = 0, \quad (23)$$

$$R_{k+1} = R_k F_k. \quad (24)$$

For a given (R_k, F_k) and $R_{k+1} = R_k F_k$, (23) is solved to find F_{k+1} . This yields a discrete map $(R_k, F_k) \mapsto (R_{k+1}, F_{k+1})$. If we parametrize the rotation matrices R and F with θ and $\Delta\theta$ and if we assume that $\Delta\theta \ll 1$, these equations are equivalent to

$$\frac{1}{h} (\theta_{k+1} - 2\theta_k + \theta_k) + \frac{hg}{l} \sin \theta_k = 0.$$

The discrete version of the Legendre transformation provides the discrete Hamiltonian map as follows.

$$F_k - F_k^T = 2h\hat{\Omega} - \frac{h^2 g}{l} \widehat{e_2^T R_k e_1}, \quad (25)$$

$$R_{k+1} = R_k F_k, \quad (26)$$

$$\Omega_{k+1} = \Omega_k - \frac{hg}{2l} e_2^T R_k e_1 - \frac{hg}{2l} e_2^T R_{k+1} e_1. \quad (27)$$

For a given (R_k, Ω_k) , we solve (25) to obtain F_k . Using this, (R_{k+1}, Ω_{k+1}) is obtained from (26) and (27). This yields a discrete map $(R_k, \Omega_k) \mapsto (R_{k+1}, \Omega_{k+1})$.

Numerical example: We compare the computational properties of the discrete equations of motion given by (25)–(27) with a 4(5)th order variable step size Runge–Kutta method. We choose $m = 1 \text{ kg}$, $l = 9.81 \text{ m}$. The initial conditions are $\theta_0 = \pi/2 \text{ rad}$, $\Omega = 0$, and the total energy is $E = 0 \text{ Nm}$. The simulation time is 1000 sec, and the step-size $h = 0.03$ of the discrete equations of motion is chosen such that the CPU times are identical. Figure 2b shows the computed total energy for both methods. The variational integrator preserves the total energy well. There is no drift in the computed total energy, and the mean variation is $1.0835 \times 10^{-2} \text{ Nm}$. But, there is a notable dissipation of the computed total energy for the Runge–

Kutta method. Note that the computed total energy would further decrease as the simulation time increases.

Example 3 (Spherical Pendulum) A spherical pendulum is a mass particle connected to a frictionless, two degree-of-freedom pivot by a rigid massless link. The mass particle acts under a uniform gravitational potential. The configuration space is the two-sphere $\mathbb{S}^2 = \{q \in \mathbb{R}^3 \mid \|q\| = 1\}$. It is common to parametrize the two-sphere by two angles, but this description of the spherical pendulum has a singularity. Any trajectory near the singularity encounters numerical ill-conditioning. Furthermore, this leads to complicated expressions involving trigonometric functions.

Here we develop equations of motion for a spherical pendulum using the global structure of the two-sphere without parametrization. In the previous example, we develop equations of motion for a planar pendulum using the fact that the one-sphere \mathbb{S}^1 is diffeomorphic to the special orthogonal group $\text{SO}(2)$. But, the two-sphere is not diffeomorphic to a Lie group. Instead, there exists a natural Lie group action on the two-sphere. That is the 3-dimensional special orthogonal group $\text{SO}(3)$, a group of 3×3 orthogonal matrices with determinant 1, i.e. $\text{SO}(3) = \{R \in \mathbb{R}^{3 \times 3} \mid R^T R = I_{3 \times 3}, \det[R] = 1\}$. The special orthogonal group $\text{SO}(3)$ acts on the two-sphere in a transitive way; for any $q_1, q_2 \in \mathbb{S}^2$, there exists a $R \in \text{SO}(3)$ such that $q_2 = Rq_1$. Thus, the discrete update for the two-sphere can be expressed in terms of a rotation matrix as (19). This is a key idea to develop a discrete equations of motion for a spherical pendulum.

Continuous equations of motion: Let $q \in \mathbb{S}^2$ be a unit vector from the pivot point to the point mass. The Lagrangian for a spherical pendulum can be written as

$$L(q, \dot{q}) = \frac{1}{2} m l^2 \dot{q} \cdot \dot{q} + m g l e_3 \cdot q, \quad (28)$$

where the gravity direction is assumed to be $e_3 = [0; 0; 1] \in \mathbb{R}^3$. The mass and the length of the pendulum are denoted by $m, l \in \mathbb{R}$, respectively. The infinitesimal variation of the unit vector q can be written in terms of the vector cross product as

$$\delta q = \xi \times q, \quad (29)$$

where $\xi \in \mathbb{R}^3$ is constrained to be orthogonal to the unit vector, i.e. $\xi \cdot q = 0$. Using this expression for the infinitesimal variation, Hamilton's principle yields the following continuous equations of motion.

$$\ddot{q} + (\dot{q} \cdot \dot{q})q + \frac{g}{l}(q \times (q \times e_3)) = 0. \quad (30)$$

Since $\dot{q} = \omega \times q$ for some angular velocity $\omega \in \mathbb{R}^3$ satisfying $\omega \cdot q = 0$, this can also be equivalently written as

$$\dot{\omega} - \frac{g}{l} q \times e_3 = 0, \quad (31)$$

$$\dot{q} = \omega \times q. \quad (32)$$

These are global equations of motion for a spherical pendulum; these are much simpler than the equations expressed in term of angles, and they have no singularity.

Discrete equations of motion: We develop a variational integrator for the spherical pendulum defined on \mathbb{S}^2 . Since the special orthogonal group acts on the two-sphere transitively, we can define the discrete update map for the unit vector as

$$q_{k+1} = F_k q_k \quad (33)$$

for $F_k \in \text{SO}(3)$. The rotation matrix F_k is not uniquely defined by this condition, since $\exp(\lambda \hat{q}_k)$, which corresponds to a rotation about the q_k direction fixes the vector q_k . Consequently, if F_k satisfies (33), then $F_k \exp(\lambda \hat{q}_k)$ does as well. We avoid this ambiguity by requiring that F_k does not rotate about q_k , which can be achieved by letting $F_k = \exp(\hat{f}_k)$, where $f_k \cdot q_k = 0$.

Define a discrete Lagrangian $L_d: \mathbb{S}^2 \times \mathbb{S}^2 \rightarrow \mathbb{R}$ to be

$$L_d(q_k, q_{k+1}) = \frac{1}{2h} m l^2 (q_{k+1} - q_k) \cdot (q_{k+1} - q_k) + \frac{h}{2} m g l e_3 \cdot q_k + \frac{h}{2} m g l e_3 \cdot q_{k+1}.$$

The variation of q_k is the same as (29), namely

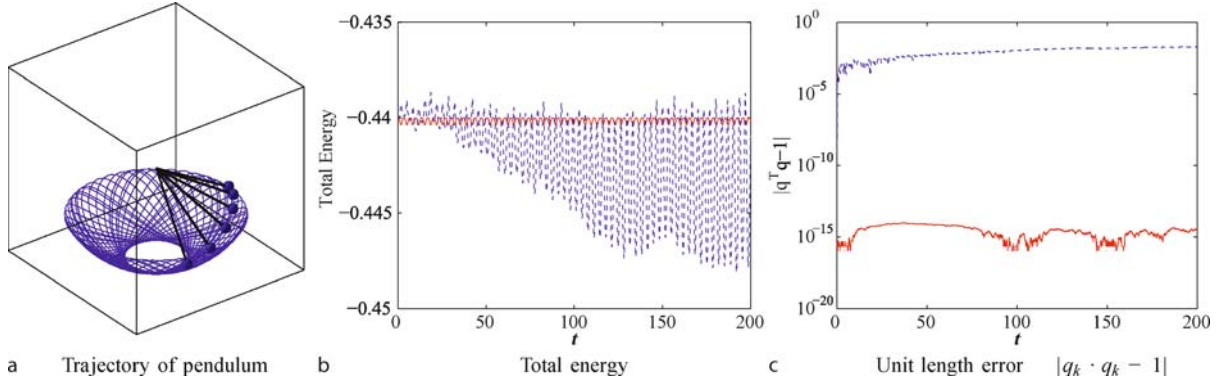
$$\delta q_k = \xi_k \times q_k \quad (34)$$

for $\xi_k \in \mathbb{R}^3$ with a constraint $\xi_k \cdot q_k = 0$. Using this discrete Lagrangian and the expression for the variation, discrete Hamilton's principle yields the following discrete equations of motion for a spherical pendulum.

$$q_{k+1} = \left(h \omega_k + \frac{h^2 g}{2l} q_k \times e_3 \right) \times q_k + \left(1 - \left\| h \omega_k + \frac{h^2 g}{2l} q_k \times e_3 \right\|^2 \right)^{1/2} q_k, \quad (35)$$

$$\omega_{k+1} = \omega_k + \frac{hg}{2l} q_k \times e_3 + \frac{hg}{2l} q_{k+1} \times e_3. \quad (36)$$

Since an explicit solution for $F_k \in \text{SO}(3)$ can be obtained in this case, the rotation matrix F_k does not appear in the equations of motion. This variational integrator on \mathbb{S}^2 exactly preserves the unit length of q_k , the constraint



Discrete Control Systems, Figure 3

Numerical simulation of a spherical pendulum (RK45: blue, dotted, VI: red, solid)

$q_k \cdot \omega_k = 0$, and the third component of the angular velocity $\omega_k \cdot e_3$ which is conserved since gravity exerts no moment along the gravity direction e_3 .

Numerical example: We compare the computational properties of the discrete equations of motion given by (35) and (36) with a 4(5)th order variable step size Runge–Kutta method for (31) and (32). We choose $m = 1$ kg, $l = 9.81$ m. The initial conditions are $q_0 = [\sqrt{3}/2, 0, 1/2]$, $\omega_0 = 0.1[\sqrt{3}, 0, 3]$ rad/sec, and the total energy is $E = -0.44$ Nm. The simulation time is 200 sec, and the step-size $h = 0.05$ of the discrete equations of motion is chosen such that the CPU times are identical. Figure 3 shows the computed total energy and the unit length errors. The variational integrator preserves the total energy and the structure of the two-sphere well. The mean total energy deviation is 1.5460×10^{-4} Nm, and the mean unit length error is 3.2476×10^{-15} . But, there is a notable dissipation of the computed total energy for the Runge–Kutta method. The Runge–Kutta method also fails to preserve the structure of the two-sphere. The mean unit length error is 1.0164×10^{-2} .

Example 4 (Rigid Body in a Potential Field) Consider a rigid body under a potential field that is dependent on the position and the attitude of the body. The configuration space is the special Euclidean group, which is a semi-direct product of the special orthogonal group and Euclidean space, i.e. $SE(3) = SO(3) \ltimes \mathbb{R}^3$.

Continuous equations of motion: The equations of motion for a rigid body can be developed either from Hamilton's principle (see [26]) in a similar way as Example 2, or directly from the generalized discrete Euler–Poincaré equation given at (5). Here, we summarize the results. Let $m \in \mathbb{R}$ and $J \in \mathbb{R}^{3 \times 3}$ be the mass and the moment of inertia matrix of a rigid body. For $(R, x) \in SE(3)$, the linear

transformation from the body-fixed frame to the inertial frame is denoted by the rotation matrix $R \in SO(3)$, and the position of the mass center in the inertial frame is denoted by a vector $x \in \mathbb{R}^3$. The vectors $\Omega, v \in \mathbb{R}^3$ are the angular velocity in the body-fixed frame, and the translational velocity in the inertial frame, respectively. Suppose that the rigid body acts under a configuration-dependent potential $U: SE(3) \rightarrow \mathbb{R}$. The continuous equations of motion for the rigid body can be written as

$$\dot{R} = R\hat{\Omega}, \quad (37)$$

$$\dot{x} = v, \quad (38)$$

$$J\dot{\hat{\Omega}} + \hat{\Omega} \times J\hat{\Omega} = M, \quad (39)$$

$$m\dot{v} = -\frac{\partial U}{\partial x}, \quad (40)$$

where the hat map $\hat{\cdot}$ is an isomorphism from \mathbb{R}^3 to 3×3 skew-symmetric matrices $\mathfrak{so}(3)$, defined such that $\hat{x}y = x \times y$ for any $x, y \in \mathbb{R}^3$. The moment due to the potential $M \in \mathbb{R}^3$ is obtained by the following relationship:

$$\hat{M} = \frac{\partial U}{\partial R}^T R - R^T \frac{\partial U}{\partial R}. \quad (41)$$

The matrix $\partial U / \partial R \in \mathbb{R}^{3 \times 3}$ is defined such that $[\partial U / \partial R]_{ij} = \partial U / \partial [R]_{ij}$ for $i, j \in \{1, 2, 3\}$, where the i, j th element of a matrix is denoted by $[\cdot]_{ij}$.

Discrete equations of motion: The corresponding discrete equations of motion are given by

$$hJ\widehat{\Omega}_k + \frac{h^2}{2}\hat{M}_k = F_k J_d - J_d F_k^T, \quad (42)$$

$$R_{k+1} = R_k F_k, \quad (43)$$

$$x_{k+1} = x_k + hv_k - \frac{h^2}{2m} \frac{\partial U_k}{\partial x_k}, \quad (44)$$

$$J\Omega_{k+1} = F_k^T J\Omega_k + \frac{h}{2} F_k^T M_k + \frac{h}{2} M_{k+1}, \quad (45)$$

$$mv_{k+1} = mv_k - \frac{h}{2} \frac{\partial U_k}{\partial x_k} - \frac{h}{2} \frac{\partial U_{k+1}}{\partial x_{k+1}}, \quad (46)$$

where $J_d \in \mathbb{R}^{3 \times 3}$ is a non-standard moment of inertia matrix defined as $J_d = 1/2 \text{tr}[J] I_{3 \times 3} - J$. For a given $(R_k, x_k, \Omega_k, v_k)$, we solve the implicit Eq. (42) to find $F_k \in \text{SO}(3)$. Then, the configuration at the next step R_{k+1}, x_{k+1} is obtained by (43) and (44), and the moment and force $M_{k+1}, (\partial U_{k+1})/(\partial x_{k+1})$ can be computed. Velocities Ω_{k+1}, v_{k+1} are obtained from (45) and (46). This defines a discrete flow map, $(R_k, x_k, \Omega_k, v_k) \mapsto (R_{k+1}, x_{k+1}, \Omega_{k+1}, v_{k+1})$, and this process can be repeated. This Lie group variational integrator on $\text{SE}(3)$ can be generalized to multiple rigid bodies acting under their mutual gravitational potential (see [26]).

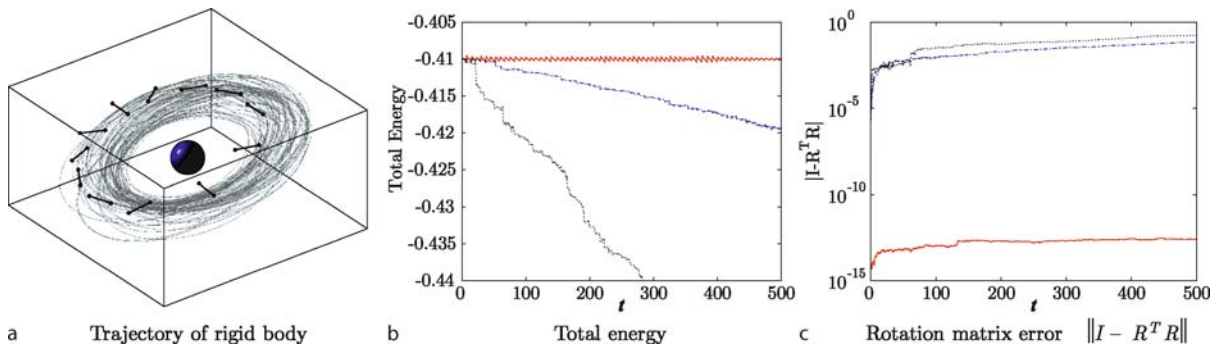
Numerical example: We compare the computational properties of the discrete equations of motion given by (42)–(46) with a 4(5)th order variable step size Runge–Kutta method for (37)–(40). In addition, we compute the attitude dynamics using quaternions on the unit three-sphere \mathbb{S}^3 . The attitude kinematics Eq. (37) is rewritten in terms of quaternions, and the corresponding equations are integrated by the same Runge–Kutta method.

We choose a dumbbell spacecraft, that is two spheres connected by a rigid massless rod, acting under a central gravitational potential. The resulting system is a restricted full two body problem. The dumbbell spacecraft model has an analytic expression for the gravitational potential, resulting in a nontrivial coupling between the attitude dynamics and the orbital dynamics.

As shown in Fig. 4a, the initial conditions are chosen such that the resulting motion is a near-circular orbit

combined with a rotational motion. Figure 4b and c show the computed total energy and the orthogonality errors of the rotation matrix. The Lie group variational integrator preserves the total energy and the Lie group structure of $\text{SO}(3)$. The mean total energy deviation is 2.5983×10^{-4} , and the mean orthogonality error is 1.8553×10^{-13} . But, there is a notable dissipation of the computed total energy and the orthogonality error for the Runge–Kutta method. The mean orthogonality errors for the Runge–Kutta method are 0.0287 and 0.0753, respectively, using kinematics equation with rotation matrices, and using the kinematics equation with quaternions. Thus, the attitude of the rigid body is not accurately computed for Runge–Kutta methods. It is interesting to see that the Runge–Kutta method with quaternions, which is generally assumed to have better computational properties than the kinematics equation with rotation matrices, has larger total energy error and orthogonality error. Since the unit length of the quaternion vector is not preserved in the numerical computations, orthogonality errors arise when converted to a rotation matrix. This suggests that it is critical to preserve the structure of $\text{SO}(3)$ in order to study the global characteristics of the rigid body dynamics.

The importance of simultaneously preserving the symplectic structure and the Lie group structure of the configuration space in rigid body dynamics can be observed numerically. Lie group variational integrators, which preserve both of these properties, are compared to methods that only preserve one, or neither, of these properties (see [27]). It is shown that the Lie group variational integrator exhibits greater numerical accuracy and efficiency. Due to these computational advantages, the Lie group variational integrator has been used to study the dynamics of the binary near-Earth asteroid 66391 (1999 KW₄) in joint work between the University of Michigan and the Jet Propulsion Laboratory, NASA (see [37]).



Discrete Control Systems, Figure 4

Numerical simulation of a dumbbell rigid body (LGVI: red, solid, RK45 with rotation matrices: blue, dash-dotted, RK45 with quaternions: black, dotted)

Optimal Control of Discrete Lagrangian and Hamiltonian Systems

Optimal control problems involve finding a control input such that a certain optimality objective is achieved under prescribed constraints. An optimal control problem that minimizes a performance index is described by a set of differential equations, which can be derived using Pontryagin's maximum principle. The equations of motion for a system are constrained by Lagrange multipliers, and necessary conditions for optimality is obtained by the calculus of variations. The solution for the corresponding two point boundary value problem provides the optimal control input. Alternatively, a sub-optimal control law is obtained by approximating the control input history with finite data points.

Discrete optimal control problems involve finding a control input for a given system described by discrete Lagrangian and Hamiltonian mechanics. The control inputs are parametrized by their values at each discrete time step, and the discrete equations of motion are derived from the discrete Lagrange–d'Alembert principle [21],

$$\delta \sum_{k=0}^{N-1} L_d(q_k, q_{k+1}) = \sum_{k=0}^{N-1} [F_d^-(q_k, q_{k+1}) \cdot \delta q_k + F_d^+(q_k, q_{k+1}) \cdot \delta q_{k+1}],$$

which modifies the discrete Hamilton's principle by taking into account the virtual work of the external forces. Discrete optimal control is in contrast to traditional techniques such as collocation, wherein the continuous equations of motion are imposed as constraints at a set of collocation points, since this approach induces constraints on the configuration at each discrete timestep.

Any optimal control algorithm can be applied to the discrete Lagrangian or Hamiltonian system. For an indirect method, our approach to a discrete optimal control problem can be considered as a multiple stage variational problem. The discrete equations of motion are derived by the discrete variational principle. The corresponding variational integrators are imposed as dynamic constraints to be satisfied by using Lagrange multipliers, and necessary conditions for optimality, expressed as discrete equations on multipliers, are obtained from a variational principle. For a direct method, control inputs can be optimized by using parameter optimization tools such as a sequential quadratic programming. The discrete optimal control can be characterized by discretizing the optimal control problem from the problem formulation stage.

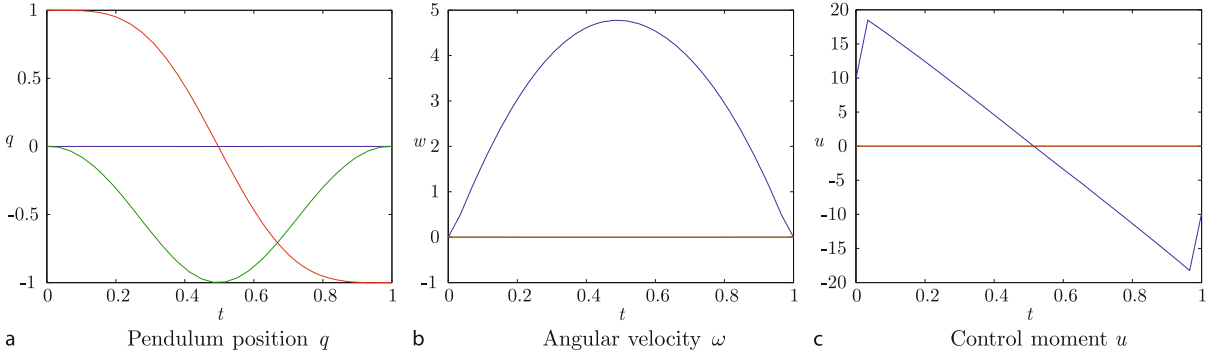
This method has substantial computational advantages when used to find an optimal control law. As discussed in the previous section, the discrete dynamics are more faithful to the continuous equations of motion, and consequently more accurate solutions to the optimal control problem are obtained. The external control inputs break the Lagrangian and Hamiltonian system structure. For example, the total energy is not conserved for a controlled mechanical system. But, the computational superiority of the discrete mechanics still holds for controlled systems. It has been shown that the discrete dynamics is more reliable even for controlled system as it computes the energy dissipation rate of controlled systems more accurately (see [31]). For example, this feature is extremely important in computing accurate optimal trajectories for long term spacecraft attitude maneuvers using low energy control inputs.

The discrete dynamics does not only provide an accurate optimal control input, but also enables us to find it efficiently. For the indirect optimal control approach, optimal solutions are usually sensitive to a small variation of multipliers. This causes difficulties, such as numerical ill-conditioning, when solving the necessary conditions for optimality expressed as a two point boundary value problem. Sensitivity derivatives along the discrete necessary conditions do not have numerical dissipation introduced by conventional numerical integration schemes. Thus, they are numerically more robust, and the necessary conditions can be solved computationally efficiently. For the direct optimal control approach, optimal control inputs can be obtained by using a larger discrete step size, which requires less computational load.

We illustrate the basic properties of the discrete optimal control using optimal control problems for the spherical pendulum and the rigid body model presented in the previous section.

Example 5 (Optimal Control of a Spherical Pendulum) We study an optimal control problem for the spherical pendulum described in Example 3. We assume that an external control moment $u \in \mathbb{R}^3$ acts on the pendulum. Control inputs are parametrized by their values at each time step, and the discrete equations of motion are modified to include the effects of the external control inputs by using the discrete Lagrange–d'Alembert principle. Since the component of the control moment that is parallel to the direction along the pendulum has no effect, we parametrize the control input as $u_k = q_k \times w_k$ for $w_k \in \mathbb{R}^3$.

The objective is to transfer the pendulum from a given initial configuration (q_0, ω_0) to a desired configuration (q^d, ω^d) during a fixed maneuver time N , while minimiz-



Discrete Control Systems, Figure 5
Optimal control of a spherical pendulum

ing the square of the weighted l_2 norm of the control moments.

$$\min_{w_k} J = \sum_{k=0}^N \frac{h}{2} u_k^T u_k = \sum_{k=0}^N \frac{h}{2} (q_k \times w_k)^T (q_k \times w_k).$$

We solve this optimal control problem by using a direct numerical optimization method. The terminal boundary condition is imposed as an equality constraint, and the $3(N+1)$ control input parameters $\{w_k\}_{k=0}^N$ are numerically optimized using sequential quadratic programming. This method is referred to as a DMOC (Discrete Mechanics and Optimal Control) approach (see [19]).

Figure 5 shows a optimal solution transferring the spherical pendulum from a hanging configuration given by $(q_0, \omega_0) = (e_3, 0_{3 \times 1})$ to an inverted configuration $(q^d, \omega^d) = (-e_3, 0_{3 \times 1})$ during 1 second. The time step size is $h = 0.033$. Experiment have shown that the DMOC approach can compute optimal solutions using larger step sizes than typical Runge–Kutta methods, and consequently, it requires less computational load. In this case, using a second-order accurate Runge–Kutta method, the same optimization code fails while giving error messages of inaccurate and singular gradient computations. It is presumed that the unit length errors of the Runge–Kutta method, shown in Example 3, cause numerical instabilities for the finite-difference gradient computations required for the sequential quadratic programming algorithm.

Example 6 (Optimal Control of a Rigid Body in a Potential Field) We study an optimal control problem of a rigid body using a dumbbell spacecraft model described in Example 4 (see [25] for detail). We assume that external control forces $u^f \in \mathbb{R}^3$, and control moment $u^m \in \mathbb{R}^3$ act on the dumbbell spacecraft. Control inputs are parametrized by their values at each time step, and the Lie group variational integrators are modified to include the effects of

the external control inputs by using discrete Lagrange–d’Alembert principle.

The objective is to transfer the dumbbell from a given initial configuration $(R_0, x_0, \Omega_0, v_0)$ to a desired configuration $(R^d, x^d, \Omega^d, v^d)$ during a fixed maneuver time N , while minimizing the square of the l_2 norm of the control inputs.

$$\min_{u_{k+1}} J = \sum_{k=0}^{N-1} \frac{h}{2} (u_{k+1}^f)^T W_f u_{k+1}^f + \frac{h}{2} (u_{k+1}^m)^T W_m u_{k+1}^m,$$

where $W_f, W_m \in \mathbb{R}^{3 \times 3}$ are symmetric positive definite matrices. Here we use a modified version of the discrete equations of motion with first order accuracy, as it yields a compact form for the necessary conditions.

Necessary conditions for optimality: We solve this optimal control problem by using an indirect optimization method, where necessary conditions for optimality are derived using variational arguments, and a solution of the corresponding two-point boundary value problem provides the optimal control. This approach is common in the optimal control literature; here the optimal control problem is discretized at the problem formulation level using the Lie group variational integrator presented in Sect. “Discrete Lagrangian and Hamiltonian Mechanics”.

$$\begin{aligned} J_a = & \sum_{k=0}^{N-1} \frac{h}{2} (u_{k+1}^f)^T W_f u_{k+1}^f + \frac{h}{2} (u_{k+1}^m)^T W_m u_{k+1}^m \\ & + \lambda_k^{1,T} \{-x_{k+1} + x_k + hv_k\} \\ & + \lambda_k^{2,T} \left\{ -mv_{k+1} + mv_k - h \frac{\partial U_{k+1}}{\partial x_{k+1}} + hu_{k+1}^f \right\} \\ & + \lambda_k^{3,T} (\log(F_k - R_k^T R_{k+1}))^\vee \\ & + \lambda_k^{4,T} \{-J\Omega_{k+1} + F_k^T J\Omega_k + h(M_{k+1} + u_{k+1}^m)\}, \end{aligned}$$

where $\lambda_k^1, \lambda_k^2, \lambda_k^3, \lambda_k^4 \in \mathbb{R}^3$ are Lagrange multipliers. The matrix logarithm is denoted by $\text{logm}: \text{SO}(3) \rightarrow \mathfrak{so}(3)$ and the vee map $\vee: \mathfrak{so}(3) \rightarrow \mathbb{R}^3$ is the inverse of the hat map introduced in Example 4. The logarithm form of (43) is used, and the constraint (42) is considered implicitly using constrained variations. Using similar expressions for the variation of the rotation matrix and the angular velocity given in (14) and (15), the infinitesimal variation can be written as

$$\begin{aligned} \delta \mathcal{J}_a = & \sum_{k=1}^{N-1} h \delta u_k^{f,T} \left\{ W_f u_k^f + \lambda_{k-1}^2 \right\} \\ & + h \delta u_k^{m,T} \left\{ W_m u_k^m + \lambda_{k-1}^4 \right\} \\ & + z_k^T \left\{ -\lambda_{k-1} + A_k^T \lambda_k \right\}, \end{aligned}$$

where $\lambda_k = [\lambda_k^1; \lambda_k^2; \lambda_k^3; \lambda_k^4] \in \mathbb{R}^{12}$, and $z_k \in \mathbb{R}^{12}$ represents the infinitesimal variation of $(R_k, x_k, \Omega_k, v_k)$, given by $z_k = [\text{logm}(R_k^T \delta R_k)^\vee; \delta x_k, \delta \Omega_k, \delta v_k]$. The matrix $A_k \in \mathbb{R}^{12 \times 12}$ is defined in terms of $(R_k, x_k, \Omega_k, v_k)$. Thus, necessary conditions for optimality are given by

$$u_{k+1}^f = -W_f^{-1} \lambda_k^2, \quad (47)$$

$$u_{k+1}^m = -W_m^{-1} \lambda_k^4, \quad (48)$$

$$\lambda_k = A_{k+1}^T \lambda_{k+1} \quad (49)$$

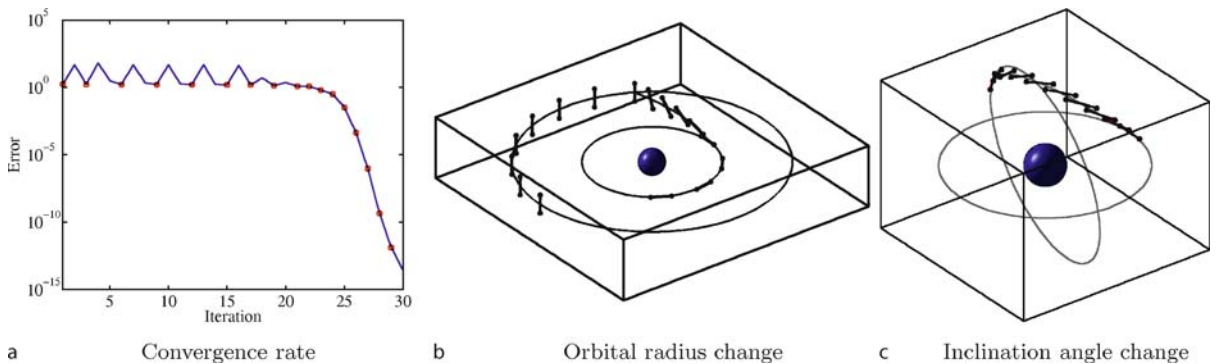
together with the discrete equations of motion and the boundary conditions.

Computational approach: Necessary conditions for optimality are expressed in terms of a two point boundary problem. The problem is to find the optimal discrete flow, multipliers, and control inputs to satisfy the equations of motion, optimality conditions, multiplier equations, and boundary conditions simultaneously. We use a neighboring extremal method (see [12]). A nominal solution sat-

isfying all of the necessary conditions except the boundary conditions is chosen. The unspecified initial multiplier is updated by successive linearization so as to satisfy the specified terminal boundary conditions in the limit. This is also referred to as the shooting method. The main advantage of the neighboring extremal method is that the number of iteration variables is small.

The difficulty is that the extremal solutions are sensitive to small changes in the unspecified initial multiplier values. The nonlinearities also make it hard to construct an accurate estimate of sensitivity, thereby resulting in numerical ill-conditioning. Therefore, it is important to compute the sensitivities accurately to apply the neighboring extremal method. Here the optimality conditions (47) and (48) are substituted into the equations of motion and the multiplier equations, which are linearized to obtain sensitivity derivatives of an optimal solution with respect to boundary conditions. Using this sensitivity, an initial guess of the unspecified initial conditions is iterated to satisfy the specified terminal conditions in the limit. Any type of Newton iteration can be applied. We use a line search with backtracking algorithm, referred to as Newton–Armijo iteration (see [22]).

Figure 6 shows optimized maneuvers, where a dumbbell spacecraft on a reference circular orbit is transferred to another circular orbit with a different orbital radius and inclination angle. Figure 6a shows the violation of the terminal boundary condition according to the number of iterations on a logarithmic scale. Red circles denote outer iterations in the Newton–Armijo iteration to compute the sensitivity derivatives. The error in satisfaction of the terminal boundary condition converges quickly to machine precision after the solution is close to the local minimum at around the 20th iteration. These convergence results are consistent with the quadratic convergence rates expected of Newton methods with accurately computed gradients.



Discrete Control Systems, Figure 6
Optimal control of a rigid body

The neighboring extremal method, also referred to as the shooting method, is numerically efficient in the sense that the number of optimization parameters is minimized. But, in general, this approach may be prone to numerical ill-conditioning (see [3]). A small change in the initial multiplier can cause highly nonlinear behavior of the terminal attitude and angular momentum. It is difficult to compute the gradient for Newton iterations accurately, and the numerical error may not converge. However, the numerical examples presented in this article show excellent numerical convergence properties. The dynamics of a rigid body arises from Hamiltonian mechanics, which have neutral stability, and its adjoint system is also neutrally stable. The proposed Lie group variational integrator and the discrete multiplier equations, obtained from variations expressed in the Lie algebra, preserve the neutral stability property numerically. Therefore the sensitivity derivatives are computed accurately.

Controlled Lagrangian Method for Discrete Lagrangian Systems

The method of controlled Lagrangians is a procedure for constructing feedback controllers for the stabilization of relative equilibria. It relies on choosing a parametrized family of controlled Lagrangians whose corresponding Euler–Lagrange flows are equivalent to the closed loop behavior of a Lagrangian system with external control forces. The condition that these two systems are equivalent results in matching conditions. Since the controlled system can now be viewed as a Lagrangian system with a modified Lagrangian, the global stability of the controlled system can be determined directly using Lyapunov stability analysis.

This approach originated in Bloch et al. [5] and was then developed in Auckly et al. [1]; Bloch et al. [6,7,8,9]; Hamberg [15,16]. A similar approach for Hamiltonian controlled systems was introduced and further studied in the work of Blankenstein, Ortega, van der Schaft, Maschke, Spong, and their collaborators (see, for example, [33,35] and related references). The two methods were shown to be equivalent in Chang et al. [13] and a nonholonomic version was developed in Zenkov et al. [40,41], and Bloch [4].

Since the method of controlled Lagrangians relies on relating the closed-loop dynamics of a controlled system with the Euler–Lagrange flow associated with a modified Lagrangian, it is natural to discretize this approach through the use of variational integrators. In Bloch et al. [10,11], a discrete theory of controlled Lagrangians was developed for variational integrators, and applied to the feedback stabilization of the unstable inverted equilibrium of the pendulum on a cart.

The pendulum on a cart is an example of an underactuated control problem, which has two degrees of freedom, given by the pendulum angle and the cart position. Only the cart position has control forces acting on it, and the stabilization of the pendulum has to be achieved indirectly through the coupling between the pendulum and the cart. The controlled Lagrangian is obtained by modifying the kinetic energy term, a process referred to as kinetic shaping. Similarly, it is possible to modify the potential energy term using potential shaping.

Since the pendulum on a cart model involves both a planar pendulum, and a cart that translates in one-dimension, the configuration space is a cylinder, $\mathbb{S}^1 \times \mathbb{R}$.

Continuous kinetic shaping: The Lagrangian has the form kinetic minus potential energy

$$L(q, \dot{q}) = \frac{1}{2} [\alpha \dot{\theta}^2 + 2\beta(\theta) \dot{\theta} \dot{s} + \gamma \dot{s}^2] - V(q), \quad (50)$$

and the corresponding controlled Euler–Lagrange dynamics is

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}} - \frac{\partial L}{\partial \theta} = 0, \quad (51)$$

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{s}} = u, \quad (52)$$

where u is the control input.

Since the potential energy is translation invariant, i. e., $V(q) = V(\theta)$, the *relative equilibria* $\theta = \theta_e, \dot{s} = \text{const}$ are unstable and given by non-degenerate critical points of $V(\theta)$. To stabilize the relative equilibria $\theta = \theta_e, \dot{s} = \text{const}$ with respect to θ , kinetic shaping is used. The controlled Lagrangian in this case is defined by

$$L^{\tau, \sigma}(q, \dot{q}) = L(\theta, \dot{\theta}, \dot{s} + \tau(\theta)\dot{\theta}) + \frac{1}{2} \sigma \gamma (\tau(\theta)\dot{\theta})^2, \quad (53)$$

where $\tau(\theta) = \kappa \beta(\theta)$. This velocity shift corresponds to a new choice of the horizontal space (see [8] for details). The dynamics is just the Euler–Lagrange dynamics for controlled Lagrangian (53),

$$\frac{d}{dt} \frac{\partial L^{\tau, \sigma}}{\partial \dot{\theta}} - \frac{\partial L^{\tau, \sigma}}{\partial \theta} = 0, \quad (54)$$

$$\frac{d}{dt} \frac{\partial L^{\tau, \sigma}}{\partial \dot{s}} = 0. \quad (55)$$

The Lagrangian (53) satisfies the simplified matching conditions of Bloch et al. [9] when the kinetic energy metric coefficient γ in (50) is constant.

Setting $u = -d(\gamma \tau(\theta) \dot{\theta})/dt$ defines the control input, makes Eqs. (52) and (55) identical, and results in controlled momentum conservation by dynamics (51)

and (52). Setting $\sigma = -1/\gamma\kappa$ makes Eqs. (51) and (54) identical when restricted to a level set of the controlled momentum.

Discrete kinetic shaping: Here, we adopt the following notation:

$$q_{k+1/2} = \frac{q_k + q_{k+1}}{2}, \quad \Delta q_k = q_{k+1} - q_k, \quad q_k = (\theta_k, s_k).$$

Then, a second-order accurate discrete Lagrangian is given by,

$$L_d(q_k, q_{k+1}) = hL(q_{k+1/2}, \Delta q_k/h).$$

The discrete dynamics is governed by the equations

$$\frac{\partial L_d(q_k, q_{k+1})}{\partial \theta_k} + \frac{\partial L_d(q_{k-1}, q_k)}{\partial \theta_k} = 0, \quad (56)$$

$$\frac{\partial L_d(q_k, q_{k+1})}{\partial s_k} + \frac{\partial L_d(q_{k-1}, q_k)}{\partial s_k} = -u_k, \quad (57)$$

where u_k is the control input. Similarly, the discrete controlled Lagrangian is,

$$L_d^{\tau, \sigma}(q_k, q_{k+1}) = hL^{\tau, \sigma}(q_{k+1/2}, \Delta q_k/h),$$

with discrete dynamics given by,

$$\frac{\partial L_d^{\tau, \sigma}(q_k, q_{k+1})}{\partial \theta_k} + \frac{\partial L_d^{\tau, \sigma}(q_{k-1}, q_k)}{\partial \theta_k} = 0, \quad (58)$$

$$\frac{\partial L_d^{\tau, \sigma}(q_k, q_{k+1})}{\partial s_k} + \frac{\partial L_d^{\tau, \sigma}(q_{k-1}, q_k)}{\partial s_k} = 0. \quad (59)$$

Equation (59) is equivalent to the *discrete controlled momentum conservation*:

$$p_k = \mu,$$

where

$$p_k = -\frac{\partial L_d^{\tau, \sigma}(q_k, q_{k+1})}{\partial s_k} = \frac{(1 + \gamma\kappa)\beta(\theta_{k+1/2})\Delta\theta_k + \gamma\Delta s_k}{h}.$$

Setting

$$u_k = -\frac{\gamma\Delta\theta_k\tau(\theta_{k+1/2}) - \gamma\Delta\theta_{k-1}\tau(\theta_{k-1/2})}{h}$$

makes Eqs. (57) and (59) identical and allows one to represent the discrete momentum equation (57) as the discrete momentum conservation law $p_k = p$.

The condition that (56)–(57) are equivalent to (58)–(59) yield the discrete matching conditions. The dynamics determined by Eqs. (56)–(57) restricted to the momentum level $p_k = p$ is equivalent to the dynamics of Eqs. (58)–(59) restricted to the momentum level $p_k = \mu$ if and only if the matching conditions

$$\sigma = -\frac{1}{\gamma\kappa}, \quad \mu = \frac{p}{1 + \gamma\kappa},$$

hold.

Numerical example: Simulating the behavior of the discrete controlled Lagrangian system involves viewing Eqs. (58)–(59) as an implicit update map $(q_{k-2}, q_{k-1}) \mapsto (q_{k-1}, q_k)$. This presupposes that the initial conditions are given in the form (q_0, q_1) ; however it is generally preferable to specify the initial conditions as (q_0, \dot{q}_0) . This is achieved by solving the boundary condition,

$$\frac{\partial L}{\partial \dot{q}}(q_0, \dot{q}_0) + D_1 L_d(q_0, q_1) + F_d^-(q_0, q_1) = 0,$$

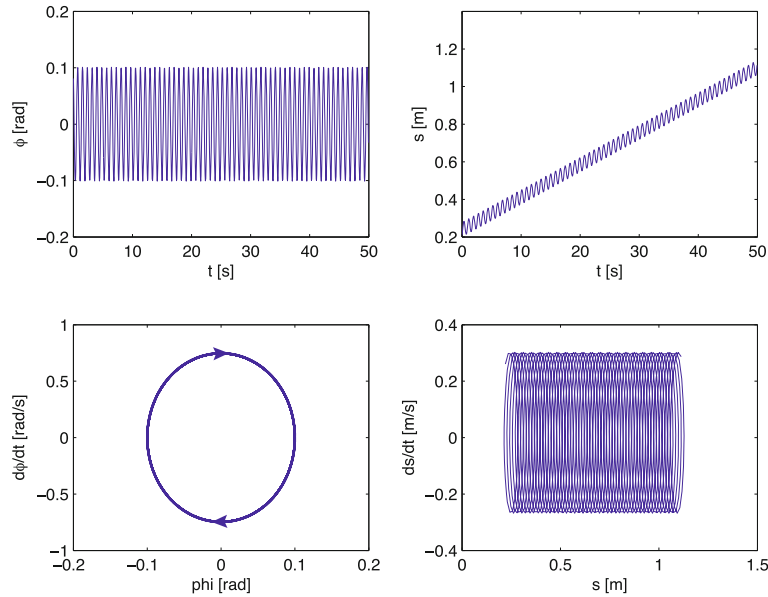
for q_1 . Once the initial conditions are expressed in the form (q_0, q_1) , the discrete evolution can be obtained using the implicit update map.

We first consider the case of kinetic shaping on a level surface, when κ is twice the critical value, and without dissipation. Here, $h = 0.05$ sec, $m = 0.14$ kg, $M = 0.44$ kg, and $l = 0.215$ m. As shown in Fig. 7, the θ dynamics is stabilized, but since there is no dissipation, the oscillations are sustained. The s dynamics exhibits both a drift and oscillations, as potential shaping is necessary to stabilize the translational dynamics.

Future Directions

Discrete Receding Horizon Optimal Control: The existing work on discrete optimal control has been primarily focused on constructing the optimal trajectory in an open loop sense. In practice, model uncertainty and actuation errors necessitate the use of feedback control, and it would be interesting to extend the existing work on optimal control of discrete systems to the feedback setting by adopting a receding horizon approach.

Discrete State Estimation: In feedback control, one typically assumes complete knowledge regarding the state of the system, an assumption that is often unrealistic in practice. The general problem of state estimation in the context of discrete mechanics would rely on good numerical methods for quantifying the propagation of uncertainty by solving the Liouville equation, which describes the evolution of a phase space distribution function advected by



Discrete Control Systems, Figure 7

Discrete controlled dynamics with kinetic shaping and without dissipation. The discrete controlled system stabilizes the θ motion about the equilibrium, but the s dynamics is not stabilized; since there is no dissipation, the oscillations are sustained

a prescribed vector field. In the setting of Hamiltonian systems, the solution of the Liouville equation can be solved by the method of characteristics (Scheeres et al. [38]). This implies that a collocational approach (Xiu [39]) combined with Lie group variational integrators, and interpolation based on noncommutative harmonic analysis on Lie groups could yield an efficient means of propagating uncertainty, and serve as the basis of a discrete state estimation algorithm.

Forced Symplectic-Energy-Momentum Variational Integrators: One of the motivations for studying the control of Lagrangian systems using the method of controlled Lagrangians is that the method provides a natural candidate Lyapunov function to study the global stability properties of the controlled system. In the discrete theory, this approach is complicated by the fact that the energy of a discrete Lagrangian system is not exactly conserved, but rather oscillates in a bounded fashion.

This can be addressed by considering the symplectic-energy-momentum [20] analogue to the discrete Lagrange-d'Alembert principle,

$$\delta \sum_{k=0}^{N_1} L_d(q_k, q_{k+1}, h_k) = \sum_{k=0}^{N-1} [F_d^-(q_k, q_{k+1}, h_k) \cdot \delta q_k + F_d^+(q_k, q_{k+1}, h_k) \cdot \delta q_{k+1}],$$

where the timestep h_k is allowed to vary, and is chosen to satisfy the variational principle. The variations in h_k

yield an Euler-Lagrange equation that reduces to the conservation of discrete energy in the absence of external forces. By developing a theory of controlled Lagrangians around a geometric integrator based on the symplectic-energy-momentum version of the Lagrange-d'Alembert principle, one would potentially be able to use Lyapunov techniques to study the global stability of the resulting numerical control algorithms.

Acknowledgments

TL and ML have been supported in part by NSF Grant DMS-0504747 and DMS-0726263. TL and NHM have been supported in part by NSF Grant ECS-0244977 and CMS-0555797.

Bibliography

Primary Literature

1. Auckly D, Kapitanski L, White W (2000) Control of nonlinear underactuated systems. *Commun Pure Appl Math* 53:354–369
2. Benettin G, Giorgilli A (1994) On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms. *J Stat Phys* 74:1117–1143
3. Betts JT (2001) *Practical Methods for Optimal Control Using Nonlinear Programming*. SIAM, Philadelphia, PA
4. Bloch AM (2003) *Nonholonomic Mechanics and Control*. In: *Interdisciplinary Applied Mathematics*, vol 24. Springer, New York

5. Bloch AM, Leonard N, Marsden JE (1997) Matching and stabilization using controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Hyatt Regency San Diego, San Diego, CA, 10–12 December 1997, pp 2356–2361
6. Bloch AM, Leonard N, Marsden JE (1998) Matching and stabilization by the method of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Hyatt Regency Westshore, Tampa, FL, 16–18 December 1998, pp 1446–1451
7. Bloch AM, Leonard N, Marsden JE (1999) Potential shaping and the method of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Crowne Plaza Hotel and Resort, Phoenix, AZ, 7–10 December 1999, pp 1652–1657
8. Bloch AM, Leonard NE, Marsden JE (2000) Controlled Lagrangians and the stabilization of mechanical systems I: The first matching theorem. *IEEE Trans Syst Control* 45:2253–2270
9. Bloch AM, Chang DE, Leonard NE, Marsden JE (2001) Controlled Lagrangians and the stabilization of mechanical systems II: Potential shaping. *IEEE Trans Autom Contr* 46:1556–1571
10. Bloch AM, Leok M, Marsden JE, Zenkov DV (2005) Controlled Lagrangians and stabilization of the discrete cart-pendulum system. In: Proceedings of the IEEE Conference on Decision and Control. Melia Seville, Seville, Spain, 12–15 December 2005, pp 6579–6584
11. Bloch AM, Leok M, Marsden JE, Zenkov DV (2006) Controlled Lagrangians and potential shaping for stabilization of discrete mechanical systems. In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 3333–3338
12. Bryson AE, Ho Y (1975) *Applied Optimal Control*. Hemisphere, Washington, D.C.
13. Chang D-E, Bloch AM, Leonard NE, Marsden JE, Woolsey C (2002) The equivalence of controlled Lagrangian and controlled Hamiltonian systems. *Control Calc Var* (special issue dedicated to Lions JL) 8:393–422
14. Hairer E, Lubich C, Wanner G (2006) *Geometric Numerical Integration*, 2nd edn. Springer Series in Computational Mathematics, vol 31. Springer, Berlin
15. Hamberg J (1999) General matching conditions in the theory of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Crowne Plaza Hotel and Resort, Phoenix, AZ, 7–10 December 1999, pp 2519–2523
16. Hamberg J (2000) Controlled Lagrangians, symmetries and conditions for strong matching. In: *Lagrangian and Hamiltonian Methods for Nonlinear Control*. Elsevier, Oxford
17. Hussein II, Leok M, Sanyal AK, Bloch AM (2006) A discrete variational integrator for optimal control problems in $SO(3)$. In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 6636–6641
18. Iserles A, Munthe-Kaas H, Nørsett SP, Zanna A (2000) Lie-group methods. In: *Acta Numerica*, vol 9. Cambridge University Press, Cambridge, pp 215–365
19. Junge D, Marsden JE, Ober-Blöbaum S (2005) Discrete mechanics and optimal control. In: *IFAC Congress*, Praha, Prague, 3–8 July 2005
20. Kane C, Marsden JE, Ortiz M (1999) Symplectic-energy-momentum preserving variational integrators. *J Math Phys* 40(7):3353–3371
21. Kane C, Marsden JE, Ortiz M, West M (2000) Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. *Int J Numer Meth Eng* 49(10):1295–1325
22. Kelley CT (1995) *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, PA
23. Lee T, Leok M, McClamroch NH (2005) Attitude maneuvers of a rigid spacecraft in a circular orbit. In: Proceedings of the American Control Conference. Portland Hilton, Portland, OR, 8–10 June 2005, pp 1742–1747
24. Lee T, Leok M, McClamroch NH (2005) A Lie group variational integrator for the attitude dynamics of a rigid body with applications to the 3D pendulum. In: Proceedings of the IEEE Conference on Control Applications. Toronto, Canada, 28–31 August 2005, pp 962–967
25. Lee T, Leok M, McClamroch NH (2006) Optimal control of a rigid body using geometrically exact computations on $SE(3)$. In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 2710–2715
26. Lee T, Leok M, McClamroch NH (2007) Lie group variational integrators for the full body problem. *Comput Method Appl Mech Eng* 196:2907–2924
27. Lee T, Leok M, McClamroch NH (2007) Lie group variational integrators for the full body problem in orbital mechanics. *Celest Mech Dyn Astron* 98(2):121–144
28. Leimkuhler B, Reich S (2004) *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics, vol 14. Cambridge University Press, Cambridge
29. Leok M (2004) *Foundations of Computational Geometric Mechanics*. Ph D thesis, California Institute of Technology
30. Marsden JE, Ratiu TS (1999) *Introduction to Mechanics and Symmetry*, 2nd edn. Texts in Applied Mathematics, vol 17. Springer, New York
31. Marsden JE, West M (2001) Discrete mechanics and variational integrators. In: *Acta Numerica*, vol 10. Cambridge University Press, Cambridge, pp 317–514
32. Marsden JE, Pekarsky S, Shkoller S (1999) Discrete Euler–Poincaré and Lie–Poisson equations. *Nonlinearity* 12(6):1647–1662
33. Maschke B, Ortega R, van der Schaft A (2001) Energy-based Lyapunov functions for forced Hamiltonian systems with dissipation. *IEEE Trans Autom Contr* 45:1498–1502
34. Moser J, Veselov AP (1991) Discrete versions of some classical integrable systems and factorization of matrix polynomials. *Commun Math Phys* 139:217–243
35. Ortega R, Spong MW, Gómez-Estern F, Blankenstein G (2002) Stabilization of a class of underactuated mechanical systems via interconnection and damping assignment. *IEEE Trans Autom Contr* 47:1218–1233
36. Sanz-Serna JM (1992) Symplectic integrators for Hamiltonian problems: an overview. In: *Acta Numerica*, vol 1. Cambridge University Press, Cambridge, pp 243–286
37. Scheeres DJ, Fahnestock EG, Ostro SJ, Margot JL, Benner LAM, Broschart SB, Bellerose J, Giorgini JD, Nolan MC, Magri C, Pravec P, Scheirich P, Rose R, Jurgens RF, De Jong EM, Suzuki S (2006) Dynamical configuration of binary near-Earth asteroid (66391) 1999 KW4. *Science* 314:1280–1283
38. Scheeres DJ, Hsiao F-Y, Park RS, Villac BF, Maruskin JM (2006) Fundamental limits on spacecraft orbit uncertainty and distribution propagation. *J Astronaut Sci* 54:505–523

39. Xiu D (2007) Efficient collocation approach for parametric uncertainty analysis. *Comm Comput Phys* 2:293–309
40. Zenkov DV, Bloch AM, Leonard NE, Marsden JE (2000) Matching and stabilization of low-dimensional nonholonomic systems. In: *Proceedings of the IEEE Conference on Decision and Control*. Sydney Convention and Exhibition Centre, Sydney, NSW Australia; 12–15 December 2000, pp 1289–1295
41. Zenkov DV, Bloch AM, Leonard NE, Marsden JE (2002) Flat nonholonomic matching. In: *Proceedings of the American Control Conference*. Hilton Anchorage, Anchorage, AK, 8–10 May 2002, pp 2812–2817

Books and Reviews

- Bloch AM (2003) *Nonholonomic Mechanics and Control*. *Interdisciplinary Appl Math*, vol 24. Springer
- Bullo F, Lewis AD (2005) *Geometric control of mechanical systems*. *Texts in Applied Mathematics*, vol 49. Springer, New York
- Hairer E, Lubich C, Wanner G (2006) *Geometric Numerical Integration*, 2nd edn. *Springer Series in Computational Mathematics*, vol 31. Springer, Berlin
- Iserles A, Munthe-Kaas H, Nørsett SP, Zanna A (2000) Lie-group methods. In: *Acta Numerica*, vol 9. Cambridge University Press, Cambridge, pp 215–365
- Leimkuhler B, Reich S (2004) *Simulating Hamiltonian Dynamics*. *Cambridge Monographs on Applied and Computational Mathematics*, vol 14. Cambridge University Press, Cambridge
- Marsden JE, Ratiu TS (1999) *Introduction to Mechanics and Symmetry*, 2nd edn. *Texts in Applied Mathematics*, vol 17. Springer
- Marsden JE, West M (2001) Discrete mechanics and variational integrators. In: *Acta Numerica*, vol 10. Cambridge University Press, Cambridge, pp 317–514
- Sanz-Serna JM (1992) Symplectic integrators for Hamiltonian problems: an overview. In: *Acta Numerica*, vol 1. Cambridge University Press, Cambridge, pp 243–286

move in response to an external force only if this force is large enough.

Scaling The fact that each of two quantities varies in a power-law relationship to the other.

Random manifold A single elastic structure (line, sheet) embedded in a random environment.

Bragg glass A periodic elastic structure embedded in a weakly disordered environment, nearly as ordered as a solid but exhibiting some characteristics normally associated with glasses.

Creep Very slow response at finite temperature of a pinned structure in response to an external force.

Definition of the Subject

Many seemingly different systems, with extremely different microscopic physics, ranging from magnets to superconductors, share the same essential ingredients and can be described under the unifying concept of disordered elastic media. In all these systems, an internal elastic structure, such as an interface between regions of opposite magnetization in magnetic systems, is subject to the effects of disorder existing in the material. A specially interesting feature of all these systems is that these disordered elastic structures can be set in motion by applying an external force on them (e.g. a magnetic field sets in motion a magnetic interface), and that motion will be drastically affected by the presence of the disorder. What properties result from this competition between elasticity and disorder is a complicated problem which constitutes the essence of the physics of disordered elastic media. The resulting physics present characteristics similar to those of glasses. This poses extremely challenging fundamental questions in determining the static and dynamic properties of these systems. Understanding both the static and dynamic properties of these objects is not only an important question from a fundamental point of view, but also has strong practical applications. Indeed, being able to write an interface between two regions of magnetization or polarization, and the speed of writing and stability of such regions, is what conditions, in particular, our ability to store information in such systems, as (for example) recordings on a magnetic hard drive. The physics pertaining to disordered elastic media directly condition how we can use these systems for practical applications.

Introduction

Understanding the statics and dynamics of elastic systems in a random environment is a long-standing problem with important applications for a host of experimental systems. Such problems can be split into two broad categories: (a)

Disordered Elastic Media

THIERRY GIAMARCHI
DPMC-MaNEP, University of Geneva,
Geneva, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Static Properties of Disordered Elastic Media
Pinning and Dynamics
Future Directions
Bibliography

Glossary

Pinning An action exerted by impurities on an object. The object has a preferential position in space, and will

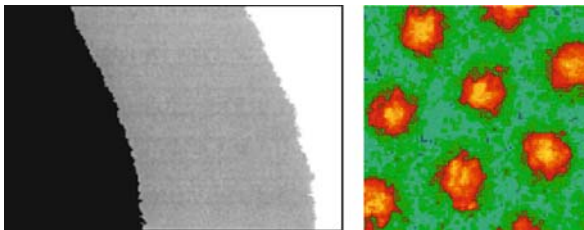
propagating interfaces such as magnetic [1,2,3,4], spintronic [5,6], or ferroelectric [7,8] domain walls, fluid invasion in porous media [9], contact lines in wetting [10], epitaxial growth [11] or crack propagation [12,13]; (b) periodic systems such as vortex lattices in type II superconductors [14,15,16], charge density waves [17,18], magnetic bubbles [19], colloids [20], Wigner crystals of classical particles [21] or of electrons [22,23].

Although all these systems have very different microscopic descriptions, one aspect of their physics is identical at a more macroscopic scale: An object exists that obeys a macroscopic elastic description. For case (a) this is an interface separating two different regions of the system, for example in a magnetic material a domain wall separating regions of opposite magnetization. An example of such an interface is shown in Fig. 1. Since creating an interface costs energy, the interface left to itself would tend to be flat, and there is an elastic cost to its deformations. Since this object lives inside a microscopic crystal with disorder, it is also subjected to potentials that tend to roughen it and pin it in specific regions of space. This interface can be set in motion by applying an external force on it, caused for example by a magnetic field for the magnetic domain wall or an electric field for a ferroelectric. For case (b), that

of periodic systems, a similar physics exists. A “crystal” of objects (lines for vortices, points for magnetic bubbles and colloids, or sheets for charge-density waves) exists inside the system. Since these objects repel each other, in the absence of disorder they would tend to form a perfect periodic crystal. In a way similar to the interfaces, such crystals can be set in motion by applying an external force (for example a current, in the case of vortices). The disorder present at the microscopic level tends to pin this crystal. It is important to note that one is dealing here with the physics of a crystal embedded in an external medium containing impurities. The disorder can thus vary at a length scale much smaller than the lattice spacing of the moving crystal, which leads to a physics radically novel compared to that of chemical impurities in a regular solid. Understanding the physics, both static and dynamic, of these objects has thus been a considerable challenge in the last 50 years or so. There are several reasons for pursuing this interest, and for tackling the challenges posed by this field of disordered elastic media.

First, at the fundamental level these systems pose difficult and important questions. It has been known since the 1970s that the presence of disorder is crucial [25] and changes the physics completely. While the resulting models are very difficult to solve, they have contributed to pushing the limits of our understanding of disordered systems, and to developing new techniques of statistical physics to deal with such issues. In particular, it is clear that from the competition between disorder and elasticity emerges a complicated energy landscape with many metastable states. This results in glassy properties [26] such as hysteresis and history dependence of the static configuration. Initially viewed as toy models of glasses, these systems have acquired their own importance and have posed their own challenging questions. Understanding the static properties of such systems has stimulated the development of sophisticated approaches such as replica theory [27], functional renormalization groups [28] and numerical methods. Much progress has recently been accomplished due to both analytical and numerical advances. If the static view allows us to improve our techniques of statistical physics, the dynamic view is even more complicated since most of our theoretical tools fail. These systems thus provide wonderful motivations to develop new techniques to tackle the out-of-equilibrium dynamics of disordered systems and to understand and unify the concepts of out-of-equilibrium physics of glasses.

Second, in addition to this theoretical motivation, the ability to apply these concepts in so many different physical systems is a tremendous motivation and challenge. The various realizations allow us to apply stringent tests to pro-



Disordered Elastic Media, Figure 1

Left: an interface in a magnetic system separating two different magnetic polarizations (dark and white). The image is $90 \times 72 \mu\text{m}^2$. The roughness of the domain wall due to the presence of disorder in the system is obvious on the image. Two positions of the interface are shown. Dark and gray correspond to two consecutive images after the interface has been pulled to the right by applying a magnetic field to the sample favoring the magnetization direction on the left of the interface. Such a magnetic field acts as a force pulling the domain wall. [From [1] (Copyright 1998 by the American Physical Society)]. **Right:** A vortex lattice image, from Scanning Tunneling Microscope, in the superconductor MgB_2 . The tips of the vortices on the surface of the sample are shown in the image, and correspond to the red parts. The image is about 250 nm^2 . In a perfectly pure system, the vortex lattice is a periodic arrangement (here in a triangular lattice) of objects of a given size (here the core size of the vortex). Disorder affects this perfectly periodic arrangement over a large distance. [From [24] (Copyright 2002 by the American Physical Society)]

posed theories and, as we will see, have very often served to kill misguided proposals or to put the theory on the right track in these quite complicated systems. Experiments in these systems can be remarkable by the range they offer. In vortex systems, for example, one can vary the vortex lattice spacing by several orders of magnitude just by changing the magnetic field applied to the sample, something impossible to do on a simple crystal. Similarly, for magnetic domain walls, measurements of the velocity in response to an external force can span about ten orders of magnitude. This interplay and exchange between theory and experiment has fueled the field and contributed greatly to its progress.

Last but not least, the phenomena studied for disordered elastic media have a potential impact for applications. Creating interfaces in magnetic or, more recently, ferroelectric materials is a way to store information (with “0” being one direction of magnetization or polarization and “1” being the reverse). This idea is at the root of information storage in a magnetic hard drive or in ferroelectric or magnetic bubble memory. How well one can store the information is thus directly related to both the static and dynamic properties of such interfaces. In particular, stability of the written information is ensured only if the interface is pinned and will not meander due to outside forces, such as thermal agitation. Similarly, there has been much interest recently on spintronic materials where the magnetic properties can be manipulated by applying electrical currents [5,6]. In the same vein, multiferroic materials [29] allow manipulation of ferroelectric properties by the application of magnetic fields. How the disordered interfaces behave in such materials will certainly determine their possible use for information technology. In a similar way, the vortex lattice in a superconductor is set in motion by the application of a current, while its motion generates a voltage. The dynamic properties of the vortex lattice, and how well it is pinned, thus directly affect the absence of resistance of a superconductor [14,15,16], hence its potential uses. Other examples, such propagation of fractures, clearly show the potential importance of such phenomena for applications.

This chapter presents basic concepts and results in this very active field. Section “[Interfaces and Basic Concepts](#)” presents the basic concepts and discusses the static properties of interfaces and domain walls. Section “[Periodic Systems and Bragg Glass](#)” deals with the periodic systems and their differences compared to interfaces. Section “[Pinning and Dynamics](#)” presents concepts and important questions for the dynamics of disordered elastic media, with focus on depinning in Sect. “[Depinning](#)”, on large velocity behavior in Sect. “[High Velocity Phase](#)” and

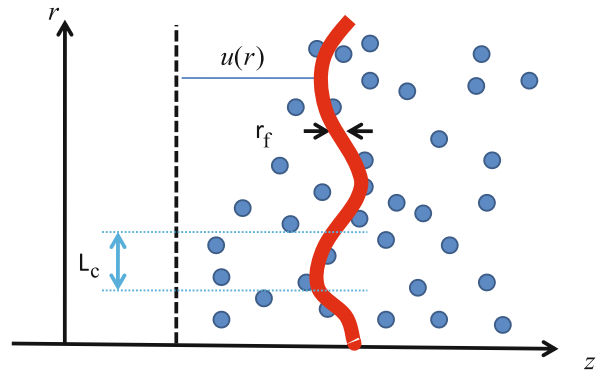
responses to small external forces in Sect. “[Small Applied Force and Creep Motion](#)”. Finally, Sect. “[Future Directions](#)” discusses future directions of and perspectives on the field.

Static Properties of Disordered Elastic Media

Interfaces and Basic Concepts

This section introduces the basic ingredients of the systems under study and discusses the specific case of interfaces. An interface is a sheet of dimension d living in a space of D dimensions. For realistic interfaces $D = d + 1$, but generalizations are of course possible. If we let r represent the internal coordinate of the interface and z all its transverse directions, the interface position is described by displacement $u(r)$ from a flat configuration. This totally determines the shape of the interface provided that u is univalued, i.e., that there are no overhangs or bubbles. The modelization of a one dimensional interface ($d = 1$) in a two dimensional film is shown in Fig. 2.

Since interface distortions cost elastic energy, a system’s zero temperature equilibrium configuration in the absence of disorder is flat. Deviations from this equilibrium position are described by a Hamiltonian $H[u]$ which is a function of the displacement u . For small displace-



Disordered Elastic Media, Figure 2

A one dimensional interface (such as a magnetic domain wall), shown in red, living in a two dimensional space (film). The position of the interface is determined (provided there are no overhangs or bubbles) by the displacement $u(r)$ from a flat configuration, indicated by the dashed line. In the absence of disorder, denoted by the blue dots, which pin the line in preferred positions in space, the line would be flat. The competition between elasticity and disorder leads to the physics of disordered elastic media and to glassy properties. The thickness of the line, denoted r_f , or the correlation length of the disorder define the Larkin length L_c for which the relative displacements are of the order of r_f , namely $u(L_c) - u(0) \sim r_f$

ments one can make the usual elastic approximation

$$H[u] = \frac{1}{2} \int \frac{d^d q}{(2\pi)^d} c(q) u_q^* u_q, \quad (1)$$

where u_q is the Fourier transform of $u(r)$ and $c(q)$ are the so-called elastic coefficients. If the elastic forces acting on the interface are short ranged then one has $c(q) = cq^2$ which corresponds to

$$H[u] = \frac{c}{2} \int d^d r (\nabla u(r))^2. \quad (2)$$

For some interfaces where long range interactions play a role, different forms of elasticity are possible. This is particularly the case when dipolar forces [30] are taken into account [8] or for the contact line in wetting [31] and crack propagation [32].

In addition to elastic energy, the interface gains some energy by coupling to the disorder. Two universality classes for the disorder exist (see Fig. 3). The random bond disorder corresponds to impurities that directly attract or repel the interface. In contrast, for random field disorder the pinning energy is affected by all the randomness that the interface has encountered in its previous motion. On a more technical level, random bond disorder couples in a symmetric way to the two order parameters on each side of the domain wall, while random field disorder introduces an asymmetry between these two inequivalent order parameters. If $V(r, z)$ denotes the random potential generated by the impurities the pinning energy:

$$H_{\text{dis}}[u] = \int d^d r \begin{cases} V(r, u(r)) & \text{random bond} \\ \int_0^{u(r)} dz V(r, z) & \text{random field.} \end{cases} \quad (3)$$

As is obvious from Fig. 3, even if the microscopic disorder is short-range correlated, as in the case of the random field disorder, the fact that the energy of the system integrates between two positions of the interfaces means that long range correlations exist if one considers the description only in terms of the interface. The full Hamiltonian given by (1) and (3) describes the properties of disordered elastic media, and despite its apparent simplicity hides an extremely rich physics.

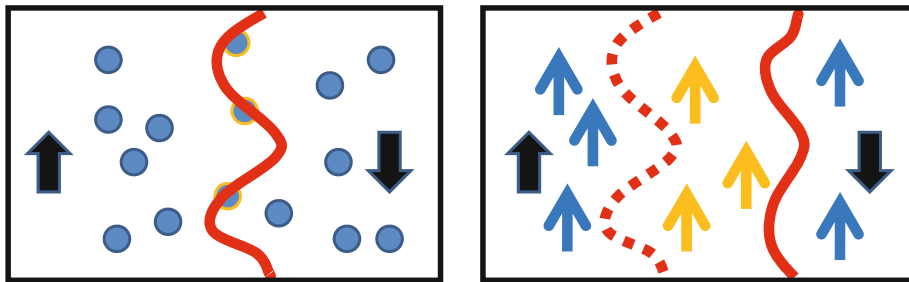
The competition between disorder and elasticity manifests itself in several properties of the interface. From an energetic point of view, this competition leads to a complicated energy landscape for the configurations of the system, with many metastable states leading to glassy properties. The competition also manifests itself in the shape of the interface. In particular, it deviates from the flat configuration and becomes rough. From the scaling of the relative displacements correlation function, a roughness exponent ζ can be defined from the correlation function of the displacements

$$B(r) = \overline{[u(r) - u(0)]^2} \propto r^{2\zeta}, \quad (4)$$

where $\langle \rangle$ denotes thermodynamic average and $\overline{\cdots}$ denotes disorder average. There are relations between the shape of the line and the energetic properties. In particular, (4) suggests that displacements would scale with distance as $u(L) \sim L^\zeta$. (2) suggests that the energy of a sample of size L fluctuates from sample to sample as

$$\Delta F(L) \sim L^{d-2+2\zeta}. \quad (5)$$

Given the complexity of the problem, several approximate methods have been put forward to portray the role



Disordered Elastic Media, Figure 3

The two universality classes of disorder (the names come from the magnetic realization of such systems). In both figures the domain wall in red separates two regions with different order parameters, denoted by the two thick black arrows. Left: random bond disorder. The impurities, denoted in blue, couple symmetrically to the two sides of the domain wall, thus only the impurities on the domain wall (denoted with the orange circle) contribute to the energy. Right: random field disorder. The impurities, denoted by the blue arrows, favor one of the two sides. Thus, all the impurities (denotes in orange) between two configurations of the domain wall contribute to the energy. This leads to a disorder seen by the domain wall which has long range correlations, even if the microscopic disorder is short-range correlated

of disorder. A remarkable model by which to probe the physics of such systems was introduced by Larkin [25], and goes by the name of the Larkin model. The idea is to focus on short length scale properties. In that case, the displacements are small and one can expand the disorder term in power of the displacements

$$H_{\text{dis}} = \int d^d r V(r, u(r)) \simeq \int d^d r [V(r, 0) + \nabla_r V(r, 0)|_{z=0} u(r)] . \quad (6)$$

The first term is a trivial constant and the second one indicates that the interface is subject to a random force

$$H_L = \int d^d r f(r) u(r) . \quad (7)$$

Although this model has several pathologies, it has the advantage of being quadratic in the displacement field u and thus of being exactly solvable. It shows that below $d = 4$, disorder plays a major role. The displacements grow as a function of distance and

$$B(r) = r^{4-d} . \quad (8)$$

This confirms that there is algebraic roughening of the interface with the displacements growing as a power law of distance. One can define the scaling $u(L) \sim L^\zeta$, with the Larkin model giving $\zeta = (4 - d)/2$. Below $d = 4$, disorder is relevant and drastically modifies the physical properties of the interface compared to those of the non-disordered one. In addition to the exponent itself, since the displacements grow unboundedly, there exists a length scale, L_c , called the Larkin length, at which the displacements become of the order of the only characteristic scale available, namely either the correlation length of the random potential or the size of the interface r_f , as shown in Fig. 2. Clearly, this is also the length where the applicability of the Larkin model breaks down, since beyond that length the potential $V(r, z)$ is not smooth anymore and thus expansion in powers of u is not justified. Beyond this length, the system will thus truly feel the effects of random potential. One can thus expect metastability, glassy effects and pinning to appear above that length. One has to determine the physics of this regime appearing above the Larkin length, which we will call the random manifold regime. The Larkin length is thus an important length scale for the static properties since it separates two different regimes for the interface. As I will discuss in Sect. “Pinning and Dynamics”, the Larkin length also has considerable consequences for the dynamics.

To solve the problem in the random manifold regime is not easy and requires greatly sophisticated techniques of statistical physics. To get a rough idea, one can simply use a scaling argument, known as the Flory argument [14]. At scale L , the elastic energy scales as $cL^{d-2}u(L)^2$. To estimate the disorder term is more complicated but one can assume that if L is large enough, one sums random variables $V(r, z)$. If one considers for example random bond disorder, because the disorder is short-range correlated, one has

$$\overline{V(r_1, u_1)V(r_2, u_2)} = D\delta^m(u_1 - u_2)\delta^d(r_1 - r_2) . \quad (9)$$

Since a $\delta(r)$ function has the dimension of $1/r^d$, this leads to the scaling

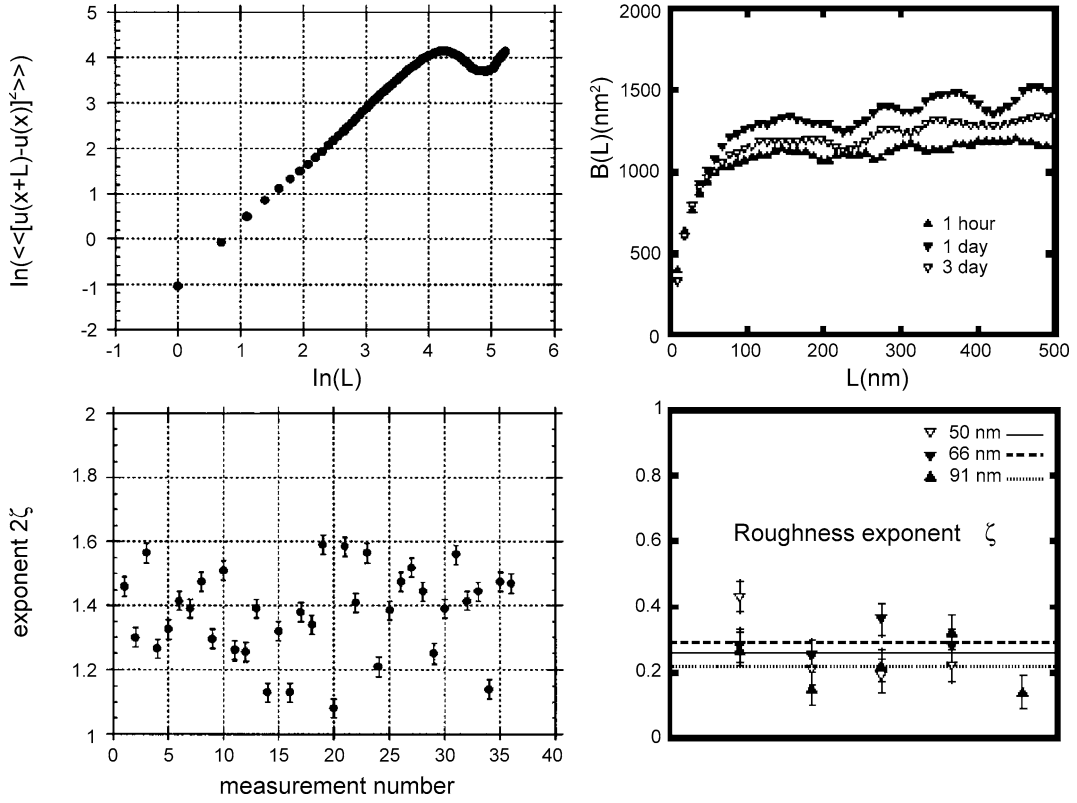
$$V(r, u) \sim D^{1/2}u(L)^{-m/2}L^{-d/2} , \quad (10)$$

where m is the number of components of u (for an interface $m = 1$). The disorder term thus scales as

$$H_{\text{dis}} = D^{1/2}u(L)^{-m/2}L^{d/2} . \quad (11)$$

Balancing the elastic and disorder terms leads to a scaling $u(L) \sim L^\zeta$ with $\zeta = \frac{4-d}{4+m}$ for the random bond case and $\zeta = \frac{4-d}{4-m}$ for the random field case. This argument suggests that even in the random manifold, the interface remains rough, with unbounded displacements growing with an *algebraic* roughness. The value of the exponent is characteristic of the universality class of the disorder, and different from the one occurring below $L < L_c$ where the Larkin model applies.

Clearly this simple argument needs to be substantiated by more rigorous calculations. Because the system is subjected to a random potential and the metastability and glassy effects matter, one can use the techniques traditionally used for disordered systems and spin glasses. For a one dimensional interface, this problem can be solved exactly and the roughness exponent $\zeta = 2/3$ obtained [33,34]. Note that this exact value is, of course, slightly different from the mean field estimate. In higher dimensions three main methods have been used to tackle this problem. The first one uses the so-called replica trick [26] to average over the disorder and then a variational approach to solve the corresponding field theory [27]. In this method the initial symmetry between replicas is broken, something familiar in spin-glasses and characteristics of glasses with many metastable minima in the energy. This approximate method gives back the Flory exponent. The second method applies the traditional renormalization technique (RG), so successful for standard critical phenom-



Disordered Elastic Media, Figure 4

Measurements of the roughness exponent ζ in two experimental system. These two examples show the algebraic roughness of the domain walls. The top figures are the correlation function (4) of the displacements $B(r)$, and the bottom ones the measured value of the roughness exponent ζ . *Left*: Magnetic domain walls in thin magnetic films. An exponent of $\zeta \sim 0.6$ is measured compatible with the value $\zeta = 2/3$ expected for a one dimensional wall in a two dimensional space. [From [1] (Copyright 1998 by the American Physical Society)]. *Right*: Ferroelectric domain wall in a ferroelectric film. An exponent of $\zeta = 0.26$ is measured. This value is compatible with the value expected for a two dimensional wall in a three dimensional space in the presence of long range dipolar interactions. [From [8] (Copyright 2005 by the American Physical Society)]

ena. This consists in looking at the problem at larger and larger length scales, eliminating degrees of freedom, while changing the Hamiltonian to ensure that the large length scale physics remain invariant. Usually one can expand the interaction potential, and only a few terms are relevant, which means that the RG consists in the flow of a small number of coupling constants. In the case of disordered elastic systems, the task is considerably more complex since all powers in the expansion of the correlator of the disorder have the same scaling dimension. During the flow the *whole correlator of the disorder* is modified. One must thus follow the renormalization of a whole function, hence the name functional renormalization group (FRG) [28]. This leads to a remarkable property: Beyond a length scale that coincides with the Larkin length, the disorder correlator, initially a smooth analytic function, becomes non-analytic and develops a cusp. The appear-

ance of this non-analyticity is, in this method, the signal of glassy physics. FRG allows us to obtain the roughening exponent in a systematic expansion in $\epsilon = 4 - d$. This has been worked out for the moment up to second order in $\epsilon = 4 - d$, leading to $\zeta = 0.20829804\epsilon + 0.0068582\epsilon^2$ and $\zeta = \epsilon/3$ for the random bond and random field disorder, respectively [35]. Note that for the random field the mean-field (Flory) exponent is exact due to the long range nature of the disorder. In addition to these analytical approaches, a very useful approach is provided by numerical studies of such systems, using either molecular dynamics simulations [36], Monte Carlo techniques [37,38], or specially designed algorithms [39,40]. Numerical approaches are of course quite challenging due to the glassy nature of the system with many metastable minima close to the ground state. However they have proven quite useful in obtaining not only the asymptotic regime but also

the full crossover between the Larkin and random manifold regimes, as well as in incorporating the effects of finite temperature.

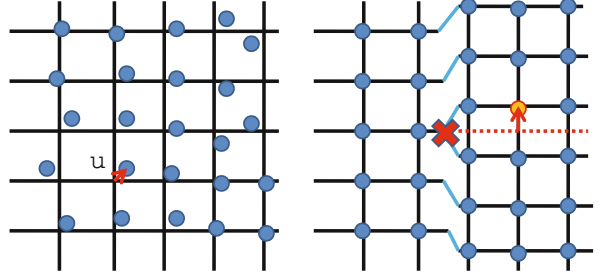
These predictions can be verified experimentally. I show in Fig. 4 the roughness exponent as measured in a magnetic and a ferroelectric film. The algebraic growth of the correlation function $B(r)$ is clearly seen. These two experimental situations correspond to two different dimensionalities for the domain walls, due to the different thicknesses of the material and different characteristics of the domain wall.

Now we can say that we have a rather good understanding of the static properties of the interfaces, at least for the simple case of local elasticity shown here. Of course even for the statics this is not the end of the story, since several microscopic systems such as the contact line of a fluid or ferroelectric systems, have long range interactions (dipolar interactions) making even the static properties quite challenging to determine. Other open questions will be discussed in Sect. “Future Directions”.

Periodic Systems and Bragg Glass

Similar concepts apply directly to the case of periodic systems. In all these systems the constituent elements (lines for vortices, points for colloids and magnetic bubbles, sheets for phase maxima in the charge density wave systems) form a solid that is embedded into the microscopic system but can have widely different characteristics and, in particular, widely different lattice spacing. For example, in the case of vortices, lattice spacing is controlled by the magnetic field and can easily be varied. An important characteristic is that, in a similar fashion to the interface, this crystal can be embedded in the “external” disorder that corresponds to the imperfections of the real microscopic lattice in which this artificial crystal lives. It is important to note that the variation of the disorder potential can thus occur at length scales much smaller than the lattice spacing.

Each point of the system can be described by an equilibrium position R_i^0 forming a perfect lattice (usually triangular for the vortex lattice), and a displacement u_i relative to this equilibrium position, as shown on Fig. 5. As for the interfaces, the interactions between the objects forming the crystal favor a perfectly ordered crystal. The energy of the system can be expanded for small deviations and lead to a quadratic expansion in u characteristic of an elastic energy. It is important to note that for such an expansion to be valid, it is only necessary for the relative displacements $u_i - u_j$ between two neighbors to be small. The displacements themselves can be arbitrary; for example translat-



Disordered Elastic Media, Figure 5

For a periodic system the ability to define a displacement $u(r)$ for the objects (blue dots) compared to the perfect lattice (here a square lattice corresponding to the intersections of the black lines), necessitates the absence of topological defects. *Left:* if there are no topological defects one can associate, for each site R_i^0 of the perfect lattice, a displacement u_i , as indicated by the red arrow. *Right:* here there is a topological defect, denoted by the red cross, corresponding to the addition of one line of particles. In this case the displacement u_i is not univalued. From the point of view of the particles on the left of the topological defect, the orange particle has a displacement u of half a lattice spacing, while one could surmise $u = 0$ looking only at particles on the right of the topological defect

ing the whole crystal by a uniform displacement does not change the energy.

$$H = \frac{1}{2} \sum_{ij} C_{ij} (u_i - u_j)^2, \quad (12)$$

where the C_{ij} are the elastic coefficients of the system. Since the interactions can be long range, the elastic coefficients are not necessarily limited to nearest neighbor only. Note that for such an expansion to be meaningful, it is necessary for the displacements to be uniquely defined. This assumes that there are no topological defects (dislocations, disclinations) in the crystal. Indeed in the presence of such defects, as shown in Fig. 5 the displacements have two different values when circling around the defect. In order to use the elastic approximation it is thus important to ascertain that topological defects are not generated. I will come back to this crucial point below.

As with the interfaces, the minimum energy configuration is the perfect crystal with all $u_i = 0$. In the absence of disorder this perfect crystal can be affected only by thermal fluctuations. If temperature becomes too large the crystal will melt. A rule of thumb for melting is when the relative displacements between two neighbors become a sizeable fraction of the lattice spacing

$$\langle (u_i - u_{i+1})^2 \rangle = C_L^2 a^2, \quad (13)$$

where $\langle \dots \rangle$ denotes the thermal average and C_L is a phenomenological constant which turns out to be of the or-

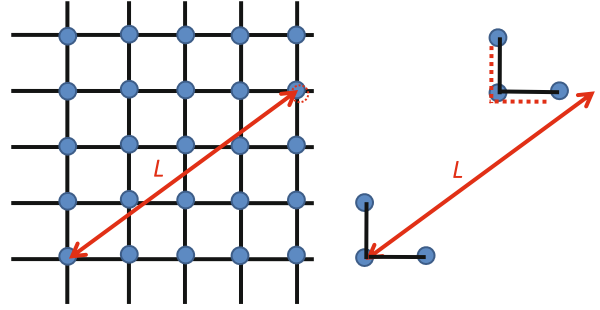
der of $C_L \sim 0.1$ to reproduce reasonable values for the observed melting of solids. This rule of thumb, called the Lindemann criterion for melting, gives, in fact, quite decent results.

In the presence of disorder, one must add to the elastic energy term (12) the energy coming from the random potential created by the disorder. This takes the form

$$\begin{aligned} H_{\text{dis}} &= \int d^d r V(r) \rho(r) \\ &= \int d^d r V(r) \sum_i \delta(r - R_i^0 - u_i) \end{aligned} \quad (14)$$

and this term will clearly tend to disorder the crystal.

The case of a periodic system constitutes a specially important and interesting situation. Indeed, the nature of order and the possible phases are more complex than they are for interfaces. An important question is thus whether these two systems are in the same universality class or not. In a general way, the order in a periodic system is characterized by a positional order, indicating that, knowing the position of a reference particle, one can find a particular particle of the solid at a given position. This positional order can be measured by the structure factor, which is the correlation function of the Fourier transform of the density $S(q) = \langle |\rho_q|^2 \rangle$. In a perfect crystal, the structure factor has *divergent* peaks at the position of the reciprocal vectors K_0 of the perfect lattice. The presence of such divergent peaks indicates a perfect positional order. The fact that one sees peaks also indicates the existence of another type of order, namely the orientational order in a solid. This is illustrated in Fig. 6. The orientational order indicates that if the bonds have a certain orientation in a region of space, this orientation is preserved in the other parts of the solid. Losing the orientational order replaces the peaks in the structure factor by a ring since the orientation of a given peak is not defined any more. In standard solids, both orders are usually lost at the same time and the solid melts to a liquid, usually by a first order phase transition. But we also know that in some cases for pure systems, such as two dimensional solids, the melting may occur as a two step process where the positional order is lost first and only then the orientational order, leading to a so called hexatic phase [41]. A summary of the various cases is shown in Fig. 7. In addition to these standard order parameters, a periodic system is also characterized by a topological order corresponding to the fact that the connectivity of the perfect crystal is preserved by small displacements. Such order is determined by a triangulation of the solid and a determination of the topology. If the topology is identical to that of the perfect lattice, it means that



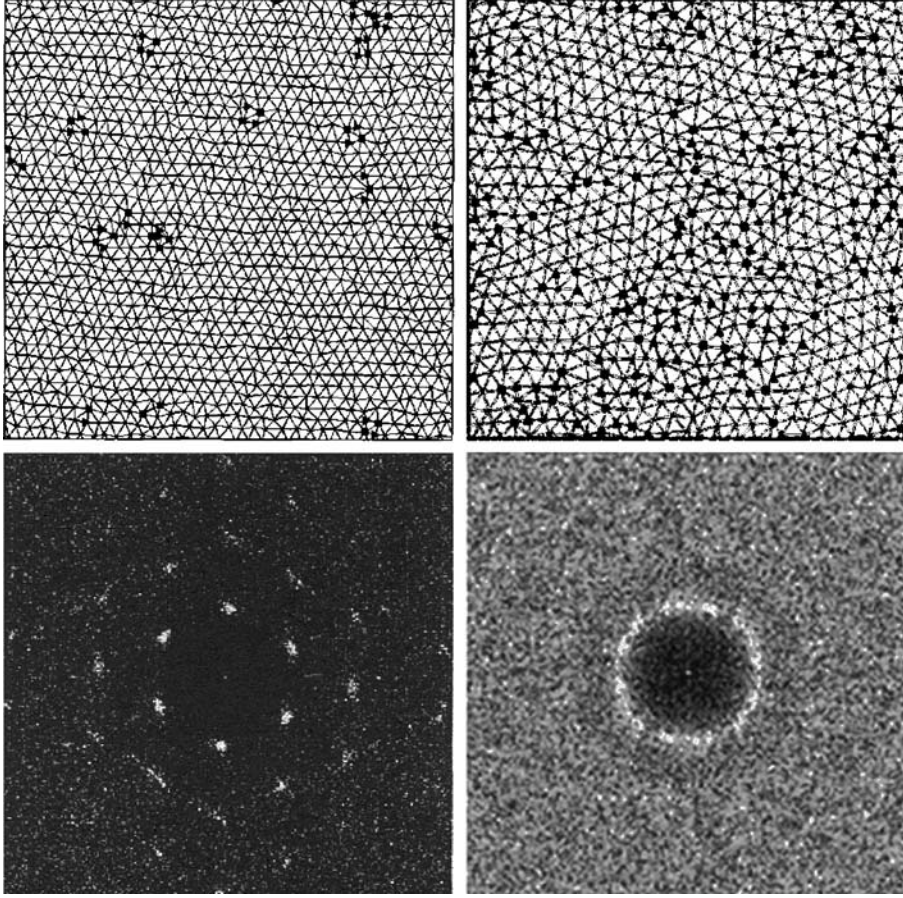
Disordered Elastic Media, Figure 6

A periodic system possesses positional, orientational and topological order. **Left:** positional order. A solid has perfect positional order if, by knowing the position of a reference particle, one can predict the position of a particle at distance L as indicated by the dashed circle and the red arrow. **Right:** orientational order. A solid has orientational order if, by knowing the orientation of the bonds in a region of space, one can predict the orientation at a distance L as indicated by the red dashed lines. Note that the system need not possess positional order for the orientational order to exist. **Topological order:** topological order exists if, after a triangulation, the topology (i. e., the number of neighbors) of each point of the solid is fixed and the displacements can be defined in a univalued manner. The picture on the left possesses perfect topological order

the displacements can be defined in a univalued way across the solid. In the liquid, topological defects such as dislocations and disclinations destroy this perfect topological order and the very concept of displacements around an equilibrium position becomes ill defined, since in that case the displacement field is no longer univalued.

As for interfaces, disorder changes the properties of the pure elastic system. In order to take into account the effect of disorder on periodic systems, it is thus important to address two different aspects of the problem: a) the effect of the disorder on an approximation of the real system given by elastic theory; b) whether the disorder is able to generate topological defects, in which case the very idea of an elastic approximation breaks down and another starting point should be found.

As was shown in the groundbreaking paper by Larkin [25], there always exists for $d \leq 4$, a characteristics length scale L_a for which the displacements become of the order of the lattice spacing a of the perfect crystal. The fact that displacements can become as large as the lattice spacing indicates that *perfect* positional order is lost. The question of *how* this destruction takes place and the nature of the resulting phase is a long standing problem. Given the complexity of the question, no solution existed until recently, but it is interesting to see that the community converged, by inference, on closely related models, to a consensus that was accepted for a long time but eventually



Disordered Elastic Media, Figure 7

Decoration images of vortex lattices, illustrating the difference between solid and liquid phases. The *top figures* are the images in real space, while the *bottom ones* are the structure factor $S(q) = \langle |\rho(q)|^2 \rangle$. *Left*: the system is in a solid-like phase (in fact a Bragg glass phase (see text)). The system possesses good positional order and orientational order. This can be seen both from the pictures in real space and from the structure factor that shows Bragg peaks at the position of the reciprocal vectors K_0 of the perfect underlying lattice. As shown by the *triangulation*, most sites have six neighbors. Topological defects where sites have five or seven neighbors (as indicated by the *triangles* and *square black marks* respectively) do exist, but are paired in 5–7 pairs, making the system free of topological defects at large length scales. *Right*: the system is in a liquid-like phase. Positional order and orientational order are lost. The Bragg peaks are gone and the structure factor has a ring-like structure (indicating the loss of orientational order). Topological defects are proliferating and are unpaired, contributing to the exponential decay of order in the system. [Images from M. Marchevsky, J. Aarts, P.H. Kes (unpublished)]

proved wrong. The route followed was to learn as much as possible from the interfaces. At short distance, one can make an expansion in powers of the displacements, and the system is described by the Larkin model. This ceases to be valid at the Larkin length L_c , for which the displacements become of the order of the size r_f of the particle in the solid. Note that L_c and L_a are in general two different length scales since the size of the particle r_f and the lattice spacing a are usually different. Naturally, one has $L_c < L_a$. For systems such as the vortex lattice or Wigner crystals, the difference can be huge, while for charge-density waves, one expects to have $r_f \sim a$ and thus $L_c \sim L_a$. Below the

Larkin length the system is described by the Larkin model and thus by an algebraic growth of the displacements with a roughness exponent of $\zeta = (4 - d)/2$. Above the Larkin length L_c , but for displacements smaller than a (i. e., for lengths smaller than L_a), one can consider that the various objects of the periodic system don't see each other except by their elastic forces. In particular, they do not sample the same random potential given the smallness of the displacements. One thus has a regime very similar to the random manifold regime of the interface. The displacement continues to grow algebraically, $u \sim L^\zeta$, albeit with a different exponent. Above L_c , the connection between the growth

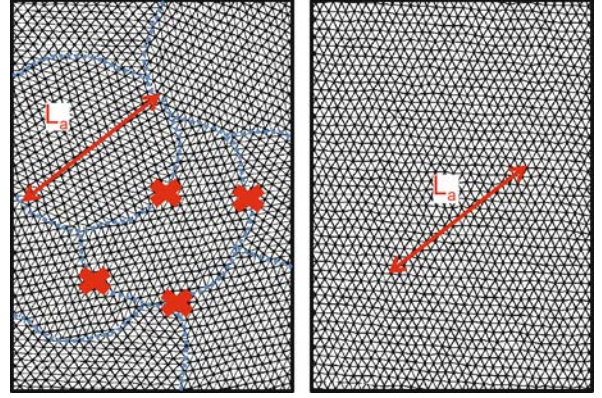
of the displacements and the structure factor (the density-density correlations) is non-trivial since the model is non-Gaussian. Indeed the structure factor is given by [42]

$$S(K_0 + q) = \int d^d q e^{iqr} C(r), \quad (15)$$

where

$$C(r) = \langle e^{iK_0 u(r)} e^{-iK_0 u(0)} \rangle. \quad (16)$$

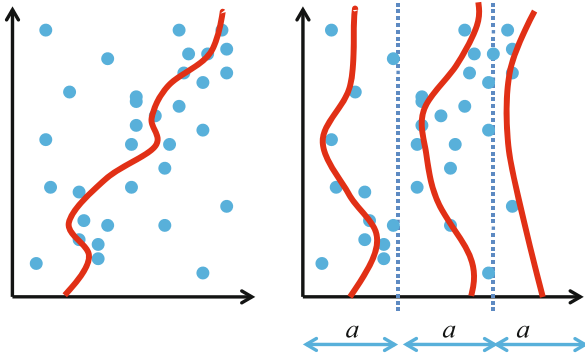
For Gaussian models such as the Larkin model, one had $C(r) = \exp[-K_0^2/2B(r)] \sim \exp[-K_0^2/2r^{2\xi}]$ indicating a stretched exponential decay of positional order. Such an exponential decay of positional order would lead to non-divergent peaks in the structure factor. The constant finding of algebraic roughening, and the existence of the length scale L_a seemed to suggest that even beyond L_a the roughening is also algebraic, with perhaps another exponent. One would thus naively expect $C(r) \simeq \exp[-K_0^2/2a^2(r/L_a)^{2\xi}]$ giving Lorentzian like non-divergent Bragg peaks with a width controlled by $1/L_a$. In addition to this exponential loss of positional order occurring even in the elastic theory, one could question the very starting point of the analysis, namely, the elastic limit and the single-valuedness of the displacements. One could indeed expect topological defects (dislocations, disclinations etc.) to be generated at the scale L_a where the displacements are of the order of the lattice spacing a . Indeed there are arguments [43] (incorrect, as we will see) “showing” that *an arbitrarily small disorder* would always generate topological defects at length scale L_a , leading definitely to an exponential loss of positional order beyond this length. All these elements thus seemed to click together to suggest the picture of a crystal broken into little crystallites of size L_a as shown in Fig. 8. A consensus was thus reached in the community that disordered periodic systems would just lose translational order, and some theories for the vortex lattices were built on this incorrect premise. However this picture crumbled on two fronts. On the experimental side, it was in direct contradiction with experiments showing, for example, extremely large regions free of defects [44,45] or a first order melting [46], which was hardly compatible with a very disordered solid in which all positional order would have been lost from the start. On the theory side, our understanding of glassy systems reached a point where better solutions of this problem can be found. The displacements were found to grow in fact only logarithmically [42,47,48,49] with distance $B(r) = A'_d \log(r)$ (or $u(L) \sim \log(L)^{1/2}$). The prefactor A'_d was computed using either a variational approach [42,48,49] or an FRG one [42,49]. Elastic disordered systems have, therefore, a completely different



Disordered Elastic Media, Figure 8

Left: the (incorrect) image of an elastic medium in presence of disorder. The system would be broken into “crystallites” of size L_a characteristic size for which the displacements become of the order of the lattice spacing a ($u(L_a) \sim a$). At the same length scale to release part of the elastic energy due to disorder, the system would prefer to create topological defects (schematically indicated by the red crosses). Beyond the size L_a indicated by the blue dashed line, positional order would be lost exponentially quickly. **Right:** the Bragg glass, describing the properties of a disordered periodic system in the presence of weak disorder. Although positional order is destroyed at large length scale, and the length scale L_a for which displacements are of order a exists, the system preserves quasi-long range positional order, and perfect topological order. No “crystallite” is thus associated with the length scale L_a , and no topological defects are generated by the disorder

roughness than interface systems. This a priori surprising behavior can be explained in a qualitative way as shown in Fig. 9. Quite interestingly the structure factor and positional order could still be computed for the full model [42,49], and it was shown that in a quite nontrivial way, the relation $C(r) = \exp[-K_0^2/2B(r)]$ remains essentially applicable, leading to a *power-law* decay of positional order $C(r) \propto (1/r)^\eta$, where the exponent η , is for all practical purposes, a number determined only by the dimension [50]. For example $\eta = 1 - 1.2$ for a three dimensional vortex lattice. The algebraic decay of positional order as well as the value of the exponent indicated that the system still retained *divergent* Bragg peaks in its structure factor and thus, although indeed losing positional order, the loss was very slow and the system was nearly as ordered as a perfect solid. Furthermore, it has been shown [42] that the argument claiming that disorder would always generate dislocations was incorrect and that, on the contrary, due to the slow algebraic decay of the positional order, a three dimensional system is *stable* to the generation of dislocations, at least when the disorder is below a certain threshold. This has led to a physics for a periodic dis-



Disordered Elastic Media, Figure 9

Schematic explanation of the difference of roughness between interfaces and periodic systems. *Left:* In an interface, a high degree of roughness is produced by the fact that there are always regions where it is energetically favorable for the line to go, and from there further to another region, endlessly increasing the displacements. *Right:* For a periodic system (here a periodic system of lines of period a), since what counts is the *total* energy of the system, there is no interest for one line to make displacements much larger than the interline distance, since it would just steal disorder from the neighbor. Thus, even with the same disorder, displacements would “saturate” (in fact still grow but very slowly with distance) when they reach the interparticle distance

ordered system radically different from the generally accepted consensus. In particular, consider the two following facts: (a) algebraic (quasi-long range) decay of the positional order, and divergent Bragg peaks; (b) absence of topological defect has led to the prediction [42] that the disordered periodic system are in fact in a new state of matter, the Bragg glass. Such a system is a disordered system with glassy properties: an energy landscape with many metastable states and the dynamics of a glass, but which “looks” nearly as ordered as a perfect solid. After the Bragg glass was first predicted, its existence has been supported by further analytical [51,52,53] and numerical [54,55] calculations.

The existence of the Bragg glass phase has important consequences and makes it possible to reconcile several apparently contradictory results on the phase diagram of the vortices. It allows us to explain that very large regions free of dislocations can be observed [44,45] while the system is obviously pinned by the disorder. It also explains [56] the narrow peaks observed in neutron scattering experiments [57,58], with a width given by the experimental resolution, which were indicating an excellent degree of positional order. The power-law nature of the peaks has been directly tested in neutron scattering experiments, proving directly the existence of the Bragg glass phase [59]. In addition to its intrinsic properties, the presence of the Bragg glass phase puts strong constraints on

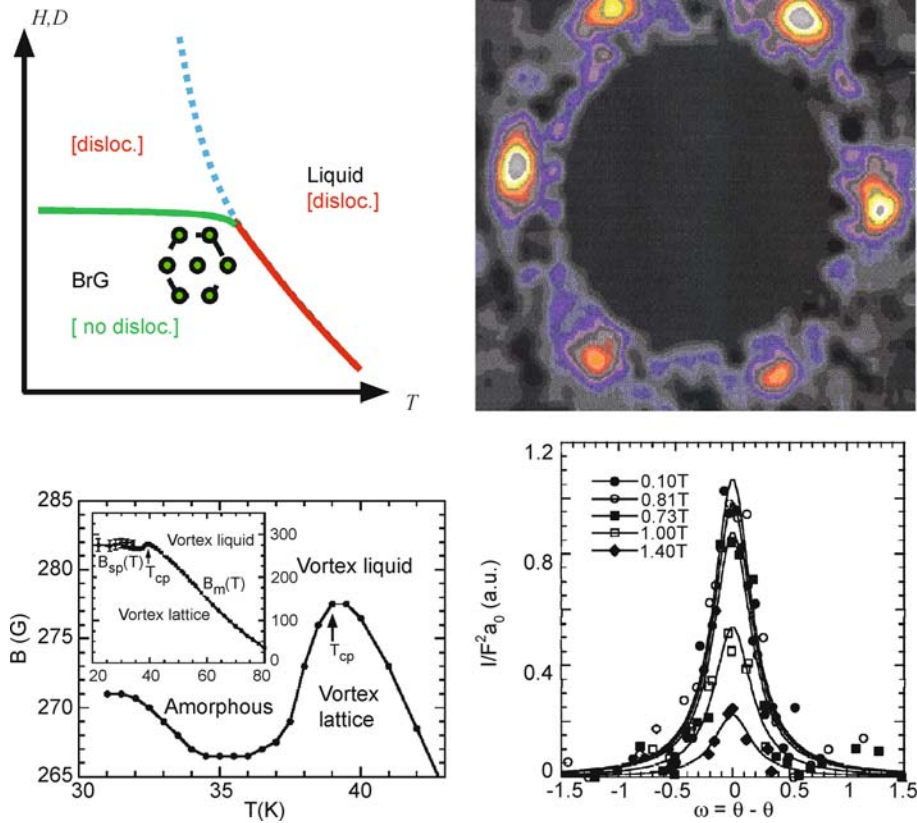
the phase diagram. Indeed, since it is a phase without free topological defects, this phase has to “melt” either when the temperature becomes too high or the disorder too strong, since topological defects have to appear. In vortex systems the latter can be done by changing the magnetic field. The Bragg glass thus provides a very natural explanation [42,60,61] for the existence of a “melting” phase transition as a function of the magnetic field [62,63], such a transition being associated with the destruction of the Bragg glass phase [64,65]. An example is shown in Fig. 10.

This section can cover only a fraction of the physics of periodic systems, and several other questions have been explored. I refer the reader to the above mentioned literature on the subject for more details.

Pinning and Dynamics

Let us now turn to dynamic properties. One of the main interests of such systems is the fact that their dynamics can easily be probed. Indeed most of these systems can be set in motion by an external force acting directly on the interface or on the crystal, and the velocity v versus force F characteristics are directly measurable. As mentioned before, this is of special importance since these characteristics are linked to paramount properties of the systems (voltage-current for vortices, current-voltage for CDW and Wigner crystals, velocity-applied magnetic field for magnetic domain walls). In addition to this practical importance, the dynamics will reflect, even in a more dramatic way than the statics, the competition between disorder and elasticity. In particular, one can expect the dynamics to be dramatically sensitive to the glassy properties and the energy landscape.

The main issues relating to the application of an external force are shown in Fig. 11. In the presence of disorder it is natural to expect that, at zero temperature, the system remains pinned and polarizes only under the action of a small applied force, i. e., it moves until it locks on a local minimum of the tilted energy landscape. At a larger drive, the system follows the force F and acquires a non-zero asymptotic velocity v . So a first set of questions is prompted by the zero temperature properties: what is F_c and how can it be computed? In addition, the $v - F$ curve at $T = 0$ is reminiscent of the curve of an order parameter in a second order phase transition. Here, the system being out of equilibrium, no direct analogy is possible, but this suggests that one could expect $v \sim (F - F_c)^\beta$ with a dynamical critical exponent β . Whether such an analogy to critical phenomena holds and what the physical consequences and calculation of such exponents might be are, of course, important questions.



Disordered Elastic Media, Figure 10

Left top: Schematic theoretical phase diagram for vortices as a function of the temperature T and the magnetic field H , or the disorder D . The Bragg glass (BrG) that has perfect topological order can melt either due to thermal fluctuations (red line) or because the disorder becomes too large (green line). The existence of the Bragg glass thus implies a single melting curve having a crossover between these two regimes. The melting of the Bragg glass thus explain the existence of a transition as a function of the magnetic field. The blue dashed line would be the melting line of the solid in the absence of disorder (After [61]). *Left bottom:* Measured phase diagram for high temperature superconductor BSCCO, showing that both melting transitions with temperature and with magnetic field are indeed the same melting curve. [From [65] (Copyright 2001 by the Nature group)]. *Right:* Neutron diffraction on the superconductor BKBO. The structure factor shows clear Bragg peaks, indicating the good degree of both positional and orientational order despite the disorder present in the sample. As shown in the bottom diagram the width of the structure factor does not change when the magnetic field is changed, since it is controlled by the experimental resolution, while the height decreases. This is in agreement with the consequences of a powerlaw divergent Bragg peak, and is thus a test of the existence of the Bragg glass. [From [59] (Copyright 2001 by the Nature group)]

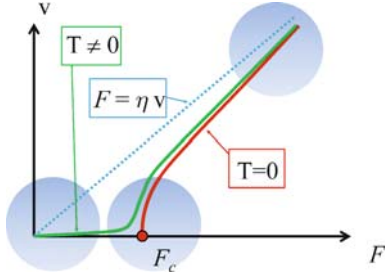
Another important set of questions pertains to the nature of the moving phase itself, and in particular to behavior at large velocity: to what extent does this moving system resemble the static one? This concerns both the positional order properties and the fluctuations in velocity such as the ones measured in noise experiments. Can we expect novel physics there, or is the system simply “surfing” over the disorder?

Finally, how does the system respond to a very small applied force? We are accustomed to the fact that a normal system when perturbed usually responds linearly to the

perturbation. We could thus naively expect $v \propto F$, with a coefficient that would define the “mobility” of the interface. Is this true, or, due to the glassy nature of the system, do we have nonlinear response and more complicated physics?

General Description of the Dynamics

Computing the dynamics is not an easy task. Let me illustrate the method using the case of the interfaces. The displacement field in each point is now a function of the



Disordered Elastic Media, Figure 11

The velocity v induced by an external force F of a disordered elastic system. In the absence of pinning and with a damping coefficient η the steady state velocity $v = F/\eta$ is reached. At zero temperature $T = 0$ the system stays pinned until a critical force F_c is reached. At finite temperature a motion can occur even for forces below the threshold $F < F_c$, since the barriers to motion can always be passed by thermal activation. One can distinguish three very different regimes in this curve: large velocity, depinning and small force response (creep)

time $u(r, t)$ and has to obey the equation of motion. The starting point is the equation of motion

$$m \frac{d^2 u(r, t)}{dt^2} + \eta \frac{du(r, t)}{dt} = \sum F, \quad (17)$$

where η is a friction coefficient that phenomenologically describes the dissipation processes that take place inside the object (interface, etc.) when there is motion. Usually one is interested in the steady state motion of the system, in which case in the long time limit, the second order derivative becomes smaller than the first order one and a good approximation is to take $m = 0$ in the above equation.

The forces are of two types. There are the forces deriving from a Hamiltonian

$$F[u(r, t)] = - \frac{\partial H[u]}{\partial u(r, t)}. \quad (18)$$

The two main contributions (elastic and disorder) lead in the equation of motion to the elastic forces trying to keep the interface flat, and to the pinning forces. In addition to these forces that would be present in equilibrium, one must add two other forces: The first one is the external force. I consider here only the simple case of a constant external force F . In the presence of this force and in the absence of pinning it is natural to expect the system to reach a steady state velocity $v = F/\eta$. Note that such a state, although time-independent, cannot be described by an equilibrium theory. In particular, the fluctuation dissipation theorem, relating in equilibrium the fluctuations in the absence of a perturbation and the response of the system to an external perturbation, is not in general obeyed any

more. The second force is needed if we want to describe the system at a finite temperature. In that case, one must add [66] a Langevin force $\zeta(z, t)$ which is a noise with correlations

$$\langle \zeta(r, t) \zeta(r', t') \rangle = \eta T \delta(r - r') \delta(t - t'). \quad (19)$$

The equation of motion thus becomes, in its simplest incarnation,

$$\eta \frac{du(r, t)}{dt} = - \frac{\partial H[u]}{\partial u(r, t)} + F + \zeta(r, t). \quad (20)$$

As is well known, the presence of Langevin noise ensures that, in the absence of an external force F , the time evolution of the system reproduces the thermodynamic ensemble average. In other words, the equal time correlations $\langle u(z, t) u(z', t) \rangle$ obtained by averaging over the thermal noises, are identical to the equilibrium correlation function $\langle u(z) u(z') \rangle_H$ that one would have obtained for a system with the Hamiltonian H at the temperature T .

In the absence of disorder the equation becomes quite simple and is known as the Edwards–Wilkinson equation

$$\eta \frac{du(r, t)}{dt} = - \nabla_r^2 u(r, t) + F + \zeta(r, t). \quad (21)$$

The system thus slides at a constant velocity $v = F/\eta$ and one can see that in the moving frame the interface is at equilibrium, since the change of variable $u(r, t) = \frac{F}{\eta} t + \delta u(r, t)$ gives for the relative displacements $\delta u(r, t)$ exactly the same equation as in equilibrium in the absence of any external force. Even in this simple case, there are several effects of the motion that need to be taken into account, in particular the presence of a cutoff in the system generates terms that would not normally have been incorporated in the original equation of motion and that can modify the behavior of the system. The most well known is the so-called Kardar–Parisi–Zhang (KPZ) term [67].

In the presence of disorder, the equation of motion becomes extremely complicated to solve since the pinning force is a random variable depending on the particular realization of the disorder, and a double averaging must be done, both on the thermal noise and on the disorder. No perfect method exists to treat such an equation. Since we are usually better equipped to deal with integrals than with differential equations, especially with stochastic terms, a convenient rewriting of this equation exists, which formally gives back the equivalent of a path integral and an action. This is the so-called Martin–Siggia–Rose (MSR) formalism [68,69]. I refer the reader to the literature on this relatively specialized method [66]. The advantage is that it allows averaging over the disorder from the start. This in particular paves the way for an FRG treatment of the problem.

Depinning

The first set of questions arises around depinning. Indeed, in the presence of disorder the naive expectation is that the interface is unable to move, at zero temperature, below a certain threshold of force F_c called the pinning force. Computing this pinning force is not easy. In a remarkable feat of physical intuition, Larkin has shown that the pinning force can be directly obtained from the static behavior of the system [70]. Indeed, the idea is that the pinning force is related to the appearance of many metastable states and the presence of random potential. Because it is quadratic in the displacements, the Larkin model does not exhibit a pinning force. The idea would thus be to relate the pinning force to the length scale at which the Larkin model ceases to pertain. As we discussed before, this is the length L_c for which the displacements are of the order of the correlation length of the random potential or the width of the elastic object. At that scale, the elastic plus disorder energy scales as $cL_c^{d-2}r_f^2$ while the additional energy due to the force scales as

$$H_F = \int d^d r F u(r) \sim F L_c^d r_f. \quad (22)$$

Balancing the two terms leads to the famous Larkin collective pinning force

$$F_c = \frac{c r_f}{L_c^2}. \quad (23)$$

This is a remarkable relation since it relates a dynamic property to purely static quantities. This intuitive result can be substantiated by considerably more complicated calculations. In the next section we will see another rough estimate based on a large velocity expansion. Finally, starting from the equation of motion (17), it is possible to obtain F_c from an FRG calculation [71,72], confirming, from this microscopic calculation, Larkin's result. Numerical methods have allowed an extremely precise calculation of F_c [39].

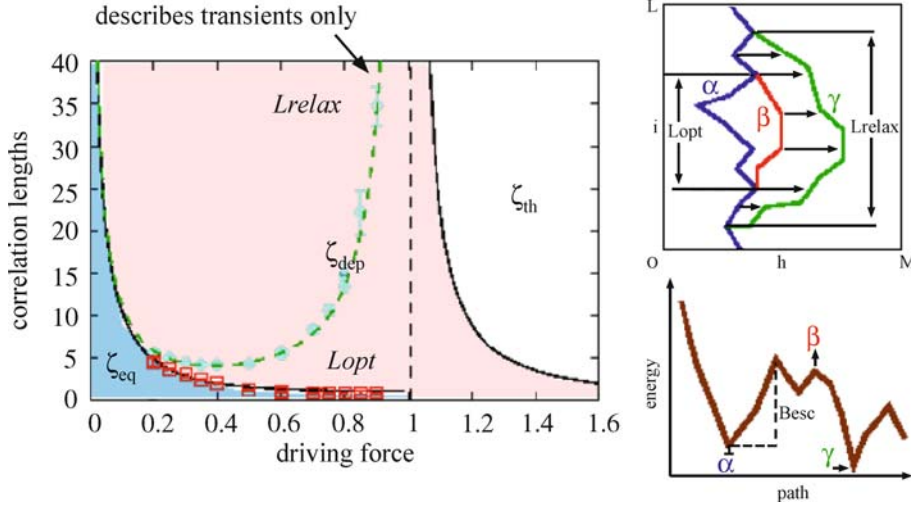
Besides the existence of the pinning force itself, the description of depinning is a considerable challenge. A very fruitful line of approach for this problem was suggested by D.S. Fisher [73]. Indeed, looking at the $v - F$ characteristics is strongly reminiscent of the curve of an order parameter as a function of temperature in a second order phase transition (zero for $T > T_c$ and non-zero for $T < T_c$). This strongly suggests using an analogy with a standard critical phenomenon to analyze the depinning. In particular, one can infer from this analogy that a divergent length scale exists at the transition, and that one can define scaling behavior and critical exponents as a function of this length

scale. One can define a critical exponent for the velocity $v \sim (F - F_c)^\beta$, for the correlation length $\xi \sim (F - F_c)^{-\nu}$ and a dynamical exponent relating space and time divergences $\tau \sim \xi^z$. These exponents are related by scaling relations, analogous to those of standard critical phenomena and that can be computed by looking at the scaling of the equation of motion. The scaling relations are

$$\nu = \frac{\beta}{2 - \zeta} = \frac{1}{z - \zeta}. \quad (24)$$

Such scaling behavior is directly confirmed from solutions of the equation of motion, either from FRG or from numerical simulations. The length scale ξ can be identified as the length scale of avalanches. Computing and measuring these exponents is a considerable challenge and sophisticated FRG [35,71,72,74] or numerical [39,75] techniques have been developed for this goal.

In addition to these quantities characterizing the motion of the line, other important physical observables are modified by the application of external force. This is in particular the case of the roughness of the line. Right at depinning $F = F_c$, the line is much more rough than when in equilibrium, since it is on the verge of depinning. There is thus a new roughness exponent ζ_{dep} which can be computed and is $\zeta_{\text{dep}} \sim 1.2$ for a one dimensional interface. This result has two very important consequences. The first one comes from the value of the roughness exponent itself. Since, at least for a line, this value is larger than one, this immediately suggests that close to depinning the elastic model will run into trouble. Indeed when u scales more than linearly with distance, the very basis of the elastic approximation $\nabla u \ll 1$ is violated at large length scales. The line will thus have to generate defects (overhangs) to heal this fact. The nature of the resulting physics when this is taken into account is a challenging and yet open question. The second observation concerns the steady state aspect of the line. At large length scales, because of finite velocity, the system will average over the disorder. We will come back in more detail to this point in the next section, but for the moment, stick with this simple vision. In this case, beyond the length ξ one can expect the disorder to be irrelevant and thus to recover the pure thermal roughness exponent. The system will thus have the depinning roughness exponent ζ_{dep} for length scales below ξ and the thermal one ζ_{th} for length scales above ξ . This is summarized in Fig. 12. One important question now is what happens at a small but finite temperature. The first effect is, of course, to smooth the $v - F$ characteristics. This leads to the important question of whether one can define a scaling with the temperature of this thermal rounding of the depinning (see e. g., [77] and references therein). Even more interest-



Disordered Elastic Media, Figure 12

Close to depinning the motion proceeds by avalanches between two configurations α and γ . Above F_c , there exists a divergent length scale (L_{opt} on the figure) below which the line is characterized by the roughness exponent ζ_{dep} and above which the line shows the thermal roughness ζ_{th} . A normal critical phenomenon would have had a similar divergent length scale for $F < F_c$. This is not the case for the depinning. A transient divergent length scale L_{relax} does exist, but does not show up in the steady state properties of the line. Contrary to naive expectations from a “standard” critical phenomenon, one observes the equilibrium roughness exponent ζ_{eq} at short distances. This shows that the analogy between depinning and “standard” critical phenomena, although very fruitful, must be taken with a grain of salt. On the right, the schematic shape of the line and energy profiles are shown. After [76]

ing is the question of the roughness of the line. The analogy with a critical phenomenon would simply suggest that a similar divergent length scale should exist for $F < F_c$, leading to the standard pattern of “critical regime”. However, as indicated in Fig. 12, such a divergent length scale does not exist [76]. This leads to a very puzzling behavior and shows that although the analogy with standard critical phenomena can be extremely fruitful, it must be taken with a grain of salt. The depinning, which is by essence a dynamical transition, possesses its own physics.

High Velocity Phase

Another important set of questions and physics occurs when the interface is moving at large velocity, i.e., for $F \gg F_c$. This is apparently the simplest regime since one could expect that at large velocity one has a control parameter on the theory and that an expansion in $1/v$ is possible. This is indeed the case for the $v - F$ characteristics. A large velocity expansion of the disorder term can be made by going into the moving frame of the elastic media. One can indeed write $u(r, t) = vt + \delta u(r, t)$, where $\delta u(r, t)$ describes the displacements in the moving frame. Because the system surfs over the disorder at very large velocity one can expect the effects of the disorder, and hence $\delta u(r, t)$, to be small. This is confirmed by a well controlled large velocity

expansion [78,79]. In particular, the correction due to the disorder to the velocity can be computed and behaves as

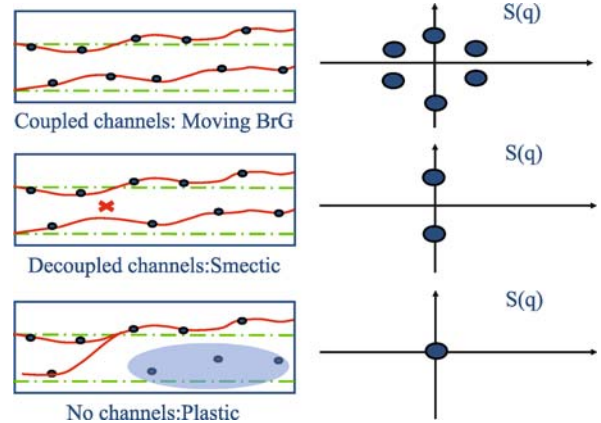
$$\frac{F - \eta v}{\eta v} \propto D \left(\frac{1}{\eta v} \right)^{\frac{4-d}{2}}. \quad (25)$$

This shows clearly that the effects of disorder are kept small at large velocity or large force and become increasingly important as the force/velocity gets smaller, in agreement with Fig. 11. The relative correction to the velocity grows in dimensions smaller than $d = 4$, confirming that disorder is relevant below this dimension. Although one cannot extrapolate the perturbative expressions, a crude way to estimate the critical force F_c is when the deviations (25) become of order one. This method gives back the estimate (23) for F_c , obtained from totally different considerations.

The large velocity expansion also allows us to address the physics of the moving phase, namely the shape and properties of the elastic system in the moving frame. A calculation of these effects was performed [80] by computing the displacements δu from the large velocity expansion. This leads to the striking result that at large velocity the effect of disorder disappears and can be absorbed in a simple modification of the temperature of the system, leading to an effective temperature T_{eff} . This has important consequences on the properties of the system in the

moving frame. For an interface, this is consistent with the idea, exposed in the previous section, that at large distance one recovers the thermal roughening exponent. For periodic systems, since there is the possibility of melting, this has the more drastic consequences that driving the system could induce a velocity controlled melting. Indeed, for large velocities, the effective temperature is small, while for smaller velocities it would be large. The system would thus be a disordered system while static, then close to depinning, where the effective temperature would be large, it would be in a melted (liquid) state, and would then recrystallize when moving at larger velocities.

Although the concept of effective temperature is extremely fruitful, in particular for this dynamic recrystallization, the properties of the moving periodic system are richer [81] than those of a simple solid subjected to temperature. Indeed, periodic systems have a structure in the direction *transverse* to the direction of motion. This structure, and the corresponding disorder structure, cannot be averaged by the motion, however large the velocity remains. This leads to the fact that disorder remains even when the system is in motion. In other words, a moving periodic system remains a glass. The way the motion takes place is quite peculiar. The system finds optimal paths [81] which are a compromise between the elastic energy, disorder and the motion. These paths are rough, similar to the way in which a static system in a disordered environment is rough. This is shown in Fig. 13. For periodic systems, pinning effects still manifest themselves in the moving system. The glassy nature and the channel motion has been confirmed both numerically [82,83,84] and experimentally [85]. The channel motion leads to an interesting consequence. Since the effects of disorder are weakening as velocity increases, the channels undergo a transition between a regime for which the particles in different channels are coupled or decoupled [86,87]. In the first case, the system is essentially moving as a “solid” (in fact a moving Bragg glass) since the topological order is perfect even if the system is still distorted by disorder. In the second case, the channels are decoupled, which means that a smectic like structure of the particles inside the channels is expected. These transitions as described in Fig. 13 have also been observed both numerically and experimentally as shown in Fig. 14. An additional consequence of the existence of such channels is the existence of a *transverse* pinning force [81,87]. Indeed, even if the particles themselves are moving along the channels, the channels themselves are pinned if one applies an additional force transverse to the direction of motion. This surprising phenomenon has been numerically confirmed [82,83,84], but observing it in classical periodic systems is still an experimental chal-



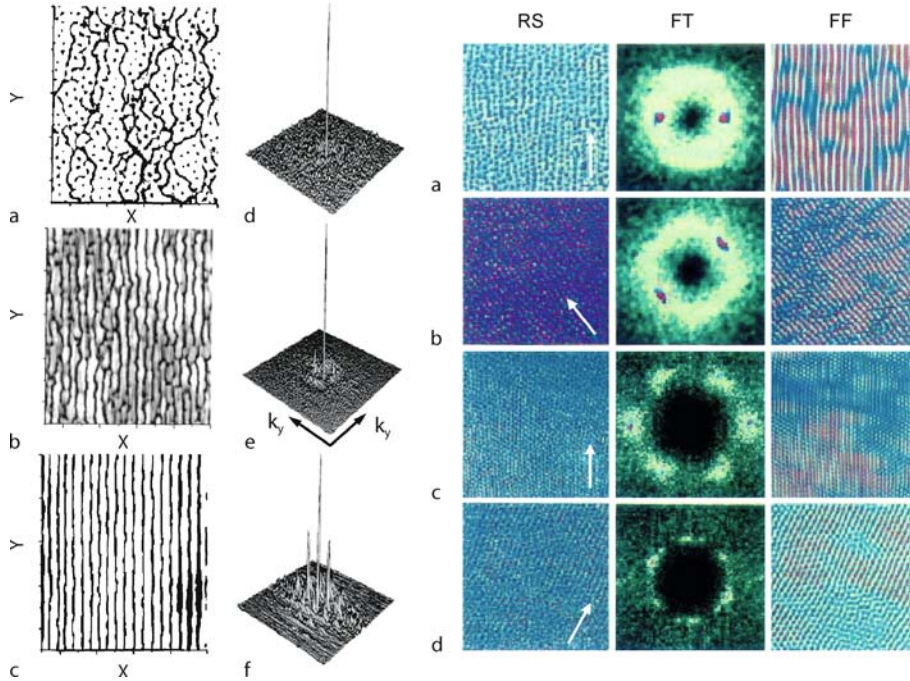
Disordered Elastic Media, Figure 13

Motion of a periodic system. The system develops rough channels which compromise between elastic energy and the transverse component of the disorder that are poorly averaged by the motion. All the particles follow on these channels like cars on highways. Depending on the velocity, the channels are increasingly coupled: *Bottom image*: close to depinning, motion can proceed through a plastic regime where unpinned and pinned regions (denoted by the blue circle) coexist; *Middle image*: topological defects can exist between the channels, so although the channels themselves are well formed, the particles in them are essentially decoupled leading to a smectic like behavior; *Top image*: the channels are coupled and the system is a moving Bragg glass with effects of both disorder and elasticity and no topological defects. On the right, the corresponding structure factors are indicated

lenge. Experiments showing the absence of Hall effect in Wigner crystal systems [88] could constitute an experimental proof of such a transverse critical force, but clearly further experimental data would be needed to unambiguously decide on that point.

Small Applied Force and Creep Motion

Finally, let us look at the response of the system to a small external force. At zero temperature, one is below the pinning force, and thus except for a transient motion the system remains pinned. Motion can thus only take place due to thermal activation over the energy barriers. The response to a small external force is thus a method of choice to probe for the nature of the energy landscape of such systems. For usual systems one expects the response to be linear. Indeed, earlier theories of such motion have found a linear response. The idea is to consider that a blob of pinned material has to move in an energy landscape with characteristic barriers Δ as shown in Fig. 15. The external force F tilts the energy landscape, thus making forward motion possible. The barriers are overcome by thermal activation [89] (hence the name: Thermally Assisted



Disordered Elastic Media, Figure 14

Left: numerical simulations confirming the presence of channels for moving periodic structures and the sequence of transitions depicted in Fig. 13. The left part of the image shows the real space trajectories, while the right part is the structure factor. The force is increasing from the top to the bottom of the image. [From [83] (Copyright 1999 by the American Physical Society)]. Right: A decoration image of moving vortices also showing the presence of channels. The direction of the applied force is indicated by an arrow. The left column is the raw decoration image, the center one is the Fourier transform giving the structure factor, the right column is a filtered version of the image showing the channels more clearly. [From [85] (Copyright 1999 by the Nature group)]

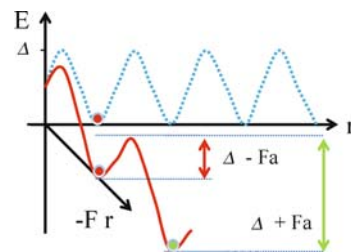
Flux Flow (TAFF)) with an Arrhenius law. If the minima are separated by a distance a the velocity is

$$v \propto e^{-\beta(\Delta - Fa/2)} - e^{-\beta(\Delta + Fa/2)} \simeq e^{-\beta\Delta} F. \quad (26)$$

The response is thus linear, but exponentially small.

However this argument is grossly inadequate for a glassy system. The reason is easy to understand if one remembers that the static system is in a glassy state. In such a state a characteristic barrier Δ does not exist, since barriers are expected to diverge as one gets closer to the ground state of the system. The TAFF formula is thus valid in systems where the glassy aspect is somehow eliminated and the barriers saturate. This could be the case, for example, for a finite size interface. When the glassy nature of the system persists up to arbitrarily large length scales, the theory should be accommodated to take into account divergent barriers. This can be done qualitatively within the framework of the elastic description using scaling arguments [47,90,91,92]. The basic idea rests on two quite strong but reasonable assumptions: (i) the motion is so slow that one can consider, at each stage, that the interface is motionless and use its static description; (ii) the scal-

ing for barriers, which is quite difficult to determine, is the same as the scaling of the minimum of energy (metastable states) that can be extracted, again, from the static calculation. If the displacements scale as $u \sim L^\xi$ then the en-



Disordered Elastic Media, Figure 15

In thermally assisted flux flow [89] a region of pinned material is considered as a particle moving in an energy landscape characterized by characteristic barriers Δ , schematized by the blue dashed periodic potential, of period a . Applying an external force tilts the energy landscape. The motion over barriers can always proceed by thermal activation. Due to tilt, the barrier to forward motion (in red) is smaller than the reverse barrier (in green). This results in an exponentially small but linear response when a small external force is applied to the system

ergy of the metastable states (see (2)) scales as given by (5): $E(L) \sim L^{d-2+2\zeta}$. Since the motion is very slow, the effect of the external force is simply to tilt the energy landscape

$$E(L) - F \int d^d r u(r) \sim L^{d-2+2\zeta} - FL^{d+\zeta}. \quad (27)$$

Thus, in order to make the motion to the next metastable state, one needs to move a piece of the pinned system of size

$$L_{\text{opt}} \sim \left(\frac{1}{F} \right)^{\frac{1}{2-\zeta}}. \quad (28)$$

The size of the optimal nucleus able to move thus grows as the force decreases. Since the barriers to overcome grow with the size of the object, the minimum barrier to overcome (assuming that the scaling of the barriers is also given by (5))

$$U_b(F) \sim \left(\frac{1}{F} \right)^{\frac{d-2+2\zeta}{2-\zeta}} \quad (29)$$

leading to the creep formula for the velocity

$$v \propto \exp \left[-\beta U_c \left(\frac{F_c}{F} \right)^\mu \right], \quad (30)$$

where F_c is the depinning force and U_c a characteristic energy scale and the *creep exponent* μ is given by,

$$\mu = \frac{d-2+2\zeta}{2-\zeta}. \quad (31)$$

Equations (30) and (31) are quite remarkable. They relate a dynamical property to *static* exponents, and show clearly the glassy nature of the system. The corresponding motion has been called creep since it is a sub-linear response. It is a direct consequence of the divergent barriers in the pinned system.

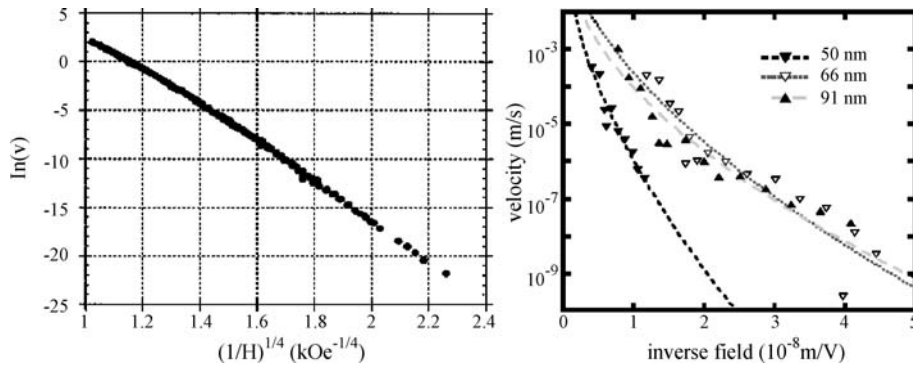
Of course the derivation above is phenomenological, so it is important to ascertain by more microscopic methods whether the results hold. Although in principle one simply has to solve the equation of motion (20), in practice this is quite complicated. A natural framework for computing perturbation theory in off-equilibrium systems is the MSR formalism. Using this formalism and an FRG analysis, one can confirm the creep formula for velocity [72,93]. Numerical simulations also show the absence of linear response and the existence of a creep response [94]. Creep has also been checked experimentally in various systems. Vortices show a creep behavior with an exponent $\mu = 1/2$ compatible with the existence of the Bragg glass ($d = 3$, $\zeta = 0$) [95]. However, the range of measurable velocities makes it difficult to unambigu-

ously check for this law. One spectacular determination of the creep law was performed in a magnetic film [1]. In such a situation the roughness exponent is known exactly ($\zeta = 2/3$) and thus the value of the creep exponent $\mu = 1/4$ is not an adjustable parameter, making it a much more stringent test. As shown in Fig. 16, the velocity was measured over ten orders of magnitude, a spectacular feat, and a remarkable confirmation of the creep law. Measurements of the creep law have also been performed for domain walls in ferroelectrics where a simultaneous measurement of the creep exponent and of the roughness exponent was performed [7,8]. As shown in Fig. 16 the stretched exponential behavior for the velocity is well verified, and the formula (31) consistent with what is expected for a two dimensional domain wall in the presence of dipolar forces, corresponding to the experimental situation.

The FRG derivation allows us to more deeply probe the physical understanding of the system. In particular it unravels a new phenomenon. Although the velocity itself is dominated by events occurring at the thermal length scale (28), interesting physics takes place beyond this length scale. Indeed, when a portion L_{opt} of the line has been able to move forward by thermal activation over the barriers, it serves as a nucleation center to trigger an avalanche [72] over a much larger length scale L_{av} . The behavior between these two length scales is thus very similar to a depinning phenomenon where temperature plays no role. Although the velocity is dominated by the first process, which is the slow one, the shape of the line reflects this much larger avalanche scale in a way which is compatible with experiments [3]. The creep, being controlled by the time to overcome divergent barriers in the system, has several other consequences, in particular on the issue of the out-of-equilibrium physics of such systems [96] and its connection to the aging of glasses [97].

Future Directions

Both because of experimental drive (no pun intended) but also because of theoretical advances and the development of the proper tools, this field has known several breakthroughs in the last decade or so. There is now a good understanding of static properties for both interfaces and for periodic systems, and most of the misconceptions and folklore have been replaced by solid results. Novel phases have emerged such as the Bragg glass phase. Steady state dynamics has also made significant progress, with the understanding of processes such as creep motion. From the point of view of methods, these systems have made it possible to perfect methods to deal with glassy systems such as replica methods, functional renormalization group as well



Disordered Elastic Media, Figure 16

Experimental verification of the creep law for magnetic and ferroelectric domain walls. *Left:* Magnetic domain walls. The film is extremely thin, thus the domain is a line in a two dimensional plane, leading to a creep exponent of $\mu = 1/4$. The creep law is observed on about ten orders of magnitude for the velocity. [From [1] (Copyright 1998 by the American Physical Society)]. *Right:* Ferroelectric films. A creep behavior is observed over several orders of magnitude giving a creep exponent of $\mu \sim 0.58$. This value together with the measured roughness exponent $\zeta = 0.26$ leads to an effective dimension of $d = 2.5$, well in agreement with a two dimensional domain wall in presence of dipolar forces. [From [8] (Copyright 2005 by the American Physical Society)]

as special numerical methods. These results have found and continue to find applications in a large variety of experimental domains. Despite these advances, it is clear that many questions remain pending, making it still a very challenging field which is yet in constant progress. Experiments regularly provide new systems and new challenges and stimulate theoretical analysis. Several lines of research are actually open and should carry the bulk of the research in that domain in the future.

From the point of view of the static, although the situation without defects is under control, we know next to nothing when elasticity, disorder and defects are included. For interfaces this means treating the overhangs and bubbles, and for periodic systems all the topological defects such as the dislocations. Although we know now that the situation without defects is robust below a certain threshold of disorder, it is clear that the ability to deal with the strong disorder situation is needed for many experimental systems. This is the case for the high field phase of vortices and strong disorder in the interfaces, both of which are dominated by defects.

In the dynamics, one of the very challenging questions that one has to face is the one of the out-of-equilibrium dynamics, when the system has not yet reached a steady state. A simple example of such a situation would be an interface relaxing slowly from a flat configuration or quenched at low temperatures from a high temperature configuration. Given that these systems are glasses, the time evolution of such cases is highly non-trivial, and should show a generic phenomenon of glasses known as aging. This is directly a situation relevant to many experiments. From the conceptual point of view this is an extremely challenging question since most of the theoretical tools that we have fail to

tackle such situations, and thus new tools or new concepts need to be invented.

Last but not least, we have dealt mainly with classical systems here. But disordered elastic systems can also be realized in the quantum worlds as briefly mentioned in the introduction. The question on how to extend the concepts of glasses to quantum systems remains largely open. In particular, one can expect the dynamics to be affected. Indeed, classical systems can pass barriers only by thermal activation, while quantum systems are good at tunneling through barriers. The extension of the above concepts to the quantum world is thus a very challenging direction.

The gold mine of disordered elastic media is thus far from being exhausted. It seems that each nugget we find is only the opening of a new vein with even richer stones. The variety of experimental realization is ever growing, as is the depth of the questions that are now within our grasp.

Bibliography

Primary Literature

1. Lemerle S et al (1998) Phys Rev Lett 80:849
2. Krusin-Elbaum L et al (2001) Nature 410:444
3. Repain V et al (2004) Europhys Lett 68:460
4. Metaxas PJ et al (2007) Phys Rev Lett 99:217208
5. Yamanouchi M et al (2006) Phys Rev Lett 96:096601
6. Yamanouchi M et al (2007) Science 317:1726
7. Tybell T, Paruch P, Giamarchi T, Triscone JM (2002) Phys Rev Lett 89:97601
8. Paruch P, Giamarchi T, Triscone JM (2005) Phys Rev Lett 94:197601
9. Wilkinsion D, Willemsen JF (1983) J Phys A 16:3365
10. Moulinet S, Guthmann C, Rolley E (2002) Eur Phys J E 8:437
11. Barabasi A-L, Stanley HE (1995) In: Fractal Concepts in Surface Growth. Cambridge University Press, Cambridge

12. Bouchaud E et al (2002) *J Mech Phys Solids* 50:1703
13. Alava M, Nukalaz PKVV, Zapperi S (2006) *Adv Phys* 55:349
14. Blatter G et al (1994) *Rev Mod Phys* 66:1125
15. Nattermann T, Scheidl S (2000) *Adv Phys* 49:607
16. Giamarchi T, Bhattacharya S (2002) In: Berthier C et al (eds) *High Magnetic Fields: Applications in Condensed Matter Physics and Spectroscopy*. Springer, Berlin, p 314. [arXiv:cond-mat/0111052](#)
17. Grüner G (1988) *Rev Mod Phys* 60:1129
18. Nattermann T, Brazovskii S (2004) *Adv Phys* 53:177
19. Seshadri R, Westervelt RM (1992) *Phys Rev B* 46:5150
20. Murray CA, Sprenger WO, Wenk R (1990) *Phys Rev B* 42:688
21. Coupier G, Guthmann C, Noat Y, Saint Jean M (2005) *Phys Rev E* 71:046105
22. Andrei EY et al (1988) *Phys Rev Lett* 60:2765
23. Giamarchi T (2004) In: *Quantum phenomena in mesoscopic systems*. IOS Press, Bologna. [arXiv:cond-mat/0403531](#)
24. Eskildsen MR et al (2002) *Phys Rev Lett* 89:187003
25. Larkin AI (1970) *Sov Phys JETP* 31:784
26. Mézard M, Parisi G, Virasoro MA (1987) *Spin Glass Theory and beyond*. World Scientific, Singapore
27. Mézard M, Parisi G (1991) *J de Phys I* 4:809
28. Fisher DS (1986) *Phys Rev Lett* 56:1964
29. Nature group (2007) *Nature Material*, vol. 6. Focus issue on Multiferroics
30. Nattermann T (1983) *J Phys C* 16:4125
31. Joanny JF, de Gennes PG (1984) *J Chem Phys* 81:552
32. Gao H, Rice JR (1989) *J Appl Mech* 56:828
33. Huse DA, Henley CL (1985) *Phys Rev Lett* 54:2708
34. Kardar M (1985) *Phys Rev Lett* 55:2923
35. Le Doussal P, Wiese K, Chauve P (2004) *Phys Rev E* 69:026112
36. Giamarchi T, Kolton AB, Rosso A (2006) In: Miguel MC, Rubi JM (eds) *Jamming, Yielding and Irreversible deformation in condensed matter*. Springer, Berlin, p 91. [arXiv:cond-mat/0503437](#)
37. Yoshino H (1998) *Phys Rev Lett* 81:1493
38. Rosso A, Krauth W (2002) *Phys Rev B* 65:12202
39. Rosso A, Krauth W (2002) *Phys Rev E* 65:025101R
40. Petaja V et al (2006) *Phys Rev E* 73:94517
41. Nelson DR (1978) *Phys Rev B* 18:2318
42. Giamarchi T, Le Doussal P (1995) *Phys Rev B* 52:1242
43. Fisher DS, Fisher MPA, Huse DA (1990) *Phys Rev B* 43:130
44. Grier DG et al (1991) *Phys Rev Lett* 66:2270
45. Kim P, Yao Z, Lieber CM (1996) *Phys Rev Lett* 77:5118
46. Schilling A, Fisher RA, Crabtree GW (1996) *Nature* 382:791
47. Nattermann T (1990) *Phys Rev Lett* 64:2454
48. Korshunov SE (1993) *Phys Rev B* 48:3969
49. Giamarchi T, Le Doussal P (1994) *Phys Rev Lett* 72:1530
50. For a scalar displacement or isotropic elasticity the exponent is universal. When the full anisotropy of the elastic constants is taken into account in the FRG equations [Bogner S, Emig T, Nattermann T (2001) *Phys Rev B* 63 174501] the possible variation of the exponent with the magnetic field is still less than a percent.
51. Carpentier D, Le Doussal P, Giamarchi T (1996) *Europhys. Lett.* 35:379
52. Kierfeld J, Nattermann T, Hwa T (1997) *Phys Rev B* 55:626
53. Fisher DS (1997) *Phys Rev Lett* 78:1964
54. Gingras MJP, Huse DA (1996) *Phys Rev B* 53:15193
55. Otterlo AV, Scalettar R, Zimányi G (1998) *Phys Rev Lett* 81:1497
56. Giamarchi T, Le Doussal P (1995) *Phys Rev Lett* 75:3372
57. Yaron U et al (1994) *Phys Rev Lett* 73:2748
58. Ling XS et al (2001) *Phys Rev Lett* 86:712
59. Klein T et al (2001) *Nature* 413:404
60. Ertas D, Nelson DR (1996) *Phys C* 272:79
61. Giamarchi T, Le Doussal P (1997) *Phys Rev B* 55:6577
62. Khaykovich B et al (1996) *Phys Rev Lett* 76:2555
63. Deligiannis K et al (1997) *Phys Rev Lett* 79:2121
64. Paltiel Y, Zeldov E, Myasoedov Y, Rappaport ML (2000) *Phys Rev Lett* 85:3712
65. Avraham N et al (2001) *Nature* 411:451
66. Zinn-Justin J (1989) *Quantum field theory and Critical Phenomena*. Clarendon Press, Oxford
67. Kardar M, Parisi G, Zhang Y (1986) *Phys Rev Lett* 56:889
68. Janssen HK (1976) *Z Phys B* 23:377
69. Martin PC, Siggia ED, Rose HA (1973) *Phys Rev A* 8:423
70. Larkin AI, Ovchinnikov YN (1979) *J Low Temp Phys* 34:409
71. Nattermann T, Stepanow S, Tang LH, Leschhorn H (1992) *J Phys* 2:1483
72. Chauve P, Giamarchi T, Le Doussal P (2000) *Phys Rev B* 62:6241
73. Fisher DS (1985) *Phys Rev B* 31:1396
74. Narayan O, Fisher D (1993) *Phys Rev B* 48:7030
75. Duemmer O, Krauth W (2005) [arXiv:cond-mat/0501467](#)
76. Kolton AB, Rosso A, Giamarchi T, Krauth W (2006) *Phys Rev Lett* 97:057001
77. Bustingorry S, Kolton AB, Giamarchi T (2008) *Europhys Lett* 81:26005
78. Larkin AI, Ovchinnikov YN (1974) *Sov Phys JETP* 38:854
79. Schmidt A, Hauger W (1973) *J Low Temp Phys* 11:667
80. Koshelev AE, Vinokur VM (1994) *Phys Rev Lett* 73:3580
81. Giamarchi T, Le Doussal P (1996) *Phys Rev Lett* 76:3408
82. Moon K et al (1997) *Phys Rev Lett* 77:2378
83. Kolton AB, Grønbech-Jensen DDN (1999) *Phys Rev Lett* 83:3061
84. Fangohr H, Cox SJ, de Groot PAJ (2001) *Phys Rev B* 64:64505
85. Pardo F et al (1998) *Nature* 396:348
86. Balents L, Marchetti C, Radzihovsky L (1998) *Phys Rev B* 57:7705
87. Le Doussal P, Giamarchi T (1998) *Phys Rev B* 57:11356
88. Perruchot F et al (2000) *Physica B* 284:1984
89. Anderson PW, Kim YB (1964) *Rev Mod Phys* 36:39
90. Ioffe LB, Vinokur VM (1987) *J Phys C* 20:6149
91. Nattermann T (1987) *Europhys Lett* 4:1241
92. Feigelman M, Geshkenbein VB, Larkin AI, Vinokur V (1989) *Phys Rev Lett* 63:2303
93. Chauve P, Giamarchi T, Le Doussal P (1998) *Europhys Lett* 44:110
94. Kolton AB, Rosso A, Giamarchi T (2005) *Phys Rev Lett* 94:047002
95. Fuchs DT et al (1998) *Phys Rev Lett* 80:4971
96. Kolton AB, Rosso A, Giamarchi T (2005) *Phys Rev Lett* 95:180604
97. Cugliandolo LF, Kurchan J, Bouchaud JP, Mezard M (1998) In: Young AP (ed) *Spin Glasses and Random fields*. World Scientific, Singapore

Books and Reviews

- Barabasi A-L, Stanley HE (1995) *Fractal Concepts in Surface Growth*. Cambridge University Press, Cambridge
- Nelson DR (2002) *Defects and Geometry in Condensed Matter Physics*. Cambridge University Press, Cambridge
- Young AP (ed) (1998) *Spin Glasses and Random fields*. World Scientific, Singapore

Dispersion Phenomena in Partial Differential Equations

PIERO D'ANCONA

Dipartimento di Matematica, Università di Roma
"La Sapienza", Roma, Italy

Article Outline

Glossary

Definition of the Subject

Introduction

The Mechanism of Dispersion

Strichartz Estimates

The Nonlinear Wave Equation

The Nonlinear Schrödinger Equation

Future Directions

Bibliography

Glossary

Notations Partial derivatives are written as u_t or $\partial_t u$, $\partial^\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}$, the Fourier transform of a function is defined as

$$\mathcal{F}f(\xi) = \widehat{f}(\xi) = \int e^{-ix \cdot \xi} f(x) dx$$

and we frequently use the mute constant notation $A \lesssim B$ to mean $A \leq CB$ for some constant C (but only when the precise dependence of C from the other quantities involved is clear from the context).

Evolution equations Partial differential equations describing physical systems which evolve in time. Thus, the variable representing time is distinguished from the others and is usually denoted by t .

Cauchy problem A system of evolution equations, combined with a set of initial conditions at an initial time $t = t_0$. The problem is *well posed* if a solution exists, is unique and depends continuously on the data in suitable norms adapted to the problem.

Blow up In general, the solutions to a nonlinear evolution equation are not defined for all times but they break down after some time has elapsed; usually the L^∞ norm of the solution or some of its derivatives becomes unbounded. This phenomenon is called *blow up* of the solution.

Sobolev spaces we shall use two instances of Sobolev space: the space $W^{k,1}$ with norm

$$\|u\|_{W^{k,1}} = \sum_{|\alpha| \leq k} \|\partial^\alpha u\|_{L^1}$$

and the space H_q^s with norm

$$\|u\|_{H^s} = \|(1 - \Delta)^{s/2} u\|_{L^q}.$$

Recall that this definition does not reduce to the preceding one when $q = 1$. We shall also use the homogeneous space \dot{H}_q^s , with norm

$$\|u\|_{\dot{H}^s} = \|(-\Delta)^{s/2} u\|_{L^q}.$$

Dispersive estimate a pointwise decay estimate of the form $|u(t, x)| \leq Ct^{-\alpha}$, usually for the solution of a partial differential equation.

Definition of the Subject

In a very broad sense, dispersion can be defined as the spreading of a fixed amount of matter, or energy, over a volume which increases with time. This intuitive picture suggests immediately the most prominent feature of dispersive phenomena: as matter spreads, its size, defined in a suitable sense, decreases at a certain rate. This effect should be contrasted with dissipation, which might be defined as an actual loss of energy, transferred to an external system (heat dissipation being the typical example).

This rough idea has been made very precise during the last 30 years for most evolution equations of mathematical physics. For the classical, linear, constant coefficient equations like the wave, Schrödinger, Klein–Gordon and Dirac equations, the decay of solutions can be measured in the L^p norms, and sharp estimates are available. In addition, detailed information on the profile of the solutions can be obtained, producing an accurate description of the evolution. For nonlinear equations, the theory now is able to explain and quantify several complex phenomena such as the splitting of solutions into the sum of a train of solitons, plus a radiation part which disperses and decays as time increases.

Already in the 1960s it was realized that precise information on the decay of free waves could be a powerful tool to investigate the effects of linear perturbations (describing interactions with electromagnetic fields) and even nonlinear perturbations of the equations. One of the first important applications was the discovery that a nonlinear PDE may admit global small solutions, provided the rate of decay is sufficiently strong compared with the concentration effect produced by the nonlinear terms. This situation is very different from the ODE setting, where dispersion is absent.

Today, dispersive and Strichartz estimates represent the backbone of the theory of nonlinear evolution equations. Their applications include local and global existence

results for nonlinear equations, existence of low regularity solutions, scattering, qualitative study of evolution equations on manifolds, and many others.

Introduction

To a large extent, mathematical physics is the study of qualitative and quantitative properties of the solutions to systems of differential equations. Phenomena which evolve in time are usually described by nonlinear evolution equations, and their natural setting is the Cauchy problem: the state of the system is assigned at an initial time, and the following evolution is, or should be, completely determined by the equations.

The most basic question concerns the local existence and uniqueness of the solution; we should require that, on one hand, the set of initial conditions is not too restrictive, so that a solution exists, and on the other hand it is not too lax, so that a unique solution is specified. For physically meaningful equations it is natural to require in addition that small errors in the initial state propagate slowly in time, so that two solutions corresponding to slightly different initial data remain close, at least for some time (continuous dependence on the data). This set of requirements is called *local well posedness* of the system. *Global well posedness* implies the possibility of extending these properties to arbitrary large times.

However, local well posedness is just the first step in understanding a nonlinear equation; basically, it amounts to a check that the equation and its Cauchy problem are meaningful from the mathematical and physical point of view. But the truly interesting questions concern global existence, asymptotic behavior and more generally the structure and shape of solutions. Indeed, if an equation describes a microscopic system, all quantities measured in an experiment correspond to asymptotic features of the solution, and local properties are basically useless in this context.

The classical tool to prove local well posedness is the energy method; it can be applied to many equations and leads to very efficient proofs. But the energy method is intrinsically unable to give answers concerning the global existence and behavior of solutions.

Let us illustrate this point more precisely. Most nonlinear evolution equations are of the form

$$Lu = N(u)$$

with a linear part L (e. g., the d'Alembert or Schrödinger operator) and a nonlinear term $N(u)$. The two terms are in competition: the linear flow is well behaved, usually with

several conserved quantities bounded by norms of the initial data, while the nonlinear term tends to concentrate the peaks of u and make them higher, increasing the singularity of the solution. At the same time, a small function u is made even smaller by a power-like nonlinearity $N(u)$.

The energy method tries to use only the conserved quantities, which remain constant during the evolution of the linear flow. Essentially, this means to regard the equation as an ODE of the form $y'(t) = N(y(t))$. This approach is very useful if the full (nonlinear) equation has a positive conserved quantity. But in general the best one can prove using the energy method is local well posedness, while a blow up in finite time can not be excluded, even if the initial data are assumed to be small, as the trivial example $y' = y^2$ clearly shows.

To improve on this situation, the strength of the linear flow must be used to a full extent. This is where dispersive phenomena enter the picture. A finer study of the operator L shows that there are quantities, different from the standard L^2 type norms used in the energy estimates, which actually *decay* during the evolution. Hence the term L has an additional advantage in the competition with N which can lead to global existence of small solutions. See Sect. “[The Nonlinear Wave Equation](#)” for more details.

This circle of ideas was initiated at the beginning of the 1970s and rapidly developed during the 1980s in several papers, (see [26,27,31,32,33,34,38,40,41,42,50] to mention just a few of the fundamental contributions on the subject). Global existence of small solutions was proved for many nonlinear evolution equations, including the nonlinear wave, Klein–Gordon and Schrödinger equations. The proof of the stability of vacuum for the Einstein equations [15,35] can be regarded as an extreme development of this direction of research.

Another early indication of the importance of L^p estimates came from the separate direction of harmonic analysis. In 1970, at a time when general agreement in the field was that the L^2 framework was the only correct one for the study of the wave equation, R. Strichartz ([44,45], see also [10,11,46]) proved the estimate

$$\begin{aligned} u_{tt} - \Delta u &= F, \quad u(0, x) = u_t(0, x) = 0 \\ \implies \|u\|_{L^p(\mathbb{R}^{n+1})} &\lesssim \|F\|_{L^{p'}(\mathbb{R}^{n+1})} \end{aligned}$$

for $p = 2(n+1)/(n-1)$. This was a seminal result, both for the emphasis on the L^p approach, and for the idea of regarding all variables including time on the same level, an idea which would play a central role in the later developments. At the same time, it hinted at a deep connection

between dispersive properties and the techniques of harmonic analysis.

The new ideas were rapidly incorporated in the mainstream theory of partial differential equations and became an important tool ([22,23,52]); the investigation of *Strichartz estimates* is still in progress. The freedom to work in an L^p setting was revealed as useful in a wide variety of contexts, including linear equations with variable coefficients, scattering, existence of low regularity solutions, and several others.

We might say that this point of view created a new important class of *dispersive equations*, complementing and overlapping with the standard classification into elliptic, hyperbolic, parabolic equations. Actually this class is very wide and contains most evolution equations of mathematical physics: the wave, Schrödinger and Klein–Gordon equations, Dirac and Maxwell systems, their nonlinear counterparts including wave maps, Yang–Mills and Einstein equations, Korteweg de Vries, Benjamin–Ono, and many others. In addition, many basic physical problems are described by systems obtained by coupling these equations; among the most important ones we mention the Dirac–Klein–Gordon, Maxwell–Klein–Gordon, Maxwell–Dirac, Maxwell–Schrödinger and Zakharov systems. Finally, a recent line of research pursues the extension of the dispersive techniques to equations with variable coefficients and equations on manifolds.

The main goal of this article is to give a quick overview of the basic dispersive techniques, and to show some important examples of their applications. It should be kept in mind however that the field is evolving at an impressive speed, mainly due to the contribution of ideas from Fourier and harmonic analysis, and it would be difficult to give a self-contained account of the recent developments, for which we point at the relevant bibliography. In Sect. “[The Mechanism of Dispersion](#)” we analyze the basic mechanism of dispersion, with special attention given to two basic examples (Schrödinger and wave equation). Section “[Strichartz Estimates](#)” is a concise exposition of Strichartz estimates, while Sects. “[The Nonlinear Wave Equation](#)” and “[The Nonlinear Schrödinger Equation](#)” illustrate some typical applications of dispersive techniques to problems of global existence for nonlinear equations. In the last Sect. “[Future Directions](#)” we review some of the more promising lines of research in this field.

The Mechanism of Dispersion

The Fourier transform puts in duality oscillations and translations according to the elementary rule

$$\mathcal{F}(f(x+h)) = e^{-ix \cdot h} \mathcal{F}f.$$

The same mechanism lies behind dispersive phenomena, and indeed two dual explanations are possible:

- in physical space, dispersion is characterized by a finite speed of propagation of the signals; data concentrated on some region tend to spread over a larger volume following the evolution flow. Correspondingly, the L^∞ and other norms of the solution tend to decrease;
- in Fourier space, dispersion is characterized by oscillations in the Fourier variable ξ , which increase for large $|\xi|$ and large t . This produces cancelation and decay in suitable norms.

We shall explore these dual points of view in two fundamental cases: the Schrödinger and the wave equation. These are the most important and hence most studied dispersive equations, and a fairly complete theory is available for both.

The Schrödinger Equation in Physical Space

The solution $u(t, x)$ of the Schrödinger equation

$$iu_t + \Delta u = 0, \quad u(0, x) = f(x)$$

admits several representations which are not completely equivalent. From spectral theory we have an abstract representation as a group of L^2 isometries

$$u(t, x) = e^{it\Delta} f; \quad (1)$$

then we have an explicit representation using the fundamental solution

$$\begin{aligned} u(t, x) &= \frac{1}{(4\pi it)^{n/2}} e^{-i\frac{|x|^2}{4t}} * f \\ &= \frac{1}{(4\pi it)^{n/2}} \int_{\mathbb{R}^n} e^{-i\frac{|x-y|^2}{4t}} f(y) dy; \end{aligned} \quad (2)$$

finally, we can represent the solution via Fourier transform as

$$u(t, x) = (2\pi)^{-n} \int e^{i(t|\xi|^2 + x \cdot \xi)} \widehat{f}(\xi) d\xi. \quad (3)$$

Notice for instance that the spectral and Fourier representations are well defined for $f \in L^2$, while it is not immediately apparent how to give a meaning to (2) if $f \notin L^1$.

From (1) we deduce immediately the *dispersive estimate* for the Schrödinger equation, which has the form

$$|e^{it\Delta} f| \leq (4\pi)^{-n/2} \frac{1}{t^{n/2}} \|f\|_{L^1}. \quad (4)$$

Since the L^2 norm of u is constant in time

$$\|e^{it\Delta} f\|_{L^2} = \|f\|_{L^2}, \quad (5)$$

this might suggest that the constant L^2 mass spreads on a region of volume $\sim t^n$ i. e. of diameter t . Apparently, this is in contrast with the notion that the Schrödinger equation has an infinite speed of propagation. Indeed, this is a correct but incomplete statement. To explain this point with an explicit example, consider the evolution of a *wave packet*

$$\pi_{x_0, \xi_0} = e^{-\frac{1}{2}|x-x_0|^2} e^{i(x-x_0) \cdot \xi_0}.$$

A wave packet is a mathematical representation of a particle localized near position x_0 , with frequency localized near ξ_0 ; notice indeed that its Fourier transform is

$$\widehat{\pi}_{x_0, \xi_0} = c e^{-\frac{1}{2}|\xi-\xi_0|^2} e^{i\xi \cdot x_0}$$

for some constant c . Exact localization both in space and frequency is of course impossible by Heisenberg's principle; however the approximate localization of the wave packet is a very good substitute, since the gaussian tails decay to 0 extremely fast. The evolution of a wave packet is easy to compute explicitly:

$$e^{it\Delta} \pi_{x_0, \xi_0} = (1+2it)^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \frac{(x-x_0-2t\xi_0)^2}{(1+4t^2)}\right) e^{i\Gamma}$$

where the real valued function Γ is given by

$$\Gamma = \frac{t(x-x_0)^2 + (x-x_0-t\xi_0) \cdot \xi_0}{(1+4t^2)}.$$

We notice that the particle “moves” with velocity $2\xi_0$ and the height of the gaussian spreads precisely at a rate $t^{n/2}$. The mass is essentially concentrated in a ball of radius t , as suspected. This simple remark is at the heart of one of the most powerful techniques to have appeared in harmonic analysis in recent years, the wave packet decomposition, which has led to breakthroughs in several classical problems (see e. g. [7,8,48,51]).

The Schrödinger Equation in Fourier Space

In view of the ease of proof of (4), it may seem unnecessary to examine the solution in Fourier space, however a quick look at this other representation is instructive and prepares the discussion for the wave equation. If we examine (3), we see that the integral is a typical oscillatory integral which can be estimated using the stationary phase method. The phase here is

$$\phi(t, x; \xi) = t|\xi|^2 + x \cdot \xi \implies \nabla_{\xi} \phi = 2t\xi + x$$

so that for each (t, x) we have a unique stationary point $\xi = -x/2t$; the n curvatures at that point are different

from zero, and this already suggests a decay of the order $t^{-n/2}$. This argument can be made precise and gives an alternative proof of (4).

We also notice that the Fourier transform in both space and time of $u(t, x) = e^{it\Delta} f$ is a temperate distribution in $S'(\mathbb{R}^{n+1})$ which can be written, apart from a constant,

$$\tilde{u}(\tau, \xi) = \delta(\tau - |\xi|^2)$$

and is supported on the paraboloid $\tau = |\xi|^2$. This representation of the solution gives additional insight to its behavior in certain frequency regimes, and has proved to be of fundamental importance when studying the existence of low regularity solutions to the nonlinear Schrödinger equation [6,9].

The Wave Equation in Physical Space

For the wave equation, the situation is reversed: the most efficient proof of the dispersive estimate is obtained via the Fourier representation of the solution. This is mainly due to the complexity of the fundamental solution; we recall the relevant formulas briefly. The solution to the homogeneous wave equation

$$w_{tt} - \Delta w = 0, \quad w(0, x) = f(x), \quad w_t(0, x) = g(x)$$

can be written as

$$u(t, x) = \cos(t|D|)f + \frac{\sin(t|D|)}{|D|}g$$

where we used the symbolic notations

$$\begin{aligned} \cos(t|D|)f &= \mathcal{F}^{-1} \cos(t|\xi|) \mathcal{F} f, \\ \frac{\sin(t|D|)}{|D|}g &= \mathcal{F}^{-1} \frac{\sin(t|\xi|)}{|\xi|} \mathcal{F} g \end{aligned}$$

(recall that \mathcal{F} denotes the Fourier transform in space variables). Notice that

$$\cos(t|D|)f = \frac{\partial}{\partial t} \frac{\sin(t|D|)}{|D|} f,$$

hence the properties of the first operator can be deduced from the second. Moreover, by Duhamel's formula, we can express the solution to the nonhomogeneous wave equation

$$w_{tt} - \Delta w = F(t, x), \quad w(0, x) = 0, \quad w_t(0, x) = 0$$

as

$$u(t, x) = \int_0^t \frac{\sin((t-s)|D|)}{|D|} F(u(s, \cdot)) ds.$$

In other words, a study of the operator

$$S(t)g = \frac{\sin(t|D|)}{|D|}g$$

is sufficient to deduce the behavior of the complete solution to the nonhomogeneous wave equation with general data.

The operator $S(t)$ can be expressed explicitly using the fundamental solution of the wave equation. We skip the one-dimensional case since then from the expression of the general solution

$$u(t, x) = \phi(t + x) + \psi(t - x)$$

it is immediately apparent that no dispersion can occur. In odd space dimension $n \geq 3$ we have

$$\frac{\sin(t|D|)}{|D|}g = c_n \left(\frac{1}{t} \partial_t \right)^{\frac{n-3}{2}} \left(\frac{1}{t} \int_{|y|=t} g(x-y) d\sigma_y \right),$$

and in even space dimension $n \geq 2$

$$\frac{\sin(t|D|)}{|D|}g = c_n \left(\frac{1}{t} \partial_t \right)^{\frac{n-2}{2}} \left(\frac{1}{t} \int_{|y|<t} \frac{g(x-y)}{\sqrt{t^2 - |y|^2}} dy \right)$$

for some constant c_n . These formulas can be expanded and the resulting integrals estimated in order to compute the explicit decay rate of the solution; this approach was followed in [50] (see also [43] for a more concise proof based on the fundamental solution). We shall prove the dispersive estimate using Fourier methods in the next section, but here it is instructive to consider the particularly simple case $n = 3$, when the above representation reduces to *Kirchhoff's formula*

$$\frac{\sin(t|D|)}{|D|}g = \frac{c}{t} \int_{|y|=t} g(x-y) d\sigma_y.$$

Using the divergence theorem it is easy to deduce that

$$\left| \frac{\sin(t|D|)}{|D|}g \right| \lesssim \frac{1}{t} \|\nabla f\|_{L^1}.$$

Recall that the rate of decay for the Schrödinger equation in dimension 3 is $\sim t^{-3/2}$; for the wave equation the rate of decay is slower, in addition the estimate shows a loss of one derivative with respect to the initial data. Actually this estimate is optimal, as it is evident from the explicit representation of the solution. In dimension n the correct rate of decay is $t^{-\frac{n-1}{2}}$ and the loss of derivative is of order $(n-1)/2$, as we shall see in the next section.

This result is in apparent contrast with our physical space picture. Indeed the L^2 norm of the first derivatives

of the solution remains constant, in view of the energy estimate

$$\|u_t(t, \cdot)\|_{L^2}^2 + \|\nabla u(t, \cdot)\|_{L^2}^2 = \|u_t(0, \cdot)\|_{L^2}^2 + \|\nabla u(0, \cdot)\|_{L^2}^2;$$

on the other hand, the velocity of propagation for the wave equation is exactly 1, so that the constant L^2 energy is spread on a volume which increases at a rate $\sim t^n$; as a consequence, we would expect the same rate of decay $\sim t^{-n/2}$ as for the Schrödinger equation. However, a more detailed study of the shape of the solution reveals that it tends to concentrate in a shell of constant thickness around the light cone. For instance, in odd dimension larger than 3 initial data with support in a ball $B(0, R)$ produce a solution which at time t is supported in the shell $B(0, t+R) \setminus B(0, t-R)$. Also in even dimension the solution tends to concentrate along the light cone $|t| = |x|$, with a faster decreasing tail far from the cone. Thus, the energy spreads over a volume of size t^{n-1} , in accordance with the rate of decay in the dispersive estimate.

The Wave Equation in Fourier Space

Since

$$2i \frac{\sin(t|D|)}{|D|}g = \frac{e^{it|D|}}{|D|}g - \frac{e^{-it|D|}}{|D|}g,$$

$$2 \cos(t|D|)f = e^{it|D|}f + e^{-it|D|}f,$$

we can unify the study of the various components of the solution by considering the operator

$$e^{it|D|}f = \mathcal{F}^{-1} e^{it|\xi|} \mathcal{F}f = (2\pi)^{-n} \int e^{i(x \cdot \xi + t|\xi|)} \widehat{f} d\xi.$$

The phase of the last integral is

$$\phi(t, x; \xi) = t|\xi| + x \cdot \xi \implies \nabla_\xi \phi = t \frac{\xi}{|\xi|} + x$$

and we notice immediately that the structure of stationary points is more complex than for Schrödinger: if $|t| = |x|$, i. e. if the point (t, x) is on the light cone, the phase has a half line of critical points

$$|t| = |x| \implies \frac{\xi}{|\xi|} = -\frac{x}{t}$$

which of course are degenerate since only $n-1$ curvatures are different from zero; on the other hand, when (t, x) is not on the light cone, the phase is nondegenerate at all points.

In the nondegenerate case $|t| \neq |x|$, we notice that

$$\nabla_\xi \phi = t \frac{\xi}{|\xi|} + x \implies |\nabla_\xi \phi| \geq ||t| - |x||$$

hence by direct integration by parts we obtain

$$\left| \int e^{i(x \cdot \xi + t|\xi|)} \widehat{f} d\xi \right| \leq C(f) |t| - |x|^{-(n-1)}. \quad (6)$$

Notice that the singularity $|\xi|$ of the phase makes it impossible to perform the integration by parts more than $n-1$ times, but this is largely sufficient for our purposes here.

In the degenerate case $|t| = |x|$, we can eliminate the degeneracy by passing to polar coordinates:

$$\begin{aligned} \int e^{i(x \cdot \xi + t|\xi|)} \widehat{f} d\xi &= \int_0^\infty e^{it\rho} \rho^{n-1} \\ &\quad \times \int_{|\omega|=1} e^{ix \cdot \rho \omega} \widehat{f}(\rho \omega) d\omega d\rho. \end{aligned}$$

Now the phase in the inner integral has exactly two isolated critical points for each value of x ; by a standard application of the stationary phase method we can obtain an estimate of the form

$$\left| \int e^{i(x \cdot \xi + t|\xi|)} \widehat{f} d\xi \right| \leq C(f) |x|^{-\frac{n-1}{2}}. \quad (7)$$

The power $(n-1)/2$ is directly related to the dimension $n-1$ of the unit sphere.

From the above computations it is possible to obtain a very detailed description of the shape of the solution (see e. g. [20]). However, here we are only interested in the rate of decay of the L^∞ norm of the solution. Thus, using estimate (6) when $|x| < |t|/2$ and (7) when $|x| > |t|/2$ we obtain

$$|e^{it|D|} f| \leq C(f) |t|^{-\frac{n-1}{2}}.$$

It is also possible to make precise the form of the constant $C(f)$ appearing in the estimate. This can be done in several ways; one of the most efficient exploits the scaling properties of the wave equation (see [10,23]). In order to state the optimal estimate we briefly recall the definition of some function spaces. The basic ingredient is the homogeneous Paley–Littlewood partition of unity which has the form

$$\begin{aligned} \sum_{j \in \mathbb{Z}} \phi_j(\xi) &= 1, \quad \phi_j(\xi) = \phi_0(2^{-j}\xi), \\ \sup \phi_j &= \{\xi : 2^{j-1} \leq |\xi| \leq 2^{j+1}\} \end{aligned}$$

for a fixed $\phi_0 \in C_0^\infty(\mathbb{R}^n)$ (the following definitions are independent of the precise choice of ϕ_0). To each symbol $\phi_j(\xi)$ we associate by the standard calculus the operator $\phi_j(D) = \mathcal{F}^{-1} \phi_j(\xi) \mathcal{F}$. Now the homogeneous Besov norm $\dot{B}_{1,1}^s$ can be defined as follows:

$$\|f\|_{\dot{B}_{1,1}^s} = \sum_{j \in \mathbb{Z}} 2^{js} \|\phi_j(D)f\|_{L^1} \quad (8)$$

and the corresponding Besov space can be obtained by completing C_0^∞ (see [4] for a thorough description of these definitions).

Then the optimal dispersive estimate for the wave equation is

$$|e^{it|D|} f| \leq C \|f\|_{\dot{B}_{1,1}^{\frac{n-1}{2}}} |t|^{-\frac{n-1}{2}}. \quad (9)$$

We might add that in odd space dimension a slightly better estimate can be proved involving only the standard Sobolev norm $\dot{W}^{\frac{n-1}{2}}$ instead of the Besov norm.

The singularity in t at the right hand side of (9) can be annoying in applications, thus it would be natural to try to replace $|t|$ by $(1+|t|)$ by some different proof. However, this is clearly impossible since both sides have the same scaling with respect to the transformation $u(t, x) \rightarrow u(\lambda t, \lambda x)$ which takes a solution of the homogeneous wave equation into another solution. Moreover, if such an estimate were true, taking the limit as $t \rightarrow 0$ we would obtain a false Sobolev embedding. Thus, the only way to get a nonsingular estimate is to replace the Besov norm of the data by some stronger norm.

Indeed, for $|t| < 1$ we can use the L^2 conservation which implies, for any $\epsilon > 0$,

$$|e^{it|D|} f| \leq \|e^{it|D|} f\|_{H^{n/2+\epsilon}} = \|f\|_{H^{n/2+\epsilon}}.$$

Combining the two estimates, we arrive at

$$|e^{it|D|} f| \leq \left(\|f\|_{\dot{B}_{1,1}^{\frac{n-1}{2}}} + \|f\|_{H^{n/2+\epsilon}} \right) (1+|t|)^{-\frac{n-1}{2}}. \quad (10)$$

Other Dispersive Equations

Dispersive estimates can be proved for many other equations; see for instance [3] for general results concerning the equations of the form

$$iu_t = P(D)u, \quad P \text{ polynomial}$$

which correspond to integrals of the form

$$\int e^{i(x \cdot \xi + P(\xi))} \widehat{f}(\xi) d\xi.$$

Here we would like to add some more details on two equations of particular importance for physics. The Klein–Gordon equation

$$u_{tt} - \Delta u + u = 0, \quad u(0, x) = f(x), \quad u_t(0, x) = g(x)$$

shows the same decay rate as the Schrödinger equation, although for most other properties it is very close to the wave

equation. This phenomenon can be easily explained. In physical space, we notice that the solution has finite speed of propagation less or equal to 1, so that for compactly supported initial data the solution spreads uniformly on a volume of order t^n , and there is no concentration along the light cone as for the wave equation.

On the other hand, the Fourier representation of the solution is

$$u(t, x) = \cos(t\langle D \rangle) f + \frac{\sin(t\langle D \rangle)}{\langle D \rangle} g$$

where we used the symbolic notations

$$\begin{aligned} \cos(t\langle D \rangle) f &= \mathcal{F}^{-1} \cos(t|\langle \xi \rangle|) \mathcal{F} f, \\ \frac{\sin(t\langle D \rangle)}{\langle D \rangle} g &= \mathcal{F}^{-1} \frac{\sin(t|\langle \xi \rangle|)}{|\langle \xi \rangle|} \mathcal{F} g \end{aligned}$$

with $\langle \xi \rangle = (1 + |\xi|^2)^{1/2}$ i.e. $\langle D \rangle = (1 - \Delta)^{1/2}$. Thus, we see that, similarly to the wave equation, the study of the solutions can be reduced to the basic operator

$$e^{it\langle D \rangle} f = \mathcal{F}^{-1} e^{it\langle \xi \rangle} \mathcal{F} f = (2\pi)^{-n} \int e^{i(t\langle \xi \rangle + x \cdot \xi)} \widehat{f}(\xi) d\xi.$$

It is easy to check that the phase has an isolated stationary point if and only if $|t| < |x|$, and no stationary point if $|t| \geq |\xi|$:

$$\phi(t, x; \xi) = t\langle \xi \rangle + x \cdot \xi, \quad \nabla_{\xi} \phi = t \frac{\xi}{\langle \xi \rangle} + x,$$

in accord with the claimed dispersion rate of order $t^{-n/2}$. However, the proof is not completely elementary (for a very detailed study, see Section 7.2 of [25]). The optimal estimate is now

$$|e^{it\langle D \rangle} f| \leq C \|f\|_{B_{1,1}^{\frac{n}{2}}} |t|^{-\frac{n}{2}} \quad (11)$$

(see the Appendix to [18]). Here we used the *nonhomogeneous Besov norm* $B_{1,1}^s$ which is defined exactly as the homogeneous one (8) by using the inhomogeneous version of the Paley–Littlewood partition of unity:

$$\begin{aligned} \phi_{-1}(\xi) + \sum_{j \in \mathbb{N}} \phi_j(\xi) &= 1, \\ \phi_j(\xi) &= \phi_0(2^{-j}\xi) \quad \text{for } j \geq 0, \\ \sup \phi_j &= \{\xi: 2^{j-1} \leq |\xi| \leq 2^{j+1}\}, \\ \sup \phi_{-1} &= \{|\xi| \leq 1\} \end{aligned}$$

for some $\phi_{-1}, \phi_0 \in C_0^\infty(\mathbb{R}^n)$.

A second important example is the *Dirac system*. From a mathematical point of view (and setting the values of the

speed of light and Planck's constant equal to 1 for simplicity of notation) this is a 4×4 constant coefficient system of the form

$$iu_t + \mathcal{D}u = 0$$

in the *massless* case, and

$$iu_t + \mathcal{D}u + \beta u = 0$$

in the *massive* case. Here $u: \mathbb{R}_t \times \mathbb{R}_x^3 \rightarrow \mathbb{C}^4$, the operator \mathcal{D} is defined as

$$\mathcal{D} = \frac{1}{i} \sum_{k=1}^3 \alpha_k \partial_k$$

and the 4×4 *Dirac matrices* can be written

$$\alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2 \end{pmatrix}, \quad k = 1, 2, 3$$

in terms of the *Pauli matrices*

$$\begin{aligned} I_2 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & \sigma_1 &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\ \sigma_2 &= \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, & \sigma_3 &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \end{aligned}$$

Then the solution $u(t, x) = e^{it\mathcal{D}} f$ of the massless Dirac system with initial value $u(0, x) = f(x)$ satisfies the dispersive estimate

$$|u(t, x)| \leq \frac{C}{t} \|f\|_{\dot{B}_{1,1}^2}$$

while in the massive case we have

$$|u(t, x)| \leq \frac{C}{t^{\frac{3}{2}}} \|f\|_{\dot{B}_{1,1}^{\frac{5}{2}}}$$

(see Sect. 7 of [16]).

Strichartz Estimates

Since the Schrödinger flow $e^{it\Delta} f$ is an L^2 isometry, it may be tempting to regard it as some sort of “rotation” of the Hilbert space L^2 . This picture is far from true: there exist small subspaces of L^2 which contain the flow for almost all times. This surprising phenomenon, which holds in greater generality for any unbounded selfadjoint operator on an abstract Hilbert space, is known as *Kato smoothing* and was discovered by T. Kato in 1966 [29], see also [39]. This fact is just a corollary of a quantitative estimate, the *Kato smoothing estimate*, which in the case of the Laplace

operator takes the following form: there exist closed unbounded operators A on L^2 such that

$$\|Ae^{it\Delta}f\|_{L^2L^2} \lesssim \|f\|_{L^2}.$$

As a consequence, the flow belongs for a. e. t to the domain of the operator A , which can be a very small subspace of L^2 . Two simple examples of smoothing estimates are the following:

$$\begin{aligned} V(x) &\in L^2 \cap L^{n-\epsilon} \cap L^{n+\epsilon} \\ \implies \|V(x)e^{it\Delta}f\|_{L^2L^2} &\lesssim \|f\|_{L^2} \end{aligned} \quad (12)$$

and

$$\| |x|^{-1} e^{it\Delta} f \|_{L^2L^2} \lesssim \|f\|_{L^2}. \quad (13)$$

On the other hand, Strichartz proved in 1977 [46] that

$$\begin{aligned} p &= \frac{2(n+2)}{n} \\ \implies \|e^{it\Delta}f\|_{L^p(\mathbb{R}^{n+1})} &\lesssim \|f\|_{L^2} \end{aligned} \quad (14)$$

thus showing that the smoothing effect could be measured also in the L^p norms. If we try to reformulate (12) in terms of an L^p norm, we see that it would follow from

$$\|e^{it\Delta}f\|_{L^2L^{\frac{2n}{n-2}}} \lesssim \|f\|_{L^2} \quad (15)$$

while (13) would follow from the slightly stronger

$$\|e^{it\Delta}f\|_{L^2L^{\frac{2n}{n-2},2}} \lesssim \|f\|_{L^2} \quad (16)$$

where $L^{p,2}$ denotes a Lorentz space. Both (15) and (16) are true, although their proof is far from elementary. Such estimates are now generally called *Strichartz estimates*. They can be deduced from the dispersive estimates, hence they are in a sense weaker and contain less information. However, for this same reason they can be obtained for a large number of equations, even with variable coefficients, and their flexibility and wide range of applications makes them extremely useful.

The full set of Strichartz estimates, with the exception of the endpoint (see below) was proved in [22], where it was also applied to the problem of global well posedness for the nonlinear Schrödinger equation. The corresponding estimates for the wave equation were obtained in [23], however they turned out to be less useful because of the loss of derivatives. A fundamental progress was made in [30], where not only the most difficult endpoint case was solved, but it also outlined a general procedure to deduce the Strichartz estimates from the dispersive estimates, with applications to any evolution equation. In the following we shall focus on the two main examples, the Schrödinger and the wave equation, and we refer to the original papers for more details and the proofs.

The Schrödinger Equation

The homogeneous Strichartz estimates have the general form

$$\|e^{it\Delta}f\|_{L^pL^q} \lesssim \|f\|_{L^2}$$

for suitable values of the couple (p, q) . Notice that we can restrict the L^pL^q norm at the right hand side to an arbitrary time interval I (i. e. to $L^p(I; L^q(\mathbb{R}^n))$). Notice also that by the invariance of the flow under the scaling $u(t, x) \mapsto u(\lambda^2 t, \lambda x)$ the possible couples (p, q) are bound to satisfy

$$\frac{2}{p} + \frac{n}{q} = \frac{n}{2}. \quad (17)$$

The range of possible values of q is expressed by the inequalities

$$2 \leq q \leq \frac{2n}{n-2} \quad \text{if } n \geq 2, \quad q \neq \infty, \quad (18)$$

$$2 \leq q \leq \infty \quad \text{if } n = 1; \quad (19)$$

the index p varies accordingly between ∞ and 2 (between 4 and 2 in one space dimension). Such couples (p, q) are called (*Schrödinger*) *admissible*. It is easy to visualize the admissible couples by plotting the usual diagram in the $(1/p, 1/q)$ space (see Fig 1–3).

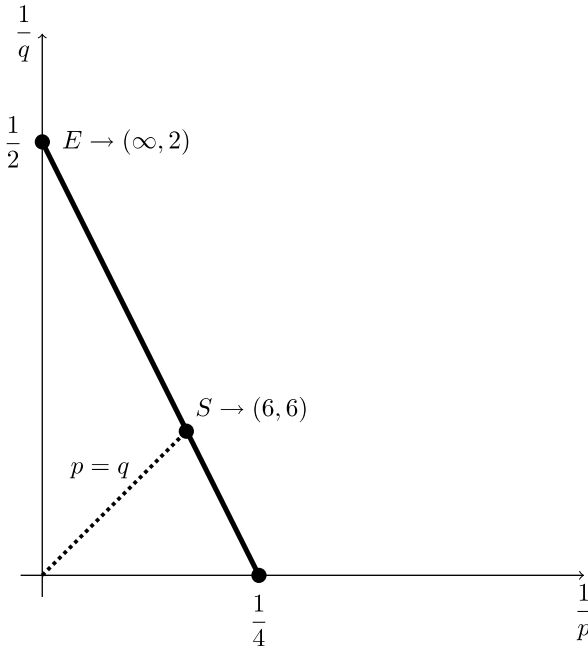
Note that the choice $(p, q) = (\infty, 2)$, the point E in the diagrams, corresponds to the conservation of the L^2 norm, while at the other extreme we have for $n \geq 2$ point P corresponding to the couple

$$(p, q) = \left(2, \frac{2n}{n-2}\right). \quad (20)$$

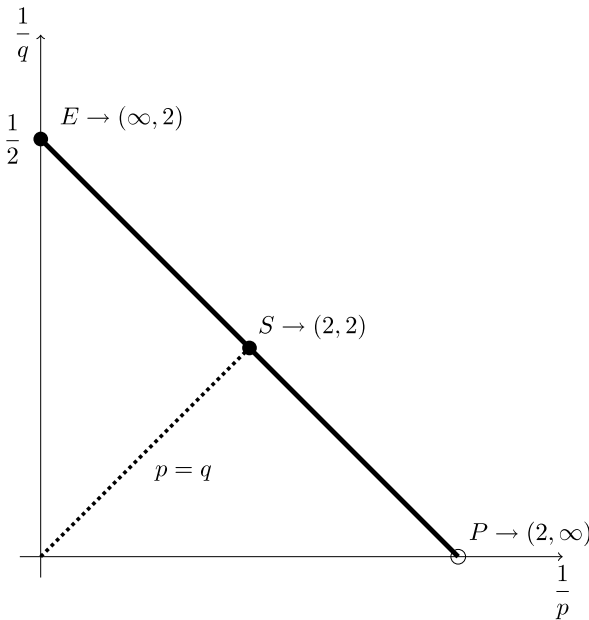
This is usually called the *endpoint* and is excluded in dimension 2, allowed in dimension ≥ 3 . The original Strichartz estimate corresponds to the case $p = q$ and is denoted by S . After the proof in [22] of the general Strichartz estimates, it was not clear if the endpoint case P could be reached or not; the final answer was given in [30] where the endpoint estimates were proved to be true for all dimension larger than three. For the two-dimensional case and additional estimates, see [47]. Notice that all admissible couples can be obtained by interpolation between the endpoint and the L^2 conservation, thus in a sense the endpoint contains the maximum of information concerning the L^p smoothing properties of the flow.

There exist an *inhomogeneous* form of the estimates, which is actually equivalent to the homogeneous one, and can be stated as follows:

$$\left\| \int_0^t e^{i(t-s)\Delta} F(s) ds \right\|_{L^pL^q} \lesssim \|F\|_{L^{p'}L^{q'}} \quad (21)$$

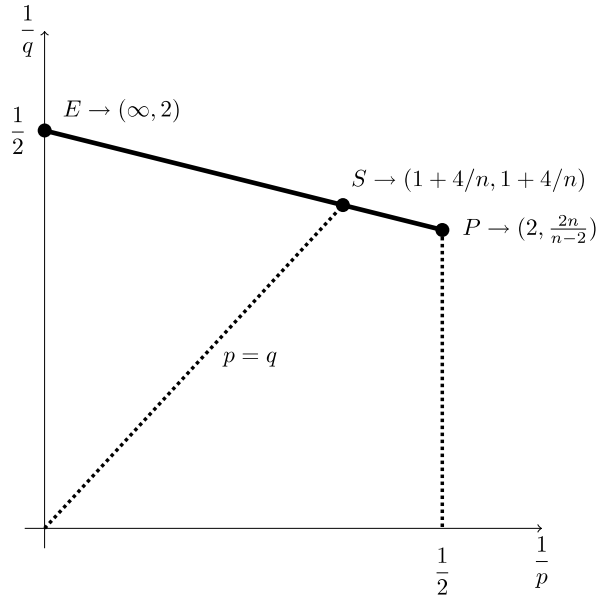


Dispersion Phenomena in Partial Differential Equations, Figure 1
Schrödinger admissible couples, $n = 1$



Dispersion Phenomena in Partial Differential Equations, Figure 2
Schrödinger admissible couples, $n = 2$

for all admissible couples (p, q) and (\tilde{p}, \tilde{q}) , which can be chosen independently. Here p' denotes the conjugate exponent to p . These estimates continue to be true in their *local* version, with the norms at both sides restricted to a fixed time interval $I \subset \mathbb{R}$. Notice that the inhomoge-



Dispersion Phenomena in Partial Differential Equations, Figure 3
Schrödinger admissible couples, $n \geq 3$

neous global estimates are optimal since other combination of indices are excluded by scale invariance; however, the local Strichartz estimates can be slightly improved allowing for indices outside the admissible region (see [21]).

We also mention that a much larger set of estimates can be deduced from the above ones by combining them with Sobolev embeddings; in the above diagrams, all couples in the region bounded by the line of admissibility, the axes and the line $1/p = 2$ can be reached, while the external region is excluded by suitable counterexamples (see [30]).

Strichartz Estimates for the Wave Equation

Most of the preceding section has a precise analogue for the wave equation. Since the decay rate for the dispersive estimates is $t^{-(n-1)/2}$ for the wave equation instead of $t^{-n/2}$ as for the Schrödinger equation, this causes a shift $n \rightarrow n - 1$ in the numerology of indices. A more serious difference is the loss of derivatives: this was already apparent in the original estimate by Strichartz [44], which can be written

$$\|e^{it|D|}f\|_{L^{\frac{2(n+1)}{n-1}}} \lesssim \|f\|_{\dot{H}^{1/2}}, \quad n \geq 2.$$

The general homogeneous Strichartz estimates for the wave equation take the form

$$\|e^{it|D|}f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{1}{p} \frac{n+1}{n-1}}};$$

now the (wave) *admissible* couples of indices must satisfy the scaling condition

$$\frac{2}{p} + \frac{n-1}{q} = \frac{n-1}{2} \quad (22)$$

and the constraints $2 \leq p \leq \infty$,

$$2 \leq q \leq \frac{2(n-1)}{n-3} \quad \text{if } n \geq 3, \quad q \neq \infty \quad (23)$$

$$2 \leq q \leq \infty \quad \text{if } n = 2, \quad (24)$$

while p varies accordingly between 2 and ∞ (between 4 and ∞ in dimension 2). Recall that in dimension 1 the solutions of the wave equation do not decay since they can be decomposed into traveling waves. We omit the graphic representation of these sets of conditions since it is completely analogous to the Schrödinger case (apart from the shift $n \rightarrow n-1$).

The inhomogeneous estimates may be written

$$\left\| \int_0^t e^{i(t-s)\Delta} F(s) ds \right\|_{L^p \dot{H}_q^{-\frac{1}{p} - \frac{n+1}{n-1}}} \lesssim \|F\|_{L^{p'} \dot{H}_{q'}^{-\frac{1}{p'} - \frac{n+1}{n-1}}} \quad (25)$$

in terms of the homogeneous Sobolev norms

$$\|g\|_{\dot{H}_q^s} = \| |D|^s g \|_{L^q}, \quad |D|^s = (1 - \Delta)^{s/2}.$$

Other Equations

The general procedure of Keel and Tao can be applied to a large variety of equations. For the convenience of the reader we list here the precise form of the homogeneous Strichartz estimates in some important cases.

For the *Klein–Gordon* equation we have

$$\|e^{it(D)} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{1}{p} - \frac{n+2}{n}}}$$

with (p, q) Schrödinger admissible (recall that the rate of dispersion is $t^{-n/2}$ for Klein–Gordon). For the *massless Dirac* system we have

$$\|e^{itD} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{2}{p}}}, \quad n = 3,$$

with (p, q) wave admissible in dimension 3, while for the *massive Dirac* system the estimate takes the form

$$\|e^{it(D+\beta)} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{5}{3p}}}, \quad n = 3$$

with (p, q) Schrödinger admissible in dimension 3.

The Nonlinear Wave Equation

In this section we shall illustrate, using the semilinear wave equation as a basic example, what kind of improvements can be obtained over the classical energy method via the additional information given by dispersive techniques.

From the linear theory it is well known that the correct setting of the Cauchy problem for the wave equation

$$u_{tt} - \Delta u = F(t, x), \quad u(t, x): \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{C}$$

requires two initial data at $t = 0$:

$$u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x).$$

For this equation the energy method is particularly simple. If we multiply the equation by \bar{u}_t we can recast it in the form

$$\partial_t (|u_t|^2 + |\nabla u|^2) = 2\Re \nabla \cdot (\bar{u}_t \nabla u) + 2\Re (\bar{u}_t F)$$

and integrating this identity on different domains we can obtain several conservation laws. Integrating on a slab $0 \leq t \leq T, x \in \mathbb{R}^n$ we arrive at an estimate of the form

$$\|u_t\|_{L_T^\infty L^2} + \|\nabla u\|_{L_T^\infty L^2} \lesssim \|u_1\|_{L^2} + \|\nabla u_0\|_{L^2} + \|F\|_{L_T^1 L^2} \quad (26)$$

where we are using the shorthand notation

$$L_T^p L^q = L^p(0, T; L^q(\mathbb{R}^n)).$$

Estimate (26) is the basic *energy estimate*, which is enough to prove the existence of a global solution $u \in C(\mathbb{R}; H^1(\mathbb{R}^n)) \cap C^1(\mathbb{R}; L^2(\mathbb{R}^n))$ of the linear wave equation for data in $(u_0, u_1) \in H^1 \times L^2$ and a forcing term in $F \in L_T^1 L^2$ for all T . Uniqueness and continuous dependence on the data are simple consequences of the energy estimate. Moreover, if we differentiate the equation k times, the same computations lead to the higher order energy estimates

$$\|u_t\|_{L_T^\infty H^k} + \|u\|_{L_T^\infty H^{k+1}} \lesssim \|u_1\|_{H^k} + \|u_0\|_{H^{k+1}} + \|F\|_{L_T^1 H^k}, \quad k = 0, 1, 2, \dots \quad (27)$$

Consider now the nonlinear wave equation

$$u_{tt} - \Delta u = F(u) \quad (28)$$

for some function $F(s)$ sufficiently smooth. The classical approach to solve it is to look for a fixed point of the mapping $u = \Phi(v)$, where $u(t, x)$ solves the linearized equation

$$u_{tt} - \Delta u = F(v). \quad (29)$$

If we apply the energy estimate to the difference of two solutions $u_1 = \Phi(v_1)$ and $u_2 = \Phi(v_2)$ with the same initial data, we obtain

$$\begin{aligned} \|\partial_t(u_1 - u_2)\|_{L_T^\infty H^k} + \|u_1 - u_2\|_{L_T^\infty H^{k+1}} \\ \lesssim \|F(v_1) - F(v_2)\|_{L_T^1 H^k}. \end{aligned}$$

Using a standard Moser-type estimate (see e. g. [25,49])

$$\|f(v_1) - f(v_2)\|_{H^k} \lesssim C(\|v_1\|_{L^\infty} + \|v_2\|_{L^\infty})\|v_1 - v_2\|_{H^k}$$

we arrive at the inequality

$$\begin{aligned} \|\partial_t(u_1 - u_2)\|_{L_T^\infty H^k} + \|u_1 - u_2\|_{L_T^\infty H^{k+1}} \\ \lesssim C(\|v_1\|_{L_T^\infty L^\infty} + \|v_2\|_{L_T^\infty L^\infty}) \cdot \|v_1 - v_2\|_{L_T^1 H^k}. \end{aligned}$$

In particular, writing $X_T = L_T^\infty H^k$, and noticing that

$$\|G\|_{L_T^1 H^k} \leq \|G\|_{X_T} \cdot T,$$

we obtain

$$\begin{aligned} \|u_1 - u_2\|_{X_T} \\ \lesssim C(\|v_1\|_{L_T^\infty L^\infty} + \|v_2\|_{L_T^\infty L^\infty}) \cdot \|v_1 - v_2\|_{X_T} \cdot T. \end{aligned}$$

We still have to handle the L^∞ norm of v_j , which can be controlled by the H^k norm provided $k > n/2$. In conclusion

$$\begin{aligned} k > \frac{n}{2} \quad \implies \quad \|u_1 - u_2\|_{X_T} \\ \lesssim C(\|v_1\|_{X_T} + \|v_2\|_{X_T}) \cdot \|v_1 - v_2\|_{X_T} \cdot T. \end{aligned}$$

At this point the philosophy of the method should be clear: if we agree to differentiate $n/2$ times the equation, we can prove a nonlinear estimate, with a coefficient which can be made small for T small. Thus, for $T \ll 1$ the mapping $v \mapsto u$ is a contraction, and we obtain a local (in time) solution of the nonlinear equation.

The energy method is quite elementary, yet it has the remarkable property that it can be adapted to a huge number of equations: it can be applied to equations with variable coefficients, and even to fully nonlinear equations. An outstanding example of its flexibility is the proof of the local solvability of Einstein's equations ([14]).

The main drawbacks of the energy method are:

- 1) A high regularity of the data and of the solution is required;
- 2) The nonlinear nature of the estimate leads to a *local* existence result, leaving completely open the question of global existence.

Actually, the two difficulties are related to each other. Indeed, from the above argument one can check that the lifespan of the local solution is a function of the H^k norm of the initial data. Most equations with a physical origin possess conserved quantities, which are essentially H^k norms of the solution with k low, e. g. $k = 0$ or 1 . Assume now that we can prove the local existence for data say in H^1 , on some interval $[0, T_0]$ with T_0 depending on the H^1 norm of the data. Since this norm is conserved, we can take T_0 as a new initial time and construct a solution on $[T_0, 2T_0]$, and so on. This is the essence of the *continuation method*. As a consequence, the local existence of low regularity solution with the addition of suitable conservation laws can lead to global existence.

We shall now apply the dispersive information to improve on the preceding result and obtain the global existence of small solution, under suitable assumptions on the nonlinear term. Consider a free wave $w(t, x)$, solution of

$$w_{tt} - \Delta w = 0, \quad w(0, x) = f(x), \quad w_t(0, x) = g(x).$$

Recalling the dispersive estimate (10), here the following simplified version

$$|w(t, x)| \leq C(1+t)^{-\frac{n-1}{2}} \|f\|_{W^{N,1}}, \quad N > n$$

will be sufficient. This suggests modification of the energy norm by adding a term that incorporates the decay information. To this end, we introduce the time-dependent quantity

$$\begin{aligned} M(t) &= \sup_{0 \leq \tau \leq t} \left\{ \|u(\tau, \cdot)\|_{H^{k+1}} + (1+\tau)^{\frac{n-1}{2}} \|u(\tau, \cdot)\|_{L^\infty} \right\}, \\ k &= \left[\frac{n}{2} \right]. \end{aligned}$$

Consider now the nonlinear Eq. (28) with initial data u_0, u_1 . We already know that a local solution exists, provided the initial data are smooth enough; we shall now try to extend this solution to a global one. Using Duhamel's principle we rewrite the equation as

$$u(t, x) = w(t, x) + \int_0^t \frac{\sin((t-s)|D|)}{|D|} F(u(s, \cdot)) ds$$

where we are using the symbolic notations of Sect. "Other Dispersive Equations". \mathcal{F} being the space Fourier transform. Note that the above dispersive estimate can be applied in particular to the last term; we obtain

$$\left| \frac{\sin((t-s)|D|)}{|D|} F(s) \right| \leq C(1+t-s)^{-\frac{n-1}{2}} \|F\|_{W^{N,1}}.$$

By this inequality and the energy estimate already proved, in a few steps we arrive at the estimate

$$M(T) \lesssim C_0 + \|F\|_{L_T^1 H^k} + (1+T)^{\frac{n-1}{2}} \int_0^T (1+T-s)^{-\frac{n-1}{2}} \|F(u(s))\|_{W^{N,1}} ds$$

where $C_0 = \|u_0\|_{W^{N,1}} + \|u_1\|_{W^{N,1}}$ and the mute constant is independent of T . By suitable nonlinear estimates of Moser type, we can bound the derivatives of the nonlinear term using only the L^∞ and the H^k norm of u , i. e., the quantity $M(t)$ itself. If we assume that $F(u) \sim |u|^\gamma$ for small u , we obtain

$$M(T) \lesssim C_0 + M(T)^\gamma (1+T)^{\frac{n-1}{2}} \int_0^T (1+T-s)^{-\frac{n-1}{2}} (1+s)^{-\frac{n-1}{2}(\gamma-2)} ds.$$

Now we have

$$\int_0^t (1+t-\tau)^{-\frac{n-1}{2}} \cdot (1+\tau)^{-\frac{n-1}{2}(\gamma-2)} d\tau \leq C(1+t)^{-\frac{n-1}{2}}$$

provided

$$\gamma > 2 + \frac{2}{n-1}$$

so that, for such values of γ , we obtain

$$M(T) \lesssim C_0 + M(T)^\gamma.$$

For small initial data, which means small values of the constant C_0 , (30) implies that $M(T)$ is uniformly bounded as long as the smooth solution exists. By a simple continuation argument, this implies that the solution is in fact global, provided the initial data are small enough.

The Nonlinear Schrödinger Equation

Strichartz estimates for the Schrödinger equation are expressed entirely in terms of L^p norms, without loss of derivatives, in contrast with the case of the wave equations. For this reason they represent a perfect tool to handle semilinear perturbations. Using Strichartz estimates, several important results can be proved with minimal effort. As an example, we chose to illustrate in this Section the critical and subcritical well posedness in L^2 in the case of power nonlinearities (see [13] for a systematic treatment).

Consider the Cauchy problem

$$iu_t - \Delta u = |u|^\gamma, \quad u(0, x) = f(x) \quad (31)$$

for some $\gamma \geq 1$. We want to study the local and global solvability in L^2 of this Cauchy problem. Recall that this is a difficult question, still not completely solved; however, at least in the subcritical range, a combination of dispersive techniques and the contraction mapping principle gives the desired result easily.

Instead of Eq. (31), we are actually going to solve the corresponding integral equation

$$u(t, x) = e^{it\Delta} f + i \int_0^t e^{i(t-s)\Delta} |u(s)|^\gamma ds.$$

Our standard approach is to look for a fixed point of the mapping $u = \Phi(v)$ defined by

$$\Phi(v) = e^{it\Delta} f + i \int_0^t e^{i(t-s)\Delta} |v(s)|^\gamma ds;$$

it is easy to check, using Strichartz estimates, that the mapping Φ is well defined, provided the function v is in the suitable $L^p L^q$. Indeed we can write, for all admissible couples (p, q) and (\tilde{p}, \tilde{q})

$$\|\Phi(v)\|_{L^p L^q} \lesssim \|f\|_{L^2} + \|\Phi(v)|^\gamma\|_{L^{\tilde{p}'} L^{\tilde{q}'}}$$

and by Hölder's inequality this implies

$$\|\Phi(v)\|_{L^p L^q} \leq C_1 \|f\|_{L^2} + C_1 \|v\|_{L^{\gamma\tilde{p}'} L^{\gamma\tilde{q}'}}^\gamma. \quad (32)$$

Thus, we see that the mapping Φ operates on the space $L^p L^q$, provided we can satisfy the following set of conditions:

$$\begin{aligned} \tilde{p}'\gamma &= p, & \tilde{q}'\gamma &= q, \\ \frac{2}{p} + \frac{n}{q} &= \frac{n}{2}, & \frac{2}{\tilde{p}} + \frac{n}{\tilde{q}} &= \frac{n}{2}, \\ p, \tilde{p} &\in [2, \infty], & \tilde{q} &\in \left[2, \frac{2n}{n-2}\right]. \end{aligned} \quad (30)$$

Notice that from the first two identities we have

$$\frac{2\gamma}{p} + \frac{n\gamma}{q} = \alpha - \frac{2}{\tilde{p}} + n - \frac{n}{\tilde{q}}$$

and by the admissibility conditions this implies

$$\frac{n}{2}\gamma = 2 + n - \frac{n}{2},$$

which forces the value of γ to be

$$\gamma = 1 + \frac{4}{n}. \quad (33)$$

This value of γ is called L^2 critical.

Let us assume for the moment that γ is exactly the critical exponent (33). Then it is easy to check that there exist choices of (p, q) and (\tilde{p}, \tilde{q}) which satisfy all the above constraints. To prove that Φ is a contraction on the space

$$X = L^p L^q$$

we take any $v_1, v_2 \in X$ and compute

$$\begin{aligned} \|\Phi(v_1) - \Phi(v_2)\|_X &\lesssim \| |v_1|^\gamma - |v_2|^\gamma \|_{L^{\tilde{p}'} L^{\tilde{q}'}} \\ &\lesssim \| |v_1 - v_2| (|v_1|^{\gamma-1} + |v_2|^{\gamma-1}) \|_{L^{\tilde{p}'} L^{\tilde{q}'}}; \end{aligned}$$

the same computation as above gives

$$\|\Phi(v_1) - \Phi(v_2)\|_X \leq C_2 \|v_1 - v_2\|_X \| |v_1| + |v_2| \|_X^{\gamma-1} \quad (34)$$

for some constant C_2 . Now assume that

$$\|v_i\|_X < \varepsilon, \quad \|f\|_{L^2} < \delta;$$

using (32) we can write

$$\|\Phi(v)\|_X \leq C_1(\delta + \varepsilon^\gamma) < \varepsilon$$

provided ε, δ are so small that

$$C_1 \delta < \frac{\varepsilon}{2}, \quad C_1 \varepsilon^\gamma < \frac{\varepsilon}{2}.$$

On the other hand, by (34) we get

$$\|\Phi(v_1) - \Phi(v_2)\|_X \leq C_2 \|v_1 - v_2\|_X (2\varepsilon)^{\gamma-1} < \frac{1}{2} \|v_1 - v_2\|_X$$

provided ε is so small that

$$C_2 (2\varepsilon)^{\gamma-1} < \frac{1}{2}.$$

In conclusion, if the initial data are small enough, the mapping Φ is a contraction on a ball $B(0, \varepsilon) \subset X$ and the unique fixed point is a global small solution of the critical L^2 equation

$$iu_t - \Delta u = |u|^{1+\frac{4}{n}}.$$

Notice that in estimate (32) the first couple can also be taken equal to $(\infty, 2)$, thus the solution we have constructed belongs to $L^\infty L^2$. A limit argument (dominated convergence) in the identity

$$u = e^{it\Delta} f + i \int_0^t e^{i(t-s)\Delta} |u(s)|^\gamma ds;$$

proves that we have actually $u \in C(\mathbb{R}; L^2(\mathbb{R}^n))$.

We briefly sketch a proof of the local existence in the subcritical case $1 < \gamma < 1 + \frac{4}{n}$. Consider the local space

$$X_T = L_T^p L^q = L^p(0, T; L^q(\mathbb{R}^n)).$$

If we follow the steps of the preceding proof and choose the same indices, we have now

$$\tilde{p}'\gamma < p, \quad \tilde{q}'\gamma < q.$$

It is easy to check that we can shift the point (\tilde{p}, \tilde{q}) so that

$$\tilde{p}'\gamma < p, \quad \tilde{q}'\gamma = q$$

and proceeding as above and using Hölder inequality in time, we obtain the inequality. Indeed, choosing the indices as above and applying Hölder's inequality in time we have

$$\|\Phi(v)\|_{X_T} \leq C_1 \|f\|_{L^2} + C_1 T^\lambda \|v\|_{X_T} \quad (35)$$

for some strictly positive $\lambda > 0$. Analogously, estimate (34) will be replaced by

$$\begin{aligned} \|\Phi(v_1) - \Phi(v_2)\|_{X_T} \\ \leq C_2 T^\lambda \|v_1 - v_2\|_{X_T} \| |v_1| + |v_2| \|_{X_T}^{\gamma-1}. \end{aligned} \quad (36)$$

Now let M be so large and T so small that

$$\|f\|_{L^2} < \frac{M}{2C_1}, \quad C_1 T^\lambda < \frac{M}{2}$$

we obtain from (35) that Φ takes the ball $B(0, M) \subset X_T$ into itself; if in addition we have also, by possibly taking T smaller,

$$C_2 T^\lambda (2M)^{\gamma-1} < \frac{1}{2}$$

we see by (36) that Φ is a contraction, and in conclusion we obtain the existence of a unique local solution to the subcritical equation. By the same argument as above this solution is in $C([0, T]; L^2(\mathbb{R}^n))$.

We notice that in the above proof the quantity M is only determined by the L^2 norm of f , and hence also the lifespan of the local solution is a function of $\|f\|_{L^2}$. If for some reason we know a priori that the L^2 norm of the solution is conserved by the evolution, this remark implies immediately that the solution is in fact global, by an elementary continuation argument. This is the case for instance of the *gauge invariant* equation

$$iu_t - \Delta u = \pm |u|^{\gamma-1} u$$

for which any solution continuous with values in L^2 satisfies $\|u(t)\|_{L^2} = \|u(0)\|_{L^2}$. The above computations apply without modification and we obtain the global well posedness for all subcritical exponents γ .

Future Directions

The dispersive properties of constant coefficient equations are now fairly well understood. In view of the wide range of applications, it would be very useful to extend the techniques to more general classes of equations, in particular for equations with variable coefficients and on manifolds. These problems have been the subject of intense interest in recent years and are actively pursued.

Strichartz estimates have been extended to very general situations, including the fully variable coefficients case, under suitable smoothness, decay and spectral assumptions on the coefficients (see e. g. [36,37] and, for singular coefficients, [18]).

The results concerning *dispersive estimates* are much less complete. They have been proved for perturbations of order zero

$$iu_t - \Delta u + V(x)u = 0, \quad u_{tt} - \Delta u + V(x)u = 0$$

with a potential $V(x)$ sufficiently smooth and decaying at infinity (see e. g. [2,19,28,53,54,55]) while for perturbation of order one

$$iu_t - \Delta u + a(x) \cdot \nabla u_{x_j} + V(x)u = 0,$$

$$u_{tt} - \Delta u + a(x) \cdot \nabla u_{x_j} + V(x)u = 0$$

very few results are available, (see [16] for the three-dimensional wave and Dirac equations, and [1,17] for the one-dimensional case where quite general results are available).

A closely connected problem is the study of the dispersive properties of equations on manifolds. This part of the theory is advancing rapidly and will probably be a very interesting direction of research in the next years (see e. g. [12] and related papers for the case of compact manifolds, and [5,24] for noncompact manifolds).

Bibliography

- Artbazar G, Yajima K (2000) The L^p -continuity of wave operators for one dimensional Schrödinger operators. *J Math Sci Univ Tokyo* 7(2):221–240
- Beals M (1994) Optimal L^∞ decay for solutions to the wave equation with a potential. *Comm Partial Diff Eq* 19(7/8):1319–1369
- Ben-Artzi M, Trèves F (1994) Uniform estimates for a class of evolution equations. *J Funct Anal* 120(2):264–299
- Bergh J, Löfström J (1976) Interpolation spaces. An introduction. *Grundlehren der mathematischen Wissenschaften*, No 223. Springer, Berlin
- Bouclet J-M, Tzvetkov N (2006) On global strichartz estimates for non trapping metrics.
- Bourgain J (1993) Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. I. Schrödinger equations. *Geom Funct Anal* 3(2):107–156
- Bourgain J (1995) Some new estimates on oscillatory integrals. In: (1991) *Essays on Fourier analysis in honor of Elias M. Stein*. Princeton Math Ser vol. 42. Princeton Univ Press, Princeton, pp 83–112
- Bourgain J (1995) Estimates for cone multipliers. In: *Geometric aspects of functional analysis*. Oper Theory Adv Appl, vol 77. Birkhäuser, Basel, pp 41–60
- Bourgain J (1998) Refinements of Strichartz' inequality and applications to 2D-NLS with critical nonlinearity. *Int Math Res Not* 5(5):253–283
- Brenner P (1975) On $L_p - L_{p'}$ estimates for the wave-equation. *Math Z* 145(3):251–254
- Brenner P (1977) $L_p - L_{p'}$ -estimates for Fourier integral operators related to hyperbolic equations. *Math Z* 152(3):273–286
- Burq N, Gérard P, Tzvetkov N (2003) The Cauchy problem for the nonlinear Schrödinger equation on a compact manifold. *J Nonlinear Math Phys* 10(1):12–27
- Cazenave T (2003) Semilinear Schrödinger equations. *Courant Lecture Notes in Mathematics*, vol 10. New York University Courant Institute of Mathematical Sciences, New York
- Choquet-Bruhat Y (1950) Théorème d'existence pour les équations de la gravitation einsteinienne dans le cas non analytique. *CR Acad Sci Paris* 230:618–620
- Christodoulou D, Klainerman S (1989) The nonlinear stability of the Minkowski metric in general relativity. In: *Bordeaux (1988) Nonlinear hyperbolic problems*. Lecture Notes in Math, vol 1402. Springer, Berlin, pp 128–145
- D'Ancona P, Fanelli L (2006) Decay estimates for the wave and dirac equations with a magnetic potential. *Comm Pure Appl Anal* 29:309–323
- D'Ancona P, Fanelli L (2006) L^p - boundedness of the wave operator for the one dimensional Schrödinger operator. *Comm Math Phys* 268:415–438
- D'Ancona P, Fanelli L (2008) Strichartz and smoothing estimates for dispersive equations with magnetic potentials. *Comm Partial Diff Eq* 33(6):1082–1112
- D'Ancona P, Pierfelice V (2005) On the wave equation with a large rough potential. *J Funct Anal* 227(1):30–77
- D'Ancona P, Georgiev V, Kubo H (2001) Weighted decay estimates for the wave equation. *J Diff Eq* 177(1):146–208
- Foschi D (2005) Inhomogeneous Strichartz estimates. *J Hyperbolic Diff Eq* 2(1):1–24
- Ginibre J, Velo G (1985) The global Cauchy problem for the nonlinear Schrödinger equation revisited. *Ann Inst Poincaré H Anal Non Linéaire* 2(4):309–327
- Ginibre J, Velo G (1995) Generalized Strichartz inequalities for the wave equation. *J Funct Anal* 133(1):50–68
- Hassell A, Tao T, Wunsch J (2006) Sharp Strichartz estimates on nontrapping asymptotically conic manifolds. *Am J Math* 128(4):963–1024
- Hörmander L (1997) Lectures on nonlinear hyperbolic differential equations. *Mathématiques & Applications (Mathematics & Applications)*, vol 26. Springer, Berlin
- John F (1979) Blow-up of solutions of nonlinear wave equations in three space dimensions. *Manuscr Math* 28(1–3):235–268
- John F, Klainerman S (1984) Almost global existence to nonlinear wave equations in three space dimensions. *Comm Pure Appl Math* 37(4):443–455
- Journé J-L, Soffer A, Sogge CD (1991) Decay estimates for Schrödinger operators. *Comm Pure Appl Math* 44(5):573–604

29. Kato T (1965/1966) Wave operators and similarity for some non-selfadjoint operators. *Math Ann* 162:258–279
30. Keel M, Tao T (1998) Endpoint Strichartz estimates. *Am J Math* 120(5):955–980
31. Klainerman S (1980) Global existence for nonlinear wave equations. *Comm Pure Appl Math* 33(1):43–101
32. Klainerman S (1981) Classical solutions to nonlinear wave equations and nonlinear scattering. In: *Trends in applications of pure mathematics to mechanics*, vol III. Monographs Stud Math, vol 11. Pitman, Boston, pp 155–162
33. Klainerman S (1982) Long-time behavior of solutions to nonlinear evolution equations. *Arch Rat Mech Anal* 78(1):73–98
34. Klainerman S (1985) Long time behaviour of solutions to nonlinear wave equations. In: *Nonlinear variational problems*. Res Notes in Math, vol 127. Pitman, Boston, pp 65–72
35. Klainerman S, Nicolò F (1999) On local and global aspects of the Cauchy problem in general relativity. *Class Quantum Gravity* 16(8):R73–R157
36. Marzuola J, Metcalfe J, Tataru D. Strichartz estimates and local smoothing estimates for asymptotically flat Schrödinger equations. To appear on *J Funct Anal*
37. Metcalfe J, Tataru D. Global parametres and dispersive estimates for variable coefficient wave equation. Preprint
38. Pecher H (1974) Die Existenz regulärer Lösungen für Cauchy- und Anfangs-Randwert-probleme nichtlinearer Wellengleichungen. *Math Z* 140:263–279 (in German)
39. Reed M, Simon B (1978) Methods of modern mathematical physics IV. In: *Analysis of operators*. Academic Press (Harcourt Brace Jovanovich), New York
40. Segal I (1968) Dispersion for non-linear relativistic equations. II. *Ann Sci École Norm Sup* 4(1):459–497
41. Shatah J (1982) Global existence of small solutions to nonlinear evolution equations. *J Diff Eq* 46(3):409–425
42. Shatah J (1985) Normal forms and quadratic nonlinear Klein–Gordon equations. *Comm Pure Appl Math* 38(5):685–696
43. Shatah J, Struwe M (1998) Geometric wave equations. In: *Courant Lecture Notes in Mathematics*, vol 2. New York University Courant Institute of Mathematical Sciences, New York
44. Strichartz RS (1970) Convolutions with kernels having singularities on a sphere. *Trans Am Math Soc* 148:461–471
45. Strichartz RS (1970) A priori estimates for the wave equation and some applications. *J Funct Anal* 5:218–235
46. Strichartz RS (1977) Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations. *J Duke Math* 44(3):705–714
47. Tao T (2000) Spherically averaged endpoint Strichartz estimates for the two-dimensional Schrödinger equation. *Comm Part Diff Eq* 25(7–8):1471–1485
48. Tao T (2003) Local well-posedness of the Yang–Mills equation in the temporal gauge below the energy norm. *J Diff Eq* 189(2):366–382
49. Taylor ME (1991) Pseudodifferential operators and nonlinear PDE. *Progr Math* 100. Birkhäuser, Boston
50. von Wahl W (1970) Über die klassische Lösbarkeit des Cauchy-Problems für nichtlineare Wellengleichungen bei kleinen Anfangswerten und das asymptotische Verhalten der Lösungen. *Math Z* 114:281–299 (in German)
51. Wolff T (2001) A sharp bilinear cone restriction estimate. *Ann Math* (2) 153(3):661–698
52. Yajima K (1987) Existence of solutions for Schrödinger evolution equations. *Comm Math Phys* 110(3):415–426
53. Yajima K (1995) The $W^{k,p}$ -continuity of wave operators for Schrödinger operators. *J Math Soc Japan* 47(3):551–581
54. Yajima K (1995) The $W^{k,p}$ -continuity of wave operators for Schrödinger operators. III. even-dimensional cases $m \geq 4$. *J Math Sci Univ Tokyo* 2(2):311–346
55. Yajima K (1999) L^p -boundedness of wave operators for two-dimensional Schrödinger operators. *Comm Math Phys* 208(1):125–152

Distributed Controls of Multiple Robotic Systems, An Optimization Approach

JOHN T. FEDDEMA, DAVID A. SCHOENWALD,
RUSH D. ROBINETT, RAYMOND H. BYRNE
Sandia National Laboratories, Albuquerque, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Cooperative Control
Communication Effects
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Multiple robotic system Multiple robotic systems are a collective of multiple robotic vehicles (e.g. ground-based, aerial, underwater, and/or some combination thereof) wherein each robot has assigned tasks designed to help the entire system achieve its overall goals.

Provably convergent cooperative controls Provably convergent cooperative controls is a distribution of control laws among the agents of a multiple robotic system with the objective that global stability is preserved and a global performance index is optimized. The control laws take into account the physics of the individual robots, the communication protocols and information exchange between the robots, the overall terrain, and the goals/constraints of the entire collective.

Connective stability A system is connectively stable if it is stable in the sense of Lyapunov for all structural perturbations of the interconnection matrix that describes the system (see [63]). These structural perturbations

include attenuation, degradation, and loss of communication between nodes in the network.

Vehicle communication protocols In a multiple robotic system, vehicle communication protocols refer to the type of communication algorithm used by the vehicles to share information amongst each other. These protocols are crucial in determining stability and performance properties.

Time division multiple access (TDMA) TDMA is a [channel access method](#) for shared medium (usually [radio](#)) networks. It allows several users to share the same [frequency channel](#) by dividing the signal into timeslots. The users transmit in rapid succession, one after the other, each using their own timeslot. This allows multiple stations (e.g. robots) to share the same transmission medium (e.g. radio frequency channel) while using only part of its [bandwidth](#).

Carrier sense multiple access (CSMA) CSMA is also a channel access method for shared medium networks where the station first listens to the channel for a predetermined amount of time so as to check for any activity on the channel. If the channel is sensed “idle” then the station is permitted to transmit. If the channel is sensed as “busy” the station defers its transmission. This access method is used in wired TCP/IP networks.

Linear TDMA broadcast The number of time slots in the TDMA network is equal to the total number of nodes in the network.

Polylogarithmic TDMA broadcast The number of time slots in the TDMA network is less than the total number of nodes because of spatial reuse. This method only applies as long as the degree of the network (the number of nodes within communication range) is small (see [66]).

Definition of the Subject

This chapter describes an integrated approach for designing communication, sensing, and control systems for mobile distributed systems. A three-step process is used to create locally optimal distributed controls for multiple robot vehicles. The first step is to define a global performance index whose extremum produces a desired cooperative result. The second step is to partition and eliminate terms in the performance index so that only terms of local neighbors are included. This step minimizes communication amongst robots and improves system robustness. The third step is to define a control law that is the gradient of the partitioned performance index. This control drives the system towards the local extremum of the

partitioned performance index. Graph theoretic methods are then used to analyze the input/output reachability and structural controllability and observability of the decentralized system. Connective stability of the resulting controls is evaluated with a vector Lyapunov technique. From this analysis, the trade-off between vehicle responsiveness and the maximum allowable communication delay is theoretically determined. The approach has been tested in simulation and robotics hardware on a number of test problems. Six examples are given. The first problem is the distribution of multiple robots along a line or curve. The second and third problems are the distribution of multiple robots in an unconstrained and constrained two-dimensional plane. The fourth problem is the distribution of multiple robots around an elliptical curve. The fifth problem is for multiple robots to converge on the source of a plume in a plane. The sixth problem is for multiple underwater robots to converge on the source of a plume in a 3D volume. Four different communication protocols are compared for minimum communication delay: a Time Division Multiple Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Multiple Access (CSMA) algorithm.

Introduction

Over the past two decades, considerable research has concentrated on the development of cooperative controls of multiple robotic systems [1]. Much of this work is originally inspired by the observation that social insects such as termites, ants, and bees perform amazing group behaviors when individually executing simple rules [2]. The hypothesis being that it must be possible to create useful “emergent” group behaviors from simple individual behaviors. While this hypothesis has yet to be proven or disproved, many people have demonstrated some interesting group behaviors. For example, researchers have shown that groups of robots can forage for objects [3], maintain formations [4], and push boxes [5].

As the field of cooperative robotics has matured, more complex cooperative behaviors have been developed to clean up hazardous waste [6], explore and map unknown environments [7,8], uniformly cover a specified area with a mobile sensor network [9], track multiple targets [10], and perform cooperative pursuit-evasion maneuvers [11,12,13,14]. Algorithms for resource allocation of heterogeneous agents has been developed [6,15], as well as distributed fault detection methods for alerting operators when individual robots are not moving as expected [16]. Measures of effectiveness for human-robot

teams have been defined [17] to evaluate the mission performance of the cooperative algorithms.

In an effort to create cooperative behaviors that meet a performance specification, researchers have begun to take a system controls perspective and analyze the stability of multiple vehicles when driving in formations. Chen and Luh [18] examined decentralized control laws that drove a set of holonomic mobile robots into a circular formation. A conservative stability requirement for the sample period is given in terms of the damping ratio and the undamped natural frequency of the system. Similarly, Yamaguchi studied line-formations [19] and general formations [20] of nonholonomic vehicles, as did Yoshida et al. [21]. Decentralized control laws using a potential field approach to guide vehicles away from obstacles can be found in [22,23]. In these studies, only continuous time analyses have been performed, assuming that the relative position between vehicles and obstacles can be measured at all time.

In the area of line formations, Stankovic [24] considered the problem of controlling a one dimensional platoon of N identical vehicles moving at a constant velocity and a fixed distance apart. The measured distance and relative velocity of each neighboring vehicle is used to regulate the formation. Seiler [25] analyzed the “string instability” problem where a disturbance in the inter-vehicular spacing is amplified as nearest neighbor feedback propagates through the leader-follower platoon. Barooah [26] showed the system performance can be improved by “mistuning” the individual controller gains.

A variety of control techniques have been used to control general formations. Orger [27] shows that multiagent coordination is possible if a vector Lyapunov function exists for the formation that is a weighted sum of Lyapunov functions for each individual robot. This same approach is used in this chapter to guarantee convergence and stability. In [28], complex maneuvers are decomposed into a sequence of maneuvers between formation patterns. The formation patterns are arranged in a bidirectional ring topology and their inter-robot distances are controlled with via passivity techniques. In [29], leader-to-formation stability gains control error amplification, relate interconnection topology to stability and performance, and offer safety bounds for different formation topologies. In [30], algebraic conditions are developed that guarantee formation feasibility given individual agent kinematics. Given the kinematics of several agents along with inter-agent constraints, [30] determine whether there exist nontrivial agent trajectories that maintain the constraints. In [31], task functions describe the desired kinematic relationship between vehicle positions and suitable variables that syn-

thetically express some feature of the platoon shape. Multiple task functions are merged using a singularity-robust task-priority inverse kinematics algorithm. The task function approach is similar to the optimization approach followed in this chapter.

Another way of analyzing stability is to investigate the convergence of a distributed algorithm. Beni and Liang [32] prove the convergence of a linear swarm of asynchronous distributed autonomous agents into a synchronously achievable configuration. The linear swarm is modeled as a set of linear equations that are solved iteratively. Their formulation is best applied to resource allocation problems that can be described by linear equations. Liu et al. [33] provide conditions for convergence of an asynchronous swarm in which swarm “cohesiveness” is the stability property under study. Their paper assumes position information is passed between nearest neighbors only and proximity sensors prevent collisions.

Also of importance is the recent research combining graph theory with decentralized controls. Most cooperative mobile robot vehicles have wireless communication, and simulations have shown that a wireless network of mobile robots can be modeled as an undirected graph [34]. These same graphs can be used to control a formation. Desai et al. [35,36] used directed graph theory to control a team of robots navigating terrain with obstacles while maintaining a desired formation and changing formations when needed. When changing formations, the transition matrix between the current adjacency matrix and all possible control graphs are evaluated.

Graph theory can also be used to maintain connectedness during maneuvers. Connectedness is the problem of ensuring that a group of mobile agents stays connected in a communication sense while achieving some performance objective. Ji [37] use a weighted graph Laplacian to generate nonlinear feedback laws that guarantee connectedness during rendezvous and formation-control maneuvers. In [38], the connectivity condition is translated to differentiable constraints on individual agent motion by considering the dynamics of the Laplacian matrix and its spectral properties. Artificial potential fields are used to drive the agents to configurations away from the undesired space of disconnected networks while avoiding collisions with each other.

The ability to scale up to very large sized systems has also been recently investigated. Belta [39] addressed the general problem of controlling a large group of robots required to move as a group. The group is abstracted to a lower dimensional manifold whose dimension is independent of the number of robots. The manifold captures the position and orientation of the ensemble and the shape

spanned by the robots. Decoupled controllers are designed for the group and shape variables, and individual controllers for each robot are designed to keep them within the manifold. In [40], the scalability is improved by representing formation structures by queues, which are smooth line segments consisting of multiple robots. Artificial potential trenches are used to control the formation of the queues. In [41], the complexity of population modeling is avoided by modeling the agent distribution as state probability density functions. The control problem becomes one of maximizing the probability of robotic presence in a given region. In [42], the large dimensional state space is reduced into a small dimensional continuous abstract space which captures essential features of the swarm. High level specifications of the continuous abstraction, given as linear temporal logic formulas over linear predicates, are automatically mapped to provably correct robot control laws.

Lastly, cooperative algorithms for mobile manipulation are also being developed. Tanner et al. [43] presented a motion planning methodology based on non-smooth Lyapunov functions that account for nonpoint articulated robots. Diffeomorphic transformations are used to map the volume of each articulated robot and obstacle into a point world and a dipolar potential field function is used to ensure asymptotic convergence of the robots. Yamashita [44] developed motion planning algorithms for controlling multiple mobile robots cooperatively transporting a large object in a three dimensional environment. The transportation portion of the problem is solved in a reduced configuration space using a potential field path planner. An A* algorithm is used to avoid local minimum caused by the constraints of object manipulation. Bonaventura [45] modeled the dynamics of complex multiple mobile manipulators constrained by concurrent contacts, including time-varying and nonholonomic constraints.

Over the past several years, the authors have been developing cooperative control systems for military missions such as perimeter surveillance [46], facility reconnaissance [47], and chemical plume tracking [48,49,50]. Multiple mobile robots perform these missions in a coordinated fashion using low bandwidth communication between nodes. We have demonstrated that by using Provably Convergent Cooperative Controls (PC^3) we can guarantee a specified level of performance even under uncertainty. PC^3 involves selecting an overall performance index and using distributed optimization techniques to minimize this performance index. The stability of the control can be proven using a vector Lyapunov function. The scheme is inherently fault tolerant to nodal failure, al-

lowing other nodes to complete a task started by a failed node.

Key to the success of PC^3 is that the communication between nodes is used to synchronize the controls. In most of our work, we have used a TDMA protocol with a small number of nodes (up to 10). To expand this PC^3 capability in larger scaled systems of 100 to 1000 nodes, we must understand the underlying communication layer and ensure that the communication delays are bounded. Therefore, an important aspect of this work is an integrated approach to both communication and controls. Given the advances being made on the bandwidth of networked systems, this is an exciting new area of research, and the interested reader may want to peruse [51,52].

PC^3 uses graph theoretic methods to analyze the input/output reachability and structural controllability and observability of a decentralized system. This analysis could be embedded in each node and be used to automatically reconfigure an ad hoc communication network for the control task at hand. The graph analysis could also be used to create the most efficient communication flow control based upon spatial distribution of the network nodes. Edge coloring algorithms in graph theory tell us that the minimum number of time slots in a planar network is equal to the maximum number of adjacent nodes plus some small number. The more spread out the nodes are, the fewer number of time slots are needed for communication, and the smaller the latency between nodes. In a coupled system, smaller latency results in a more responsive control system. In this chapter, network protocols that propagate this information and distributed algorithms that automatically adjust the number of time slots available for communication were evaluated. For real-time implementation, these protocols and algorithms must be extremely efficient and only updated as network nodes move.

The next section provides a common mathematical framework that can be used to develop cooperative behaviors among mobile robot vehicles. Six examples of the application of this approach are given. The proceeding section describes the effects of communication on the sampling period of a cooperative system. The trade-off between communication delay, interaction gain, and vehicle responsiveness is evaluated using a vector Lyapunov technique. Four different communication protocols: a Time Division Multiple Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Multiple Access (CSMA) algorithm are compared. The final section summarizes this work and lists areas of suggested future research.

Cooperative Control

In this section, a common mathematical framework is used to describe several cooperative behaviors. This mathematical framework is applied to three generic behaviors: containment, coverage, and converging. The authors believe that this same framework could be applied to the other behaviors, although this has not been proven at the time of this publication.

This mathematic framework is motivated by the fact that most of the laws in physics and mechanics can be derived by finding a stationary point [53] that may be a maximum or minimum of some performance index, in this case, called the Lagrangian integral [54]. In physics, the Lagrangian integral is a function of energy, and the goal is often to find the action that minimizes the total energy or action integral. It should be noted that the Lagrangian of each phenomenon is a function of some gradient term squared. In the analysis that follows, the reader will notice that each behavior's performance index is also a function of some gradient term squared.

Following this same optimization approach, a three-step process for developing cooperative control algorithms has been developed [55]. These three steps are as follows:

1. Define a global performance index as a function of parameters from all entities.
2. Partition and eliminate terms in the performance index so that only terms of local neighbors are included.
3. The local control law is the gradient (or Newton's method) of the partitioned performance index.

The first step requires that one understands the problem well enough that it can be posed as a global optimization problem. This step can be relatively difficult, but as the examples in the remainder of this section will show, with the right simplifying assumptions, rather simple equations can be used to generate complex cooperative behaviors.

The second step, partitioning the performance index, is often used in parallel optimization to reduce the computation time for large-scale problems [56]. In this case, the second step is used to reduce communications between robots and to increase robustness of the distributed system. The control law that would result from step 1 would require that every robot be able to communicate with all the other robots. As the number of robots increase to 100 s and 1000 s, the time delay necessary for communication would make the resulting control infeasible. Instead, partitioning the performance index and eliminating terms that include interaction between distant neighbors results in a control law that only requires communication with nearest neighbors, thus greatly reducing communication de-

lay. Also, using nearest neighbors that change throughout the motion adds an element of robustness. The mathematical formulation of the partition does not specify that robot number 10 must communicate with robot number 6. Instead, the mathematical formulation specifies a group of nearest neighbors that can change based on external forces and environmental conditions. This creates an element of self-organization that allows the system to change and evolve. If a robot fails, a new set of nearest neighbors is formed.

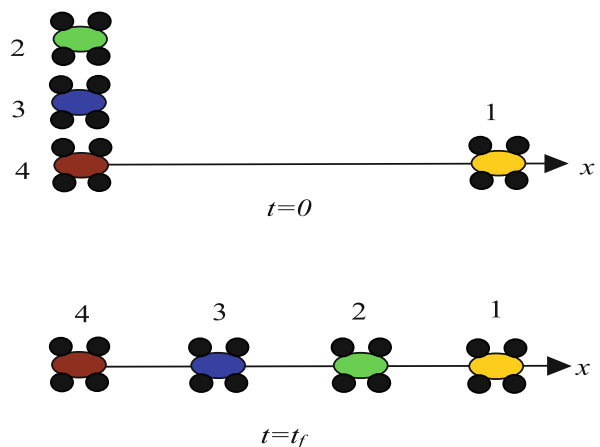
The third step is to solve for the extremum of the partitioned performance index using either a first-order steepest descent or second order method such as the Newton's Method [57].

The remainder of this section shows how these three steps have been used in practice. Six examples are given with details on the problem formulation and the task that was performed.

Example 1: Spreading Apart Along a Line – A Containment Behavior

This first example is a simple one-dimensional problem. The goal is for multiple robots to evenly spread apart along a straight line using only information from the neighboring robots on the right and left. In Fig. 1, the first and last robots are assumed to be stationary while the ones in between are to spread apart a distance d away from each other.

The optimization steps are as follows.



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 1

One-dimensional control problem. The *top line* is the initial state. The *second line* is the desired final state. Vehicles 1 and 4 are boundary conditions. Vehicles 2 and 3 spread out along the line using only the position of their left and right neighbors

Step 1. Specified as an optimization problem, the objective is to

$$\min_{\bar{x}} v(\bar{x}) \quad (1)$$

where the global performance index is

$$v(\bar{x}) = \frac{1}{2} \sum_{i=1}^{N-1} (d - |x_{i+1} - x_i|)^2, \quad (2)$$

x_i is the position of robot i , $\bar{x} = [x_1 \dots x_N]^T$ are the positions of all the robots, and d is the desired distance between each robot. The goal is to minimize the sum of squared errors in distances between every robot.

Step 2. This problem is easily partitioned amongst the interior $N - 2$ robots. The distributed objective is to

$$\min_{x_i} v_i(\bar{x}) \quad \forall i = 2, \dots, N - 1 \quad (3)$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2} (d - |x_i - x_{i-1}|)^2 + \frac{1}{2} (d - |x_i - x_{i+1}|)^2. \quad (4)$$

Because of the additive form of Eq. (2), simultaneously solving Eq. (3) for each robot is the same as minimizing the global performance index in Eq. (2). Therefore, in this case, no terms were eliminated. This is not necessarily true for the other example problems below.

Step 3. A steepest descent control law for the partitioned performance index is given by

$$x_i(k+1) = x_i(k) - \alpha \nabla v_i(\bar{x}(k)), \quad 0 < \alpha \leq 1, \quad (5)$$

where

$$\nabla v_i(\bar{x}) = 2x_i - (x_{i+1} + x_{i-1}) \quad \text{if } x_{i-1} < x_i < x_{i+1}. \quad (6)$$

Note that $\nabla v_i(\bar{x}) = 0$ when $x_i = \frac{1}{2}(x_{i+1} + x_{i-1})$. Therefore, the vehicles will disperse along the line until they have reached a position that is exactly in the middle of its nearest neighbors. In [47], it is shown that α is actually more constrained than $0 < \alpha \leq 1$ depending on the responsiveness of the vehicle and the communication sample period.

The control law in Eqs. (5)–(6) have been used to spread robot vehicles apart along a perimeter [46] as shown in Fig. 2, as well as to spread out hopping minefield robots [58] as shown in Fig. 3.

Example 2: Coverage of a Two-Dimensional Space

Next, we consider the example of dispersing robots in a plane in a specified pattern. In Fig. 4, the robots are to move from the configuration on the left to the configuration on the right. The configuration on the right is specified by the distances d_{ij} between robot i and robot j .

Step 1. The objective is to

$$\min_{\bar{x}} v(\bar{x}) \quad (7)$$

where the global performance index is

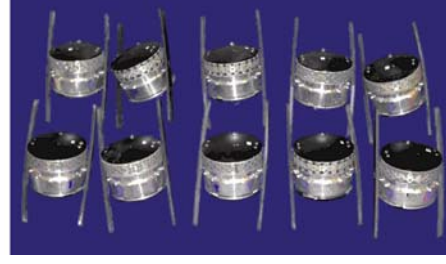
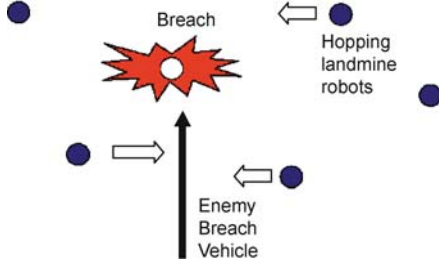
$$v(\bar{x}) = \frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right)^2, \quad (8)$$

$$\bar{x}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \in \mathbb{R}^2$$



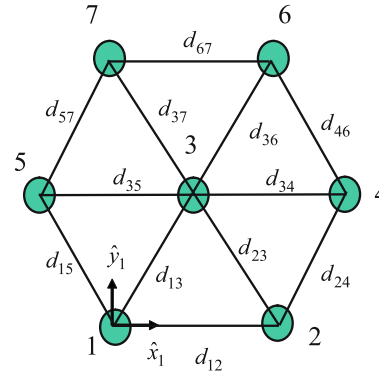
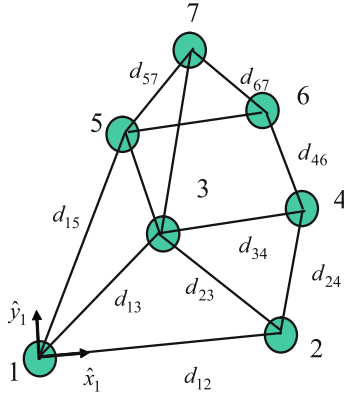
Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 2

Left: Four robot vehicles are shown guarding a perimeter denoted by the blue line segments. When an intrusion detection sensor denoted by the numbered circles alarms, one robot vehicle attends to the alarm (vehicle near sensor 33) while the others spread apart along the perimeter so that each vehicle is midway between its neighbors. **Right:** Robot vehicles used to perform perimeter surveillance task



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 3

Left: Hopping landmine robots are filling breach left by enemy vehicle. When a robot is breached, the robots will hop towards the missing robot and settle when each robot is midway between its neighbors. **Right:** Hopping landmine robots used in self-healing minefield tests



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 4

Left: Initial configuration of robots. **Right:** Desired final configuration

is the position of robot i in the xy plane, and $\bar{x} = [\bar{x}_1^T \dots \bar{x}_N^T]^T$ is the position of all the robots in the xy plane. By minimizing the error between the squared desired distance and the squared measured distance between every pair-wise combination of robots, we can drive the robots from an initial pattern to the desired specified pattern. Notice that the global performance index does not specify the orientation or final absolute position of the group of robots.

Step 2. The global performance index is over constrained since it is possible to achieve the same minimum solution without having to minimize the error between every pair-wise combination. The same minimum solution can be achieved by only minimizing the error between neighboring robots. The distributed objective is to

$$\min_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \dots, N \quad (9)$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2} \sum_{j \in \text{NN}} \left(d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right)^2 \quad (10)$$

and NN stands for nearest neighbor.

Step 3. The steepest descent control law for the partitioned performance index is given by

$$\bar{x}_i(k+1) = \bar{x}_i(k) - \alpha \nabla v_i(\bar{x}(k)), \quad 0 < \alpha \leq 1 \quad (11)$$

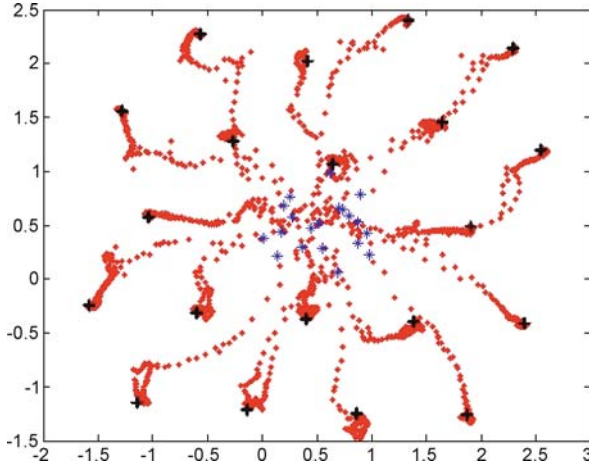
where

$$\nabla v_i = \begin{bmatrix} \frac{\partial v_i}{\partial x_i} \\ \frac{\partial v_i}{\partial y_i} \end{bmatrix} \in \mathbb{R}^2, \quad (12)$$

$$\begin{aligned} \frac{\partial v_i(\bar{x})}{\partial x_i} &= -2 \sum_{j \in \text{NN}} \left[d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right] (x_i - x_j), \\ & \quad (13) \end{aligned}$$

$$\begin{aligned} \frac{\partial v_i(\bar{x})}{\partial y_i} &= -2 \sum_{j \in \text{NN}} \left[d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right] (y_i - y_j). \\ & \quad (14) \end{aligned}$$

Note that $\nabla v_i = 0$ when $d_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2$



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 5

Plot of 20 vehicles' trajectories started from a clustered position with the goal of spreading out uniformly through the space (blue * indicates initial position, red marks indicate trajectory, and black + indicate final position)

for $j \in \text{NN}$. In [59], the connective stability of this control law is proven using a vector Lyapunov technique. The control law in Eqs. (11)–(14) has been used to spread apart the hopping minefield robots as shown in Fig. 5. In this case, the specified distances d_{ij} are all equal and the number of nearest neighbors used for control is three.

Example 3: Coverage of a Two-Dimensional Space with Constraints

Next, consider the same problem as in the previous example, except that the robots are constrained to stay within

a region that is bounded by line segments as shown in Fig. 6.

Step 1. The objective is to

$$\min_{\bar{x}} v(\bar{x}) \quad (15)$$

where the global performance index is

$$v(\bar{x}) = \frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(d_{ij}^2 - |\bar{x}_i - \bar{x}_j|^2 \right)^2 \quad (16)$$

subject to

$$A\bar{x}_i \leq b \quad \forall i = 1, \dots, N \quad (17)$$

where $A \in \mathbb{R}^{m \times 2}$ and $b \in \mathbb{R}^m$. Equation (17) specifies the boundary conditions of m straight-line segments.

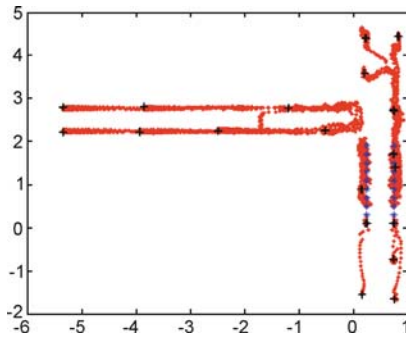
Step 2. The distributed objective is to

$$\min_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \dots, N \quad (18)$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2} \sum_{j \in \text{NN}} \left(d^2 - |\bar{x}_i - \bar{x}_j|^2 \right)^2 + \frac{1}{2} \Lambda \sum_{l \in \text{NO}} (A_l \bar{x}_i - b_l)^{-2}. \quad (19)$$

Here, the inequality constraints in Eq. (17) have been added as a weighted penalty function that is the sum of the inverse squared perpendicular distances between robot i and the nearest obstacle (NO) line segments l . The Λ is a scalar used to vary the importance of obstacle avoidance. As before, NN stands for the set of nearest neighbors. Similarly, the set NO is the set of nearest obstacles.



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 6

Left: Plot of 20 vehicles' trajectories started from a clustered position with the goal of spreading apart uniformly through a hallway with a side corridor (blue * indicates initial position, red marks indicate trajectory, and black + indicate final position). Right: Robot vehicles used in an indoor communication/navigation network

Step 3. The steepest descent control law is

$$\tilde{x}_i(k+1) = \tilde{x}_i(k) - \alpha \nabla v_i(\tilde{x}(k)), \quad 0 < \alpha \leq 1 \quad (20)$$

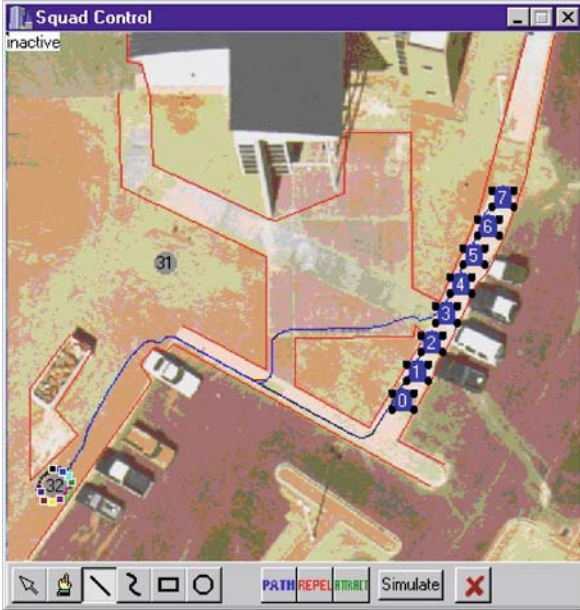
where

$$\begin{aligned} \nabla v_i(\tilde{x}) = & -2 \sum_{j \in \text{NN}} (d^2 - |\tilde{x}_i - \tilde{x}_j|^2)(\tilde{x}_i - \tilde{x}_j) \\ & - \Lambda \sum_{l \in \text{NO}} A_l^T (A_l \tilde{x}_i - b_l)^{-3}. \end{aligned} \quad (21)$$

The control law in Eqs. (20) and (21) has been used to spread out the robot vehicles in a hallway as shown in Fig. 6. The nearest obstacles are determined from IR proximity sensors. The specified distance d between vehicles was chosen to be within the 10 meter acoustic range of the sensors on top of the vehicle (see Fig. 6) [60]. Again, the number of nearest neighbors used for control is three.

Example 4. Forming an Ellipse with Constraints – A Containment Behavior

Next, consider a path following/formation problem where multiple vehicles are to 1) travel towards and spread apart on an ellipse, 2) not drive into each other, and 3) stay away from obstacle line segments. This is shown in Fig. 7.



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 7

Robot path planner drives vehicles towards ellipse while staying away from obstacles, denoted by red line, and the other vehicles

Step 1. The objective is to

$$\min_{\tilde{x}} v(\tilde{x}) \quad (22)$$

where the global performance index is

$$v(\tilde{x}) = \frac{1}{2} \sum_{i=1}^N \left((\tilde{x}_i - \tilde{x}_o)^T \begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} (\tilde{x}_i - \tilde{x}_o) - 1 \right)^2 \quad (23)$$

subject to

$$|\tilde{x}_{i+1} - \tilde{x}_i| > d \quad \forall i = 1, \dots, N-1$$

$$A \tilde{x}_i \leq b \quad \forall i = 1, \dots, N. \quad (24)$$

The global performance index is the squared error of the robot's position from the ellipse. The position of the center of the ellipse is \tilde{x}_o , and ρ and σ are the elliptical parameters along the x and y -axes. The first constraint ensures that the vehicles stay a distance d apart from each other. The second constraint ensures that the vehicles stay away from the line constraints as in the previous example.

Step 2. The distributed objective is

$$\min_{\tilde{x}_i} v_i(\tilde{x}) \quad \forall i = 1, \dots, N \quad (25)$$

where the partitioned performance index is

$$\begin{aligned} v_i(\tilde{x}) = & \frac{1}{2} \left((\tilde{x}_i - \tilde{x}_o)^T \begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} (\tilde{x}_i - \tilde{x}_o) - 1 \right)^2 \\ & + \frac{1}{2} \sum_{j \in \text{NN}} (d^2 - |\tilde{x}_i - \tilde{x}_j|^2)^2 \\ & + \frac{1}{2} \Lambda \sum_{l \in \text{NO}} (A_l \tilde{x}_i - b_l)^{-2}. \end{aligned} \quad (26)$$

The two constraints are implemented as penalty functions. The equations are the same as in the previous example.

Step 3. The steepest descent control law is

$$\tilde{x}_i(k+1) = \tilde{x}_i(k) - \alpha \nabla v_i(\tilde{x}(k)) \quad (27)$$

where

$$\begin{aligned} \nabla v_i(\tilde{x}) = & 2 \left((\tilde{x}_i - \tilde{x}_o)^T \begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} (\tilde{x}_i - \tilde{x}_o) - 1 \right) \\ & \cdot \begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} (\tilde{x}_i - \tilde{x}_o) \\ & - 2 \sum_{j \in \text{NN}} (d^2 - |\tilde{x}_i - \tilde{x}_j|^2) (\tilde{x}_i - \tilde{x}_j) \\ & - \Lambda \sum_{l \in \text{NO}} A_l^T (A_l \tilde{x}_i - b_l)^{-3}. \end{aligned} \quad (28)$$

The control law in Eqs. (27)–(28) has been implemented on a path planner as shown in Fig. 7. The number of nearest neighbors and number of nearest obstacles can be one if the time step is small. The nearest neighbor and obstacle will continually change throughout the motion.

Example 5. Converging on the Source of a Plume – 2D Case

The next example is a plume localization problem. The objective is for multiple vehicles to locate and converge on a source, which could either be acoustic, radio frequency, temperature, or chemical (see Fig. 8). It is assumed that the spatial signature of the source can be approximated by a quadratic surface. The form of this second order equation allows us to easily formulate a convergent control to the extremum of the surface. If the data were fit to a higher order surface with many local extremum, then it would not be possible to guarantee convergence to a single solution.

Step 1. The objective is

$$\max_{\tilde{x}} v(\tilde{x}) \quad (29)$$

where the global performance index is

$$v(\tilde{x}) \cong \sum_{i=1}^N a_0 + A_1^T (\tilde{x}_i - \tilde{x}_0) + \frac{1}{2} (\tilde{x}_i - \tilde{x}_0)^T A_2 (\tilde{x}_i - \tilde{x}_0). \quad (30)$$

The parameters of the quadratic surface are $a_0 \in \mathbb{R}$, $A_1 \in \mathbb{R}^2$, and $A_2 \in \mathbb{R}^{2 \times 2}$. The center of the source is located at $\tilde{x}_0 \in \mathbb{R}^2$.

Step 2. The distributed objective is

$$\max_{\tilde{x}_i} v_i(\tilde{x}) \quad \forall i = 1, \dots, N \quad (31)$$

where the partitioned performance index is

$$v_i(\tilde{x}) \cong \sum_{j \in \text{NN}} a_{0i} + A_{1i}^T (\tilde{x}_j - \tilde{x}_i) + \frac{1}{2} (\tilde{x}_j - \tilde{x}_i)^T A_{2i} (\tilde{x}_j - \tilde{x}_i). \quad (32)$$

Each vehicle determines its own estimate of the quadratic surface using information from its nearest neighbors. An alternative approach is to use data from as many neighbors as possible and calculate a least-squares estimate of the quadratic coefficients. References [48,49,50] and [61] describe the least squares fitting algorithm in more detail.

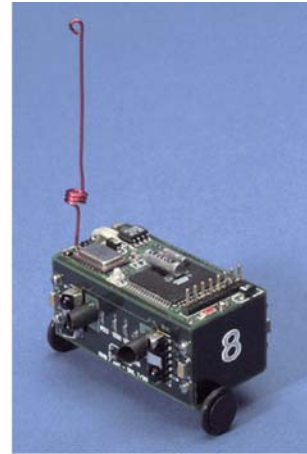
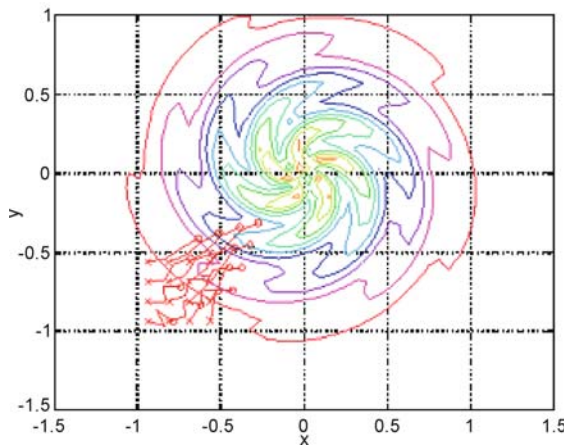
Step 3. The second order Newton's method control law is

$$\tilde{x}_i(k+1) = \tilde{x}_i(k) - \alpha A_{2i}^{-1} \big|_{\tilde{x}(k)} A_{1i} \big|_{\tilde{x}(k)} \quad (33)$$

where the quadratic coefficients are determined from the solution to the nearest neighbor equations

$$v_i(\tilde{x}_j) = a_{0i} + A_{1i}^T (\tilde{x}_j - \tilde{x}_i) + \frac{1}{2} (\tilde{x}_j - \tilde{x}_i)^T A_{2i} (\tilde{x}_j - \tilde{x}_i) \quad \forall j \in \text{NN}. \quad (34)$$

The control law in Eqs. (33)–(34) has been implemented on RATLER vehicles (see Fig. 2) that locate an acoustic source and on a set of miniature robotic vehicles (see Fig. 8) that locate a block of dry ice [48,49,50,61]. In both cases, the number of nearest neighbors is six because seven measurements (including itself) are needed to uniquely determine the quadric coefficients A_{1i} and A_{2i} .



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 8

Left: Multiple vehicles converging on a rotating plume. Right: Miniature robot used in the plume localization experiment that located a block of dry ice

Example 6. Converging on the Source of a Plume – 3D Case

The last example is a three-dimensional plume localization problem. The objective is for multiple vehicles to locate and converge on a source, which could either be acoustic, temperature, or chemical. It is assumed that the spatial signature of the source can be approximated by a quadratic surface.

Step 1. The objective is

$$\max_{\bar{x}} v(\bar{x}) \quad (35)$$

where the global performance index is

$$v(\bar{x}) \cong \sum_{i=1}^N a_0 + A_1^T (\bar{x}_i - \bar{x}_0) + \frac{1}{2} (\bar{x}_i - \bar{x}_0)^T A_2 (\bar{x}_i - \bar{x}_0). \quad (36)$$

The parameters of the quadratic surface are $a_0 \in \mathbb{R}$, $A_1 \in \mathbb{R}^3$, and $A_2 \in \mathbb{R}^{3 \times 3}$. The center of the source is located at $\bar{x}_0 \in \mathbb{R}^3$.

Step 2. The distributed objective is

$$\max_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \dots, N \quad (37)$$

where the partitioned performance index is

$$v_i(\bar{x}) \cong \sum_{j \in \text{NN}} a_{0i} + A_{1i}^T (\bar{x}_j - \bar{x}_i) + \frac{1}{2} (\bar{x}_j - \bar{x}_i)^T A_{2i} (\bar{x}_j - \bar{x}_i). \quad (38)$$

Each vehicle determines its own estimate of the quadratic surface using information from its nearest neighbors.

Step 3. The second order Newton's method control law is

$$\bar{x}_i(k+1) = \bar{x}_i(k) - \alpha A_{2i}^{-1} \big|_{\bar{x}(k)} A_{1i} \big|_{\bar{x}(k)} \quad (39)$$

where the quadratic coefficients are determined from the solution to the nearest neighbor equations

$$v_i(\bar{x}_j) = a_{0i} + A_{1i}^T (\bar{x}_j - \bar{x}_i) + \frac{1}{2} (\bar{x}_j - \bar{x}_i)^T A_{2i}^T (\bar{x}_j - \bar{x}_i) \quad \forall j \in \text{NN}. \quad (40)$$

For the 3D case, the number of nearest neighbors is nine because ten measurements (including itself) are needed to uniquely determine the quadratic coefficients A_{1i} and A_{2i} . Most recently, this algorithm has been implemented on underwater vehicles that locate and converge on a 3D

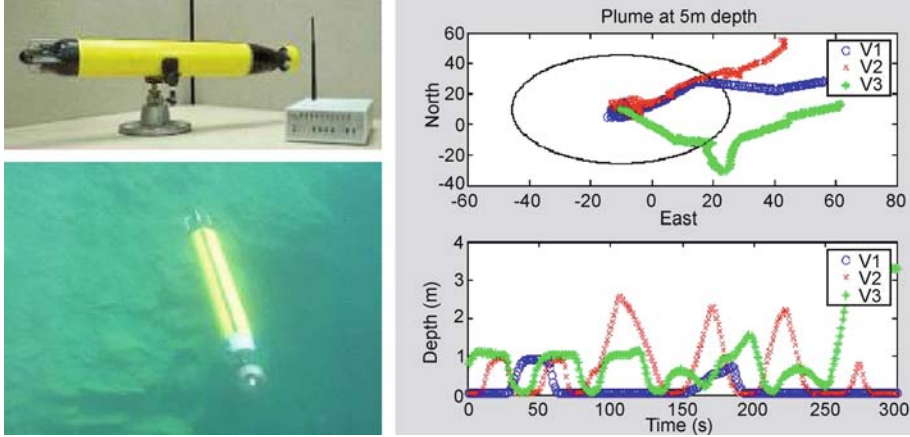
plume [62]. Preliminary tests were conducted with a synthetic plume. Synthesized sensor data was calculated as a function of position to debug the algorithm. The underwater robots are shown in Fig. 9, as well as the results of a typical test run.

These six examples demonstrate the utility of this three-step process for creating locally optimal distributed controls for multiple robotic vehicles. The resulting control laws are robust and only require sharing of information between nearest neighbors. The robustness is the result of the self-organizing nature of the control where nearest neighbors are continually changing throughout the motions. If a vehicle is lost or dies, another set of nearest neighbors can be used to complete the task. By using penalty functions to approximate constraints, the control laws are in a form that is identical to the potential field control laws often used for controlling single and multi-robots. The main difference is the switching of potential fields based on the nearest neighbors and the nearest obstacles.

Communication Effects

In this section, previous analysis regarding stable control of multiple vehicles using large-scale decentralized control techniques [47] is extended to include the communications aspects of the problem. A stability analysis shows that the local feedback control gains of the robotic vehicles must be decreased if the communication sample period is increased. Therefore, there is a tight coupling between communications and controls that cannot be ignored. In general, a system will be more responsive and have shorter settling times if the feedback control gains are as large as possible and the communication sample period is as short as possible. This section evaluates the resulting communication sample period of four different communication protocols: a Time Division Multiple Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Multiple Access (CSMA) coloring algorithm. The selection of the best protocol depends on the density of the robot vehicles and the communication radius of each vehicle.

Throughout this section, the one-dimensional dispersion example from the previous section is used to illustrate the design methodology. In [47], it is shown that α in Eq. (5) is actually more constrained than $0 < \alpha \leq 1$ depending on the responsiveness of the vehicle and the communication sample period. Therefore, the next question to ask is that of connective stability. Under what conditions will the overall system be globally asymptotically stable



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 9

Left: Nekton Research underwater vehicle used to locate synthetic plume source. Right: Underwater synthetic plume test results

ble even under structural perturbations? Analysis of connective stability is based upon the concept of vector Lyapunov functions, which associates several scalar functions with a dynamic system in such a way that each function guarantees stability in different portions of the state space. The objective is to prove that there exist Lyapunov functions for each of the individual subsystems and then prove that the vector sum of these Lyapunov functions is a Lyapunov function for the entire system.

Stability Analysis

To simplify matters, we will assume that the control function has already been chosen and the closed loop dynamics of the discrete time system can be written as

$$\mathbf{S}: x_i(k+1) = g_i(k, x_i) + \tilde{g}_i(k, \bar{x}), \quad i \in \{1, \dots, N\}, \quad (41)$$

where $\bar{x}(k) \in \mathbb{R}^n$ is the state of \mathbf{S} (e.g., x, y position, orientation, and linear and angular velocities of all vehicles) at time $k \in T$, $x_i(k) \in \mathbb{R}^{n_i}$ is the state of the i th subsystem \mathbf{S}_i at time $k \in T$. The function $g_i: T \times \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_i}$ describes the local dynamics of \mathbf{S}_i , and the function $\tilde{g}_i: T \times \mathbb{R}^n \rightarrow \mathbb{R}^{n_i}$ represents the dynamic interaction of \mathbf{S}_i with the rest of the system \mathbf{S} . The interconnection function can be written as

$$\tilde{g}_i(k, \bar{x}) = \tilde{g}_i(k, \bar{e}_{i1}x_1, \bar{e}_{i2}x_2, \dots, \bar{e}_{iN}x_N), \quad i \in \{1, \dots, N\} \quad (42)$$

where $\bar{e}_{ij} \in B^{n_i \times n_j}$, and the elements of the fundamental interconnection matrix $\bar{E} = (\bar{e}_{ij})$ are

$$(\bar{e}_{ij})_{pq} = \begin{cases} 1, & (x_j)_q \text{ occurs in } (\tilde{g}_i(t, \bar{x}))_p \\ 0, & (x_j)_q \text{ does not occur in } (\tilde{g}_i(t, \bar{x}))_p, \end{cases} \quad (43)$$

where $q \in \{n_j\}$ and $p \in \{n_i\}$.

The structural perturbations of \mathbf{S} are introduced by assuming that the elements of the fundamental interconnection matrix that are one can be replaced by any number between zero and one, i.e.

$$e_{ij} = \begin{cases} [0, 1], & \bar{e}_{ij} = 1 \\ 0, & \bar{e}_{ij} = 0. \end{cases} \quad (44)$$

Therefore, the elements e_{ij} represent the strength of coupling between the individual subsystems. A system is connectively stable if it is stable in the sense of Lyapunov for all possible $E = (e_{ij})$ [63]. In other words, if a system is connectively stable, it is stable even if an interconnection becomes decoupled, i.e. $e_{ij} = 0$, or if interconnection parameters are perturbed, i.e. $0 < e_{ij} < 1$. This is potentially very powerful, as it proves that the system will be stable even if an interconnection is lost through communication failure.

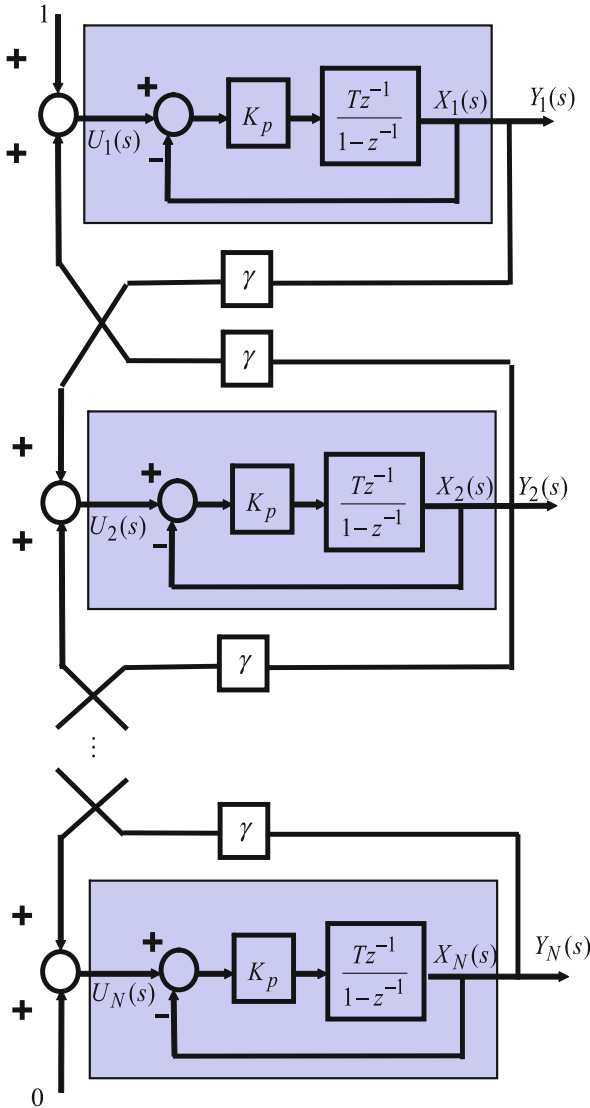
For linear systems, the discrete time dynamics may be written as

$$\mathbf{S}: x_i(k+1) = A_{ii}x_i(k) + \sum_{j=1}^N e_{ij}A_{ij}x_j(k), \quad i \in \{1, \dots, N\}, \quad (45)$$

and the Lyapunov function for each individual subsystems is $v_i(x_i) = (x_i^T H_i x_i)^{1/2}$ where H_i is a positive definite matrix. For the system S to be connectively stable, the following test matrix $W = (w_{ij})$ must be an M -matrix (i. e., all leading principal minors must be positive) [64]:

$$w_{ij} = \begin{cases} \xi_i, & i = j \\ -e_{ij}\xi_{ij}, & i \neq j \end{cases} \quad (46)$$

where $\xi_i = 1 - \sqrt{1 - \frac{1}{\lambda_M(H_i^*)}}$, $\xi_{ij} = \lambda_M^{1/2}(A_{ij}^T A_{ij})$, and



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 10

Discrete time control block diagram of N -vehicle interaction problem

$A_{ii}^T H_i^* A_{ii} - H_i^* = -I$, $\lambda_M(\bullet)$ is the maximum eigenvalue of the corresponding matrices, and the superscript $*$ denotes the Hermitian operator. For the linear dispersion example, we will model the vehicle dynamics as a discrete time integrator with a position feedback loop (see Fig. 10). The proportional control gain is K_p , and the sampling period is T .

The sampling period is both the communication and position update sample time. The state equations of the system are

$$\begin{aligned} S: x_1(k+1) &= (1 - K_p T) x_1(k) + \gamma K_p T x_2(k) \\ x_i(k+1) &= (1 - K_p T) x_i(k) + \gamma K_p T x_{i-1}(k) \\ &\quad + \gamma K_p T x_{i+1}(k), \quad i \in \{2, \dots, N-1\} \\ x_N(k+1) &= (1 - K_p T) x_N(k) + \gamma K_p T x_{N-1}(k). \end{aligned} \quad (47)$$

Note that when comparing Eq. (47) to Eqs. (5) and (6), it is evident that $2\alpha = K_p T$ and $\alpha = \gamma K_p T$. If Eq. (47) is forced to be exactly equivalent to Eqs. (5) and (6), then $\gamma = 1/2$ and $\alpha = K_p T/2$. The following stability test is less restrictive, and the interaction gain γ is less constrained. If $0 < K_p T \leq 1$, the resulting test matrix is

$$W = \begin{bmatrix} K_p T & -K_p T \gamma & 0 & \dots & 0 \\ -K_p T \gamma & K_p T & -K_p T \gamma & & \vdots \\ 0 & -K_p T \gamma & K_p T & & 0 \\ \vdots & & & \ddots & -K_p T \gamma \\ 0 & \dots & 0 & -K_p T \gamma & K_p T \end{bmatrix}, \quad (48)$$

and if $1 < K_p T \leq 2$, the test matrix is

$$W = \begin{bmatrix} (2 - K_p T) & -K_p T \gamma & 0 & & \\ -K_p T \gamma & (2 - K_p T) & -K_p T \gamma & & \\ 0 & -K_p T \gamma & (2 - K_p T) & & \\ \vdots & & & \ddots & \\ 0 & \dots & 0 & & 0 \\ & & & \dots & 0 \\ & & & & \vdots \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & -K_p T \gamma \\ & & & & & (2 - K_p T) \end{bmatrix}. \quad (49)$$

For $N = 2$, the test matrix is an M -matrix, and the system is connectively stable if

$$|\gamma| < \begin{cases} 1, & 0 < K_p T \leq 1 \\ \frac{2}{K_p T} - 1, & 1 < K_p T \leq 2. \end{cases} \quad (50)$$

Figure 11 illustrates the stability region for the case of $N = 2$. The dark region represents stable combinations of the interaction gain γ and $K_p T$ (proportional control gain multiplied by the sampling period). The white region represents unstable combinations of γ and $K_p T$. We refer to the dark region as a stability “house” due to the shape of the stable zone. The size of this stability house varies only with N . As N is increased, the house gets smaller in width but maintains the same height and shape. The size of the stability house is a measure of the robustness of the closed-loop system to parameter variations in interaction gain γ , sampling period T , and proportional control gain K_p . Figure 11 also shows the stability region for $N = 10\,000$.

For this particular example, another way to check the stability of this linear system is to check that the eigenvalues of the system matrix A are within the unit circle. There is a special formula (p. 59 in [65]) for the eigenvalues of A given by

$$\lambda_i(A) = 1 - K_p T + 2K_p T \gamma \cos\left(\frac{i\pi}{N+1}\right), \quad i = 1, \dots, N. \quad (51)$$

From this formula, we can see that as $N \rightarrow \infty$, the cosine term becomes unity. This implies that γ must stay between -0.5 and 0.5 for $K_p T$ less than one in order to maintain stability. For $K_p T$ greater than one, the admissible γ values taper off parabolically (the sloped “roof”) until $K_p T = 2$.

Several conclusions can be drawn from this stability analysis. First, asymptotic stability of vehicle positions depends on vehicle responsiveness K_p , communication sampling period T , and vehicle interaction gain γ . If the vehicle is too fast (large K_p) or the sample period is too long (large T) then the vehicles will go unstable. There is a dependence on interaction gain for stability as well. Second,

the interaction gains can be used to bunch the vehicles closer together or spread them out. Third, the stability region shrinks as the number of vehicles, N , increases but only to a defined limit.

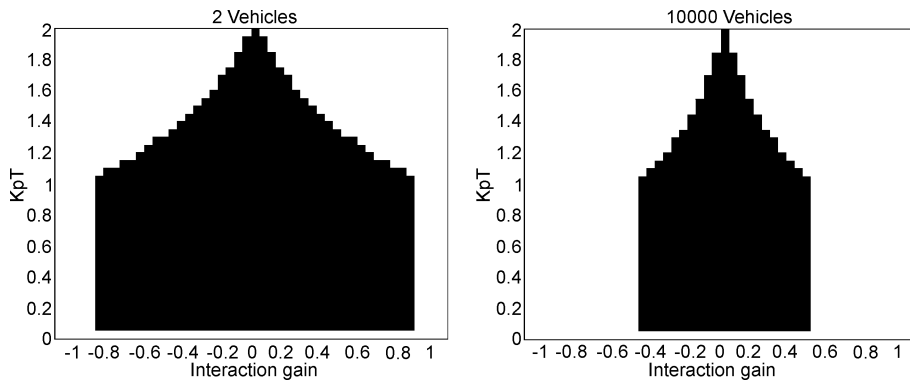
Communications Sample Period

As noted, the communication sample period greatly affects the stability of the system. As defined in the equations above, this sample period is the time it takes for every node to communicate once. In this section, we will evaluate the communication sample period of four different communication schemes: a Time Division Medium Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Medium Access (CSMA) coloring algorithm. All of these schemes assume that each node has a unique identification number. The TDMA schemes also assume that each node has a synchronized clock that is used to notify each node when it may transmit a message. The CSMA scheme first checks the communication channel for a collision before transmitting a packet.

In order to determine this sample period, one important parameter associated with an ad hoc communication network is the degree of the network. The degree of the network is defined as the maximum number of nodes that any node can communicate with given a limited communication range.

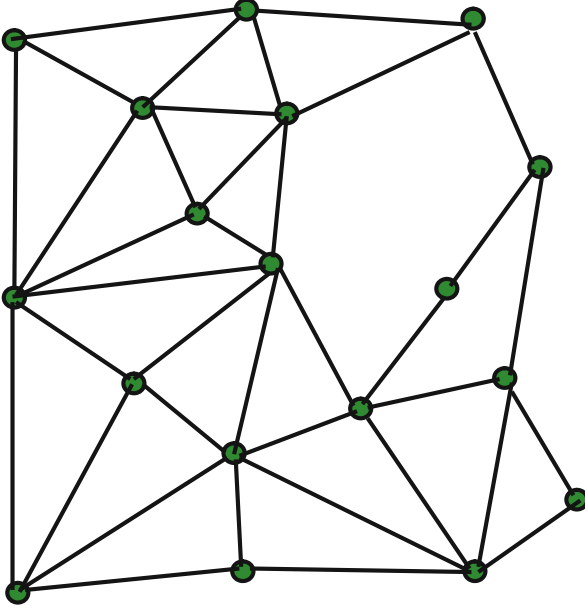
For the network shown in Fig. 12, the degree of the network is

$$\Delta = \max_{i \in \{1, \dots, N\}} \left\{ \sum_{j=1, j \neq i}^N \text{range}(x_i - x_j, R_c) \right\} \quad (52)$$



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 11

Left: Stability region for the $N = 2$ vehicle case. Right: Stability region for the $N = 10\,000$ vehicle case



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 12

Graph of ad hoc communication network. Nodes with connecting lines can communicate with each other

where

$$\text{range}(x_i - x_j, R_c) = \begin{cases} 1 & \text{if } |x_i - x_j| < R_c \\ 0 & \text{else} \end{cases} \quad (53)$$

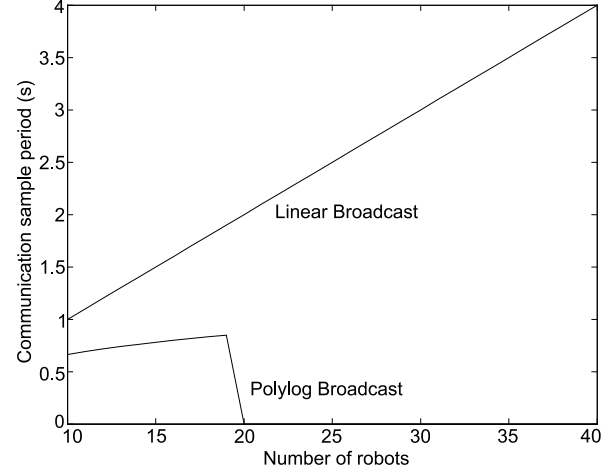
and R_c is the communication radius of each node. Assuming all the robots are evenly spaced along a line of length L and have a density $\delta = \frac{N}{L}$, then the degree of the resulting network is

$$\Delta = 2 \lfloor \delta R_c \rfloor = 2 \left\lfloor \frac{N R_c}{L} \right\rfloor. \quad (54)$$

For the TDMA linear broadcast where every node is assigned a unique identification number, the communication sample period time required for every node to send a message is

$$T_{\text{linear}} = \tau N \quad (55)$$

where τ is the time period associated with each communication time slot. Notice that the above expression is proportional to N . This delay time can be shortened by using a polylogarithmic broadcast scheme [66] where each node communicates during multiple time slots. Even though multiple messages are being broadcast at the same time, the message is guaranteed a successful broadcast during



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 13

Communication sample period for TDMA linear and polylogarithmic broadcasts when $R_c/L = 0.1$ and $\tau = 0.1$ s

one of the time slots as long as the degree of the network is below a certain value. The communication period for a polylogarithmic broadcast is

$$T_{\text{polylog}} = \tau (2 \log_2 N)^h \quad (56)$$

when

$$\Delta \leq 2^{h+1} - 1 \quad (57)$$

and where

$$h = \lfloor (\log_2 N) / (\log_2 (\log_2 N) + 1) \rfloor. \quad (58)$$

Notice that this expression is proportional to the $\log_2 N$ instead of N . Figure 13 compares the linear broadcast to the polylogarithmic broadcast for a network spread out over a line and with each node having a communication radius that is one-tenth the length of the line. At 20 robots, the average degree of the network becomes too large and the polylogarithmic broadcast will no longer work.

Even better than the linear broadcast and the polylogarithmic broadcast, a color scheme allows multiple nodes to communicate at the same time by using spatial reuse of time slots. Time slots (called colors) are assigned so that each node has a different color than its first and second nearest neighbors. By using different colors, the hidden node problem, where two nodes speak to an intermediate node at the same time, is eliminated. In graph theory, the minimum number of colors can range from the maximum degree of the network plus one to the square of the

maximum degree of the network plus one.

$$\Delta + 1 \leq k \leq \Delta^2 + 1. \quad (59)$$

However, in a typical planar wireless network the number of colors is typically bounded by

$$k \leq \Delta + \varepsilon \quad (60)$$

where ε is a small number, typically 1 to 5. For a TDMA colored network, the communication sample period is given by

$$T_{\text{TDMA}} = \tau (\Delta + \varepsilon) \quad \text{where } \varepsilon \text{ is small.} \quad (61)$$

For a CSMA network, the actual communication time is non-deterministic because the packets often collide and a random back-off is used before retransmitting. The average communication time depends on the network utilization, i. e. the percentage of time the network is being used. Modeling the CSMA network as a M/D/1 queue [67], the average communication time is

$$\tau_{\text{CSMA}} = \frac{\rho T_m}{2(1-\rho)} + T_m \quad (62)$$

where T_m is the time to send a single message (or service time of the queue). For simplicity of comparison with the TDMA network, we will assume that $T_m = \tau$. In reality, the message length T_m in a TDMA network must be slightly less than the time slot τ . The utilization factor is

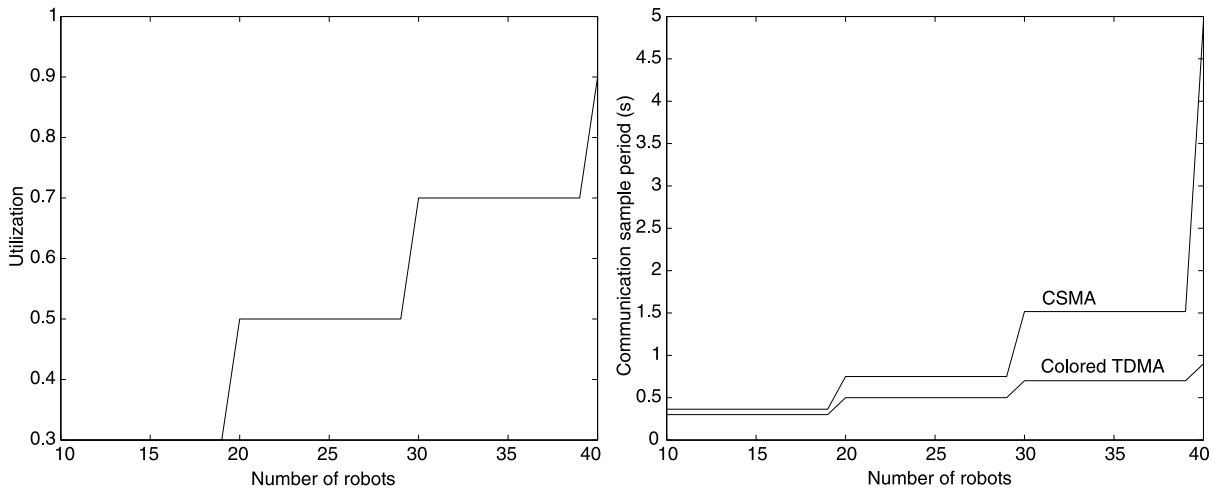
$$\rho = \frac{T_m (\Delta + \varepsilon)}{T_d} \quad (63)$$

where T_d is the delay time between each new message that a node initiates. For a CSMA network with degree Δ , the communication sample period is approximately given by

$$T_{\text{CSMA}} = \tau_{\text{CSMA}} (\Delta + \varepsilon). \quad (64)$$

Figure 14 shows how the network utilization changes as the number of robots increases in our one-dimensional dispersion example. The resulting communication sampling period for both the TDMA and CSMA colored networks are also shown in Fig. 14. Notice that for smaller numbers of vehicles both the colored TDMA and CSMA networks have a shorter communication sample period than both linear and polylogarithmic broadcasts. However, when the utilization of the CSMA network reaches 0.9, the CSMA network starts to substantially degrade in performance. This is caused by flooding the network with messages as the density of the robots increase while the back-off time of communication stays the same. It should be noted that this flooding can be alleviated if the nodes were to adjust how often they send out messages based on the maximum degree of the network. Ideally, the nodes should adjust their communication sample period so that $T_d = T_{\text{CSMA}}$.

These results show that a colored TDMA network appears to perform the best. However, the disadvantage of the colored TDMA network is that there is an initialization time that is required to determine the color of each node whenever the network topology changes. The other algorithms have the advantage that they do not require a network initialization time. Using an algorithm by [68], this initialization time is the time required to broadcast their



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 14

Left: Utilization for a CSMA network and right: communication sample period for TDMA and CSMA reconfigurable coloring when $R_c/L = 0.1$, $\tau = 0.1$ s, $\varepsilon = 1$, and $T_d = 1$ s

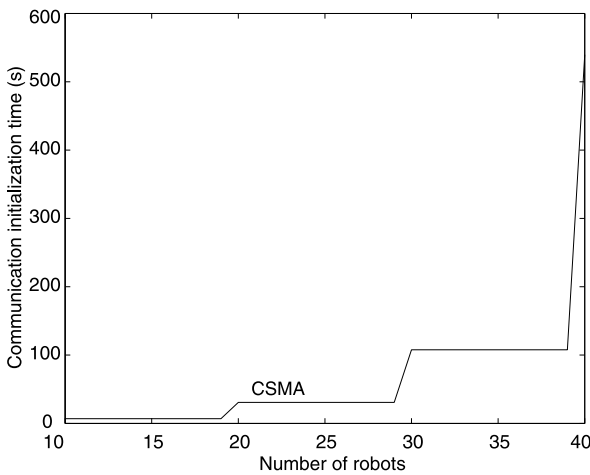
own identification number, their 1st nearest neighbor lists (after which each node can determine 2nd nearest neighbor), and $k + 2$ additional messages for coloring. In addition to these messages, each neighboring node must acknowledge the messages containing the neighbor lists and coloring information to ensure that the messages were received. Since time slots are typically not assigned before hand, this initialization process occurs using CSMA protocols, and it should be performed whenever the topology of the network changes, i. e. when robots move. Assuming that all messages are the same length, the resulting initialization time is given by

$$\begin{aligned} t_{\text{init}} = & \tau_{\text{CSMA}} [2(\Delta + \varepsilon) + (\Delta + \varepsilon + 2)(\Delta + \varepsilon)] \\ & + \tau_{\text{CSMA}} [(\Delta + \varepsilon)(\Delta + \varepsilon - 1) \\ & + (\Delta + \varepsilon)(\Delta + \varepsilon + 2)(\Delta + \varepsilon - 1)] \end{aligned} \quad (65)$$

where ε is small. This initialization time is plotted as a function of the number of robots in Fig. 15. This figure shows that this initialization time can be substantial. Adding the initialization time in Fig. 15 to the communication sample period in Fig. 14, we see that the linear, polylogarithmic, and CSMA networks have a shorter sample period than the colored TDMA network. In general, it is best to use the TDMA mode only if the network does not reconfigure; otherwise, it is best to use a CSMA mode of communications.

Utilization Versus Delay in CSMA Networks

If a control system is implemented with a CSMA communications scheme, there is a tradeoff between network



Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 15

Communication overhead for reconfigurable coloring when $R_c/L = 0.1$, $\tau = 0.1$ s, $\varepsilon = 1$, and $T_d = 1$ s

utilization and message delay. Ideally, one would like to maximize network utilization while minimizing the delay seen by each message. Modeling the CSMA network as an M/D/1 queue, the normalized delay (where T_m is the message length) is given by

$$\frac{\tau_{\text{CSMA}}}{T_m} = \frac{\rho}{2(1-\rho)} + 1 \quad (66)$$

where τ_{CSMA} is the average communication time. Using the cost function $J(\rho)$ below, the gains K_1 and K_2 may be used to weight the penalty associated with utilization and delay.

$$J(\rho) = \frac{K_1}{\rho} + \frac{K_2 \rho}{2(1-\rho)} + K_2. \quad (67)$$

The optimal solution is then obtained by minimizing the cost function over $0 \leq \rho < 1$

$$J(\rho^*) = \min_{0 \leq \rho < 1} J(\rho) = \min_{0 \leq \rho < 1} \left\{ \frac{K_1}{\rho} + \frac{K_2 \rho}{2(1-\rho)} + K_2 \right\}. \quad (68)$$

Differentiating Eq. (67) with respect to ρ yields

$$\frac{\partial J(\rho)}{\partial \rho} = -\frac{K_1}{\rho^2} + \frac{2(1-\rho)K_2 + 2K_2\rho}{4(1-\rho)^2}. \quad (69)$$

Solving Eq. (69) for the minimum gives the following optimal value for network utilization ρ^* based on the weighting values K_1 and K_2 .

$$\rho^* = \frac{-2K_1 \pm \sqrt{2K_1K_2}}{(K_2 - 2K_1)}, \quad 0 \leq \rho^* < 1. \quad (70)$$

If utilization and average delay are equally weighted, the optimal value of ρ is $\rho^* = 0.5858$. Optimal utilization values for several values of K_1 and K_2 are summarized in Table 1. These results are intuitive – if average message delay is critical, then network utilization should be kept to a minimum.

This section illustrates the tight coupling that exists between communications and controls when designing large-scale cooperative robotic systems. A connective stability analysis shows that local feedback control gains and

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Table 1

Optimal utilization for different values of K_1 and K_2

Weighting gains	Optimal utilization, ρ^*
$K_1 = K_2 = 1$	$\rho^* = 0.5858$
$K_1 = 1, K_2 = 10$	$\rho^* = 0.309$
$K_1 = 0.1, K_2 = 10$	$\rho^* = 0.124$

communication sample periods are inversely related. If the communication sample period increases, then the local feedback control gains must decrease. The communication sample period is a function of the protocol, and the protocol with the shortest communication sample period depends on the density of robots and the communication radius. By assuming worst-case conditions for robot density and communication range, this analysis can be used off-line to determine conservative control gains required for stable control. In the future, it might also be possible to use this analysis on-line to adjust control gains and/or communication range as the robot density changes.

Conclusions

This chapter described an integrated approach to designing communication, sensing, and control systems for fixed and mobile distributed systems. The analysis built upon provably convergent cooperative controls techniques and upon concepts from graph theory as applied to communication networks. A common mathematical framework consisting of three steps was developed for creating decentralized cooperative control laws. The first step is to define a global performance index for the cooperative behavior. The second step is to partition the performance index so that only local interactions are included. The third step is to create a first or second order gradient control law that minimized the partitioned performance index. After these three steps, a vector Lyapunov technique is used to determine the stability constraints on the individual subsystem control gains, interaction control gains, and the communication sampling period.

The communication protocols were evaluated to ensure that they could meet the stability constraint on the communication sampling period. Coloring algorithms from graph theory were used to compare the required performance of TDMA or CSMA communication networks. In general, we found that TDMA networks will outperform CSMA networks if the network is stationary, and there is sufficient time to perform the network coloring algorithms. However, if the network is moving, then a CSMA network will outperform the TDMA network as long as the network utilization stays below 58 percent. Queuing theory was used to determine a closed form solution of the time response of a CSMA network, and Monte Carlo simulations have been used to verify the solution. The vector Lyapunov analysis shows that the collective CSMA networked system will still remain stable as long as the maximum communication time is below the maximum sampling period determined from the vector Lyapunov analysis.

Future Directions

Future work should focus in several directions. First, more complex vehicle dynamics should be considered. This may include the effects of nonlinearities in the vehicle locomotion and/or higher order friction effects. Second, more complicated vehicle formation problems should be considered which may include obstacles and varied terrain features. Third, other collective control applications should be addressed such as distributed electric power grids and other swarm applications. Finally, the communications issue should continue to be studied to look at the effects of sample periods, delays, and interrupted communications on the stability of collectives for other communication protocols. Eventually, real-time algorithms should be developed that evaluate the stability of a networked system, and appropriately adjust the communication protocol and vehicle responsiveness.

Acknowledgments

The authors greatly appreciate the help of Steven Eskridge, John Hurtado, Chris Lewis, John Harrington, and Nekton Research, LLC, in implementing and testing these algorithms on a variety of robot platforms. This work was supported in part by the Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000. In addition, this research was partially funded by the Information Processing Technology Office and Microsystems Technology Office of the Defense Advanced Research Projects Agency.

Bibliography

1. Arai T, Pagello E, Parker PE (2002) Guest Editorial: Advances in Multirobot Systems. *IEEE Trans Robot Autom* 18(5):655–659
2. Bonabeau E, Dorigo M, Theraulaz G (1999) *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, New York
3. Arkin RC (1992) Cooperation Without Communication: Multiagent Schema-Based Robot Navigation. *J Robotic Syst* 9(3): 351–364
4. Balch T, Arkin RC (1998) Behavior-Based Formation Control for Multirobot Teams. *IEEE Trans Robot Autom* 14(6):926–939
5. Kube RC, Zhang H (1993) *Collective Robotics: From Social Insects to Robots*. *Adapt Behav* 2(2):189–218
6. Parker LE (1998) ALLIANCE: An Architecture for Fault Tolerant Multirobot Cooperation. *IEEE Trans Robot Autom* 14(2): 220–240
7. Di Marco M, Carulli A, Giannitrapani A, Vicino A (2003) Simultaneous Localization and Map Building for a Team of Cooperating Robots: A Set Membership Approach. *IEEE Trans Robot* 19(2):238–249

8. Burgard W, Moors M, Stachniss C, Schneider FE (2005) Coordinated Multi-Robot Exploration. *IEEE Trans Robot* 21(3):376–386
9. Cortes J, Martinez S, Karatas T, Bullo F (2004) Coverage Control for Mobile Sensing Networks. *IEEE Trans Robot Autom* 20(2):865–875
10. Tang Z, Ozguner U (2005) Motion Planning for Multitarget Surveillance With Mobile Sensor Agents. *IEEE Trans Robot* 21(5):998–908
11. Vidal R, Shakernia O, Kim HJ, Shim DH, Sastry S (2002) Probabilistic Pursuit-Evasion Games: Theory, Implementation, and Experimental Evaluation. *IEEE Trans Robot Autom* 18(5):662–669
12. Isler V, Kannan S, Khanna S (2005) Randomized Pursuit-Evasion in a Polygonal Environment. *IEEE Trans Robot* 21(5):875–894
13. Cao Z, Tan M, Li L, Gu N, Wang S (2006) Cooperative Hunting by Distributed Mobile Robots Based on Local Interaction. *IEEE Trans Robot* 22(2):403–407
14. Bopardikar SD, Bullo F, Hespanha JP (2007) Cooperative Pursuit with Sensing Limitations. In: *Proceedings of the 2007 American Control Conference*. IEEE, Minneapolis
15. Fua CH, Ge SS (2005) COBOS: Cooperative Backoff Adaptive Scheme for Multirobot Task Allocation. *IEEE Trans Robot* 21(6):1168–1178
16. Daigle MJ, Koutsoukos XD, Biswas G (2007) Distributed Diagnosis in Formations of Mobile Robots. *IEEE Trans Robot* 23(2):353–369
17. Crandall JW, Cummings ML (2007) Identifying Predictive Metrics for Supervisory Control of Multiple Robots. *IEEE Trans Robot* 23(5):942–951
18. Chen Q, Luh JYS (1994) Coordination and Control of a Group of Small Mobile Robots. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, vol 3. pp 2315–2320. IEEE, San Diego
19. Yamaguchi H, Arai T (1994) Distributed and Autonomous Control Method for Generating Shape of Multiple Mobile Robot Group. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, vol 2. pp 800–807. IEEE, Munich
20. Yamaguchi H, Burdick JW (1998) Asymptotic Stabilization of Multiple Nonholonomic Mobile Robots Forming Group Formations. In: *Proceedings of the 1998 Conference on Robotics & Automation*. Leuven, Belgium, pp 3573–3580
21. Yoshida E, Arai T, Ota J, Miki T (1994) Effect of Grouping in Local Communication System of Multiple Mobile Robots. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, vol. 2. pp 808–815. IEEE, Munich
22. Molnar P, Starke J (2000) Communication Fault Tolerance in Distributed Robotic Systems. In: Parker LE, Bekey G, Barhen J (eds) *Distributed Autonomous Robotic Systems 4*. Springer, Berlin, pp 99–108
23. Schneider FE, Wildermuth D, Wolf HL (2000) Motion Coordination in Formations of Multiple Robots Using a Potential Field Approach. In: Parker LE, Bekey G, Barhen J (eds) *Distributed Autonomous Robotic Systems 4*. Springer, Berlin, pp 305–314
24. Stankovic SS, Stanojevic MJ, Siljak DD (2000) Decentralized Overlapping Control of a Platoon of Vehicles. *IEEE Trans Control Syst Technol* 8:816–832
25. Seiler P, Pant A, Hedrick K (2004) Disturbance Propagation in Vehicle Strings *IEEE Trans Autom Control* 29:1835–1841
26. Barooah P, Mehta PG, Hespanha JP (2007) Control of Large Vehicular Platoons: Improving Closed Loop Stability by Mistuning. In: *Proceedings of the 2007 American Control Conference*. IEEE, Minneapolis
27. Ogren P, Egerstedt M, Hu X (2002) A Control Lyapunov Function Approach to Multiagent Coordination. *IEEE Trans Robot Autom* 18(5):847–851
28. Lawton JRT, Beard RW (2003) A Decentralized Approach to Formation Maneuvers. *IEEE Trans Robot Autom* 19(6):933–941
29. Tanner HG, Pappas GJ, Kumar V (2004) Leader-to-Formation Stability. *IEEE Trans Robot Autom* 20(3):443–455
30. Tabuada P, Pappas GJ, Lima P (2005) Motion Feasibility of Multi-Agent Formations. *IEEE Trans Robot* 21(3):387–392
31. Antonelli G, Chiaverini S (2006) Kinematic Control of Platoons of Autonomous Vehicles. *IEEE Trans Robot* 22(6):1285–1292
32. Beni G, Liang P (1996) Pattern Reconfiguration in Swarms – Convergence of a Distributed Asynchronous and Bounded Iterative Algorithm. *IEEE Trans Robot Autom* 12(3):485–490
33. Liu Y, Passino K, Polycarpou M (2001) Stability Analysis of One-Dimensional Asynchronous Swarms. *American Control Conference*, Arlington, pp 25–27, 716–721
34. Winfield A (2000) Distributed Sensing and Data Collection via Broken Ad Hoc Wireless Connected Networks of Mobile Robots. In: Parker LE, Bekey G, Barhen J (eds) *Distributed Autonomous Robotic Systems 4*. Springer, Berlin, pp 273–282
35. Desai JP, Ostrowski J, Kumar V (1998) Controlling Formations of Multiple Mobile Robots. In: *Proceedings of the 1998 IEEE Conference on Robotics & Automation*. IEEE, Leuven, pp 2864–2869
36. Desai JP, Kumar V, Ostrowski J (2001) Modeling and Control of Formations of Nonholonomic Mobile Robots. *IEEE Trans Robot Autom* 17(6):905–908
37. Ji M, Egerstedt M (2007) Distributed Coordination Control of Multiagent Systems While Preserving Connectedness. *IEEE Trans Robot* 23(4):693–703
38. Zavlanos MM, Pappas GJ (2007) Potential Fields for Maintaining Connectivity of Mobile Networks. *IEEE Trans Robot* 23(4):812–816
39. Belta C, Kumar V (2004) Abstraction and Control of Groups of Robots. *IEEE Trans Robot Autom* 20(5):865–875
40. Ge SS, Fua CH (2005) Queues and Artificial Potential Trenches for Multirobot Formations. *IEEE Trans Robot* 21(4):646–656
41. Milutinovic D, Lima P (2006) Modeling and Optimal Centralized Control of a Large-Size Robotic Population. *IEEE Trans Robot* 22(6):1280–1285
42. Kloetzer M, Belta C (2007) Temporal Logic Planning and Control of Robotic Swarms by Hierarchical Abstractions. *IEEE Trans Robot* 23(2):320–330
43. Tanner HG, Loizou SG, Kyriakopoulos KJ (2003) Nonholonomic Navigation and Control of Cooperating Mobile Manipulators. *IEEE Trans Robot Autom* 19(1):53–64
44. Yamashita A, Arai T, Ota J, Asama H (2003) Motion Planning of Multiple Mobile Robots for Cooperative Manipulation and Transportation. *IEEE Trans Robot Autom* 19(2):223–237
45. Bonaventura CS, Jablokow KW (2005) A Modular Approach to the Dynamics of Complex Multirobot Systems. *IEEE Trans Robot* 21(1):26–37
46. Lewis C, Feddema JT, Klarer P (1998) Robotic Perimeter Detection System. In: *Proceedings of SPIE*, vol 3577. Boston, pp 14–21
47. Feddema JT, Lewis C, Schoenwald DA (2002) Decentralized

Control of Cooperative Robotic Vehicles: Theory and Application. *IEEE Trans Robot Autom* 18(5):852–864

48. Hurtado JE, Robinett III RD (2005) Convergence of Newton's Method via Lyapunov Analysis. *AIAA J Guidance Control Dyn* 28(2):363–365
49. Hurtado JE, Robinett III RD, Dohrmann CR, Goldsmith SY (2004) Decentralized Control for a Swarm of Vehicles Performing Source Localization. *J Intell Robot Syst* 41:1–18
50. Robinett RD, Hurtado JE (2004) Stability and Control of Collectives Systems. *J Intell Robot Syst* 39:43–55
51. Clauzet A, Tanner B, Byrne R, Abdallah CT (2007) Controlling Across Complex Networks. *Proceedings of the IFAC Time-delay Symposium, Nantes*, pp 17–19
52. Hespanha JP, Naghshtabrizi P, Xu Y (2007) A Survey of Recent Results in Networked Control Systems. *Proc IEEE Special Issue: Technol Netw Control Syst* 95(1):138–162
53. Gelfand IM, Fomin SV (1963) *Calculus of Variations*. Prentice-Hall, New Jersey
54. Frieden BR (1998) *Physics from Fischer Information*. Cambridge University Press, Cambridge
55. Feddema JT, Robinett RD, Byrne RH (2003) An Optimization Approach to Distributed Controls of Multiple Robot Vehicles, *Workshop on Control and Cooperation of Intelligent Miniature Robots. IEEE/RSJ Int Conf Intell Robot Syst. IEEE, Las Vegas*
56. Censor Y, Zenios SA (1997) *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, Oxford
57. Luenberger DG (1984) *Linear and Nonlinear Programming*, 2nd edn. Addison Wesley, Reading
58. Schmitt DJ et al (2003) *Intelligent Mobile Land Mine (IMLM) System*, Sandia National Laboratories Report. SAND 2003–1186
59. Feddema JT, Schoenwald DA (2002) Stability Analysis of Decentralized Cooperative Controls. In: Shultz AC, Parker LE (eds) *Multi-Robot Systems: From Swarms to Intelligent Automata*. Kluwer, Boston, pp 133–122
60. Feddema JT, Schoenwald DA (2002) Distributed Communication/Navigation Robot Vehicle Network. In: *Proceedings of World Automation Congress, Orlando*, pp 9–13. TSI Press, Albuquerque
61. Byrne RH, Adkins DR, Eskridge SE, Harrington JJ, Heller EJ, Hurtado JE (2002) Miniature Mobile Robots for Plume Tracking and Source Localization Research. *J Micromechatronics* 1(3): 253–261
62. Byrne RH, Eskridge SE, Hurtado JE, Salvage EL (2003) Algorithms and Analysis of Underwater Vehicle Plume Tracing, Sandia National Laboratories Report. SAND pp 2003–2643
63. Siljak DD (1991) *Decentralized Control of Complex Systems*. Academic Press, San Diego
64. Sezer ME, Siljak DD (1988) Robust Stability of Discrete Systems. *Int J Control* 48(5):2055–2063
65. Smith GD (1985) *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, 3rd edn. Oxford University Press, Oxford
66. Basagni S, Bruschi D, Chlamtac I (1999) A Mobility-Transparent Deterministic Broadcast Mechanism for Ad Hoc Networks. *IEEE/ACM Trans Netw* 7(6):799–807
67. Kleinrock L (1975) *Queueing Systems, Volume 1: Theory*, 1st edn. Wiley, New York
68. Chlamtac I, Pinter SS (1987) Distributed Nodes Organization Algorithm for Channel Access in a Multihop Dynamic Radio Network. *IEEE Trans Comput C* 36(6):728–737

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling

MICHAEL JANSSEN, ANDREW DRENNER,
NIKOLAOS PAPANIKOLOPOULOS
Center for Distributed Robotics,
Department of Computer Science and Engineering,
University of Minnesota,
Minneapolis, USA

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Robotic Team Design](#)
[Software Modeling and Simulation](#)
[Tasks Well-Suited to Distributed Robotics](#)
[Tasks for Optimizing Distributed Robotics](#)
[Future Directions](#)
[Acknowledgment](#)
[Bibliography](#)

Glossary

Marsupial robot A robot that can pick up and store other (usually smaller) robots and then possibly move or recharge those robots.

Marsupial pair One specific robot which can be picked up by another specific robot, as a pair.

Marsupial team Two or more robots which can form one or more marsupial pairs, i. e. one marsupial robot and multiple carrier robots.

Marsupial mechanism A hardware device which enables a robot to carry one or more robots. May be active or passive.

Dead-reckoning A type of robot localization accomplished by measuring the distance of wheel rotation. Prone to drifting error.

Bluetooth Wireless networking with low bandwidth and a short range, typically less than 10 meters. Used in distributed robots because of its low power requirements.

Homogeneous team A group of robots which are all of the same model, including all of the same sensors. Each robot is replaceable with any other from the team.

Heterogeneous team A group of robots which includes members that have unique sensors. More types of sensors can be included in a heterogeneous team as each robot can have a unique sensor load.

Reverse auction A type of auctioning where the lowest bidder is the winner. Used for task allocation as the lowest cost represents the lowest amount of resources expended for the task.

Definition of the Subject

The field of robotics covers devices that are in use in a wide variety of applications from interplanetary exploration [56] to performing common household tasks (such as vacuuming your floor) [23]. Definitions of what is and is not a robot can vary wildly. Generally speaking, a robot is a device with the ability to sense and interact with its environment. Usually there is some degree of intelligence or autonomy in a robotic system, but this can vary as well from a *tele-robotic* device which is completely controlled by operators from a remote location [58] to a *fully autonomous robot* which can be given a goal and will reach that goal without any human intervention [5]. The phrase *distributed robotics* deals with teams of robots used to accomplish tasks that a single robot either could not achieve or could not achieve within some constraint (cost of the system, time allotted, etc.).

Robotics is an important field because it allows for the collection of information while minimizing potential hazards to humans. Robots are already in use in war zones, searching for survivors in collapsed buildings, and operating in areas which are inhospitable to humans. Distributed robotics are often used to accomplish these tasks faster and at lower cost.

Introduction

As technology pushes toward smaller, faster, and cheaper devices, it is reasonable to expect multiple robots to work together to solve a common goal. The field of distributed robotics aims to solve the problems which appear when robots work together. It work based on robotics, electrical engineering, mechanical engineering, psychology, intelligent agents, and game theory. Certain distributed robotic approaches are inspired by biological systems as well.

Distributed robotics has many advantages over single-robot systems. Having multiple autonomous systems makes it possible to complete complex and large tasks more quickly than a single robot could. Also, building a single robot with the combined sensing of a robotic team may be cost-prohibitive. The many sensors available on the multiple smaller robots present a significantly larger amount of data about the environment than a single robot alone would produce. Having multiple robots is preeminently useful when there is more than one task to complete, as it is possible to split the team to perform both

tasks at once, and even reassign team members from one task to another. Ideally, each member of the team can be built at a relatively low cost. This affordability allows for increased redundancy throughout the robotic team.

The advantages brought by a robot team do not come without drawbacks. Having more robots in use multiplies costs, in both time of maintenance, materials cost, and potentially operator-related expenses. Some of these costs can be overcome by building many simple devices rather than a single complex one as well as developing an increased degree of autonomy. The introduction of a “team mechanic” introduces new problems not found in a monolithic robot, specifically in the areas of team coordination, task assignment, and communication. Communication itself may require additional hardware on each device, making the design of such robots more complex, requiring engineering experience in wireless and RF communication. When a team is tele-operated as opposed to autonomous, usually one operator is required for each robot on the team. This makes the cost of operating such a team of robots much higher, especially in situations where those operators could perform other tasks.

In the rest of this chapter, we will present an overview of distributed robotics. In Sect. “[Robotic Team Design](#)”, design of robotic teams and views on the composition of such teams is considered. Attention is given to the concerns relating to individual members of a team in Sect. “[Design Considerations](#)” and the makeup of a full distributed robot system in Sect. “[Robotic Team Composition](#)”. Section “[Software Modeling and Simulation](#)” covers some of the software used currently to simulate distributed teams and presents some of the challenges specific to distributed teams. Particular attention is presented to modeling a marsupial system in Sect. “[Marsupial Modeling](#)”. We then cover two robotic tasks that are well-suited to robotic teams, along with an approach for each that takes advantage of a team in Sect. “[Tasks Well-Suited to Distributed Robotics](#)”. Problems which then arise in optimizing distributed robot teams are covered in Sect. “[Tasks for Optimizing Distributed Robotics](#)”. Finally in Sect. “[Future Directions](#)”, we consider briefly the future direction of distributed robotics.

Robotic Team Design

Designing a robotic platform requires the balancing of many factors, many of which are at odds. Cost is a major factor in the design, and may be prohibitive to all others. Sensing modalities and on-board processing are also major in design. In addition to these factors, distributed robotics adds the complexity of multiple robots.

Even when not designing the individual robots, some design decisions must be made for each team. The question of homogeneity of the robots is particularly interesting, as choosing to be homogeneous or heterogeneous has differing strengths. Further, marsupial teams should also be considered.

Design Considerations

Each robot in a distributed team would ideally be designed specific to the purpose of the team. In a research setting, this is often not possible due to the team possibly being used in many different situations for many differing goals. There are many advantages which could be exploited, but would cause other areas of the platform to be neglected. Each of these dimensions of robot design is somewhat dependent on the others.

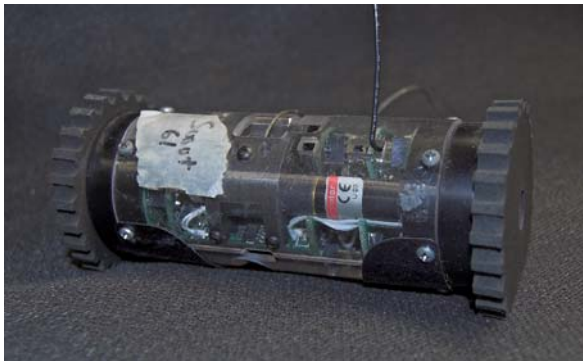
There are a number of important criteria in designing a distributed robotic team. Many of the trade-offs between these criteria are based on mission specific goals and will vary between implementations. Designing a “general purpose” distributed robotic team is a difficult if not impossible task as some of the criteria below will create interdependencies which may not be easily satisfied.

- *Physical Characteristics* – The physical size of the robot, its shape, and its mass will have impact on how the robots can be deployed, traverse terrain, or withstand impacts if it were to fall in its environment. Smaller robots may be able to negotiate cramped spaces allowing exploration of areas that larger robots may be unable to reach. However, smaller robots may also have more difficulty in traversing terrain, less volume for computation, communication, and power storage, and require increased precision in manufacturing.
- *System Longevity* – Running time is crucial to some applications, especially in situations where retrieval of the robots may not be possible and the distributed team must travel to and from the area of activity. The energy density of current battery technology requires that significant percentage of the robot’s volume be dedicated to energy storage. This constraint can be hard to satisfy with respect to the robot’s physical size. Some robots overcome this limitation by tethering while others are able to self-recharge, either from the environment in the form of solar energy [52] or auto-docking [53], or by being recharged from another robot [42]. The inclusion of these technologies requires additional electronics on the robot.
- *Communication* – Communication is a key component of distributed robotic teams. Wireless networking in the form of the 802.11, 802.15, Bluetooth, and newer emerging protocols such as 802.16 and 802.20 allow for the team to communicate without wires and without rendezvous. Research in ad-hoc networks and mesh networking is very applicable here as many robots in a distributed robot team may not be directly within communication range of each other. When this is the case, there must be a routing protocol in order to get a message to a particular robot. This can be accomplished with simple broadcast of all messages, but it does not scale well to large numbers of robots as it will eventually saturate the available bandwidth. Smarter routing protocols [46] must be used in order to minimize the amount of bandwidth used. Recent networking research in the areas of sensor networks [28] and ad-hoc networking impact this area as well. There are also other methods of communication which are in use. Some robotic teams also communicate by visual manipulation through LED lights [22], or can recognize each other through vision [12]. Usually a single network type is used over all robots in the system. Some technologies require additional electronics or are simply too large or expensive to integrate into smaller robots. In addition to the electronic components of the communication, the size, type, directionality, and number of antennas used greatly impacts the range and therefore usefulness of the wireless communication. Proper selection of communication methods can greatly extend the robot’s operational lifetime. Another aspect of the communication of distributed robot teams is connectivity. It is preferable to keep all robots in the distributed team be able to communicate with all others at all times. Due to the mobile nature of each robot in the team, it may be that a robot’s movement needs to be limited in order to maintain the link between the other members of the team. This problem has been addressed in research in a number of ways. One way is to define a set of robots to be the network nodes, and have their only goal to keep the network available. This is undesirable because it sets aside resources which could otherwise be used to perform tasks. Another option is to compute the connectivity for the entire group and detect disconnects, and make movements to reestablish connectivity.
- *Mobility Methods* – The mobility of the robot platform is one dimension that should be considered as well, especially if the robot team is going to be deployed in a real-world environment or a simulated real-world environment [25]. Larger wheel bases and stronger motors are required in these situations to effectively navigate, but are at odds with a small robot design and require more energy for movement. Some type of dead-

reckoning sensors may also be included in the motors allowing for localization. While wheels are not exclusively used in mobile robotics [1,45,47], they are the most prevalent. Stairs and multi-level obstacles are also a mobility consideration, as they are quite common in real-world environments. There are some interesting approaches that have been presented in this area, including wheels that change their size [14].

- *Sensing Capabilities* – Finally, sensing is a key component of any robot's design. Many sensors which are high-fidelity such as laser range-finders, sonar sensors, or camera platforms are also fairly large in size. Some locomotion sensing would be desired for the locomotion, although recent advances in visual odometry [9,34,43] may be used. Sensors also require power, meaning that a large number of sensors may reduce the run time significantly. Sensor expansion is an important design consideration as well. Processing of sensor data may require significant computational power. This computational power is yet another trade-off. If it is on-board, it may require a larger chassis or more energy reserves. If it is done off-board, then there is more reliance on network connectivity and bandwidth.

The Scout 2000, shown in Fig. 1, is a small robotic platform designed for distributed robotics. At 40 mm in diameter and 110 mm in length, the Scout was compact and had little volume for expensive computation or sensing capabilities. The speed of the platform is minimal compared to other larger platforms. However, it also contained an innovative jumping mechanism which allows it to climb stairs [57] which increased both its mobility and cost. The Scout 2000's primary locomotion was accomplished using two wheels, although it could be ballistically deployed from a larger robot, the Ranger. These two robots are

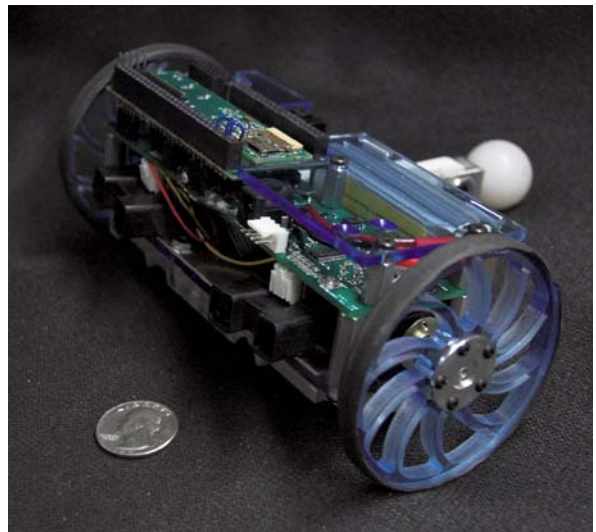


Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 1
A Scout 2000 Robot

part of the Scout team [20]. There were many variations on the Scout platform, designed to give teams of Scouts increased usefulness through innovative sensing and locomotion methods. Some of these methods included the “Actuating Wheel Scout” which was a specially modified Scout that had the ability to increase its wheel diameter from 40 mm to around 120 mm. Other variations included a deployable grappling hook, a miniature pan-tilt camera head, geiger counters, or IR emitters which could illuminate a room for the whole team. The small size of the Scout allowed only minimal onboard processing —instead, relying on a wireless video broadcasts to be processed by a remote monitoring station. Coordination of teams of Scouts was accomplished using a CORBA-based architecture [37] which would dynamically load balance mission objectives across a number of systems.

The Explorer robots, shown in Fig. 2 are a newer evolution of the Scout 2000 robot platform which is designed primarily for educational use and research purposes. They are designed with expandability in mind as well as communication and mobility.

The Explorer includes a larger amount of processing in the form of an ARM processor, and expansions for more sensing connections to the device. It also has an articulated tail in order to change the angle of sensors mounted on-board the robot. The Explorer can broadcast analog video as well as utilize an integrated Bluetooth device, allowing Explorer robots to relay information between robots. While larger than the Scout 2000, the Explorer is still small



Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 2
An Explorer robot



Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 3

A TiTAN robot

at 17 cm long and 7 cm in diameter. The Bluetooth compatibility also allows control of the Explorer through standards compliant devices such as the Nintendo Wii-mote.

The TiTAN, pictured in Fig. 3, is a platform developed between the Scout 2000 and the Explorer robots. It is of significantly larger size, sacrificing the small form factor for more features in all of the other dimensions of robot design. TiTAN robots measure 37 cm long and have a 13 cm diameter, 35 times the volume of the original Scout 2000, and 7 times the Explorer robots. In exchange for this increased size, the TiTAN has significantly better sensing, allowing for swappable bays of sensors which include independent tilt control. The increased size allows for payloads on the back and top.

Mobility is also increased, as the much more powerful motors of the TiTAN allow it to travel at almost 6 m/s over flat terrain, and even handle towing a payloads in excess of 100 kg. The TiTAN, like the Scout 2000, also has a set of actuating wheels which increase its wheel base and allow it to traverse over rocks and obstacles which would otherwise stop it due to its low ground clearance. Communication is achieved through 802.11 wifi with a diversity antenna system which provide significantly longer range than bluetooth both indoors and outside. TiTANs also have an optional 900 MHz antenna, allowing them to receive video from the Scout 2000 and other robots on this band. Processing onboard the TiTAN uses a Pentium based PC104 stack. The selection of this stack enables onboard vision processing and a path for future upgrades.

Robotic Team Composition

Whether the team will be homogeneous is often the first question when presented with designing a distributed team. The answer will impact what can actually be done with the team, as many task-allocation methods and distributed robotics methods require the use of a team which is homogeneous. Having a homogeneous team may

make tasks more efficient, as any robot's sensing information can be substituted, and information sharing between robots may be more prevalent. Having a homogeneous team also makes it easier to keep a team of robots running – extra robots can be kept to be used while some are being fixed.

Heterogeneous teams, alternately, may be easier and cheaper to create, as existing robots that are used for other purposes can simply be combined into a single robot team. Such teams may also include multiple sensing modalities for a single area, enhancing the ability of the entire team by sharing information between members. Differing sensors on each member of the team may also be an advantage if one sensor is cost-prohibitive but may be only needed in some areas of an environment.

Having a heterogeneous team may also enable marsupial team mechanics, allowing for the benefits of small robotic platforms as well as large robotic platforms. The smaller, carried robots of a marsupial team generally have a smaller footprint and can traverse smaller confined spaces. They are also usually as many or more carried robots as carrier robots. This enables the task at the mission site to be completed in a more timely fashion than would be attainable. The smaller robot teams are slower over long distances as well, which can cause problems for very distant mission areas. Because they are limited in size, they also carry smaller energy stores and may not be able to even complete a round trip to the mission area and back without assistance.

In contrast, the larger robots in a marsupial team can travel very long distances due to their increased wheel base, and may be able to traverse more difficult terrain. Carrying the smaller robots, in total the energy cost of the entire mission may be lowered because of the efficiency of the carrier robot. Additionally, the marsupial device may allow for the recharging of the energy of the smaller carried systems, allowing for a much longer lifetime of the entire system. Having extra space and operating power can also be useful in processing the large amount of data. Powerful computing resources on the larger carrier robot may be able to coordinate communication or store large amounts of sensor data, and also provide a general overview of the scene. More costly and high-fidelity sensors as well as sensors with a higher demand on computing resources may be kept on the carrier robot. Another option is for the larger robot to relay communication back to an operator which would not be in the range of the smaller carried robots' communication systems.

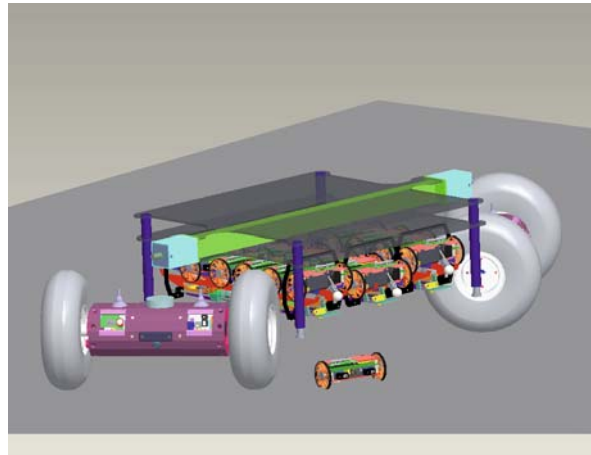
Marsupial teams have been researched by many universities and experiments have been performed. One of the earliest examples of a marsupial team are the team of Sil-

ver Bullet and Bujold [40], which was used to evaluate the usefulness of a marsupial team. In this team, Bujold was tethered to Silver Bullet for all of its power needs, significantly reducing the ability of the marsupial pair. Further research was done with the Yellowjacket team [13], which used a static arm mounted on the marsupial carrier, allowing for the carried robots to dock and undock themselves. It allowed for six carried robots to be docked on the carrier at once.

Marsupial teams need to deal with challenges which are not present in either of the single robot systems alone. The most important of these problems is the storage and deployment process. Some marsupial pairs handle this issue by placing a beacon of some sort on the carrier robot, allowing it to be recognized and docked with. There are many examples of carrier robots which use vision along with markers on the carrier robot in order to either line up or dock [38]. A team of Scout 2000 robots was used in at the University of Minnesota which took advantage of their unique jumping ability to both store and deploy themselves onto a carrier robot [27]. More recent work focuses on using an earlier version of the Explorer robot in order to dock with a prototype robot transportation mechanism referred to as the “Saddlepack”. The Explorer-like robots use orange markers which are situated around the location which needs to be approached in order to store the robot in order to position itself in the correct orientation and approach the docking station. Once close enough, dead reckoning can be used to drive onto the forklift and be stored [38].

One example of a complete marsupial system is the Saddlepack system developed by the University of Minnesota. The SaddlePack is a modular system of interconnected robotic platforms. Three basic building blocks make up the SaddlePack system; two relatively large, two-wheeled robotic platforms, the TiTANs, which cooperatively maneuver the system, and a single immobile superstructure that suspends between the TiTANs and serves as communication base, power reserve, and attachment point for modular systems. The TiTANs themselves are actually specially-equipped TiTAN robots. A concept drawing of the SaddlePack can be seen in Fig. 4.

The SaddlePack’s ability to support modular functional devices is an expansion of capability. The first of these modular components to be developed is a Modular Robotic Docking Station, MRDS, which is capable of carrying, deploying, retrieving and re-energizing up to six relatively small robots, the Explorers, for expanded search and explore missions at a lesser energy use rate. The SaddlePack can hold up to three of these stations for a total of eighteen extra Explorers for added utility.



Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 4
Concept drawing of the SaddlePack

Ideally, the SaddlePack system will operate autonomously for a minimum of 10 hours due to its large battery store. This means that it must be capable of continuous movement, communication, and computation for that time. The docking station must also have adequate battery power for recharging the deployable systems within the Modular Robotic Docking Station and any future modular functional unit demands.

The docking station itself lacks mobility. The design of this system allows for two pre-existing robots, the TiTANs, to dock with the docking station. This mechanism is general enough that a variety of robots can provide the requisite mobility. The MRDSs are designed to lift robots to and from the ground plane, which is 10 cm below the base of the station to allow for ground clearance, and shuffle the robots within the secure space to replenish their power supply and transport them efficiently between mission areas.

The Saddlepack system is a good example of a marsupial robotic system which includes both the smaller Explorer robots which can be deployed and retrieved by the MRDS. It also includes the possibility of additional mobility of the TiTAN robots which carry the MRDS to the mission area and have a large number of sensors and processing power. This provides in whole a large number of different mobility scenarios. The robots within the MRDS can be carried to the mission zone by the TiTANs, where they can then be deployed to complete tasks which require mobility in cramped spaces. The TiTANs themselves can detach and use their higher processing power and sensing capabilities to further explore the environment. In addi-

tion, the MRDS acts as a relay point. The coordination of all of the separate members of this robotic team is a daunting task.

The Saddlepack concept's core tenet is modularity. The individual MRDS system can be transported by the TiTANs or mounted to a number of other larger commercially available or custom platforms. This enables flexible design and reconfiguration of the team. For example, a modified iRobot Packbot could be used to transport an MRDS system up a flight of stairs, a task that could not be completed with the TiTANs.

Software Modeling and Simulation

Simulating distributed robotics has its own unique set of challenges and advantages. Some of the advantages of simulation are particularly useful in distributed robotics where the complexity of experiments increases. However, simulating such experiments also includes challenges that are not present when simulating a single robot platform in an environment. In addition, simulating marsupial systems presents its own set of unique issues in addition to the issues caused by simulating a multi-robot system.

Simulation is used extensively by robotics researchers for a variety of reasons. First, simulation reduces the cost of algorithmic development for distributed robotic teams. Additionally, simulation removes many of the issues associated with maintaining hardware, noise in sensors, difficult environments, and presents an ideal world that allows for a "theoretically fair" comparison of approaches. It is even more appealing in the field of distributed robotics – the addition of multiple other robots that now must all work simultaneously causes maintenance to take even more time than with single robot platforms. The time for setup of experiments and programming of the robot systems is also multiplied by additional robots. In light of these expanded issues, it is reasonable to be drawn to the use of software modeling tools.

Distributed robotics brings its own set of challenges to modeling. Communication, which is nonexistent with a monolithic robot, is very important when working with multiple coordinated robots. In addition, software modeling of wireless communication is especially tricky due to the physical properties of wireless signals. Many simulation systems do not include a communication component, and many papers may simply assume the existence of a network which can communicate a set of shared data, which can cause distributed behaviors that work in simulation to have unexpected issues in real-world environments. Even in this case, the complexity required to exchange data between multiple running programs may not

be reduced in a simulation environment. Along with these modeling issues, simulation software must be made more complex in order to handle multiple robots at once, possibly including concurrent programming when none was used before. As the number of robots used in simulation is raised, the computing power required to simulate the aforementioned robots in an environment increases minimally in a linear fashion. This makes it difficult to simulate large numbers of robots such as those deployed in a large building or over a large area.

There are a number of software packages available for researchers to use. One of the most prevalent in use today is the Player/Stage software [8], which is in active development. Player presents a set of interfaces to the programmer of a robot, which can then be used to control a variety of robots which would support that interface. The `position2d` interface, for example, represents a robot's position on a 2d plane, which can then be manipulated and updated through commands to the interface. The commands are then translated by the Player server software, contacting the driver for the specific hardware the server is running on in order to move the robot. The same program can be used unchanged for any robot which supports the `position2d` interface. In addition to the Player interfaces, a server which represents a simulated set of robots and environment exists in the Stage software. The Stage software simulates 2 1/2 dimensions in relatively low fidelity. This level of fidelity is used specifically in order to lower the amount of processing needed in order to simulate large numbers of robots at once. In this way a robotics researcher can write a Player client program and test it in the Stage simulator before running it on the actual hardware. The Player/Stage software is also easy to extend due to its open source, allowing specific drivers to be written for custom devices.

Player also can communicate with the Gazebo simulator, which provides a full 3-D environment simulation for robots. This provides a higher fidelity than the Stage software, allowing for much more complex interactions. Unfortunately, it also takes much more computing resources than Gazebo in order to perform its physics calculations. It also allows for marsupial actions to occur even without explicit support in the simulator – marsupial carriers actually physically pick up the carried robots and the interactions are modeled through the simulated physics. As the simulator also connects with Stage, the same programs which are programmed for Gazebo or real robots can be used when connecting to Gazebo. Gazebo has a level of complexity which makes it harder to create new simulated devices, but it is still open source and therefore extendable by the programmers.

The Microsoft Robotics Studio (MSRS) was introduced in 2007, and it also contains a simulation system. The simulation used in MSRS is of much higher fidelity than the Stage simulation software, including a full 3-D physics simulation of the environment. MSRS uses a commercial physics simulator which is highly accurate. This increases the accuracy of the situation but also causes a significant increase in computing resources used for the simulation similar to Gazebo. However, specialized hardware is available which the commercial simulator can take advantage of. This may allow for a larger number of objects to be simulated and make the MSRS more suitable to simulating robotic teams. The MSRS also presents a set of interfaces to the programmer through its service-based architecture, but they currently are at a much lower level of abstraction than the Player interfaces. Users can build their own services on top of existing services, as well as program new drivers which provide access through existing services. MSRS communicates through a REST-like HTTP-based protocol.

These are only three of the many simulation packages used for robotics available. They are the most popular, in part because they are both free to use. A number of other simulation systems are used as well, with various advantages and disadvantages [11]. Some researchers also prefer instead to create simple simulations using the MATLAB commercial mathematical programming language.

Marsupial Modeling

Marsupial systems present all of the challenges present in distributed robot simulation, but also have unique challenges which arise when coordinating two robots. Each marsupial system handles the retrieval and deployment of the stored robots in a different way, causing problems when trying to create a generalized solution to a problem in which marsupials are present. It is desirable then to create a high-level abstraction which can be used to deploy and store robots in a marsupial situation. We present one such model in this section.

There are very few robotic simulation systems which include marsupial actions. The easiest example from this class of simulators are full physics simulators such as Gazebo [8] and the one included in the MSRS. However, in current simulators with physics, there are still no high-level marsupial abstractions – the entire sequence of manipulators or motors would have to be activated individually in sequence each time a marsupial action was desired. In addition to these issues involving the modeling of a marsupial mechanism, the implementation of full physics simulation requires much more computing power

than the relatively low fidelity environment provided by software such as Stage. This can make it difficult and sometimes impossible to simulate marsupial systems, which rely on the physics of the manipulators in order to store and retrieve robots.

When considering a general model for marsupial mechanisms, very few mechanical considerations need to be made. It should be general enough to possibly describe many different systems while being specific to marsupial actions, as opposed to movement, manipulation, or general sensing actions. While it may be the case that zero or more actuators are used to perform a marsupial action, the desired action is more high-level. Simulating this general device allows us to separate the marsupial actions from actual hardware, removing some of the difficulties in performing marsupial actions. At the same time, it allows for the desirable portability of algorithms by providing an abstraction that may be used by multiple different marsupial hardware systems.

The level of abstraction present here is fairly high, so there are few actions which should be possible for the marsupial abstraction. The first and most important action defines the group of mechanisms, that is *storing* and *deploying*. In our abstract model, we do not care how this is accomplished, only that it can occur. Therefore we consider two states of the model: it is either *open*, ready to store or deploy, or *closed* and it is not ready.

Because we are not interested in the specifics of storage, we focus on what can be stored. All physical systems have a maximum size of object which can be stored due to the physical limitations of the device that we name its *door size*. Objects larger than the door size cannot be stored by the mechanism. Marsupial mechanisms also store a limited amount of objects due to physical constraints on the storage area, which we call the *storage capacity*. Some systems also can sense how much of the storage capacity is currently occupied.

This general model for a marsupial mechanism because it covers a wide range of possible marsupial devices. By limiting the requirements for a marsupial system, we expand the set of devices which can be represented, hoping to impose nothing beyond the definition of a marsupial robotic system. This model was implemented in the Player/Stage software [26] and is currently being used to simulate the prototype SaddlePack marsupial system described in Sect. “[Robotic Team Composition](#)”.

Tasks Well-Suited to Distributed Robotics

There are a number of tasks which are well-suited to a distributed robot team. Some of these tasks may be possi-

ble to complete with a static deployment of sensors, but they may be suboptimal. Others may be able to be performed in a shorter amount of time with a set of distributed robots, but the addition of more than one robot may require communication throughout the team or distribution of the work load, introducing new issues which must be addressed.

Here we discuss two different tasks which can be accomplished without a distributed robot team. The first is mapping or coverage of an area. These two problems are considered together because they are very closely related. Mapping and coverage are both completable by a single mobile robot, but utilizing a distributed robot team potentially decreases the time to complete both of these tasks. The other task which is considered is optimal placement of sensors with respect to observing activity. This task is possible to complete by placing static sensors, but static sensors cannot respond to dynamic changes in the activity being observed. When considering either of these tasks, we propose improvements and challenges which must be considered specifically for a marsupial team as well.

Mapping/Coverage

Mapping is particularly well suited to distributed robotics, and has been well researched in the field [10,32,48]. Mapping is of particular use in situations where the environment has recently changed and an exhaustive search is desired, for example with in a collapsed building or in an unknown structure. It is generally considered a basic problem in the field of robotics. A distributed robot team can finish the task in a shorter amount of time by either coordination of search areas, or by a distributed behavior which naturally causes the team to explore different areas independently.

Using a distributed team also brings new challenges to the table when mapping. Because multiple robots and therefore multiple sensors are now being used to sense the environment, some type of relationship between those sensors is needed in order to combine the maps into a full map is required. This is generally referred to as the *map merging* problem. Also, exploring an area to be mapped now also contains the other members of the team themselves, which can corrupt the map with spurious obstacles. The problem of conflict management must also be addressed, as robots may need to travel to different areas but may block each other in doing so – consider the case where one robot needs to search a room at the end of a hallway, and another robot is exiting the same hallway. The methods used must employ a protocol to resolve these deadlocks.

In the most basic type of method for distributed mapping, the occupancy grid which is used in mapping with a single robot is used and simply extended to the multi-robot situation. Assuming that global communication is available, the global map with the occupancy grid can be filled in independently by each of the exploring robots and each robot can select the closest longest frontier in order to expand the amount of cells which are explored. This method has problems if the robots start in a similar location and orientation, because the same frontier cells will be chosen by each robot. This may cause the exploration to take the same amount of time as a single robot which is undesirable. Some methods can be used to introduce randomness or choose target frontiers in a different way to reduce this behavior, but it cannot be removed without communication of the target frontiers also. This method does not scale well because it requires communication of the current task at regular intervals which will eventually overload the communications medium if enough robots are added to the team.

Methods which communicate the frontiers which are being traversed have been improved upon from there. Maximizing the information gain of the relative robot's assigned frontiers is one improvement that can be made [6]. The relative robot localization loss in each area is considered [30] in order to facilitate map merging at the end of the method. Limitations can also be added to improve localization, such as incrementally approaching the frontier in order to maintain visual contact [21]. Task assignment methods such as auctioning and market economies can be used as well [60].

Another type of mapping which is considered for single robots is the semantic place map. This type of map usually doesn't take the form of a traditional map, but is more similar to a subway or large train system map in that it conveys the connections between places which can be easily recognized. This type of map is usually built specific to a sensor such as a omnidirectional camera or a high-resolution laser range-finder which has a high density of data. This high volume of data is required in order to sufficiently disambiguate the specific locations in the case of highly similar locations. The locations are connected in a directed graph representing the travel required by a robot in order to travel from one place to another. This method is extended to the distributed robotic team by ensuring that the required sensor for identifying the nodes of the semantic map is present on all of the members of the team. As such, this method is more well-suited to a homogeneous team of robots than a heterogeneous team. The information required in order to recognize each place is then distributed throughout the team in order to

facilitate localization within the semantic map. Some issues may persist if the locations are very similar to each other, but they can be disambiguated at the very worst by their location within the map itself. This may make localization of a new member of the team a longer process.

Recent developments in distributed exploration involve the expansion of random trees of traversable areas and combining and optimizing the trees instead of explicitly mapping the environment. This results in a map which is not only easily usable by the humans observing the exploration, but also by the robots themselves in order to use later for path planning and safe traversal of the mission space. The methods which are in use are based on Rapidly-expanding Random Trees, with new research being done in Sensor-based Random Trees (SRT) [16]. These methods were first developed for single robot exploration, but have been parallelized in order to increase the efficiency of the search by cooperation and modified in order to avoid conflicts between differing members of the robot team. These methods also employ a frontier region similar to the occupancy grid methods, with similar selection problems being handled in much the same way as those methods. The robots build a SRT and then make that information available for others to use while creating their own trees. When the method completes, the trees can be combined into a usable map for other robotic teams which may need to explore the area.

While distributed mapping may construct the map after the exploration is complete, simultaneous localization and mapping (SLAM) methods construct the map while the robot is exploring the area. There are many different approaches to SLAM, and some have been extended to work in multi-robot situation in real time. Some methods are based on the initial work [15] simply combine all landmarks and localization together assuming adequate communication between all robots. This causes a problem as the set of landmarks increases in a large environment, because as some landmarks are removed in order to lower the complexity of the computation required, the landmarks may be sparse and a robot may be out of range. It also does not scale well to a large number of robots. Other methods focus on improving the local SLAM methods by using the other robots in the team as landmarks that are mobile. The simplest way to extend the traditional SLAM method [33] is to only have a single robot move at once, using the stationary robot as a landmark. This again is not reasonable in situations where the robot team is large, and some methods that allow for multiple or all robots in the team to move have been developed [50].

Search is a problem closely related to mapping, but not exactly the same. While mapping focuses on sensing one area of an environment at least once, coverage focuses on sensing over all of the environment. This may be performed in an unknown environment with a single robot, and can be thought of as a search method which will eventually find an object no matter where it is located in the environment. Coverage can be used trivially to construct a map such as the ones discussed with occupancy grids, but a search task may not require the creation of a map, and a complete map may not be sufficient to guarantee that an item searched for may be found. Coverage may also be thought of in terms of painting where a robot is required to paint an entire surface of an object. Optimizing the coverage results in a robot which crosses its path a minimal number of times.

Interestingly, communication is not a required element for many search strategies. Some methods [36] use potential fields along with recognition of other robots in order to search effectively areas that have not been searched by having other robots and obstacles repel the robot and the frontier of the search attract the robot. However, this method itself acknowledges that the lack of communication degrades the performance of the search activity. Additionally, implicit cooperation methods are used, with some shared state which is discernible through local sensors. Some implicit communication involves using environmental markers such as those used by ants while exploring [31]. Random strategies can also be used within a large team of robots to perform search, and can be shown to be highly scalable without issues to run on a large team of robots [18].

Communication has been shown to increase the performance of search tasks however [51], so utilizing the communication available is desired. Simple methods use a global map or grid which can be communicated to all team members when areas have been searched [59]. The grid size of completed work may be at a much larger granularity as an occupancy grid, and some methods allow for two or more robots to cooperate in coverage of a single cell [24]. Yet another method [55] allows the robot to assign tasks for search only to itself as it discovers new areas, but other robots may be able to remove a task if the area has already been searched. This limited communication allows for robots to make search choices in parallel independently, increasing the scalability of this approach.

These limited communication methods may eliminate some duplicate coverage, but they do not completely eliminate duplicate coverage. More complex methods will share the current task which is being completed as well. This allows for each area of a map to be searched only once.

An early way of doing this was to assign a global coordinator which would collect the possible actions of each robot in the team, and assign each task to each robot in order to maximize global utility [54]. Recent advances however favor a distributed approach, with tasks being generated when new areas are discovered, and can be completed by any robot. This may take the form of an assistance request [17] when the area is particularly large, or auctioning using a method as described in Sect. “[Team Coordination/Task Allocation](#)”. This task allocation process lowers the amount of communication required in contrast to the coordinator method which requires state and intent information to be communicated continuously. However, many of these methods do not scale well with the size of the robot team because of the large amount of communication needed. Some researchers are addressing this problem by improving on the current strategies by lowering the amount of communication which is required for each of them. One such improvement [49] involves the robots agreeing before the search starts on a strategy for search if communication is lost or is otherwise unavailable due to traffic constraints.

Another possibility which can be considered is static coverage. This is when the entire area is covered by the sensors of at least one robot. It is related in some ways to the art gallery problem [44]. This is a task which is almost impossible to consider for a single robot – it is specifically the domain of distributed robotics. Static coverage is useful in real-world situations for expanding a network of communication to its largest available network range while continuing to provide routing ability even in situations with failing sensors. It is also useful when the environment of the robots is changing rapidly and therefore must be mapped at a frequency which is higher. The placement of robots evenly across an entire environment also allows the completion time of randomly placed tasks to be lower than if the set of robots were initially concentrated on a single point.

Static coverage can also be achieved without communication by utilizing a potential field which repels the robots from each other and from large obstacles such as walls. This results in the robot team being spread throughout the environment. If the environment is a priori known, a coordinator robot may place the location of the robots ahead of time in order to maximize static coverage. The placement methods which have been developed for sensor networks are analogous to the placement needed here.

It is interesting to consider the implications that a marsupial team may have on exploration and search methods. Exploration may be enhanced specifically by a mar-

supial robotic team. One method would be similar to the centralized assignment method, with the marsupial carrier robot coordinating the actions of the smaller robots in order to efficiently explore the frontiers of the region. This method would effectively let the smaller carried robots explore the area faster than possible before because of the increased mobility of the carrier robots. An extension of this method may shift the control of the method to the carrier robots, having the small robots request a marsupial robot in order to travel distances which would take a long amount of time when using normal locomotion. This would allow any of the distributed exploration or coverage methods to be used without change on the network of carried robots, simply increasing the mobility over long distances.

Another method could be considered which attempts to center the larger marsupial carrier in the space and then a deployment phase happens, and the carried robots carry out their own exploration of the room, returning to the carrier robot when finished. This process would then be repeated in other spaces, or in parallel by multiple carriers with the set of robots they are carrying. In this method, the carrier robot may be used as a landmark for localization, reducing the total error in the map and decreasing the computational requirements for map merging. It is also possible that a higher-level survey of a room may be acquired when the marsupial carrier is entering the room, and subsequently the smaller carried robots would increase the information density in the map acquired by filling in the smaller spaces which can be sensed by a smaller platform.

The marsupial team also brings with it challenges which are not present in the general distributed robot exploration problem. The differing sensors on the carrier and carried robots in a marsupial pair must be considered in an extension of these methods to a marsupial team. The smaller robots also have a level of resolution for mapping and localization which varies by a large amount compared to the larger carrier robots. This may be exploited as in the second method as described above. In addition, the sensor levels of the smaller robots will be at a lower level than the carrier robots. This may be an advantage because two differing heights of maps can be created, but it may also be a detriment – maps which signal the openness of a particular spot for a carried robot may not be accessible for a larger robot. A good example of an obstacle which would pose a risk for the carrier robot but not for the smaller carried robots is a coffee table, which has a wide open space underneath and a solid top. This would not be traversable for the larger carrier robot but would be easy to navigate for the smaller robots.

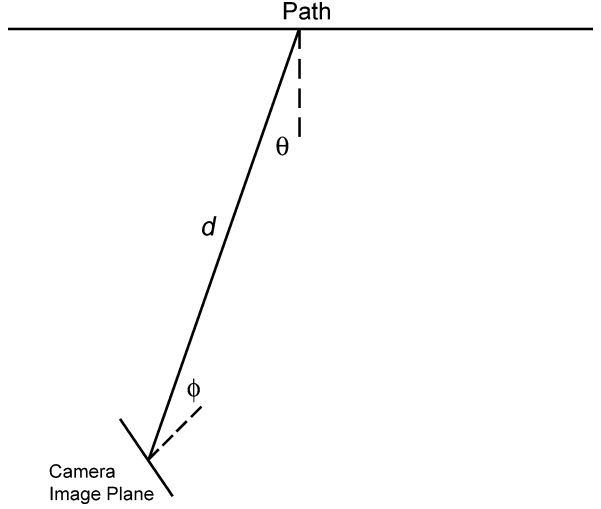
Coordinating Teams to Optimize Observability of a Situation

Sensor networks are useful for monitoring an area for a variety of purposes. Static camera networks are utilized for monitoring traffic or security applications. Chemical sensors may be deployed throughout a production facility to monitor for chemical leaks. However, the effectiveness of the sensors is a function of how well placed they are. Traditionally, these problems can be characterized as variations on the “Art Gallery Problem” [44] where the task is to determine the number of cameras necessary and where to place them in order to monitor art exhibits in a gallery.

However, in dynamic situations, statically placed networks may miss critical pieces of information. This is especially critical when utilizing directional sensors such as cameras. Camera systems can be utilized for automated facial recognition purposes, human activity monitoring [3], and abandoned object detection [2]. However, if the cameras are not deployed in an effective configuration, the systems will operate poorly, potentially misclassifying activity in a scene or missing an abandoned object.

Distributed robotic teams offer an alternative to solving this type of camera placement problem. Robotic teams offer the ability to relocate the cameras as the situation changes, thus providing a continuously optimizing observation of the situation. The selection of the locations for the cameras first requires monitoring of movement within the scene. When the goal of the observation is to characterize human motion (walking, running, jumping, etc.), the best vantage points for the cameras may be perpendicular to the motion of the individuals in the scene. This can be complicated by the number of individuals moving throughout the scene in a non-uniform fashion. When the goal of the system is facial recognition, the cameras must be relocated to observe the trajectories of individuals moving along the focal axis of the camera.

The framework developed by Bodor et al. [4] offers one method of solving the camera placement problem with a distributed robotic team. In this approach, the goal is to place the robots perpendicular to the path that individuals moving throughout a scene take. However, in observing the path, there are additional constraints, first that the entire path be in the field of view, and second that the camera be as close as possible to the subject. In a perspective projection camera with a fixed focal length, the size of an object is proportional to the distance from the object. By placing the camera as close as possible, the object being observed is larger, and thus more information can be gained. A minimum distance d_0 can be determined based upon the characteristics of the camera and the paths that



Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 5
Angles to minimize for each path

are being monitored. The distance between the i th camera and the j th path can be given by d_{ij} and must be at least d_0 away from any path in the distribution being observed.

In two dimensional space, the angle ϕ can be used to represent the angle between the focal axis of the camera and the line connecting the center of the camera’s image plane to the midpoint of the j th path (X_{midp_j} , Y_{midp_j}) while θ represents the angle between a perpendicular axis run through the midpoint of the path and the image plane of the camera. These angles are depicted in Fig. 5. Considering the entire path distribution and the multi-camera system, this results in the following function which must be optimized with respect to the camera placement and orientation:

$$O_P = \sum_i^{\text{cams}} \sum_j^{\text{paths}} \frac{d_0^2}{d_{ij}^2} \cos(\theta_{ij}) \cos(\phi_{ij}) . \quad (1)$$

The individual robots are able to relocate themselves to the positions resulting from the optimization. These positions are derived by changing the variables from (d, θ, ϕ) to absolute positioning and orientation of the camera and path in the real-world (X_{cam} , Y_{cam} , pan_{cam}) using Eqs. (2), (3), and (4) for each camera-path pair. When Eqs. (2), (3), and (4) are substituted into (1), the value of the observability objective function can be determined for a given camera position and pose in world coordinates:

$$d_{ij} = \sqrt{(X_{\text{midp}_j} - X_{\text{cam}_i})^2 + (Y_{\text{midp}_j} - Y_{\text{cam}_i})^2} \quad (2)$$

$$\theta_{ij} = \cos^{-1} \left(\frac{T_1 + T_2}{T_3 \cdot T_4} \right) \quad (3)$$

where

$$\begin{aligned} T_1 &= (Y_{\text{startp}_j} - Y_{\text{midp}_j}) (X_{\text{midp}_j} - X_{\text{cam}_i}) \\ T_2 &= (X_{\text{startp}_j} - X_{\text{midp}_j}) (Y_{\text{midp}_j} - Y_{\text{cam}_i}) \\ T_3 &= \sqrt{(X_{\text{midp}_j} - X_{\text{startp}_j})^2 + (Y_{\text{midp}_j} - Y_{\text{startp}_j})^2} \\ T_4 &= \sqrt{(X_{\text{midp}_j} - X_{\text{cam}_i})^2 + (Y_{\text{midp}_j} - Y_{\text{cam}_i})^2} \\ \phi_{ij} &= \cos^{-1} \left(\frac{0.167 [P_1 - P_2]}{P_3 \cdot P_4} \right) \end{aligned} \quad (4)$$

with

$$\begin{aligned} P_1 &= (Y_{\text{midp}_j} - Y_{\text{cam}_i}) \cos(\text{pan}_{\text{cam}_i}) \\ P_2 &= (X_{\text{midp}_j} - X_{\text{cam}_i}) \sin(\text{pan}_{\text{cam}_i}) \\ P_3 &= \sqrt{0.028 [\cos^2(\text{pan}_{\text{cam}_i}) + \sin^2(\text{pan}_{\text{cam}_i})]} \\ P_4 &= \sqrt{(X_{\text{midp}_j} - X_{\text{cam}_i})^2 + (Y_{\text{midp}_j} - Y_{\text{cam}_i})^2}. \end{aligned}$$

Methods of constrained nonlinear optimization allow for the parameters $(X_{\text{cam}}, Y_{\text{cam}}, \text{pan}_{\text{cam}})$ to be derived. This framework can be extended to operation in three dimensions by including two additional angles in the optimization.

Teams of robots can also be used to identify the sources of potential biological, chemical, or radiological contaminants. In [7], a framework is developed where a team of distributed robots can be deployed to explore an environment. As the robots explore the environment, the individual robots take readings of airborne chemical contaminants. These readings are used to estimate the source location and quantity of the chemical. In doing so, these measurements can be used to predict where the plume will travel in order to warn and evacuate those that may be in the path of the plume.

Tasks for Optimizing Distributed Robotics

While many tasks can benefit from a distributed robotic team, the team itself introduces issues and optimizations must be tackled. Many of these tasks are directly related to other fields, such as the issues of communication and routing through a set of mobile communication nodes dispersed throughout an environment. We discuss two tasks which arise from the use of a distributed robotic team. The first is related to the longevity of a robotic team and

utilizes the specific properties of a marsupial team in order to achieve the optimization of the team. The other is task allocation, which is related to work in the intelligent agents domain in auctions. The impact of task assignment is considered in the context of marsupial distributed robotic teams as well.

Coordinating Teams to Maximize Longevity

Extended coverage of an operational environment by a distributed robotic team is essential in minimizing risks to humans that have to enter a hazardous area. In order to do this, the robots in the area must have access to sufficient power to remain functional for the mission duration. While this could be as simple as integrating a larger battery system, batteries still offer only a finite power supply and it may be impossible to predict how long a robotic team may be required. There are three options for distributed teams of this nature to remain effective:

1. *Infinite Power* – One way to avoid this difficulty is to utilize tethering to provide “infinite power.” This option is attractive for long term deployment, however, tethers can be difficult to manage, require extra coordination/larger motors/power expenditure for movement and relocation, and can snag on objects in the environment trapping the robot in a specific place.
2. *Energy from the Environment* – In extreme conditions, it is necessary for the robot to be designed such that it is able to extract all necessary energy from the environment. This is very common in cases of interplanetary exploration. Here, solar panels are utilized to obtain energy for recharging on board battery systems. This is extremely effective in these scenarios, but may not be effective when responding to hazardous incidents where chemicals in the air, ash from fires, or other debris may coat and block solar panels. Alternatively, robots such as Slugbot [29] are able to extract energy from the environment by identifying sources of sugar which it can use to create a chemical reaction to generate power. This approach while interesting, also has limitations, in that energy is expended foraging for sugar and sugar must be deployed in the environment.
3. *Resupply* – The third approach for extending operational longevity is for the robots to autonomously resupply themselves. This can be done by autonomously docking with fixed charging stations or with mobile docking stations. Fixed docking stations can reduce the complexity of the problem and are found in a number of environments [19]. However, when responding to emergency situations, it is unlikely that there will ex-

ist fixed docking stations, and if they are present, that they will be functional.

Given that the distributed robotic team must be able to operate in arbitrary environments, the problem of longevity can be reduced to the following. Given n distributed robots and d mobile docking stations, how can those d mobile docking stations be distributed in order to maximize the operational longevity of the n robots?

In order to solve this problem, first we define a robot as a system that can be represented as a location (X, Y) , orientation θ , velocity V and available energy E . A robot's position at a given time (t) can be represented by the following equations:

$$X(t) = X(t - \delta t) + V\delta t \cos(\theta) \quad (5)$$

$$Y(t) = Y(t - \delta t) + V\delta t \sin(\theta) . \quad (6)$$

Thus, the possible locations the robot could be at in a given time can be represented by a circle centered at the robot, with a radius equal to the distance travelled (velocity \times time). A mobile docking station utilizes a similar representation to determine where it can be within a given time. The problem then becomes an optimization one. Where can the docking station move so that the most robots in need of power can arrive there at a minimum power expenditure?

Each robot R_i , maintains an estimate for its position $(R_{iX}(t), R_{iY}(t))$, velocity R_{iV} , and battery life R_{iE} . The cost to return to a docking station D_i located $(D_{iX}(t), D_{iY}(t))$ at time t can be given by the following equation:

$$C(R_i(t), D(t)) = xe^{x-f} \quad (7)$$

where:

$$x = \frac{d_i}{R_{iV}}$$

$$d_i = \text{dist}(R_i(t), D_i(t))$$

$$f = 1 .$$

Here dist is assumed to be the Euclidean distance, however dist can be any distance function which in general will be the shortest known traversable path.

Let S_r represent the set of robots that are in need of recharge. There is a number of ways in which this set can be derived. The most straightforward way is for each robot to maintain an estimate of the location of the nearest docking stations that is periodically relayed through a wireless network. The robots can then estimate the amount of energy that is required to return to a docking station. As their onboard reserve is reduced, they can periodically broad-

cast a message that they are reaching a critical level of power.

This reduces the problem to identifying which movements of the docking station will minimize Eq. (8) at time $t + \delta t$:

$$\sum_{i \in S_r} C(R_i(t), [D_X(t) + D_V\delta t \cos(D_\theta) , D_Y(t) + D_V\delta t \sin(D_\theta)]) . \quad (8)$$

However, simply minimizing this function will result in a global solution that may not be optimal. In order to increase the utility of this function, a clustering method is applied to the members of S_r in order to divide it into reasonably sized spatial groupings. A prioritization is then applied to these clusters and the docking station will then only concern itself with the members of that cluster.

Team Coordination/Task Allocation

Coordination of a team of robots is a key component to creating a usable distributed robot team. Significant research in the field is based on methods developed in the area of intelligent agents, as each robot in a team can be thought of as an independent agent. Task allocation can be done a priori by a single scheduler for all tasks, but this is not desirable as it does not allow for changing conditions such as the loss of a robot or obstacles getting in the way resulting in a task being handled by a robot which is unable to complete the task. Methods have been developed based on auctioning which distribute the tasks fairly to all robots and attempt to approach optimality by minimizing path length, energy cost, or total time of the tasks.

Using auctions to distribute tasks among a group of robots is a common theme in the research in task allocation [39]. In this type of framework, the robots act as agents which are bidding on a task for completion. It is setup as a first-price reverse auction. One task may be auctioned, or multiple tasks may be bid on collectively. This allows for tasks which have good synergy to be completed together, for example two tasks which have completion points very close to each other, to be more optimally assigned. In this auction environment, tasks may be ordered by priority, auctioning off the most important tasks first. These tasks will receive more of the available resources of the entire team of distributed robots because less robots are committed to tasks already. Bids are computed based on the type of optimization preferred: for energy cost, a hilly terrain may require a higher bid because of the difficulty in traversing.

One example of task allocation in a distributed robot team is presented [41]. In this auction-based task allocation

tion method, one robot is committed to all tasks at the beginning of the auction, and serves as an auctioneer for the tasks. One task at a time is auctioned to the other robots in the team. Tasks are ordered by priority in order to give the first tasks more attention. In this method, after any task is completed, all robots then rebid for tasks. This allows tasks which may be easy to complete after one task is completed. If no bids are received for a task, it may be dropped because no robot can complete the task. This method is particularly useful in dynamic environments, because as tasks are completed, some areas of the environment may be blocked off by other robots or moving obstacles. The re-bidding process allows for those tasks to be completed by having them switched to another robot. The re-auctioning rounds include tasks that did not receive bids in the previous rounds, as it may later be possible for these tasks to be completed. In this framework, new tasks can be entered into the system by assigning the task to any of the robots as tasks which had no bid in the previous round, and they will be automatically completed by the robot which is most well suited to complete the task.

It has been shown that even in the simple case of homogeneous tasks and homogeneous robots, task assignment is a NP-hard problem [35]. This makes task allocation especially important when a heterogeneous team of robots is considered. Some members of the team may not be able to complete all of the tasks which need to be completed by the group. More able or powerful members of the group may be able to complete multiple tasks with a lower cost than other members. This makes the problem of assigning tasks much more complex, because the tasks cannot be simply assigned to any robot.

Marsupial robotics adds again this complexity, because tasks can be assigned to heterogeneous robots of the team, and the hybrid and changing nature of the team must be taken into account. One method would be to consider each robot separately and use a previously discussed method, but this is very suboptimal because it does not exploit the advantages which emerge when a marsupial pair is considered. The marsupial pair has enhanced mobility due to its changing size and sensing abilities. The advantages of the marsupial pair should be exploited in order to complete the tasks assigned in a more optimal way than the standard methods. There are two simple extensions to the protocol presented here which may be considered and are presented here.

With communication, each marsupial pair can negotiate as a team in order to place bids for task in the auction systems. This allows for the marsupial team to allocate its resources better. The complicating factor of the consideration that each marsupial carrier as well as each car-

ried robot are potentially part of multiple marsupial pairs, causing there to be *carriers * carried* bidding entities on top of the individual entities for each carrier and carried robot. There is also an issue that must be addressed considering the order of the bidding. If a marsupial pair completes a bid, there is a significant change in the resources available to the entire group. This changes the bids of all of the marsupial pair combinations significantly and many must be recomputed, at least all of the pairs which include the marsupial carrier which has bid and secured the task for completion. This computation may cause the ability of the team to complete the task allocation within a reasonable amount of time to be hampered significantly, which must be considered when proposing such a system.

Another method to be considered is for the bidding carried robots to be able to generate tasks, which would include being deployed at a particular spot. This method would require the carried robots to be able to compute the traversable area for the larger carrier robot. However, this method may be suboptimal as these tasks may be specific to a marsupial pair. It also may be advantageous for a carrier to hand-off the carried robot to another carrier in order to complete the task faster. This makes this method unreasonable because of the large number of possible generated tasks grows without bound, and similarly to the previously discussed method may increase the computational requirements of the bidding process unreasonably.

Describing a task which is to be completed by a group of robots is also an ongoing problem in distributed robotics. In the simplest case, a task which is destined for a team of robots may be decomposable into tasks which are assignable to single robots. However, there may be tasks which are not decomposable, such as two pressure switches being activated at the same time or simultaneously carrying a large object. A task that requires more than one robot to perform needs special consideration. Tasks which may need an unknown number of robots are even more complex. It may be desirable to generate a task which only requires a single robot, but which can be amended to require multiple robots in the future.

Future Directions

Mobile robots are in use today around the world in order to solve problems in areas where it is not safe for humans to enter. They are used to disarm and detonate improvised explosive devices, and in collapsed buildings and natural disasters. Unmanned aerial vehicles are used in the field by soldiers to get a bird's-eye view of the battlefield. The advancement of the field of distributed robotics shows

promise in accomplishing these types of tasks with more efficiency than a single robot alone.

Distributed robotics is a fertile field of research with many open problems still to solve. It also gains from research done in related areas such as networking, intelligent agents, game theory, and control. Robots which are part of a team may require special hardware considerations. The team must be composed and communication between each of the team members must be considered.

Marsupial teams as a subfield of distributed robotics presents its own challenges and rewards. While each marsupial team can take advantage of two separate robot designs in order to maximize the potential of the distributed team, complex hardware and coordination issues must be considered. Marsupials must be considered specially for simulation of the behaviors which they exhibit, and in order to study the effects of differing parameters on the marsupial setup. A marsupial team also greatly increases the challenge of task allocation among the members of a distributed robot team. In exploration and coverage, the marsupial robot can be exploited to utilize the specific features of the multi-resolution sensing. Marsupial teams are a deep subfield of distributed robotics with possibly wide-reaching results yet to be researched.

As hardware technology advances, the improvements will have a large impact on distributed robotics. The trend toward smaller sensors and cheaper designs make it much more feasible to create a large team of robots which must then work together to complete a goal. Advances in batteries which increase energy density and operational lifetime will enhance the functioning lifetime of distributed teams. At the same time, wireless technologies such as WiMax and (other wireless technology) which both increase the bandwidth and range of communications for devices will increase the operational range of robotic teams. These advances will make distributed robotics even more applicable as robots become easier to produce in quantity and can communicate over a longer range. The ubiquity of robotics in the world today may be changed to make distributed robot teams ubiquitous in the near future.

One problem which must be considered is the issue of control. While increasingly tasks which are given to distributed robotic teams are autonomous and teams may be able to generate their own tasks through programming for a specific purpose, the creation and assignment of tasks in a short time frame by a non-expert needs to be addressed. Currently the most common form of control in the field by robot technicians is tele-operation, either in direct view of the robot or aided by the robots sensors or both. This form of control will not scale to distributed robot teams very well. While it is possible for robots being tele-op-

erated to cooperate, it is very unwieldy for the operators and fails at a much higher rate of incidence. In most cases, each single robot must be controlled by an operator which is dedicated to the control of the specific robot in the field. In cases where a large environment must be searched and other tasks must be performed in a short time frame, this is simply not possible. The development of algorithms and interfaces which allow for the control of a large number of robots in a distributed team by a single operator must be developed. The development of these interfaces requires a combination of many of the elements which have been presented here; a level of autonomy of the robots after tasks have been assigned must be expected. The longevity of such a deployed team should not be a burden considered by the operator of a team. The tasks which are assigned by the operator may need to be shifted to a robot which is available and able to perform the task – the interface itself may be task-centric with tasks being defined by the operator and the robot team deciding which robot is best suited to the task. It is certain that this issue with distributed robots' teams will be present in the future, as robots become smaller and are produced in higher quantities.

Acknowledgment

This work has been supported in part by National Science Foundation through grants #IIS-0219863, #CNS-0224363, #CNS-0324864, #CNS-0420836, #IIP-0443945, #IIP-0726109, and #CNS-0708344

Bibliography

Primary Literature

1. Albro JV, Bobrow J (2004) Motion generation for a tumbling robot using a general contact model. In: Proceedings of the 2004 IEEE International Conference on Robotics and Automation, IEEE, New Orleans, pp 3270–3275
2. Bird N, Atev S, Caramelli N, Martin R, Papanikolopoulos N (2006) Real-time, online detection of abandoned objects in public areas. In: Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, pp 3775–3780
3. Bird N, Masoud O, Papanikolopoulos N, Isaacs A (2005) Detection of loitering individuals in public transportation areas. *IEEE Trans Intell Transp Syst* 6(2):167–177
4. Bodor R (2005) Multi-camera human activity recognition in unconstrained indoor and outdoor environments. Ph D thesis, University of Minnesota
5. Broggi A, Caraffi C, Fedriga RI, Grisleri P (2005) Obstacle detection with stereo vision for off-road vehicle navigation. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, vol 3, p 65

6. Burgard W, Moors M, Stachniss C, Schneider FE (2005) Coordinated multi-robot exploration. *IEEE Trans Robotics* 21(3):376–386
7. Christopoulos VN, Roumeliotis S (2005) Adaptive sensing for instantaneous gas release parameter estimation. In: *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Barcelona, Spain, pp 4450–4456
8. Collett TH, MacDonald BA, Gerkey BP (2005) Player 2.0: Toward a practical robot programming framework. In: *Proceedings of the Australasian Conference on Robotics and Automation*. <http://www.araa.asn.au/acra/acra2005/papers/collett.pdf>. Accessed 6 June 2008
9. Comport A, Malis E, Rives P (2007) Accurate quadrifocal tracking for robust 3D visual odometry. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 40–45
10. Correll N, Martinoli A (2007) Robust distributed coverage using a swarm of miniature robots. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 379–384
11. Craighead J, Murphy R, Burke J, Goldiez B (2007) A survey of commercial & open source unmanned vehicle simulators. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 852–857
12. Das AK, Fierro R, Kumar V, Ostrowski JP, Spletzer J, Taylor CJ (2002) A vision-based formation control framework. *IEEE Trans Robotics Autom* 18(5):813–825
13. Dellaert F, Balch T, Kaess M, Ravichandran R, Alegre F, Berhaut M, McGuire R, Merrill E, Moshkina L, Walker D (2002) The Georgia Tech yellow jackets: A marsupial team for urban search and rescue. In: *AAAI Mobile Robot Competition Workshop*, Edmonton, Alberta, pp 44–49
14. Drenner A, Burt I, Kratochvil B, Nelson BJ, Papanikolopoulos N, Yessin KB (2002) Communication and mobility enhancements to the scout robot. In: *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Lausanne, Switzerland
15. Fenwick JW, Newman PM, Leonard JJ (2002) Cooperative concurrent mapping and localization. In: *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, IEEE, Washington, DC, pp 1810–1817
16. Franchi A, Freda L, Oriolo G, Vendittelli M (2007) A randomized strategy for cooperative robot exploration. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 768–774
17. Gage A, Murphy RR (2004) Affective recruitment of distributed heterogeneous agents. In: *Proceedings of the 19th National Conference on Artificial Intelligence*, San Jose, pp 14–19
18. Gage DW (1993) Randomized search strategies with imperfect sensors. In: *Proceedings of SPIE Mobile Robots VIII*, Boston, vol 2058, pp 270–279
19. Hada Y, Yuta S (2000) A first experiment of long term activity of autonomous mobile robot – result of repetitive base-docking over a week. In: *Proceedings of the ISER 2000 7th International Symposium on Experimental Robotics*, Waikiki, pp 235–244
20. Hougen DF, Bonney JC, Budenske JR, Dvorak M, Gini M, Krantz DG, Malver F, Nelson B, Papanikolopoulos N, Rybski PE, Stoeter SA, Voyles R, Yesin KB (2000) Reconfigurable robots for distributed robotics. In: *Government Microcircuit Applications Conf.*, Anaheim, CA, pp 72–75
21. Howard A, Mataric M, Sukhatme G (2002) An incremental deployment algorithm for mobile robot teams. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, EPFL, Switzerland, vol 3, pp 2849–2854
22. Hu H, Kelly I, Keating D, Vinagre D (1998) Coordination of multiple robots via communication. In: *Proceedings of SPIE*, Boston, pp 94–103
23. iRobot corporation (2008) iRobot Roomba® Vacuuming Robot. <http://irobot.com/sp.cfm?pageid=122>. Accessed 6 June 2008
24. Iv DL, Srinivasa S, Lee-Shue V (2002) Towards sensor based coverage with robot teams. In: *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, Washington, DC, vol 1, pp 961–967
25. Jacoff A, Messina E, Evans J (2000) A standard test course for urban search and rescue robots. In: *Proceedings of the 2000 performance metrics for intelligent system workshop*, Gaithersburg, August 2000
26. Janssen M, Papanikolopoulos N (2007) Enabling complex behavior by simulating marsupial actions. In: *Proceedings of the 15th Mediterranean Conference on Control and Automation*, Athens, Greece
27. Kadioglu E, Papanikolopoulos N (2003) A method for transporting a team of miniature robots. In: *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, NV, vol 3, pp 2297–2302
28. Kamath S, Meisner E, Isler V (2007) Triangulation based multi target tracking with mobile sensor networks. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 3283–3288
29. Kelly I, Holland O, Melhuish C (2000) Slugbot: A robotic predator in the natural world. In: *Proceedings of the 5th International Symposium on Artificial Life and Robotics for Human Welfare and Artificial Liferobotics*, Oita, Japan, pp 470–475
30. Ko J, Stewart B, Fox D, Konolige K, Limketkai B (2003) A practical, decision-theoretic approach to multi-robot mapping and exploration. In: *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Las Vegas, NV, vol 4, pp 3232–3238
31. Koenig S, Szymanski B, Liu Y (2001) Efficient and inefficient ant coverage methods. *Ann Math Artif Intell* 31(1):41–76
32. Konolige K, Fox D, Limketkai B, Ko J, Stewart B (2003) Map merging for distributed robot navigation. In: *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Las Vegas, NV, pp 212–217
33. Kurazume R, Hirose S (2000) An experimental study of a cooperative positioning system. *Autonomous Robots* 8(1):43–52
34. Kwolek B (2007) Visual odometry based on gabor filters and sparse bundle adjustment. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 3573–3578
35. Lagoudakis MG, Markakis E, Kempe D, Keskinocak P, Kleywegt A, Koenig S, Tovey C, Meyerson A, Jain S (2005) Auction-based multi-robot routing. In: *Proceedings of Robotics: Science and Systems*, Cambridge, USA
36. Lau H (2003) Behavioural approach for multi-robot exploration. In: *Proceedings of 2003 Australasian Conference on Robotics and Automation*, Brisbane, Australia
37. McMillen C, Stubbs K, Rybski PE, Stoeter SA, Gini M, Papanikolopoulos N (2002) Resource scheduling and load balancing in distributed robotic control systems. In: *The 7th international conference on intelligent autonomous systems*, Marina del Rey, pp 223–230

38. Min HJ, Drenner A, Papanikolopoulos N (2007) Autonomous docking for an erosi robot based on a vision system with points clustering. In: *Proceedings of the 15th Mediterranean Conference on Control and Automation*, Athens, Greece
39. Mosteo AR, Montano L (2007) Comparative experiments on optimization criteria and algorithms for auction based multi-robot task allocation. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 3345–3350
40. Murphy RR (2000) Marsupial and shape-shifting robots for urban search and rescue. *IEEE Intell Syst* 15(2):14–19
41. Nanjanath M, Gini M (2006) Dynamic task allocation for robots via auctions. In: *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, Orlando, FL, pp 2781–2786
42. Ngo TD, Raposo H, Schioler H (2007) Being sociable: Multi-robots with self-sustained energy. In: *Proceedings of the 15th Mediterranean Conference on Control and Automation*, Athens, Greece
43. Nistér D, Naroditsky O, Bergen J (2004) Visual odometry. In: *Proceedings of the 2004 IEEE International Conference on Computer Vision and Pattern Recognition*, IEEE, vol 1, Washington DC, pp 652–659
44. O'Rourke J (1987) *Art Gallery Theorems and Algorithms*. Oxford University Press, New York
45. Ota Y, Kuga T, Yoneda K (2006) Deformation compensation for continuous force control of a wall climbing quadruped with reduced-dof. In: *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, IEEE, Orlando, FL, pp 468–474
46. Perkins CE, Royer EM (1999) Ad-hoc on-demand distance vector routing. In: *Proceedings of the 2nd IEEE Workshop on Mobile Computer Systems and Applications*, New Orleans, p 90
47. Pongas D, Mistry M, Schaaf S (2007) A robust quadruped walking gait for traversing rough terrain. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 1474–1479
48. Pugh J, Martinoli A (2007) The cost of reality: Effects of real-world factors on multi-robot search. In: *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Roma, Italy, pp 397–404
49. Rekleitis I, Lee-Shue V, New AP, Choset H (2004) Limited communication, multi-robot team based coverage. In: *Proceedings of the 2004 IEEE International Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, pp 3462–3468
50. Roumeliotis SI, Bekey GA (2002) Distributed multi-robot localization. *IEEE Trans Robotics Autom* 18(5):781–795
51. Rybski PE, Larson A, Veeraraghavan H, LaPoint M, Gini M (2004) Communication strategies in multi-robot search and retrieval: Experiences with minDART. In: *DARS 2004*, Toulouse, France, pp 301–310
52. Shillcutt KJ (2000) Solar based navigation for robotic explorers. Ph D thesis, Carnegie Mellon University
53. Silverman MC, Nies D, Jung B, Sukatme GS (2002) Staying alive: A docking station for autonomous robot recharging. In: *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, Washington, DC, pp 1050–1055
54. Simmons R, Apfelbaum D, Burgard W, Fox D, Moors M, Thrun S, Younes H (2000) Coordination for multi-robot exploration and mapping. In: *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, 0-262-51112-6, Austin, pp 852–858
55. Singh K, Fujimura K (1993) Map making by cooperating mobile robots. In: *Proceedings of the 1993 IEEE International Conference on Robotics and Automation*, IEEE, Atlanta, pp 254–259
56. Southard L, Hoeg TM, Palmer DW, Antol J, Kolacinski RM, Quinn RD (2007) Exploring mars using a group of tumbleweed rovers. In: *IEEE international conference on robotics and automation*, Roma, Italy, pp 775–780
57. Stoeter SA (2003) Vision-based control of miniature jumping scout robots. Ph D thesis, University of Minnesota
58. Whittaker W, Champeny L (1988) Conception and development of two mobile teleoperated systems for TMI-2. In: *Proceedings of the international meeting and topicam meeting TMI-2 accident*. American Nuclear Society, Washington DC
59. Yamauchi B (1998) Frontier-based exploration using multiple robots. In: *Proceedings of the 2nd international conference on autonomous agents*, Minneapolis/St. Paul, pp 47–53
60. Zlot R, Stentz A, Dias MB, Thayer S (2002) Multi-robot exploration controlled by a market economy. In: *Proceedings of the 2002 IEEE international conference on robotics and automation*, IEEE, Washington DC, pp 3016–3023

Books and Reviews

- Bräunl T (2006) *Embedded robotics: Mobile robot design and applications with embedded systems*. Springer, Berlin
- Dudek G, Jenkin M (2000) *Computational principles of mobile robotics*. Cambridge University Press, Cambridge
- Schultz AC, Parker LE (eds) (2002) *Multi-robot systems: From swarms to intelligent automata*. Kluwer Academic, Dordrecht

DNA Computing

MARTYN AMOS

Manchester Metropolitan University, Manchester, UK

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The DNA Molecule](#)

[The First DNA Computation](#)

[Models of DNA Computation](#)

[Subsequent Work](#)

[Assessment](#)

[Future Directions](#)

[Bibliography](#)

Glossary

DNA Deoxyribonucleic acid. Molecule that encodes the genetic information of cellular organisms.

Enzyme Protein that catalyzes a biochemical reaction.

Nanotechnology Branch of science and engineering dedicated to the construction of artifacts and devices at the nanometer scale.

RNA Ribonucleic acid. Molecule similar to DNA, which helps in the conversion of genetic information to proteins.

Satisfiability (SAT) Problem in complexity theory. An instance of the problem is defined by a Boolean expression with a number of variables, and the problem is to identify a set of variable assignments that makes the whole expression true.

Definition of the Subject

DNA computing (or, more generally, biomolecular computing) is a relatively new field of study that is concerned with the use of biological molecules as fundamental components of computing devices. It draws on concepts and expertise from fields as diverse as chemistry, computer science, molecular biology, physics and mathematics. Although its theoretical history dates back to the late 1950s, the notion of computing with molecules was only physically realized in 1994, when Leonard Adleman demonstrated in the laboratory the solution of a small instance of a well-known problem in combinatorics using standard tools of molecular biology. Since this initial experiment, interest in DNA computing has increased dramatically, and it is now a well-established area of research. As we expand our understanding of how biological and chemical systems process information, opportunities arise for new applications of molecular devices in bioinformatics, nanotechnology, engineering, the life sciences and medicine.

Introduction

In the late 1950s, the physicist Richard Feynman first proposed the idea of using living cells and molecular complexes to construct “sub-microscopic computers.” In his famous talk *“There’s Plenty of Room at the Bottom”* [18], Feynman discussed the problem of “manipulating and controlling things on a small scale”, thus founding the field of nanotechnology. Although he concentrated mainly on information storage and molecular manipulation, Feynman highlighted the potential for biological systems to act as small-scale information processors:

The biological example of writing information on a small scale has inspired me to think of something that should be possible. Biology is not simply writing information; it is doing something about it. A biological system can be exceedingly small. Many of the cells are very tiny, but they are very active; they manufacture various substances; they walk around;

they wiggle; and they do all kinds of marvelous things – all on a very small scale. Also, they store information. Consider the possibility that we too can make a thing very small which does what we want – that we can manufacture an object that maneuvers at that level! [18].

Early Work

Since the presentation of Feynman’s vision there has been an steady growth of interest in performing computations at a molecular level. In 1982, Charles Bennett [8] proposed the concept of a “Brownian computer” based around the principle of reactant molecules touching, reacting, and effecting state transitions due to their random Brownian motion. Bennett developed this idea by suggesting that a Brownian Turing Machine could be built from a macromolecule such as RNA. “Hypothetical enzymes”, one for each transition rule, catalyze reactions between the RNA and chemicals in its environment, transforming the RNA into its logical successor.

In the same year, Conrad and Liberman developed this idea further in [15], in which the authors describe parallels between physical and computational processes (for example, biochemical reactions being employed to implement basic switching circuits). They introduce the concept of molecular level “word processing” by describing it in terms of transcription and translation of DNA, RNA processing, and genetic regulation. However, the paper lacks a detailed description of the biological mechanisms highlighted and their relationship with “traditional” computing. As the authors themselves acknowledge, “our aspiration is not to provide definitive answers ... but rather to show that a number of seemingly disparate questions must be connected to each other in a fundamental way.” [15]

In [14], Conrad expanded on this work, showing how the information processing capabilities of organic molecules may, in theory, be used in place of digital switching components. Particular enzymes may alter the three-dimensional structure (or *conformation*) of other *substrate* molecules. In doing so, the enzyme switches the *state* of the substrate from one to another. The notion of *conformational computing* (q.v.) suggests the possibility of a potentially rich and powerful computational architecture. Following on from the work of Conrad et al., Arkin and Ross show how various logic gates may be constructed using the computational properties of enzymatic reaction mechanisms [5] (see Dennis Bray’s article [10] for a review of this work). In [10], Bray also describes work [23,24] showing how chemical “neurons” may be constructed to form the building blocks of logic gates.

Motivation

We have made huge advances in machine miniaturization since the days of room-sized computers, and yet the underlying computational framework (the *von Neumann architecture*) has remained constant. Today's supercomputers still employ the kind of sequential logic used by the mechanical "dinosaurs" of the 1940s [13].

There exist two main barriers to the continued development of "traditional", silicon-based computers using the von Neumann architecture. One is inherent to the machine architecture, and the other is imposed by the nature of the underlying *computational substrate*. A computational substrate may be defined as "*a physical substance acted upon by the implementation of a computational architecture*." Before the invention of silicon integrated circuits, the underlying substrates were bulky and unreliable. Of course, advances in miniaturization have led to incredible increases in processor speed and memory access time. However, there is a limit to how far this miniaturization can go. Eventually "chip" fabrication will hit a wall imposed by the *Heisenberg Uncertainty Principle* (HUP). When chips are so small that they are composed of components a few atoms across, quantum effects cause interference. The HUP states that the act of observing these components affects their behavior. As a consequence, it becomes impossible to know the exact state of a component without fundamentally changing its state.

The second limitation is known as the *von Neumann bottleneck*. This is imposed by the need for the central processing unit (CPU) to transfer instructions and data to and from the main memory. The route between the CPU and memory may be visualized as a two-way road connecting two towns. When the number of cars moving between towns is relatively small, traffic moves quickly. However, when the number of cars grows, the traffic slows down, and may even grind to a complete standstill. If we think of the cars as units of information passing between the CPU and memory, the analogy is complete. Most computation consists of the CPU fetching from memory and then executing one instruction after another (after also fetching any data required). Often, the execution of an instruction requires the storage of a result in memory. Thus, the speed at which data can be transferred between the CPU and memory is a limiting factor on the speed of the whole computer.

Some researchers are now looking beyond these boundaries and are investigating entirely new computational architectures and substrates. These developments include *quantum computing* (q.v.), *optical computing* (q.v.), *nanocomputers* (q.v.) and bio-molecular computers.

In 1994, interest in molecular computing intensified with the first report of a successful non-trivial molecular computation. Leonard Adleman of the University of Southern California effectively founded the field of DNA computing by describing his technique for performing a massively-parallel random search using strands of DNA [1]. In what follows we give an in-depth description of Adleman's seminal experiment, before describing how the field has evolved in the years that followed. First, though, we must examine more closely the structure of the DNA molecule in order to understand its suitability as a computational substrate.

The DNA Molecule

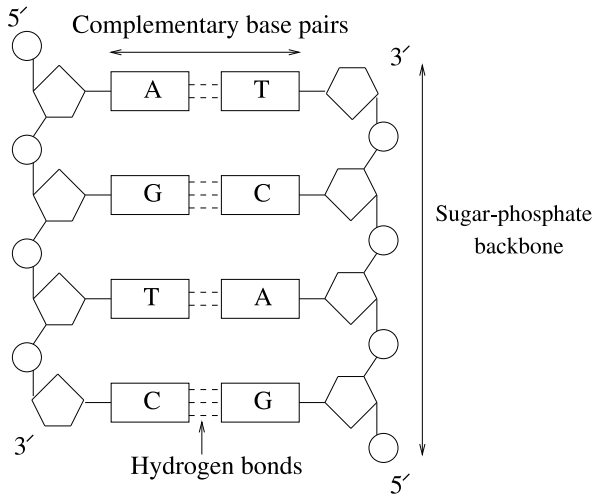
Ever since ancient Greek times, man has suspected that the features of one generation are passed on to the next. It was not until Mendel's work on garden peas was recognized [38] that scientists accepted that both parents contribute material that determines the characteristics of their offspring. In the early 20th century, it was discovered that *chromosomes* make up this material. Chemical analysis of chromosomes revealed that they are composed of both *protein* and *deoxyribonucleic acid*, or *DNA*. The question was, which substance carries the genetic information? For many years, scientists favored protein, because of its greater complexity relative to that of DNA. Nobody believed that a molecule as simple as DNA, composed of only four subunits (compared to 20 for protein), could carry complex genetic information.

It was not until the early 1950s that most biologists accepted the evidence showing that it is in fact DNA that carries the genetic code. However, the physical structure of the molecule and the hereditary mechanism was still far from clear.

In 1951, the biologist James Watson moved to Cambridge to work with a physicist, Francis Crick. Using data collected by Rosalind Franklin and Maurice Wilkins at King's College, London, they began to decipher the structure of DNA. They worked with models made out of wire and sheet metal in an attempt to construct something that fitted the available data. Once satisfied with their double helix model, they published the paper [42] (also see [41]) that would eventually earn them (and Wilkins) the Nobel Prize for Physiology or Medicine in 1962.

DNA Structure

DNA (deoxyribonucleic acid) [43] encodes the genetic information of cellular organisms. It consists of *polymer chains*, commonly referred to as *DNA strands*. Each strand may be viewed as a chain of *nucleotides*, or *bases*, attached



DNA Computing, Figure 1
Structure of double-stranded DNA (from [3])

to a sugar-phosphate “backbone”. An n -letter sequence of consecutive bases is known as an n -mer or an *oligonucleotide* (commonly abbreviated to “oligo”) of length n . Strand lengths are measured in *base pairs* (b.p.)

The four DNA nucleotides are adenine, guanine, cytosine, and thymine, commonly abbreviated to A, G, C, and T respectively. Each strand, according to chemical convention, has a 5' and a 3' end; thus, any single strand has a natural orientation (Fig. 1). The classical double helix of DNA is formed when two separate strands bond. Bonding occurs by the pairwise attraction of bases; A bonds with T and G bonds with C. The pairs (A, T) and (G, C) are therefore known as *complementary base pairs*. The two pairs of bases form *hydrogen bonds* between each other, two bonds between A and T, and three between G and C.

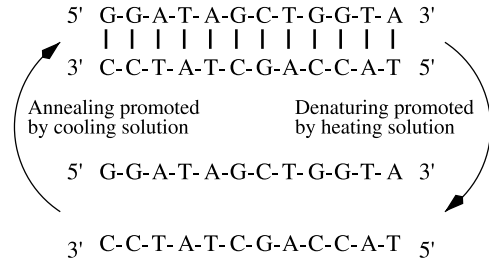
The bonding process, known as *annealing*, is fundamental to our implementation. A strand will only anneal to its complement if they have opposite polarities. Therefore, one strand of the double helix extends from 5' to 3', and the other from 3' to 5', as depicted in Fig. 1.

Operations on DNA

Some (but not all) DNA-based computations apply a specific sequence of biological operations to a set of strands. These operations are all commonly used by molecular biologists, and we now describe them in more detail.

Synthesis

Oligonucleotides may be synthesized to order by a machine the size of a microwave oven. The synthesizer is supplied with the four nucleotide bases in solution, which are



DNA Computing, Figure 2
DNA melting and annealing (from [3])

combined according to a sequence entered by the user. The instrument makes millions of copies of the required oligo and places them in solution in a small vial.

Denaturing, Annealing, and Ligation

Double-stranded DNA may be dissolved into single strands (or *denatured*) by heating the solution to a temperature determined by the composition of the strand [11]. Heating breaks the hydrogen bonds between complementary strands (Fig. 2). *Annealing* is the reverse of melting, whereby a solution of single strands is cooled, allowing complementary strands to bind together (Fig. 2).

In double-stranded DNA, if one of the single strands contains a discontinuity (i. e., one nucleotide is not bonded to its neighbor) then this may be repaired by DNA *ligase* [12]. This particular enzyme is useful for DNA computing, as it allows us to create a unified strand from several strands bound together by their respective complements. Ligase therefore acts as a molecular “cement” or “mortar”, binding together several strands into a single strand.

Separation of Strands

Separation is a fundamental operation, and involves the extraction from a test tube of any *single* strands containing a specific short sequence (e. g., extract all strands containing the sequence GCTA). For this, we may use a “molecular sieving” process known as *affinity purification*. If we want to extract from a solution single strands containing the sequence x , we may first create many copies of its complement, \bar{x} . We attach to these oligos biotin molecules (a process known as “biotinylation”) which in turn bind to a fixed matrix. If we pour the contents of the test tube over this matrix, strands containing x will anneal to the anchored complementary strands. Washing the matrix removes all strands that *did not* anneal, leaving only strands containing x . These may then be removed from the matrix.

Gel Electrophoresis

Gel electrophoresis is an important technique for sorting DNA strands by size [12]. Electrophoresis is the movement of charged molecules in an electric field. Since DNA molecules carry a negative charge, when placed in an electric field they tend to migrate toward the positive pole. The rate of migration of a molecule in an *aqueous* solution depends on its shape and electric charge. Since DNA molecules have the same charge per unit length, they all migrate at the same speed in an aqueous solution. However, if electrophoresis is carried out in a *gel* (usually made of agarose, polyacrylamide, or a combination of the two), the migration rate of a molecule is also affected by its *size*. This is due to the fact that the gel is a dense network of pores through which the molecules must travel. Smaller molecules therefore migrate faster through the gel, thus sorting them according to size. In order to sort strands, the DNA is placed in a well cut out of the gel, and a charge applied. After running the gel, the results are visualized by staining the DNA with dye and then viewing the gel under ultraviolet light. Bands of DNA of a specific length may then be cut out of the gel and soaked to free the strands (which may then be used again in subsequent processing steps).

PCR

The DNA *polymerases* perform several functions, including the repair and duplication of DNA. Given a short *primer* (or “tag”) oligo, in the presence of nucleotide triphosphates (i. e., “spare” nucleotides), the polymerase extends the primer if and only if the primer is bound to a longer *template* strand.

Primer extension is fundamental to the *Polymerase Chain Reaction*, or PCR [28]. PCR is a process that quickly amplifies the amount of DNA in a given solution. PCR employs polymerase to make copies of a specific region (or *target sequence*) of DNA that lies between two *known* sequences. Note that this target sequence (which may be up to around 3,000 b.p. long) can be unknown ahead of time. In order to amplify template DNA with known regions (perhaps at either end of the strands), we first design forward and backward primers (i. e. primers that go from 5' to 3' on each strand. We then add a large excess (relative to the amount of DNA being replicated) of primer to the solution and heat it to denature the double-stranded template. Cooling the solution then allows the primers to anneal to their target sequences. We then add the polymerase, which extends the primers, forming an identical copy of the template DNA.

If we start with a single template, then of course we now have two copies. If we then repeat the cycle of heating, annealing, and polymerizing, it is clear that this approach yields an exponential number of copies of the template (since the number of strands doubles after each cycle). A typical number of cycles would be perhaps 35, yielding (assuming a single template) around 68 billion copies of the *target sequence* (for example, a gene).

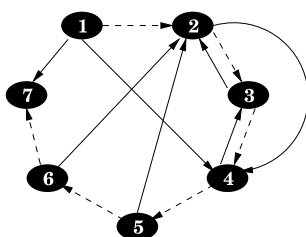
The First DNA Computation

We now describe in detail the first ever successful computation performed using strands of DNA. This experiment was carried out in 1994 by Leonard Adleman of the University of Southern California. Adleman was already a highly distinguished computer scientist prior to this work, having been awarded, along with Rivest and Shamir, the 2002 ACM Turing Award for the development of the RSA public-key encryption scheme [32].

Adleman utilized the incredible storage capacity of DNA to implement a brute-force algorithm for the directed Hamiltonian Path Problem (HPP) [1]. The HPP involves finding a path through a graph (or “network”) that visits each vertex (“node”) exactly once. For an example application, consider a salesperson who wishes to visit several cities connected by rail links. In order to save time and money, she would like to know if there exists an itinerary that visits every city precisely once. We model this situation by constructing a graph where vertices represent cities and edges the rail links. The HPP is a classic problem in the study of *complex networks and graph theory* (q.v.) [20], and belongs to the class of problems referred to as *NP-complete* [19]. To practitioners in the field of *computational complexity* (q.v.), the NP-complete problems are most interesting, since they are amongst the hardest known problems, and include many problems of great theoretical and practical significance, such as network design, scheduling, and data storage.

The instance of the HPP that Adleman solved is depicted in Fig. 3, with the unique Hamiltonian Path (HP) highlighted by a dashed line. Adleman’s approach was simple:

1. Generate strands encoding random paths such that the Hamiltonian Path (HP) is represented with high probability. The quantities of DNA used far exceeded those necessary for the small graph under consideration, so it is likely that *many* strands encoding the HP were present.
2. Remove all strands that do not encode the HP.
3. Check that the remaining strands encode a solution to the HPP.



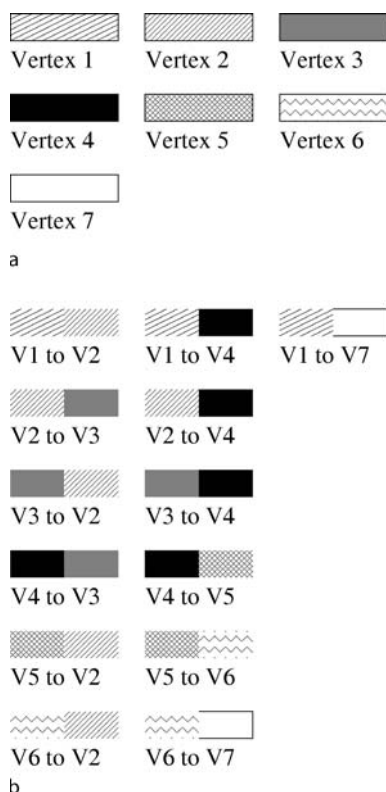
DNA Computing, Figure 3

Instance of the HPP solved by Adleman (from [3])

The individual steps were implemented as follows:

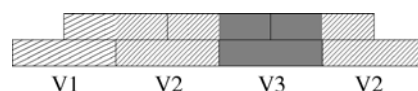
Stage 1: Each vertex and edge was assigned a distinct 20-base sequence of DNA (Fig. 4a). This implies that strands encoding a HP were of length 140 b.p., since there are seven vertices in the problem instance. Sequences representing edges act as ‘splints’ between strands representing their endpoints (Fig. 4b).

In formal terms, the sequence associated with an edge $i \rightarrow j$ is the 3' 10''-mer of the sequence representing v_i followed by the 5' 10''-mer of the sequence representing v_j . These oligos were then combined to form strands encod-



DNA Computing, Figure 4

Adleman's scheme for encoding paths-schematic representation of oligos (from [3])



DNA Computing, Figure 5

Example path created in Adleman's scheme (from [3])

ing random paths through the graph. An (illegal) example path ($v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4$) is depicted in Fig. 5.

Fixed amounts (50 pmol) of each oligo were synthesized, mixed together and then ligase was added to seal any backbone nicks and form completely unified strands from the shorter “building blocks”. At the end of this reaction, it is assumed that a complete strand representing the HP is present with high probability. This approach solves the problem of generating an exponential number of different paths using a polynomial number of initial oligos.

Stage 2: PCR was first used to massively amplify the population of strands encoding paths starting at v_1 and ending at v_7 . These strands were amplified to such a massive extent that they effectively “overwhelmed” any strands that *did not* start and end at the correct points.

Next, strands that do not encode paths containing exactly n visits were removed. The product of the PCR amplification was run on an agarose gel to isolate strands of length 140 b.p. A series of affinity purification steps was then used to isolate strands encoding paths that visited each vertex exactly once.

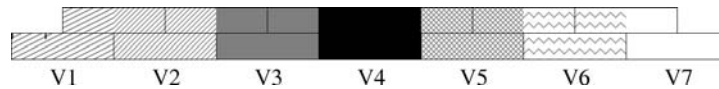
Stage 3: PCR was used to identify the strand encoding the unique HP that this problem instance provides. For an n -vertex graph, we run $n - 1$ PCR reactions, with the strand representing v_1 as the left primer and the complement of the strand representing v_i as the right primer in the i th lane. The presence of molecules encoding the unique HP depicted in Fig. 3 should produce bands of length 40, 60, 80, 100, 120, and 140 b.p. in lanes 1 through 6, respectively. This is exactly what Adleman observed.

Models of DNA Computation

Although Adleman provided the impetus for subsequent work on DNA computing, his algorithm was not expressed within a formal model of computation. Richard Lipton realized that Adleman's approach, though seminal, was limited in that it was specific to solving the HPP. Lipton proposed extending Adleman's algorithm ‘in a way that allows biological computers to potentially radically change the way we do all computations, not just HPPs’ [25].

Satisfiability Model

Lipton's article described a methodology for solving the *satisfiability problem* (SAT) [16] using DNA. In terms of



DNA Computing, Figure 6

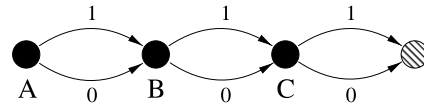
Unique Hamiltonian path (from [3])

computational complexity (q.v.), SAT is the “benchmark” NP-complete problem, and may be phrased as follows: given a finite set $V = \{v_1, v_2, \dots, v_n\}$ of logical variables, we define a *literal* to be a variable, v_i , or its complement, \bar{v}_i . If v_i is *true* then \bar{v}_i is *false*, and vice-versa. We define a *clause*, C_j , to be a set of literals $\{v_1^j, v_2^j, \dots, v_l^j\}$. An instance, I , of SAT consists of a set of clauses. The problem is to assign a Boolean value to each variable in V such that at least one variable in each clause has the value *true*. If this is the case we may say that I has been *satisfied*.

By showing how molecular computers could be applied to this archetypal NP-complete problem, Lipton hoped to show how *any* difficult problem may be solved using this approach. The underlying principle was the same as Adleman’s approach: generate all possible solutions to the problem, then gradually *filter out* strands until any remaining strands *must* be a solution to the problem. Lipton proposed solving an instance of SAT by starting with a tube containing strands representing all possible assignments to its variables.

Lipton’s main contribution lay in his method for encoding *arbitrary* binary strings as strands of DNA. If we only need a small number different strands as the foundation of our computer, then it would be fairly cheap and quick to specify and order them to be individually synthesized. However, Lipton realized that this approach would quickly become infeasible as the number of variables grew. One of the characteristics of the NP-complete problems is that for even a small increase in the problem size (in this case, the number of variables), the number of possible solutions rises exponentially. For any non-trivial problem, Lipton would require a prohibitively expensive number of ‘one-off’ strands. For example, if Lipton wanted to solve a problem with ten variables, he would need to order $2^{10} = 1,024$ individual strands.

Lipton needed a way of encoding an *exponential*-sized pool of starting strands, using only a polynomial number of ‘building-block’ strands. Just as Adleman encoded a large number of possible paths through a graph, using a small number of node and edge strands, Lipton wanted to build a large number of assignment strands, using a small number of ‘variable’ strands. His key insight was that an arbitrary binary string (in this case encoding the values of a set of variables) could be represented as



DNA Computing, Figure 7

Lipton’s graph representation for a three-bit binary string

a path through a graph, where each node represented one particular bit (or variable).

Figure 7 shows an example of a graph to encode a string of three bits (variables), each represented by a node labeled A, B or C. Each node has two edges emanating from it, labeled either 1 or 0 (the edges leading out of node C lead to a ‘dummy’ node, which just acts like a railway buffer to end the paths). Any three-bit string can be represented by a path built by taking one of the two possible paths at each node, the value of each bit being defined by the label of the edge that is taken at each step. For example, if we only ever take the ‘top’ path at each branch, we encode the string 111, and if we only ever take the ‘bottom’ path, we end up with 000. A path that goes ‘top, bottom, top’ encodes the string 101, ‘top, top, bottom’ encodes 110, and so on.

The power of this encoding is that, just like Adleman’s approach, we only have to generate one sequence for each node (including the dummy node) and one for each edge. With the correct Watson–Crick encoding, random paths through the graph form spontaneously, just like in Adleman’s Hamiltonian Path experiment. Using this approach, Lipton believed he could be confident of starting off with a tube containing all possible binary strings of a given length. Each strand in the tube would encode a possible assignment of values for a set of variables. The next stage was to remove strands encoding those assignments that did *not* result in satisfaction of the formula to be solved.

Because the general form of a SAT formula is a set of clauses combined with the AND operation, it follows that each clause *must* be satisfied: if evaluating only a single clause results in a value of 0, then the *whole* formula evaluates to 0, and the formula is not satisfied. Because the variables in each clause are separated by the OR operation, it only takes a single variable to take the value 1 for the whole clause to be satisfied. Lipton’s algorithm was very straightforward: he would proceed one clause at a time, looking at each component of it in turn, and keeping only the strands encoding a sequence of bits that satisfied that

particular clause. The ‘winning’ strands would then go forward to the *next* clause, where the process would repeat, until there were no more clauses to examine. At the end of this procedure, if Lipton had *any* strands left in his tube, then he would know with certainty that the formula was satisfiable, since only strings satisfying every clause would have survived through the successive filters.

We now show how Lipton’s method can be *formally* expressed. In all *filtering* models of DNA computing, a computation consists of a sequence of operations on finite *multi-sets* of strings. Multi-sets are sets that may contain more than one copy of the same element. It is normally the case that a computation begins and terminates with a single multi-set. Within the computation, by applying legal operations of a model, several multi-sets may exist at the same time. We define operations on multi-sets shortly, but first consider the nature of an *initial set*.

An initial multi-set consists of strings which are typically of length $O(n)$ where n is the problem size. As a subset, the initial multi-set should include all possible solutions (each encoded by a string) to the problem to be solved. The point here is that the superset, in any implementation of the model, is supposed to be relatively easy to generate as a starting point for a computation. The computation then proceeds by *filtering out* strings which cannot be a solution.

Within one possible model [2], the following operations are available on sets of strings over some alphabet α :

- *separate*(T, S). Given a set T and a substring S , create two new sets $+(T, S)$ and $-(T, S)$, where $+(T, S)$ is all strings in T containing S , and $-(T, S)$ is all strings in T not containing S .
- *merge*(T_1, T_2, \dots, T_n). Given set T_1, T_2, \dots, T_n , create $\cup(T_1, T_2, \dots, T_n) = T_1 \cup T_2 \cup \dots T_n$.
- *detect*(T). Given a set T , return *true* if T is nonempty, otherwise return *false*.

For example, given $\alpha = \{A, B, C\}$, the following algorithm returns *true* only if the initial multi-set contains a string composed entirely of “A”s:

```
Input( $T$ )
 $T \leftarrow -(T, B)$ 
 $T \leftarrow -(T, C)$ 
Output(detect( $T$ ))
```

Although Lipton does not explicitly define his operation set in [25], his solution to SAT may be phrased in terms of the the operations above, described by Adleman in [2]. Lipton employs the *merge*, *separate*, and *detect* operations described above. The initial set T contains many strings, each encoding a single n -bit sequence. All possible n -bit

sequences are represented in T . The algorithm proceeds as follows:

- (1) Create initial set, T
- (2) For each clause do begin
- (3) For each literal v_i do begin
- (4) if $v_i = x_j$ extract from T strings encoding $v_i = 1$ else extract from T strings encoding $v_i = 0$
- (5) End for
- (6) Create new set T by merging extracted strings
- (7) End for
- (8) If T nonempty then I is satisfiable

The pseudo-code algorithm may be expressed more formally thus:

- (1) Input(T)
- (2) **for** $a = 1$ to $|I|$ **do begin**
- (3) **for** $b = 1$ to $|C_a|$ **do begin**
- (4) **if** $v_b^a = x_j$ **then** $T_b \leftarrow +(T, v_b^a = 1)$ **else** $T_b \leftarrow +(T, v_b^a = 0)$
- (5) **end for**
- (6) $T \leftarrow \text{merge}(T_1, T_2, \dots, T_b)$
- (7) **end for**
- (8) Output(*detect*(T))

Step 1 generates all possible n -bit strings. Then, for each clause $C_a = \{v_1^a, v_2^a, \dots, v_l^a\}$ (Step 2) we perform the following steps. For each literal v_b^a (Step 3) we operate as follows: if v_b^a computes the positive form then we extract from T all strings encoding 1 at position v_b^a , placing these strings in T_b ; if v_b^a computes the negative form we extract from T all strings encoding 0 at position v_b^a , placing these strings in T_b (Step 4); after l iterations, we have satisfied every variable in clause C_a ; we then create a new set T from the union of sets T_1, T_2, \dots, T_b (Step 6) and repeat these steps for clause $C_a + 1$ (Step 7). If any strings remain in T after all clauses have been operated upon, then I is satisfiable (Step 8). Although Lipton did not report an experimental verification of his algorithm, the biological implementation would be straightforward, using affinity purification to implement extraction, pouring of tubes for *merge* and running the solution through a gel to *detect*.

Soon after Lipton’s paper, the Proceedings of the Second Annual Workshop on DNA Based Computers contained several papers describing significant developments in molecular computing. These included the parallel filtering model, the sticker model, DNA self-assembly, RNA computing and surface-based computing.

We describe the last four developments (or variants thereof) in the next section. Here, we briefly introduce the parallel filtering model, as it motivates some of the later discussion.

Parallel Filtering Model

The first description of the parallel filtering model appeared in [4], with further analysis in [3]. This model was the first to provide a formal framework for the easy description of DNA algorithms for *any* problem in the complexity class *NP*. The main difference between the parallel filtering model and those previously proposed lies in the *implementation* of the removal of strings. All other models propose *separation* steps, where strings are *conserved*, and may be used later in the computation. Within the parallel filtering model, however, strings that are removed are *discarded*, and play no further part in the computation. This model is the first exemplar of the so-called “mark and destroy” paradigm of molecular computing, and was motivated in part by discussion of the error-prone nature of biotin-based separation techniques. The new operation introduced within this model is:

- $remove(U, \{S_i\})$. This operation removes from the set U , in parallel, any string which contains at least one occurrence of any of the substrings S_i .

The proposed implementation of the *remove* operation is as follows: for a given substring S_i , add “tag” strands composed of its Watson–Crick complement to the tube U , so that the tags anneal only to strands containing the target subsequence. We then add polymerase to make tagged strands double-stranded from the point of annealing (“mark”). We assume that working strands contain a specific restriction site embedded along them at fixed intervals, so adding the appropriate restriction enzyme removes only marked strands by repeated digestion (“destroy”).

Subsequent Work

In this section we describe several other successful laboratory implementations of molecular-based computing, each illustrating a novel approach or new technique. Our objective is not to give an exhaustive description of each experiment, but to give a high-level treatment of the general methodology, so that the reader may approach with confidence the fuller description in the literature.

DNA Addition

One of the first successful experiments reported after Adleman’s result was due to Guarnieri et al. in 1996 [21], in which they describe a DNA-based algorithm for binary addition. The method uses single-stranded DNA reactions to add together two nonnegative binary numbers. This application, as the authors note, is very different from previ-

ous proposals, which use DNA as the substrate for a massively parallel random search.

Adding binary numbers requires keeping track of the position of each digit and of any “carries” that arise from adding 1 to 1 (remembering that $1 + 1 = 0$ plus carry 1 in binary). The DNA sequences used represent not only binary strings but also allow for carries and the extension of DNA strands to represent answers. Guarnieri et al. use sequences that encode a digit in a given position and its significance, or position from the right. For example, the first digit in the first position is represented by two DNA strands, each consisting of a short sequence representing a “position transfer operator”, a short sequence representing the digit’s value, and a short sequence representing a “position operator”.

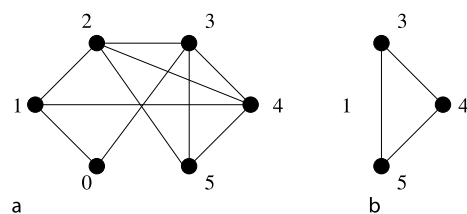
DNA representations of all possible two bit binary integers are constructed, which can then be added in pairs. Adding a pair involves adding appropriate complementary strands, which then link up and provide the basis for strand extension to make new, longer strands. This is termed a “horizontal chain reaction”, where input sequences serve as templates for constructing an extended result strand. The final strand serves as a record of successive operations, which is then read out to yield the answer digits in the correct order.

The results obtained confirmed the correct addition of $0 + 0$, $0 + 1$, $1 + 0$, and $1 + 1$, each calculation taking between 1 and 2 days of bench work. Although limited in scope, this experiment was (at the time) one of the few experimental implementations to support theoretical results.

Maximal Clique Computation

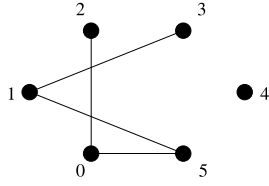
The problem of finding a Maximal Clique using DNA was addressed by Ouyang et al. in 1997 [30]. A *clique* is a fully connected subgraph of a given graph (Fig. 8). The *maximal clique* problem asks: given a graph, how many vertices are there in the largest clique? Finding the size of the largest clique is an NP-complete problem.

The algorithm proceeds as follows: for a graph G with n vertices, all possible vertex subsets (subgraphs) are rep-



DNA Computing, Figure 8

a Ouyang graph. b Example subgraph (from [3])



DNA Computing, Figure 9
Complement of Ouyang graph (from [3])

represented by an n -bit binary string $b_{n-1}, b_{n-2}, \dots, b_0$. For example, given the six-vertex graph used in [30] and depicted in Fig. 8a, the string 111000 corresponds to the subgraph depicted in Fig. 8b, containing v_5, v_4 , and v_3 .

Clearly, the largest clique in this graph contains v_5, v_4, v_3 , and v_2 , represented by the string 111100.

The next stage is to find pairs of vertices that are not connected by an edge (and, therefore, by definition, cannot appear together in a clique). We begin by taking the complement of G , \bar{G} , which contains the same vertex set as G , but which only contains an edge $\{v, w\}$ if $\{v, w\}$ is not present in the edge set of G . The complement of the graph depicted in Fig. 8 is shown in Fig. 9.

If two vertices in \bar{G} are connected by an edge then their corresponding bits cannot both be set to 1 in any given string. For the given problem, we must therefore remove strings encoding $***1*1$ (v_2 and v_0), $1****1$ (v_5 and v_0), $1***1*$ (v_5 and v_1) and $**1*1*$ (v_3 and v_1), where $*$ means either 1 or 0. All other strings encode a (not necessarily maximal) clique.

We must then sort the remaining strings to find the largest clique. This is simply a case of finding the string containing the largest number of 1s, as each 1 corresponds to a vertex in the clique. The string containing the largest number of 1s encodes the largest clique in the graph.

The DNA implementation of the algorithm goes as follows. The first task is to construct a set of DNA strands to represent all possible subgraphs. There are two strands per bit, to represent either 1 or 0. Each strand associated with a bit i is represented by two sections, its position P_i and its value V_i . All P_i sections are of length 20 bases. If $V_i = 1$ then the sequence representing V_i is a restriction site unique to that strand. If $V_i = 0$ then the sequence representing V_i is a 10 base “filler” sequence. Therefore, the longest possible sequence is 200 bases, corresponding to the string 000000, and the shortest sequence is 140 bases, corresponding to 111111.

The computation then proceeds by digesting strands in the library, guided by the complementary graph \bar{G} . To remove strands encoding a connection i, j in \bar{G} , the current tube is divided into two, t_0 and t_1 . In t_0 we cut strings en-

coding $V_i = 1$ by adding the restriction enzyme associated with V_i . In t_1 we cut strings encoding $V_j = 1$ by adding the restriction enzyme associated with V_j . For example, to remove strands encoding the connection between V_0 and V_2 , we cut strings containing $V_0 = 1$ in t_0 with the enzyme Afl II, and we cut strings containing $V_2 = 1$ in t_1 with Spe I. The two tubes are then combined into a new working tube, and the next edge in \bar{G} is dealt with.

In order to read the size of the largest clique, the final tube was simply run on a gel. The authors performed this operation, and found the shortest band to be 160 bp, corresponding to a 4-vertex clique. This DNA was then sequenced and found to represent the correct solution, 111100.

Although this is another good illustration of a DNA-based computation, the authors acknowledge the lack of scalability of their approach. One major factor is the requirement that each vertex be associated with an individual restriction enzyme. This, of course, limits the number of vertices that can be handled by the number of restriction enzymes available. However, a more fundamental issue is the exponential growth in the problem size (and thus the initial library), which we shall encounter again.

Chess Games

In [17], Faulhammer et al. describe a solution to a variant of the satisfiability problem that uses RNA rather than DNA as the computational substrate. They consider a variant of SAT, the so-called “Knight problem”, which seeks configurations of knights on an $n \times n$ chess board, such that no knight is attacking any other [40].

The authors prefer the “mark and destroy” strategy rather than the repeated use of extraction to remove illegal solutions. However, the use of an RNA library and ribonuclease (RNase) H digestion gives greater flexibility, as one is not constrained by the set of restriction enzymes available. In this way, the RNase H acts as a “universal restriction enzyme”, allowing selective marking of virtually any RNA strands for parallel destruction by digestion.

The particular instance solved in [17] used a 3×3 board, with the variables $a - i$ representing the squares. If a variable is set to 1 then a knight is present at that variable’s square, and 0 represents the absence of a knight. The 3×3 knight problem may therefore be represented as the following instance of SAT:

$$\begin{aligned} & ((\neg h \wedge \neg f) \vee \neg a) \wedge ((\neg g \wedge \neg i) \vee \neg b) \\ & \wedge ((\neg d \wedge \neg h) \vee \neg c) \wedge ((\neg c \wedge \neg i) \vee \neg d) \\ & \wedge ((\neg a \wedge \neg g) \vee \neg f) \wedge ((\neg b \vee \neg f) \vee \neg g) \\ & \wedge ((\neg a \wedge \neg c) \vee \neg h) \wedge ((\neg d \wedge \neg b) \vee \neg i) \end{aligned}$$

which, in this case, simplifies to

$$\begin{aligned} & ((\neg h \wedge \neg f) \wedge \neg a) \wedge ((\neg g \wedge \neg i) \vee \neg b) \\ & \wedge ((\neg d \wedge \neg h) \vee \neg c) \wedge ((\neg c \wedge \neg i) \vee \neg d) \\ & \wedge ((\neg a \wedge \neg g) \vee \neg f) . \end{aligned}$$

This simplification greatly reduces the number of laboratory steps required. The experiment proceeds by using a series of RNase H digestions of “illegal” board representations, along the lines of the parallel filtering model [4].

Board representations are encoded as follows: the experiment starts with all strings of the form x_1, \dots, x_n , where each variable x_i takes the value 1 or 0; then, the following operations may be performed on the population of strings:

- Cut all strings containing any pattern of specified variables p_i, \dots, p_k
- Separate the “test tube” into several collections of strings (molecules) by length
- Equally divide (i. e., split) the contents of a tube into two tubes
- Pour (mix) two test tubes together
- Sample a random string from the test tube.

The first stage of the algorithm is the construction of the initial library of strands. Each strand sequence follows the template depicted in Fig. 10.

The prefix and suffix regions are included to facilitate PCR. Each variable is represented by one of two unique sequences of length 15 nucleotides, one representing the fact that the variable is set to 1, and the other the fact that it is set to 0. Variable regions are separated by short (5 nucleotide) spacer regions. In order to avoid having to individually generate each individual sequence, a “mix and split” strategy (described in more detail in [17]) is used. The RNA version of the library is then generated by *in vitro* transcription.

The algorithm proceeds as follows:

1. For each square, sequentially, split the RNA library into two tubes, labeled 1 and 2. After digestions have taken place, tube 1 will contain strands that contain a knight at that square, and tube 2 will contain strands that do not have knights at that square

2. In tube 1, digest with RNase H strands that have no knight at position **a**, as well as strands that describe a knight at attacking positions **h** and **f**. This implements the logical statement $((\neg h \wedge \neg f) \vee \neg a)$
3. In tube 2, digest strands that have a knight present at position **a**
4. Remove the DNA oligos used to perform the above digestions
5. Go to step 1, repeating with square **b**.

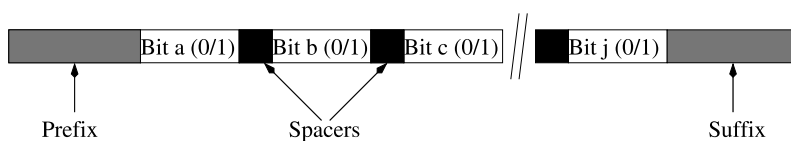
Steps 1 through 4 implement the following: “There may or may not be a knight in square **a**: if there is, then it is attacking squares **h** and **f**, so disallow this.” The algorithm only needs to be performed for squares **a**, **b**, **c**, **d**, and **f**, as square **e**, by the rules of chess, cannot threaten or be threatened on a board this size, and any illegal interactions that squares **g**, **h**, and **i** may have are with **a**, **b**, **c**, **d**, and **f**, and have already been dealt with. At the conclusion of this stage, any remaining full-length strands are recovered, as they should encode legal boards.

The “mark and destroy” digestion operation is implemented as follows. If we wish to retain (i. e., select) strands encoding variable a to have value 1, DNA oligonucleotides corresponding to the complement of the $a = 0$ sequence are added to the tube, and anneal to all strands encoding $a = 0$. RNase H is then added to the solution. Ribonuclease H (RNase H) is an endoribonuclease which specifically hydrolyzes the phosphodiester bonds of RNA hybridized to DNA. RNase H does not digest single or double-stranded DNA, so his operation therefore leaves intact only those strands encoding $a = 1$, in a fashion similar to the removal operation of the parallel filtering model [4].

The results obtained (described in [17]) were extremely encouraging: out of 43 output strands sampled, only one contained an illegal board. Given that the population sampled encoded 127 knights, this gave an overall knight placement success rate of 97.7%.

Computing on Surfaces

Another experiment that makes use of the “mark and destroy” paradigm is described in [26] (although early work was performed from 1996). The key difference between



DNA Computing, Figure 10
Template for RNA strands (from [3])

this and previous experiments is that the DNA strands used are tethered to a support rather than being allowed to float freely in solution. The authors argue that this approach greatly simplifies the automation of the (potentially very many) repetitive chemical processes required during the performance of an experiment.

The authors report a DNA-based solution to a small instance of SAT. The specific problem solved is

$$(w \vee x \vee y) \wedge (w \vee \neg y \vee z) \wedge (\neg x \vee y) \wedge (\neg w \vee \neg y).$$

16 unique DNA strands were synthesized, each one corresponding to one of the $2^4 = 16$ combinations of variable values. The actual encodings are given in Table 1 (taken from [26]).

Each of the 16 sets of strands was then affixed to a specific region of a gold coated surface, so that each solution to the SAT problem was represented as an individual cluster of strands.

The algorithm then proceeds as follows. For each clause of the problem, a cycle of “mark”, “destroy”, and “unmark” operations is carried out. The goal of each cycle is to destroy the strands that do *not* satisfy the appropriate clause. Thus, in the first cycle, the objective is to destroy strands that do not satisfy the clause $(w \vee x \vee y)$. Destruction is achieved by “protecting”, or *marking* strands that *do* satisfy the clause by annealing to them their complementary strands. *E. coli* exonuclease I is then used to digest unmarked strands (i. e., any single-stranded DNA).

By inspection of Table 1, we see that this applies to only two strands, S_0 (0000) and S_1 (0001). Thus, in cycle 1 the

complements of the 14 other strands

$$(w = 1(S_8, S_9, S_{10}, S_{11}, S_{12}, S_{13}, S_{14}, S_{15});$$

$$x = 1(S_4, S_5, S_6, S_7, S_{12}, S_{13}, S_{14}, S_{15});$$

$$y = 1(S_2, S_3, S_6, S_7, S_{10}, S_{11}, S_{14}, S_{15}))$$

were combined and hybridized to the surface before the exonuclease I was added. The surface was then regenerated (the unmark operation) to return the remaining surface-bound oligos ($S_2 - S_{15}$) to single-stranded form. This process was repeated three more times for the remaining three clauses, leaving a surface containing only strands encoding a legal solution to the SAT problem.

The remaining molecules were amplified using PCR and then hybridized to an addressed array. The results of fluorescence imaging clearly showed four spots of relatively high intensity, corresponding to the four regions occupied by legal solutions to the problem (S_3, S_7, S_8 , and S_9).

Although these results are encouraging as a first move toward more error-resistant DNA computing, as the authors themselves acknowledge, there remain serious concerns about the scalability of this approach (1,536 individual oligos would be required for a 36-bit, as opposed to 4-bit, implementation).

Computing with Hairpins

The tendency of DNA molecules to self-anneal was exploited by Sakamoto et al. in [35] for the purposes of solving a small instance of SAT. The authors encode the given formula in “literal strings” which are conjunctions of the literals selected from each SAT clause (one literal per clause). A formula is satisfiable if there exists a literal string that does not contain any variable together with its negation. If each variable is encoded as a DNA subsequence that is the Watson–Crick complement of its negation then any strands containing a variable and its negation self-anneal to form “hairpin” structures. These can be distinguished from non-hairpin structure-forming strands, and removed. The benefit of this approach is that it does not require physical manipulation of the DNA, only temperature cycling. The drawback is that it requires 3^m literal strings for m clauses, thus invoking once again the scalability argument.

Gel-Based Computing

A much larger (20 variable) instance of 3-SAT was successfully solved by Adleman’s group in an experiment described in 2002 [9]. This is, to date, the largest known problem instance successfully solved by a DNA-based

DNA Computing, Table 1

Strands used to represent SAT variable values

Strand	Sequence	wxyz
S_0	CAACCCAA	0000
S_1	TCTCAGAG	0001
S_2	GAAGGCAT	0010
S_3	AGGAATGC	0011
S_4	ATCGAGCT	0100
S_5	TTGGACCA	0101
S_6	ACCATTGG	0110
S_7	GTTGGGTT	0111
S_8	CCAAGTTG	1000
S_9	CAGTTGAC	1001
S_{10}	TGGTTTGG	1010
S_{11}	GATCCGAT	1011
S_{12}	ATATCGCG	1100
S_{13}	GGTTCAAC	1101
S_{14}	AACCTGGT	1110
S_{15}	ACTGGTCA	1111

computer; indeed, as the authors state, “this computational problem may yet be the largest yet solved by non-electronic means” [9].

The architecture underlying the experiment is related to the Sticker Model described by Roweis et al. [34]. The difference here is that only separation steps are used – the application of stickers is not used. Separations are achieved by using oligo probes immobilized in polyacrylamide gel-filled glass modules, and strands are pulled through them by electrophoresis. Strands are removed (i.e., retained in the module) by virtue of their hybridizing to the immobilized probes, with other strands free to pass through the module and be subject to further processing. Captured strands may be released and transported (again via electrophoresis) to other modules for further processing.

The potential benefits of such an approach are clear; the use of electrophoresis minimizes the number of laboratory operations performed on strands, which, in turn, increases the chance of success of an experiment. Since strands are not deliberately damaged in any way, they, together with the glass modules, are potentially reusable for multiple computations. Finally, the whole process is potentially automatable.

The problem solved was a 20-variable, 24-clause 3-SAT formula Φ , with a unique satisfying truth assignment. These are

$$\begin{aligned}\Phi = & (\neg x_{13} \vee x_{16} \vee x_{18}) \wedge (x_5 \vee x_{12} \vee \neg x_9) \\ & \wedge (\neg x_{13} \vee \neg x_2 \vee x_{20}) \wedge (x_{12} \vee x_9 \vee \neg x_5) \\ & \wedge (x_{19} \vee \neg x_4 \vee x_6) \wedge (x_9 \vee x_{12} \vee \neg x_5) \\ & \wedge (\neg x_1 \vee x_4 \vee \neg x_{11}) \wedge (x_{13} \vee \neg x_2 \vee \neg x_{19}) \\ & \wedge (x_5 \vee x_{17} \vee x_9) \wedge (x_{15} \vee x_9 \vee \neg x_{17}) \\ & \wedge (\neg x_5 \vee \neg x_9 \vee \neg x_{12}) \wedge (x_6 \vee x_{11} \vee x_4) \\ & \wedge (\neg x_{15} \vee \neg x_{17} \vee x_7) \wedge (\neg x_6 \vee x_{19} \vee x_{13}) \\ & \wedge (\neg x_{12} \vee \neg x_9 \vee x_5) \wedge (x_{12} \vee x_1 \vee x_{14}) \\ & \wedge (x_{20} \vee x_3 \vee x_2) \wedge (x_{10} \vee \neg x_7 \vee \neg x_8) \\ & \wedge (\neg x_5 \vee x_9 \vee \neg x_{12}) \wedge (x_{18} \vee \neg x_{20} \vee x_3) \\ & \wedge (\neg x_{10} \vee \neg x_{18} \vee \neg x_{16}) \wedge (x_1 \vee \neg x_{11} \vee \neg x_{14}) \\ & \wedge (x_8 \vee \neg x_7 \vee \neg x_{15}) \wedge (\neg x_8 \vee x_{16} \vee \neg x_{10})\end{aligned}$$

with a unique satisfying assignment of

$$\begin{aligned}x_1 = F, \quad x_2 = T, \quad x_3 = F, \quad x_4 = F, \quad x_5 = F, \\ x_6 = F, \quad x_7 = T, \quad x_8 = T, \quad x_9 = F, \quad x_{10} = T, \\ x_{11} = T, \quad x_{12} = T, \quad x_{13} = F, \quad x_{14} = F, \quad x_{15} = T, \\ x_{16} = T, \quad x_{17} = T, \quad x_{18} = F, \quad x_{19} = F, \quad x_{20} = F.\end{aligned}$$

As there are 20 variables, there are $2^{20} = 1,048,576$ possible truth assignments. To represent all possible assignments, two distinct 15 base *value sequences* were assigned

to each variable x_k ($k = 1, \dots, 20$), one representing true (T), X_k^T , and one representing false (F), X_k^F . A mix and split generation technique similar to that of Faulhammer et al. [17] was used to generate a 300-base *library sequence* for each of the unique truth assignments. Each library sequence was made up of 20 value sequences joined together, representing the 20 different variables. These library sequences were then amplified with PCR.

The computation proceeds as follows: for each clause, a glass clause module is constructed which is filled with gel and contains covalently bound probes designed to capture only those library strands that *do satisfy* that clause; strands that do not satisfy the clause are discarded.

In the first clause module ($\neg x_{13} \vee x_{16} \vee x_{18}$) strands encoding X_3^F , X_{16}^F , and X_{18}^T are retained, while strands encoding X_3^T , X_{16}^T , and X_{18}^F are discarded. Retained strands are then used as input to the next clause module, for each of the remaining clauses. The final (24th) clause module should contain only those strands that have been retained in all 24 clause modules and hence encode truth assignments satisfying Φ .

The experimental results confirmed that a unique satisfying truth assignment for Φ was indeed found using this method. Impressive though it is, the authors still regard with scepticism claims made for the potential superiority of DNA-based computers over their traditional silicon counterparts. In a separate interview, Len Adleman stated that “DNA computers are unlikely to become stand-alone competitors for electronic computers. We simply cannot, at this time, control molecules with the deftness that electrical engineers and physicists control electrons” [31]. However, “they [DNA computers] enlighten us about alternatives to electronic computers and studying them may ultimately lead us to the true ‘computer of the future’” [9]. In the next Section, we consider how the focus of DNA computing has since shifted away from the solution of “traditional” problems.

Assessment

In February 1995, one of the founding fathers of the theory of *computational complexity* (q.v.) published a short paper on DNA computing. In his article, titled “On the Weight of Computations” [22], Juris Hartmanis sounded a cautionary note amidst the growing interest surrounding DNA computing. By calculating the smallest amount of DNA required to encode *all possible paths* through a graph, he calculated that Adleman’s experiment, if scaled up from 7 cities to 200 cities, would require an initial set of DNA strands that would weigh more than the Earth. Hartmanis was quick to point out the value of the initial work –

“Adleman’s molecular solution of the Hamiltonian path problem is indeed a magnificent achievement and may initiate a more intensive exploration of molecular computing and computing in biological systems”. However, he also emphasised the long-term futility of hoping that DNA-based computers could ever beat silicon machines in this particular domain. Soon after Adleman’s experiment, hopes were expressed that a lot of the promise of molecular computers could be derived from their massive inherent parallelism – each operation is performed at the same time on trillions of DNA strands. However, as Turing showed, computers are not limited to any particular physical construction, and a DNA computer will suffer just as much as its silicon counterpart from the “exponential curse”. If we require a molecular algorithm for a difficult problem to run in reasonable *time* then there is a fundamental requirement (unless P is shown to be equivalent to NP) for an exponential amount of *space* (in this case, the amount of DNA required). As Hartmanis observed, “... the exponential function grows too fast and the atoms are a bit too heavy to hope that the molecular computer can break the exponential barrier, this time the weight barrier”.

This view would now seem to largely reflect the community consensus. As Ogihara and Ray argued in 2000, “It is foolish to attempt to predict the future of technology, but it may be that the ideal application for DNA computation does not lie in computing large NP problems” [29]. In an interview in 2002, nanotechnologist Ned Seeman stated that “I am not a computer scientist, but I suspect it [molecular computing] is not well suited to traditional problems. I certainly feel that it is better suited to algorithmic self-assembly and biologically-oriented applications” [36]. That is not to say that molecular-based computations do not have a future, and in the final Section we briefly discuss some possible future directions that molecular computing may take (and offer pointers to other articles in the current volume).

Future Directions

If molecular computations are to be scalable and/or sustainable then it is clear that they should be performed with a minimum of human intervention. The type of experiment originally performed by Adleman was feasible because it required a relatively small number of biological steps (although it still took a week of bench time to carry out). However, the number of manipulations (and the amount of material) required for a non-trivial computation would quickly render this approach infeasible. Although attempts have been made to automate molec-

ular computations using, for example, microfluidics [39], this does not allow us to avoid the fundamental issue of scalability.

One promising subfield concerns molecular computing using machines (or “automata”) constructed from DNA strands and other biomaterials. The construction of *molecular automata* (q.v.) was demonstrated by Benenson et al. in [6]. This experiment builds on the authors’ earlier work [7] on the construction of biomolecular machines. In [6], the authors describe the construction of a molecular automaton that uses the process of DNA backbone hydrolysis and strand hybridization, fuelled by the potential free energy stored in the DNA itself.

One aim of researchers in this field is to build automata that may operate within living cells. Such devices may offer possibilities in terms of novel therapeutics, drug synthesis, bio-nanotechnology or *amorphous computing* (q.v.) Theoretical studies in DNA computing based on abstract models of the cell are already well-established (see the article on *membrane computing*), but recent work on *bacterial computing* (q.v.) has illustrated the feasibility of engineering living cells for the purposes of human-defined computation.

Finally, the notion of *biomolecular self-assembly* (q.v.) suggests ways in which the tendency of molecules to form spontaneously ordered structures may be harnessed, either for the purposes of computation or for engineering-based applications (for example, *DNA-templated self-assembly of proteins and nanowires* (q.v.)). *Algorithmic self-assembly* has been demonstrated in the laboratory by Mao et al. [27]. This builds on work done on the self-assembly of periodic two-dimensional arrays (or “sheets”) of DNA tiles connected by “sticky” pads [44,45]. The authors of [27] report a one-dimensional algorithmic self-assembly of DNA triple-crossover molecules (tiles) to execute four steps of a logical XOR operation on a string of binary bits.

Triple-crossover molecules contain four strands that self-assemble through Watson–Crick complementarity to produce three double helices in roughly a planar formation. Each double helix is connected to adjacent double helices at points where their strands cross over between them. The ends of the core helix are closed by hairpin loops, but the other helices may end in sticky ends which direct the assembly of the macrostructure. The tiles then self-assemble to perform a computation. The authors of [27] report successful XOR computations on pairs of bits, but note that the scalability of the approach relies on proper hairpin formation in very long single-stranded molecules, which cannot be assumed.

In 2006, Paul Rothemund produced a ground-breaking piece of work, which he called “DNA origami” [33].

His breakthrough was to describe a completely novel method for constructing *arbitrary* two-dimensional DNA-based structures. “When it comes to making shapes out of DNA”, explained Lloyd Smith in a commentary article [37], “the material is there, and its properties are understood. What was missing was a convincing, universal design scheme to allow our capabilities to unfold to the full”. Rothemund demonstrated the generality of his approach by showing how it could build structures as diverse as a triangle, a five-pointed star, and even a “smiley” face and a map of the Americas. The shapes that he managed to construct were ten times more complex than anything that had been constructed from DNA before. Rothemund’s scheme was different to previous tile-based methods in that it used only a *single* long DNA strand at its foundation. By repeatedly folding a long strand of DNA around in a maze-like pattern, a scaffold was formed that traced the outline of the desired object. This structure was then “clipped” into place by short DNA strands, which Rothemund referred to as “staples”. He used as the main building block a 7,000 base sequence of DNA taken from the M13 virus. He developed a computer program that would take a human-designed folding path (the “shape”), map it onto the sequence of the DNA strand and then generate the sequences of the staples that would be required to hold it all together. In order to demonstrate the power of his method, he built a DNA map of the Americas, where a single nanometer (one billionth of a meter) represented 200 kilometers in real life (the map was 75 nanometers across). “The results that emerge are stunning”, praised Lloyd Smith, after seeing the work. The possible impact of this work may well be huge, with one possible application being a “nanoworkbench”. In 2003, John Reif and his team had shown how a simple DNA scaffold could cause proteins to self-assemble into regular, periodic two-dimensional grids [46]. By using his approach to build much more complex holding structures, Rothemund argued that it could now be used to carry out biological experiments at the nanoscale, studying the behavior of complex protein assemblies, such as drugs, in a spatially-controlled environment. As Lloyd Smith observes, “Thus equipped not only with DNA building materials and an understanding of their structural and chemical properties, but also with a versatile general approach to weaving them together, we are arriving at a new frontier in our pursuit of ever-smaller structures. The barrier we have to surmount next is to deploy our knowledge to develop structures and devices that are really useful. Happily, in that endeavor we are now perhaps limited more by our imagination than our ability” [37].

Bibliography

Primary Literature

1. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. *Science* 266:1021–1024
2. Adleman LM (1995) On constructing a molecular computer. Draft, University of Southern California
3. Amos M (2005) Theoretical and Experimental DNA Computation. Springer, Berlin
4. Amos M, Gibbons A, Hodgson D (1996) Error-resistant implementation of DNA computations. In: Landweber LF, Baum EB (eds) 2nd Annual Workshop on DNA Based Computers. Princeton University, NJ, 10–12 June 1996. American Mathematical Society
5. Arkin A, Ross J (1994) Computational functions in biochemical reaction networks. *Biophys J* 67:560–578
6. Benenson Y, Adam R, Paz-Livneh T, Shapiro E (2003) DNA molecule provides a computing machine with both data and fuel. *Proc Natl Acad Sci* 100:2191–2196
7. Benenson Y, Paz-Elizur T, Adar R, Keinan E, Livneh Z, Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. *Nature* 414:430–434
8. Bennett CH (1982) The thermodynamics of computation – a review. *Int J Theor Phys* 21:905–940
9. Braich RS, Chelyapov N, Johnson C, Rothemund PWK, Adleman L (2002) Solution of a 20-variable 3-SAT problem on a DNA computer. *Science* 296:499–502
10. Bray D (1995) Protein molecules as computational elements in living cells. *Nature* 376:307–312
11. Breslauer KJ, Frank R, Blocker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci* 83(11):3746–3750
12. Brown TA (1993) *Genetics: A Molecular Approach*. Chapman and Hall, New York
13. Campbell-Kelly M, Aspray W (2004) *Computer: A History of the Information Machine*, 2nd edn. Westview Press, Colorado
14. Conrad M (1985) On design principles for a molecular computer. *Commun ACM* 28:464–480
15. Conrad M, Liberman EA (1982) Molecular computing as a link between biological and physical theory. *J Theor Biol* 98: 239–252
16. Cook S (1971) The complexity of theorem proving procedures. *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing*, pp 151–158
17. Faulhammer D, Cukras AR, Lipton RJ, Landweber LF (2000) Molecular computation: RNA solutions to chess problems. *Proc Nat Acad Sci* 97:1385–1389
18. Feynman RP (1961) There’s plenty of room at the bottom. In: Gilbert D (ed) *Miniaturization*. Reinhold, New York, pp 282–296
19. Garey MR, Johnson DS (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. WH Freeman and Company, New York
20. Gibbons AM (1985) *Algorithmic Graph Theory*. Cambridge University Press, Cambridge
21. Guarnieri F, Fliss M, Bancroft C (1996) Making DNA add. *Science* 273:220–223
22. Hartmanis J (1995) On the weight of computations. *Bull Euro Assoc Theor Comput Sci* 55:136–138
23. Hjelmsfelt A, Schneider FW, Ross J (1993) Pattern recognition in coupled chemical kinetic systems. *Science* 260:335–337

24. Hjeltnelt A, Weinberger ED, Ross J (1991) Chemical implementation of neural networks and Turing machines. *Proc Nat Acad Sci* 88:10983–10987
25. Lipton RJ (1995) DNA solution of hard computational problems. *Science* 268:542–545
26. Liu Q, Wang L, Frutos AG, Condon AE, Corn RM, Smith LM (2000) DNA computing on surfaces. *Nature* 403:175–179
27. Mao C, LaBean TH, Reif JH, Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature* 407:493–496
28. Mullis KB, Ferré F, Gibbs RA (eds) (1994) *The Polymerase Chain Reaction*. Birkhauser, Boston
29. Ogihara M, Ray A (2000) DNA computing on a chip. *Nature* 403:143–144
30. Ouyang Q, Kaplan PD, Liu S, Libchaber A (1997) DNA solution of the maximal clique problem. *Science* 278:446–449
31. Regalado A (2002) DNA computing. MIT Technology Review. <http://www.technologyreview.com/articles/00/05/regalado0500.asp>. Accessed 26 May 2008
32. Rivest R, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public key cryptosystems. *Comm ACM* 21:120–126
33. Rothmund PWK (2006) Folding DNA to create nanoscale patterns. *Nature* 440:297–302
34. Roweis S, Winfree E, Burgoyne R, Chelyapov NV, Goodman MF, Rothmund PWK, Adleman LM (1996) A sticker based architecture for DNA computation. In: Landweber LF, Baum EB (eds) 2nd Annual Workshop on DNA Based Computers. Princeton University, NJ, 10–12 June 1996. American Mathematical Society
35. Sakamoto K, Gouzu H, Komiya K, Kiga D, Yokoyama S, Yokomori T, Hagiya M (2000) Molecular computation by DNA hairpin formation. *Science* 288:1223–1226
36. Smalley E (2005) Interview with Ned Seeman. *Technology Research News*, May 4
37. Smith LM (2006) Nanostructures: The manifold faces of DNA. *Nature* 440:283–284
38. Stubbe H (1972) History of Genetics – from Prehistoric times to the Rediscovery of Mendel's Laws. MIT Press, Cambridge
39. van Noort D, Gast F-U, McCaskill JS (2002) DNA computing in microreactors. In: Jonoska N, Seeman NC (eds) *DNA Computing: 7th International Workshop on DNA-Based Computers*. LNCS, vol 2340. Springer, Berlin, pp 33–45
40. Watkins JJ (2004) *Across the Board: The Mathematics of Chess Problems*. Princeton University Press, Princeton
41. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribose nucleic acid. *Nature* 171:964
42. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171:737–738
43. Watson JD, Hopkins NH, Roberts JW, Steitz JA, Weiner AM (1987) *Molecular Biology of the Gene*, 4th edn. Benjamin/Cummings, Menlo Park
44. Winfree E, Liu F, Wenzler L, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystals. *Nature* 394:539–544
45. Winfree E (1998) Algorithmic self-assembly of DNA. Ph D thesis, California Institute of Technology
46. Yan H, Park SH, Finkelstein G, Reif JH, LaBean TH (2003) DNA-templated self-assembly of protein arrays and highly conductive nanowires. *Science* 301:1882–1884

Books and Reviews

- Adleman L (1998) Computing with DNA. *Sci Am* 279:54–61
- Amos M (2006) *Genesis Machines: The New Science of Biocomputing*. Atlantic Books, London
- Forbes N (2004) *Imitation of Life: How Biology is Inspiring Computing*. MIT Press, Cambridge
- Gonick L, Wheelis M (1983) *The Cartoon Guide to Genetics*. Harper Perennial, New York
- Jones R (2004) *Soft Machines: Nanotechnology and Life*. Oxford University Press, Oxford
- Păun G, Rozenberg G, Salomaa A (1998) *DNA Computing: New Computing Paradigms*. Springer, Berlin
- Pool R (1995) A boom in plans for DNA computing. *Science* 268:498–499
- Watson J (2004) *DNA: The Secret of Life*. Arrow Books, London

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires

BAOQUAN DING¹, YAN LIU², SHERRI RINKER²,
HAO YAN²

¹ Molecular Foundry, Lawrence Berkeley National Lab, Berkeley, USA

² Department of Chemistry and Biochemistry and Biodesign Institute, Arizona State University, Tempe, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 DNA Self-Assembly
 DNA-Templated Assembly of Protein Arrays
 DNA-Templated Highly Conductive Nanowires
 Future Directions
 Bibliography

Glossary

DNA Deoxyribonucleic acid, is a nucleic acid molecule that contains the genetic instructions used in the development and functioning of all living organisms. Chemically, DNA is a long polymer of simple units called nucleotides, which are held together by a backbone made of alternating sugars and phosphate groups. Attached to each sugar is one of four types of molecules called bases: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C).

Double-crossover (DX) motif The DX motif consists of two DNA double helices linked in two different places.

These molecules are related to intermediates in genetic recombination but, in the molecules used in DNA self-assembly, the linkages are usually between strands of opposite polarity, rather than the same polarity.

Paranemic-crossover DNA A DNA motif that can be formed by reciprocal exchange between strands of the same polarity on two DNA double helices at every possible position.

Definition of the Subject

DNA self-assembly is a very useful and powerful tool for bottom-up nanofabrication. It consists of combining unusual DNA motifs by specific structurally well-defined cohesive interactions (sticky ends) to produce target materials with predictable 3D structure. This method has generated versatile DNA nanostructures including polyhedral catenanes, robust nanomechanical devices, and a variety of periodic and aperiodic arrays in two dimensions. DNA self-assembled structures have been used as the template for different guest functional molecules such as proteins, metallic nano-particles, DNA based nano-devices and highly conductive nanowires. Regular lattices made of DNA could hold copies of large biological molecules in a highly ordered array for x-ray crystallography to determine their structure, an important step in the “rational” design of drugs. Alternatively, the lattices could serve as scaffolding for nanoelectronic components, either as a working device or as a step in the manufacture of a device. Materials could be constructed-either made of DNA or templated by DNA-with precisely designed structures in the nanometer scale. The organizational capabilities of DNA nanotechnology are just beginning to be explored, and the field is expected to be able to organize a variety of species that will lead to exciting and possibly revolutionary materials.

Introduction

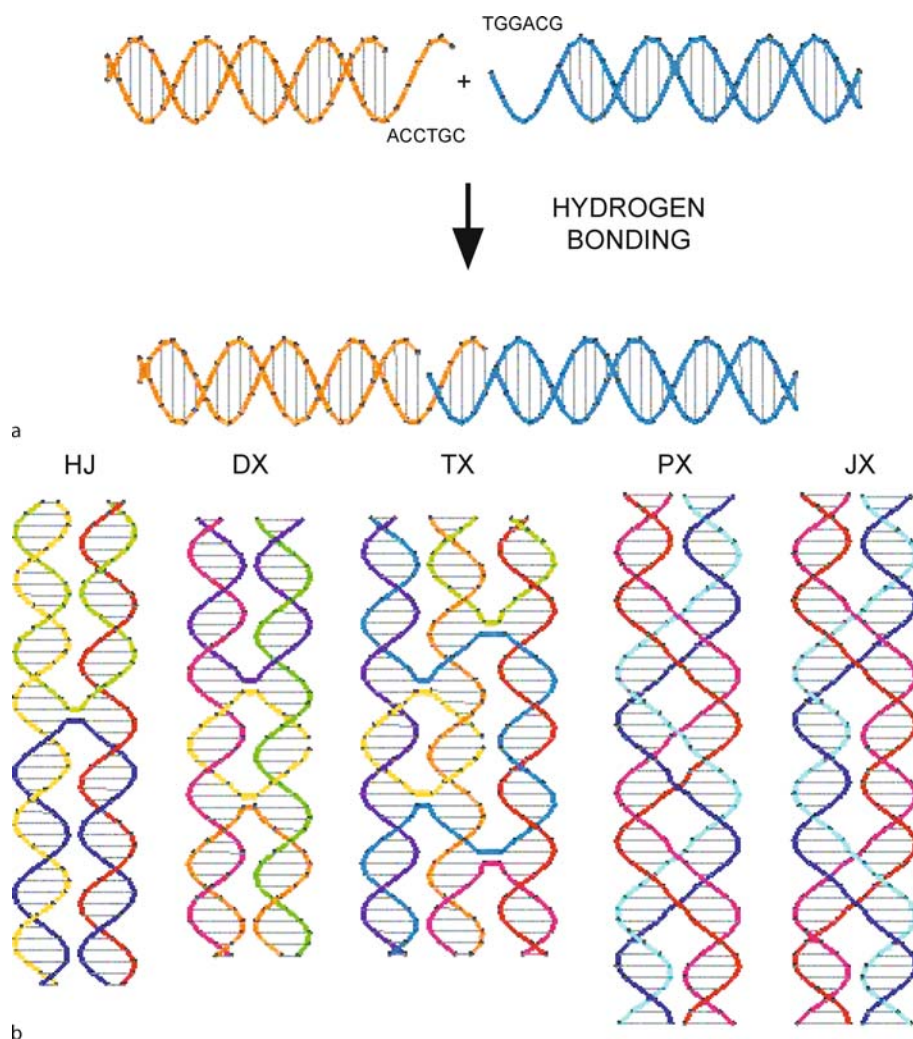
DNA is well known as the genetic material for information storage and duplication in biology. But, the applications of DNA are not restricted to biological science. DNA's molecular structure and chemical properties make it centrally important in biology. These properties make it to be highly useful as an engineering material for nanoscale constructions. The diameter of DNA double helix is about 2 nm and its helical repeat is about 3.5 nm. The key feature of DNA that makes it useful as a nanoscale building block is its specificity and programmability in inter-molecular interactions. The famous Watson-Crick base pairing consists of Adenine (A) paired specifically with Thymine (T) and Guanine (G) paired specifically with Cy-

tosine (C) through hydrogen bonds within the context of the DNA double helical backbone. It is also possible to get two different double helices to cohere end-to-end by having short overhangs at the ends of the strands; these overhangs are called ‘sticky ends’. It has been shown that sticky ends cohere to form the conventional structure of B-DNA. The structure of sticky-ended DNA complexes is well-defined and specifically predictable. Given that the persistence length of DNA is about 50 nm, sticky ends can be used to assemble DNA molecules. DNA can be manipulated using commercially available enzymes for site-selective DNA cleavage, ligation, labeling, transcription, replication, kination, and methylation. DNA nanotechnology is further empowered by well-established methods for purification and structural characterization and by solid-phase synthesis, so that any designed DNA strand can be constructed. The above advantages of DNA make DNA based nanotechnology a rapidly progressing and exciting research field [1,2,3], especially in self-assembled nanostructures [4,5,6], nanorobotics [7,8,9,10,11,12], and nanocomputation [13,14,15].

DNA Self-Assembly

The simple hybridization of single strand DNA to assemble into linear duplex structures is not sufficient for nanoconstruction purposes, because it can not produce precise complex systems in 2D or 3D structures. To do that, it is necessary to work with branched DNA molecules. Branched DNA occurs naturally in living systems. For example, ephemeral branch points are found as Holliday junction intermediates in the process of genetic recombination. Branching can occur when exchange occurs between two DNA strands. A single exchange event can produce a Holliday junction-like molecule. It is a 4-arm branch, which can be stabilized by ensuring that there is no twofold symmetry around its branch point; minimizing sequence symmetry is key to the design of all unusual DNA motifs. A double exchange between DNA double helices produces double crossover (DX) molecules [4], and double exchange between three successive helices produced triple crossover (TX) molecules [5]. Some examples are shown in Fig. 1. Synthetic DNA complexes can be designed to have fixed branch points. Thus, using DNA as the building blocks for complex materials with nanoscale features is simple: Take synthetic branched DNA molecules with programmed sticky ends, and get them to self-assemble into the desired structures.

The first artificial DNA structure is a stick-cube, whose edges are double helices. More complex polyhedra and



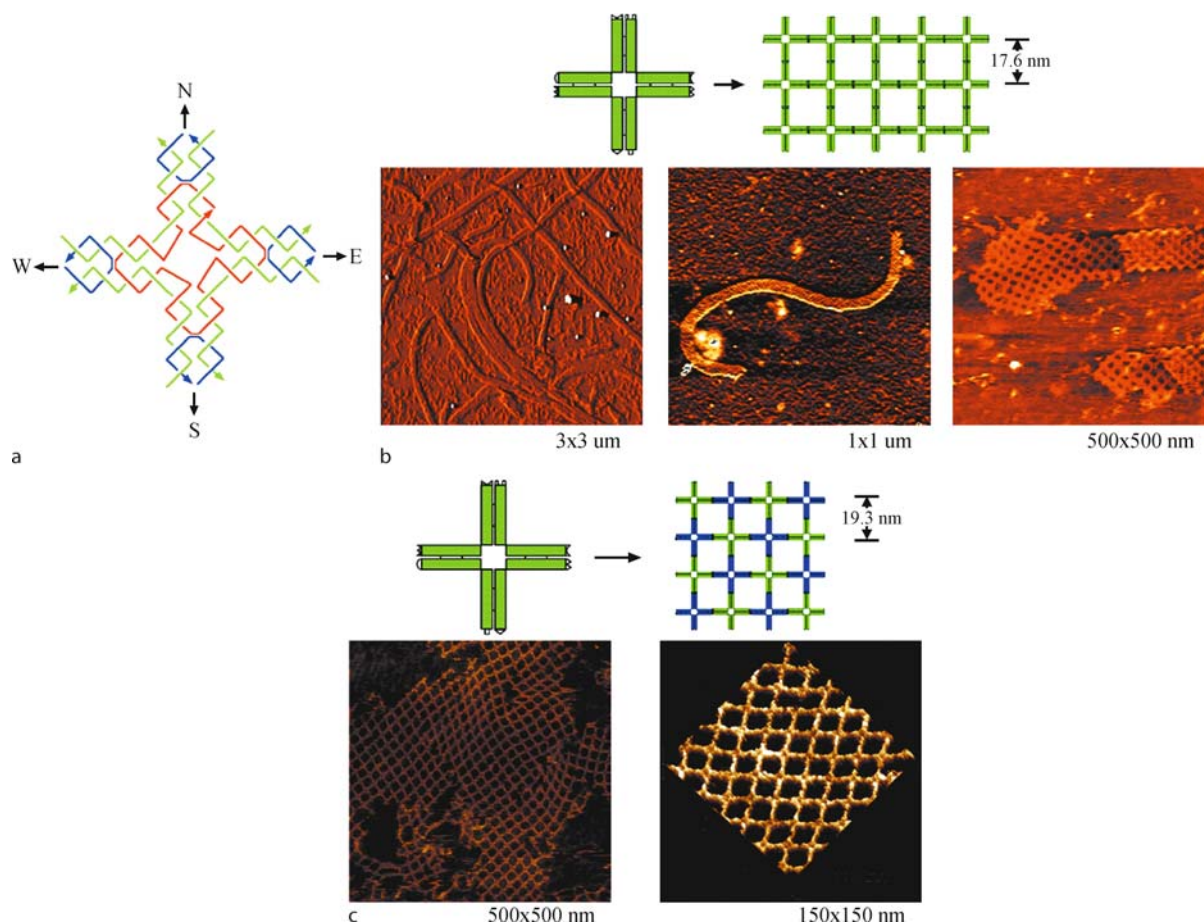
DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 1

Component of DNA self-assembly. **a** Sticky Ended Cohesion. Two linear double helical molecules of DNA are shown at the *top of panel a*. The *right end* of the *orange* molecule and *left end* of the *blue* molecule have single-strand extension 'sticky ends' that are totally complementary to each other. The *middle* portion shows that, under the proper conditions, these bind to each other specifically by hydrogen bonding. (5' and 3' ends of DNA molecule are not shown). **b** Key Motifs in DNA self-assembly. On the *left* is a Holliday Junction (HJ), a 4-arm junction that results from a single strand exchange between double helices. To its *right* is a double crossover (DX) molecule, resulting from a double exchange. To the *right* of the DX is a triple crossover (TX) molecule that results from two successive double reciprocal exchanges. The HJ, the DX and the TX molecule all contain exchange between strands of opposite polarity. To the *right* of the TX molecule is a paranemic crossover (PX) molecule, where two double helices exchange strands at every possible points where the helices come into proximity. To the *right* of the PX molecule is a JX₂ molecule that lacks two of the crossover of the PX molecule. The exchange in the PX and JX₂ molecule are between strands of the same polarity

topological structures, followed [2]. DX molecules were the initial stable and rigid building blocks for the construction of periodic assemblies and the formation of the first 2D DNA array. DX, TX and other DNA motifs have been used to produce algorithmic assemblies displaying well designed patterns. These different DNA lattices provide multiple attachment sites for complex programmable struc-

tures and have lead to diverse possibilities for templating useful components and interesting chemistry.

Since DX and TX tiles are designed with their helices parallel and coplanar, their 2D lattices grow well in the direction parallel to the helix axis and not as well in the direction perpendicular to it. H. Yan et al. have designed a 4×4 cross tile shown in Fig. 2 to get rid of these problem

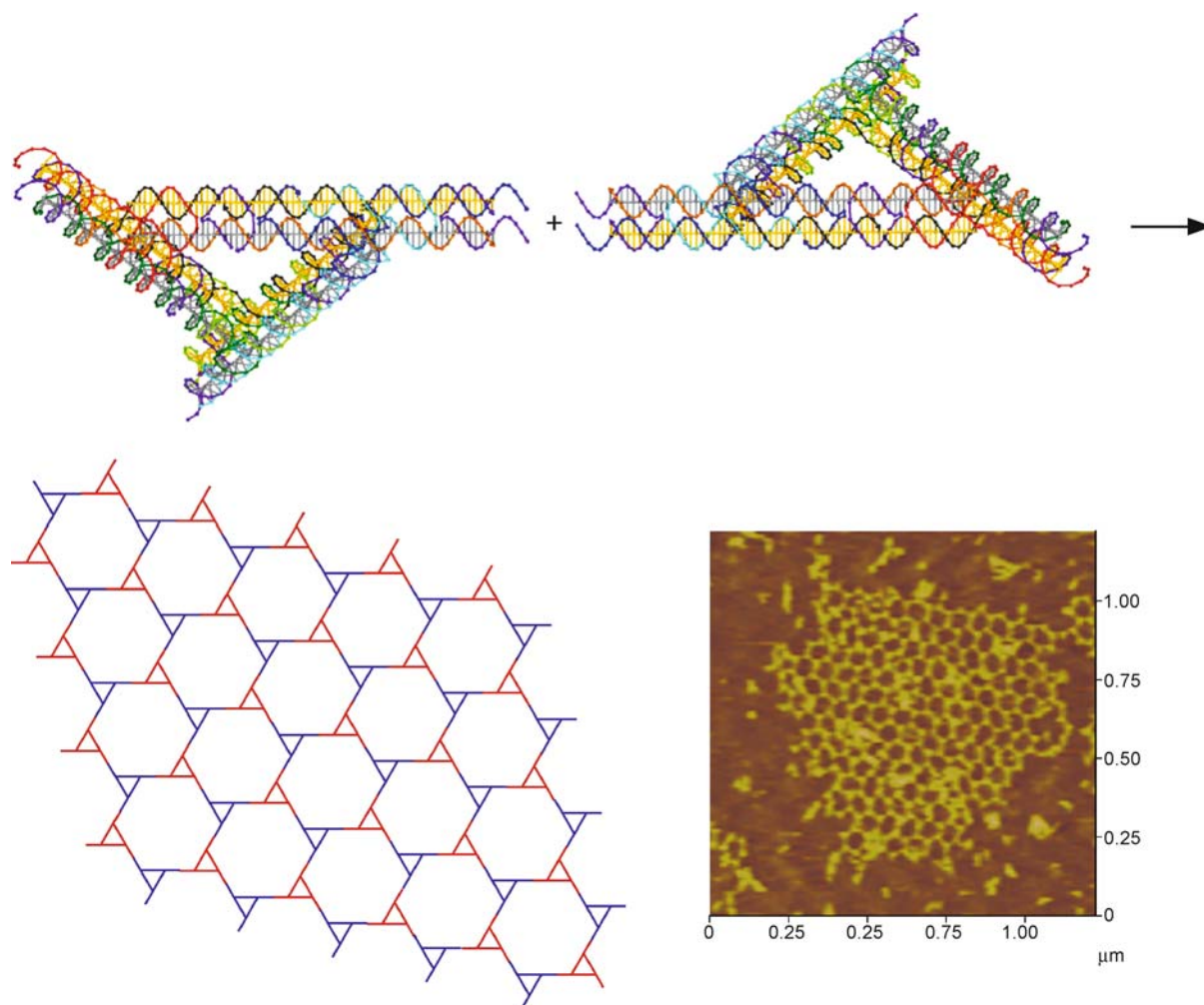


DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 2

Self-assembly of DNA nanoribbons and nanogrids using 4×4 DNA tiles. **a** The 4×4 tile strand structure. The tile contains nine oligonucleotides, shown as simplified backbone traces. One four-arm junction is oriented in each direction (N, S, E, W); the red strand participates in all four junctions and contain T_4 loops connecting adjacent junctions. **b** Self-assembly of nanoribbons with original design. Double-helical domains are illustrated as rectangles, and paired rectangles represent four-arm junctions. Complementary sticky ends are shown as matching geometric shapes. (Upper right) Designed structure of self-assembled lattice. (Bottom) AFM images of the nanoribbons. The left panel shows an amplitude-mode image and the right two panels are AFM images in height mode. **c** Self-assembly of 2D nanogrids with corrugated design. (Upper left) The component tile is drawn similar to that in **b**; positions of sticky ends are changed. The tile has two surfaces; one faces out of the plane (green) and the other faces into the plane (blue). (Upper right) Corrugated self-assembly. (Bottom) AFM images of the 2D lattice formed from the corrugated design. The right panel is a surface plot of a magnified region from the left panel

and grow a lattice with a square aspect ratio [16]. By programming the sticky-end association to vary the assembly strategy, they could control the preferred lattice formation. One strategy produced a high preponderance of uniform-width ribbon structure (Fig. 2b). Their second strategy, referred to as the corrugated design, cause tiles to associate with one another such that the same face of each tile is oriented up and down alternately in neighboring tiles; therefore, the surface curvature inherent in each tile should be canceled out within the assembly (Fig. 2c).

Natural objects have hexagonal 2D packing, but making hexagonal DNA 2D array was long an unfulfilled goal. The simple three-arm junction is floppy and even stiffer motifs were incapable of producing a hexagonal array. In addition, the simple sticky end is susceptible to variations of twist. B. Ding et al. designed a double layer triangle molecule to produce a trigonal-pseudo-hexagonal array [17]. Figure 3 show the structures and AFM image. The honey-comb nature of the arrangement is obvious. A key control experiment was to remove the sticky ends



DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 3

The assembly of a trigonal lattice by DX Cohesion. The *upper part* of the panel shows two triangles with each arm containing Double Crossover structures. They combine in the way shown in the *lower left* to produce a pseudo-hexagonal lattice. The *lower right* shows an AFM image of the array

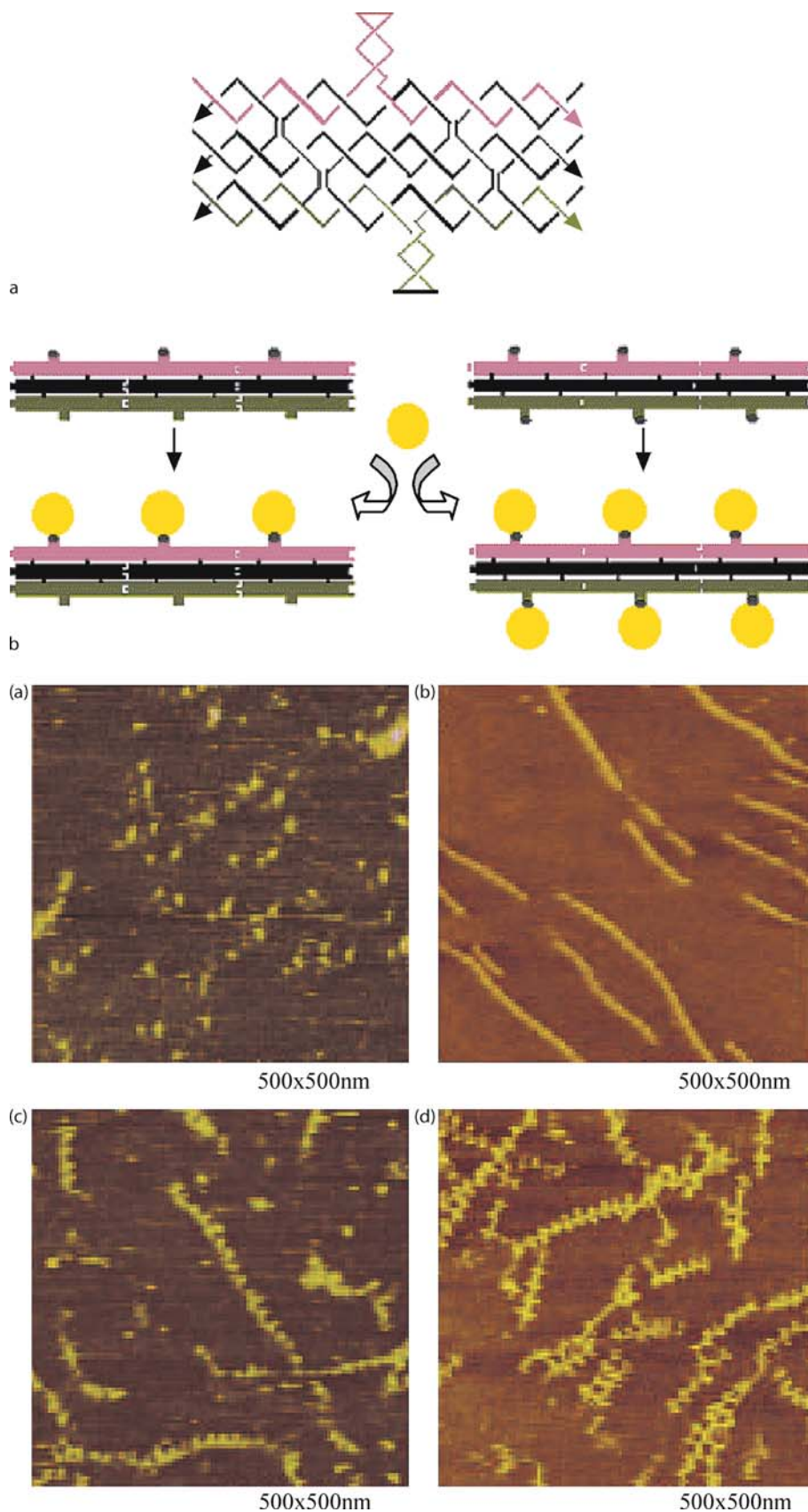
from one of the two helices in each direction: No array was found, demonstrating that the presence of DX cohesion helps to solve geometrical problems as exist when only a single helix is used.

P. Rothemund illustrated a new method of DNA self-assembly [18], known as DNA origami by folding a long single DNA strand (7249 bases) into well designed shapes such as smiley faces and many other complex 2D patterns. His design is performed in several steps. The first is to build a geometric model of a DNA structure that will approximate the desired shape. The shape is filled by a certain number of parallel helices. Then a single long scaffold strand is folded back and forth so that it comprised one of the two strands in every helix. Short staple strands are then

► DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 4

a Schematic drawing of the TX DNA-templated self-assembly of streptavidin linear array. **b** AFM images of the DNA-templated protein array. (a) streptavidin alone. (b) bare TX DNA tile arrays. (c) single-layer streptavidin array. (d) double-layer streptavidin array

designed, which can hybridize to segments of the scaffold strand and cross over from one row of helix to another like DX molecules. Thus, the staple strands hold the designed pattern together rigidly. Recently, W. Shih extended this ideal and has successfully constructed 3D origami structures [19].



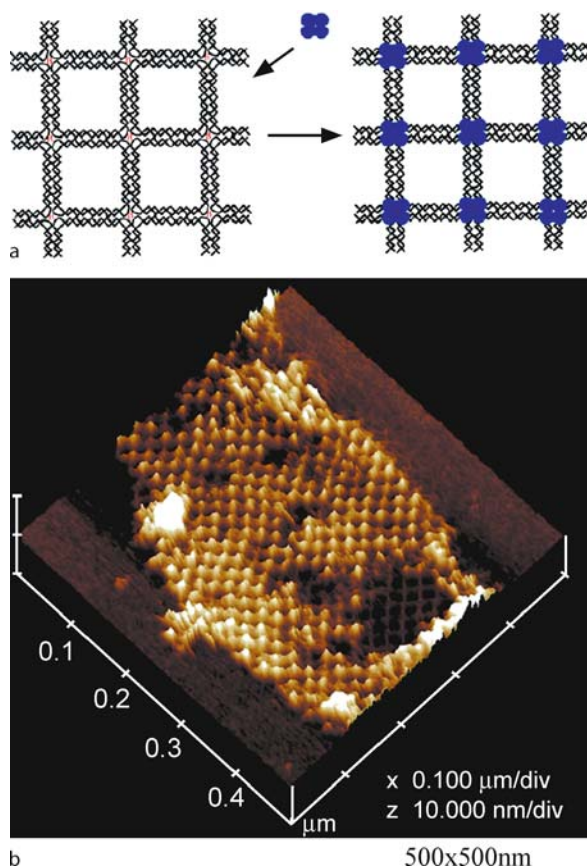
DNA-Templated Assembly of Protein Arrays

One of the key goals of DNA self-assembly is to build lattices and other rigid well defined structures from DNA in a controlled way so that DNA can form a designed template on which biological macromolecules can be orderly arrayed. DNA lattice could also bring enzymes and organic molecules into certain proximity, thus catalyze reactions.

H. Li et al. have demonstrated the use of a linear triple crossover (TX) DNA array for the assembly of two forms (single-layer or double-layer) streptavidin arrays through biotin-streptavidin interaction [20]. Figure 4a shows the design. The TX molecule used here consists of seven oligonucleotides hybridized to form three double-stranded helix domains lying in a plane and linked by strand exchange at four immobile crossover points. The TX molecule contains two stem loops protruding on two opposite sides of the TX tile plane. To obtain a single-layer streptavidin array, only the stem loop on one side of each TX tile is modified with biotin; to obtain a double-layer streptavidin array, stem loops on both sides of each TX tiles are modified with biotin. Streptavidin can bind to biotin-modified stem loops specifically and thus self-assemble into single or double-layer arrays respectively. Atomic force microscopy images in Fig. 4b demonstrate the DNA self-assembly TX array templates single-layer and double layer streptavidin particles. The distance between each pair of adjacent particles is about 17 nm, matching the designed parameter.

A 4×4 cross tile lattice was used to template the assembly of protein to form 2D array as shown in Fig. 5 [16]. The 4×4 DNA tile is modified by incorporating a biotin group into one of the T4 loops at the tile center. When streptavidin was added to the solution of the self-assembled DNA lattice, the interaction of streptavidin with biotin leads to the formation of periodic streptavidin array. Individual protein molecules are visible as separate peaks in the AFM image (Fig. 5). Further design evolution of the 4×4 cross tile system to a two tile type (a and b) allowed for more complex structures and patterns [21].

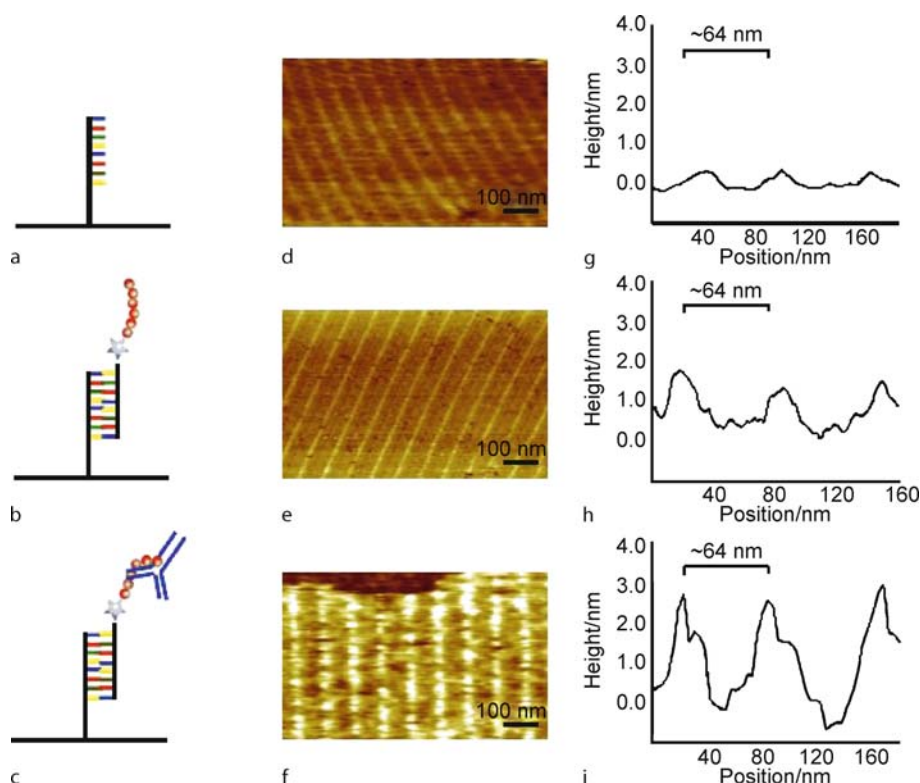
One type of guest molecule attachment has been shown: Biotin-streptavidin interaction is useful to create patterned protein arrays. The design of more diverse surfaces containing different molecular components, each at well-defined and addressable locations remains a challenging problem. As a possible solution, B. Williams et al. have developed a general method to produce high-density peptide arrays that rely on the addressable information encoded in the nucleic acid portion of a DNA-tagged peptide [22]. The idea is to template a capture stand that is protruding out from the DNA self-assembled lattice. Then



DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 5

Self-assembly of protein arrays templated by 4×4 DNA nanogrids. **a** Schematic drawing of the DNA nanogrids assembly of streptavidin. (Left) The DNA nanogrids, a biotin group labeled as a red letter B, are incorporated into one of the loops at the center of each tile. (Right) Binding of streptavidin (represented by a blue tetramer) to biotin will lead to protein nanoarrays on DNA lattices. **b** AFM image of the self-assembled protein arrays

DNA-tagged peptide will bind to the pre-designed capture strand through DNA hybridization. The peptide portion will serve as the probe for other protein such as antibody recognition. The DNA template was assembled from ABCD four tile DX motifs. D tile was modified to contain a DNA capturing tag for positioning DNA-peptide fusion at a specific addressable location on the DNA array. The DNA-peptide fusion was constructed by standard peptide coupling chemistry to covalently link the myc-epitope peptide to the 5'-amino-modified DNA strand. The DNA portion of the DNA-tagged peptide is complementary in the sequence to the capture probe on the D tile. Then the anti-myc antibody can be assembled to form 2D structure. The AFM images in Fig. 6 have clearly shown



DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 6

AFM imaging of the peptide nanoarray. a–c Schematic illustration showing the DNA capture probe on the DNA surface, annealed to the myc-peptide fusion, and immunocaptured by the anti-myc antibody, respectively. d–f AFM images were collected for the array before hybridization of the myc-peptide fusion, after hybridization of the myc peptide, and following incubation with the anti-myc antibody, respectively. g–i Corresponding height profiles of the arrays

the binding of DNA-peptide tag and antibody molecules to the DNA template.

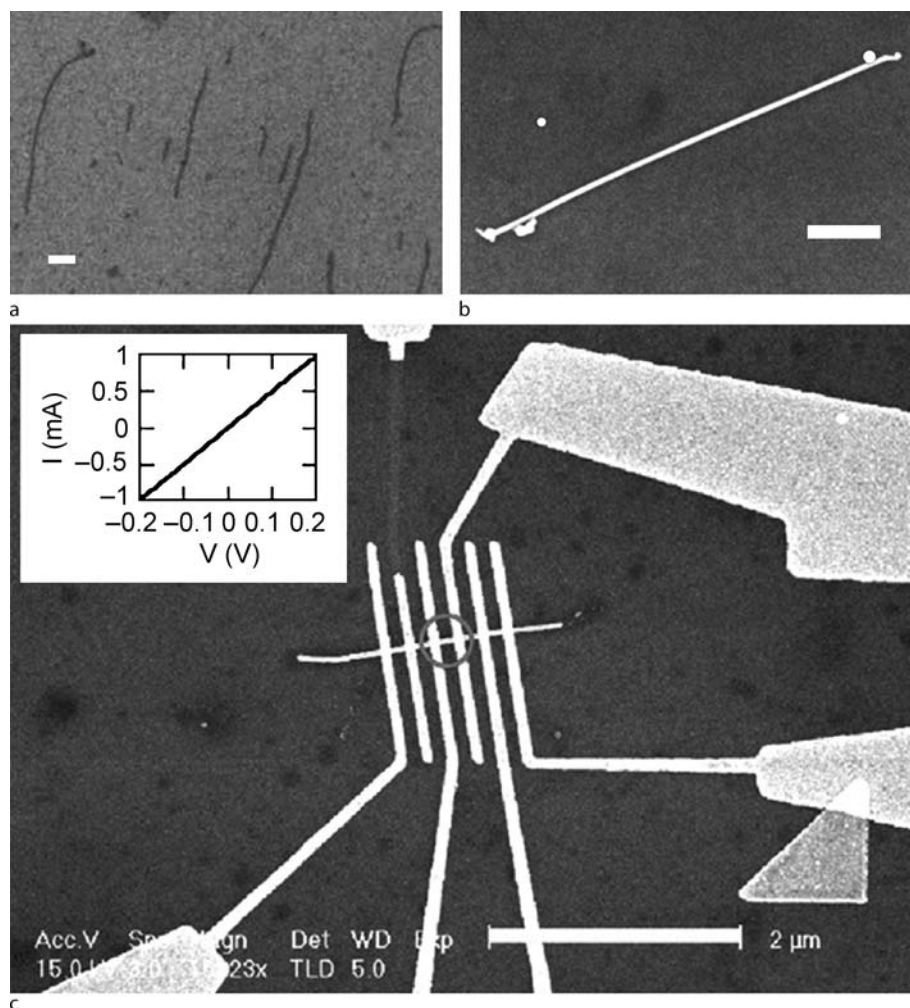
DNA-Templated Highly Conductive Nanowires

One potential application of DNA self-assembly is the use of DNA lattices to template assembly of nanoelectronic components, such as metallic nanoparticles [23,24,25,26] and even metallic nanowires. Several groups have demonstrated the successful organization of metallic nanoparticles through DNA template. DNA templated highly conductive nanowired was illustrated by H. Yan et al. [16]. The self-assembled 4×4 nanoribbons serve as the template, then was metalized with silver. The resulting nanowires have been characterized by AFM and SEM (Fig. 7). The metalized nanoribbons have heights of around 35 nm, widths of around 43 nm, and lengths of up to $5 \mu\text{m}$. Metal leads were patterned to the wire by electro beam lithography (5 nm Cr followed by 28 nm Au). An SEM image of an actual device is shown in Fig. 7c. A two-terminal current-voltage (I – V) measurement was

conducted on this device at room temperature. The I – V curve (Fig. 7c, inset) is linear, demonstrating Ohmic behavior in the range of -0.2 to 0.2 V, and the resistance of this sample is 200Ω as measured between the two central contacts. This number corresponds to a bulk resistivity of $2.4 \times 10^{-6} \Omega\text{m}$ for the silver nanowire. This nanowire is easily reproducible and has markedly higher conductivity than previously reported nanowires [27].

Future Directions

This article has reviewed the current development of structural DNA nanotechnology and DNA templated self-assembly. What is the future direction of this research field? The achievement of several key near-term goals will move DNA nanotechnology to a system with more practical capabilities. First among these goals is the extension of array-making capabilities from 2D to higher order 3D. Also, new DNA templates—to organize guest components in a higher controlled fashion—should be developed for the application of nanoelectronics and nanophotonics. Other



DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 7

Metalization and conductivity measurements of metalized 4 × 4 ribbons. **a** SEM images of nonmetalized 4 × 4 DNA nanoribbons (scale bar: 500 nm). **b** SEM image of silver-seeded silver nanoribbon (scale bar: 500 nm). The change in the signal contrast between **a** and **b** is apparent. **c** SEM image of the actual device (scale bar: 2 μm). (Inset) Current-voltage curve of the silver-seeded silver 4 × 4 nanoribbon

goals include further development of binding and release of biological molecules, molecular detection, diagnosis and drug release. It is also a challenge to combine DNA bottom-up assemble with top-down lithography to generate more complex and practical nano-device systems. There are undoubtedly many other directions, and DNA will continue playing its important role in complex nanoscale systems.

Bibliography

1. Seeman NC (2003) DNA in a material world. *Nature* 421:427–431
2. Seeman NC, Lukeman PS (2005) Nucleic acid nanostructures: Bottom-up control of geometry on the nanoscale. *Rep Prog Phys* 68:237–270
3. Reif JH (2002) Introduction to self-assembling DNA nanostructures for computation and nanofabrication In: Wang JTL, Wu CH, Wang PP (eds) *Computational biology and genome informatics*. World Scientific, River Edge
4. Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystal. *Nature* 394:539–544
5. LaBean T, Yan H, Kopatsch J, Liu F, Winfree E, Reif JH, Seeman NC (2000) The construction, analysis, ligation and self-assembly of DNA triple crossover complexes. *J Am Chem Soc* 122:1848–1860
6. Mao C, Sun W, Seeman NC (1999) Two-dimensional DNA hol-

liday junction arrays visualized by atomic force microscopy. *J Am Chem Soc* 121:5437–5443

7. Mao C, Sun W, Shen Z, Seeman NC (1999) A DNA nanomechanical device based on the B-Z transition. *Nature* 397:144–146
8. Yurke B, Turberfield AJ, Mills AP Jr, Simmel FC, Newmann JL (2000) A DNA-fuelled molecular machine made of DNA. *Nature* 406:605–608
9. Yan H, Zhang X, Shen Z, Seeman NC (2002) A robust DNA mechanical device controlled by hybridization topology. *Nature* 415:62–65
10. Li JJ, Tan W (2002) A single DNA molecule nanomotor. *Nano Lett* 2(4):315–318
11. Sherman WB, Seeman NC (2004) A precisely controlled DNA biped walking device. *Nano Lett* 4:1203–1207
12. Ding B, Seeman NC (2007) Operation of a DNA robot arm inserted into a 2D DNA crystalline substrate. *Science* 314:1583–1585
13. Adleman LM (1994) Molecular computation of solution to combinatorial problems. *Science* 266:1021–1024
14. Liu Q, Wang L, Frutos AG, Condon AE, Corn RM, Smith LM (2000) DNA computing on surfaces. *Nature* 403:175–179
15. Mao C, Labean TH, Reif JH, Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple crossover molecules. *Nature* 407:493–496
16. Yan H, Park SH, Finkelstein G, Reif JH, LaBean TH (2003) DNA-templated self-assembly of protein arrays and highly conductive nanowires. *Science* 301:1882–1884
17. Ding B, Sha R, Seeman NC (2004) Pseudo-hexagonal 2D DNA crystals from double crossover cohesion. *J Am Chem Soc* 126:10230–10231
18. Rothmund PWK (2006) Folding DNA to create nanoscale shapes and patterns. *Nature* 440:297–302
19. Douglas SM, Chou JJ, Shih WM (2007) DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. *Proc Nat Acad Sci USA* 104:6644–6648
20. Li H, Park SH, Reif J, LaBean TH, Yan H (2004) DNA templated self-assembly of protein and nanoparticle linear arrays. *J Am Chem Soc* 126:418–419
21. Park SH, Yin P, Liu Y, Reif JH, LaBean TH, Yan H (2005) Programmable DNA self-assemblies for nanoscale organization of ligands and proteins. *Nano Lett* 5(4):729–733
22. William BAR, Lund K, Liu Y, Yan H, Chaput JC (2007) Self-assembled peptide nanoarrays: An approach to studying protein-protein interactions. *Angew Chem Int Ed* 46:3051–3054
23. Mirkin CA, Letsinger RL, Mucic RC, Storhoff JJ (1996) A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* 382:607–609
24. Alivisatos AP, Johnsson KP, Peng X, Wilson TE, Loweth TCJ, Bruchez MP, Schultz PG (1996) Organization of 'nanocrystal molecules' using DNA. *Nature* 382:609–611
25. Pinto YY, Le JD, Seeman NC, Musier-Forsyth K, Taton TA, Kiehl RA (2005) Sequence-encoded self-assembly of multiple-nanocomponent arrays by 2D DNA scaffolding. *Nano Lett* 5(12):2399–2402
26. Zheng J, Constantinou PE, Micheel C, Alivisatos AP, Kiehl RA, Seeman NC (2006) Two-dimensional nanoparticle arrays show the organizational power of robust DNA motifs. *Nano Lett* 6(7):1502–1504
27. Braun E, Eichen Y, Sivan U, Ben-Yoseph G (1998) DNA-templated assembly and electrode attachment of a conducting silver wire. *Nature* 391:775–778

Drug Design with Artificial Intelligence Methods

OVIDIU IVANCIUC

Department of Biochemistry and Molecular Biology,
University of Texas Medical Branch, Galveston, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Genetic Algorithms](#)

[Ant Colony Optimization](#)

[Particle Swarm Optimization](#)

[Artificial Immune Systems](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Ant colony optimization Ant colony optimization (ACO) is an agent-based algorithm procedure inspired by the function of ant colonies and the ants search for the optimum path to food sources. The virtual agents are called artificial ants or ants, and the optimization problem is represented as a trial and error search for the optimum path on a weighted graph. The pheromone that is deposited by ants on the trail is represented as weights for graph components (vertices or edges). Each ant generates a solution by moving on the graph and by selecting the next step based on the pheromone level. The pheromone level is updated after each cycle (when all ants found a solution) by adding a pheromone quantity proportional to the quality of the solutions to which it belongs.

Antigen An antigen is a molecule (chemical compound, protein or polysaccharide) that induces an immune response. Each pathogen contains specific antigens that are recognized by the immune system. The antigen region that is recognized by the immune system is called an epitope.

Antibody An antibody (or immunoglobulin) is a protein used by the immune system to identify bacteria, viruses and other pathogens or foreign molecules. The antibody region that binds antigens is extremely variable, thus allowing the immune system to recognize a large diversity of pathogens. The ability to recognize antigens is improved through successive cycles of antigen presentation, antibody cloning, and hypermutation of the variable region of the antibody.

Artificial immune systems Artificial immune systems (AIS) represent a class of optimization algorithms inspired by the components and mechanisms of the biological immune system. AIS simulate the learning and memory capabilities of the immune system to develop computational algorithms for pattern recognition, function optimization, classification, process control, and intrusion detection.

Genetic algorithms Genetic algorithms (GA) solve high-dimensional problems through a Darwinian evolution of a population of individuals, in which each individual (chromosome) represents a possible solution. Depending on the type of the optimization problem, chromosomes may represent the solution in a binary, continuous, or hybrid encoding. Each chromosome has a fitness value that measures the quality of the solution. A population of parents evolves to a generation of children by crossover and mutation.

Particle swarm optimization Swarm intelligence (SI) represents a group of distributed intelligence algorithms that solve optimization problems by applying processes inspired by swarming, herding, and flocking of various species. Particle swarm optimization (PSO) simulates the swarming behaviors observed in swarms of bees, flocks of birds, or schools of fish. PSO considers a swarm of particles that start from a random position and have a random velocity. At each step a particle moves to a new position that is determined by its own experience (the best past position) and by the memory of the best particle in the swarm. PSO may be applied to both binary and continuous optimization problems, and its main strength is a fast convergence.

Quantitative structure-activity relationships

Quantitative structure-activity relationships (QSAR) represent regression models that define quantitative correlations between the chemical structure of molecules and their physical properties (boiling point, melting point, aqueous solubility), chemical properties and reactivities (chromatographic retention, reaction rate), or biological activities (cell growth inhibition, enzyme inhibition, lethal dose). The fundamental hypotheses of QSAR are that similar chemicals have similar properties, and that small structural changes result in small changes in property values. The general form of a QSAR equation is $P(i) = f(\mathbf{SD}_i)$, where $P(i)$ is a physical, chemical, or biological property of compound i , \mathbf{SD}_i is a vector of structural descriptors of i , and f is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. A QSAR model for a property P is based on a dataset of chemical compounds

with known values for the property P , and a matrix of structural descriptors computed for all chemicals. The learning (training) of the QSAR model is the process of determining the optimum parameters of the regression function f . After the training phase, a QSAR model may be used to predict the property P for novel compounds that are not present in the learning set of molecules.

Structural descriptor A structural descriptor (SD) is a numerical value computed from the chemical structure of a molecule, which is invariant to the numbering of the atoms in the molecule. Structural descriptors may be classified as constitutional (counts of molecular fragments, such as rings, functional groups, or atom pairs), topological indices (computed from the molecular graph), geometrical (volume, surface, charged-surface), quantum (atomic charges, energies of molecular orbitals), and molecular field (such as those used in CoMFA, CoMSIA, or CoRSA).

Structure-activity relationships Structure-activity relationships (SAR) represent classification models that can discriminate between sets of chemicals that belong to different classes of biological activities, usually active/inactive towards a certain biological receptor. The general form of a SAR equation is $C(i) = f(\mathbf{SD}_i)$, where $C(i)$ is the activity class of compound i (active/inactive, inhibitor/non-inhibitor, ligand/non-ligand), \mathbf{SD}_i is a vector of structural descriptors of i , and f is a classification function such as k -nearest neighbors, linear discriminant analysis, random trees, random forests, Bayesian networks, artificial neural networks, or support vector machines.

Definition of the Subject

Drug design and development represents a complex and expensive process that is based on the creative application of scientific results from various disciplines, including genomics, chemistry, biology, computational chemistry, pharmacology, toxicology, and clinical studies. The average cost of bringing a new drug to market is currently around US\$800 million, with a large part of the cost coming from chemical compounds that fail in different stages of development. Computational simulation of biochemical processes may guide the drug discovery process through reliable *in silico* models of biochemical properties (aqueous solubility, octanol-water partition, intestinal absorption, blood-brain barrier transport, excretion), prediction of enzyme-ligand interactions, simulations of cells, tissues and organisms. In this chapter we review the most important applications of artificial intelligence

in structure-activity relationships (SAR) and quantitative structure-activity relationships (QSAR). These techniques are used in different stages of drug design, including large scale screening of chemical libraries, optimization of protein-ligand interactions, modeling the drug transport through membranes, prediction of drug metabolism, mutagenicity, and carcinogenicity. The common goal of artificial intelligence applications in computer-assisted drug design is to identify the best candidates in each step, which may eventually lead to reduced costs for the development of new drugs.

Introduction

Biology is a rich source of inspiration for developing algorithms that solve complex problems by emulating mechanisms and functions of biological systems. Well-known examples of biologically inspired algorithms are artificial neural networks, genetic algorithms, ant colony optimization, DNA computing, particle swarm optimization, and artificial immune systems.

Evolutionary algorithms represent a family of stochastic methods that solve optimization problems by evolving solutions based on Darwinian evolution and concepts of DNA genetics (for details on GA and evolutionary algorithms, see ► [Genetic and Evolutionary Algorithms and Programming: General Introduction and Application to Game Playing](#)). The main algorithms from this class are genetic algorithms (GA), genetic programming (GP), and evolutionary programming (EP). The major principles of genetic algorithms were developed by Holland [1], and then further developed by Goldberg [2]. Many applications of chemoinformatics and computational chemistry have a large search space that must be explored to locate the solution. Usually, the brute-force grid search approach cannot be applied but for small systems, and various stochastic methods were developed to find near-optimal solutions. Several examples of high-dimensional problems are the prediction of the biopolymer structure from sequence (peptides, proteins, DNA, RNA), protein-protein docking, protein-ligand docking, conformational search, geometry optimization, design of chemical libraries, and design of chemical compounds with special physico-chemical and biological properties. For other GA applications in chemistry and biology see the reviews by Jones [3], Terfloth [4], and von Homeyer [5]. The most important GA applications in drug development are reviewed in Sect. “[Genetic Algorithms](#)”.

Dorigo and co-workers developed the ant colony optimization (ACO) algorithm to mimic the foraging behavior of some ant species [6,7,8,9,10]. The main feature modeled

in ACO is the ability of an ant population to find the shortest path to a food source using as guide the pheromone trace that is deposited on the path explored by each individual ant. The pheromone accumulates on paths explored more frequently by ants, which indicates that those paths are shorter routes to the food source. ACO has numerous applications, mainly in combinatorial optimization, when their ability to explore large solution spaces is a clear advantage. For theoretical details and applications of agent based simulation, see ► [Agent Based Modeling and Simulation](#), ► [Agent Based Modeling, Mathematical Formalism for](#), and ► [Agent Based Modeling and Artificial Life](#). In Sect. “[Ant Colony Optimization](#)” we present an overview of ACO applications in drug design.

The particle swarm optimization (PSO) algorithm proposed by Kennedy and Eberhart is inspired by the social behavior of large groups of individuals, such as bird flocking, fish schooling, and animal herding [11]. Each individual of the group, represented as a particle that moves with a particular velocity through the search space, is a solution for the optimization problem. The movement of each particle is determined by the best position visited by the particle, and by the best position found by the group. The balance between a local and a global search is introduced by weighting the attraction of the best solution of the particle and the best solution of the swarm (for more details on the PSO algorithms, see ► [Swarm Intelligence](#)). PSO converges fast and may be used with success to explore high dimensional spaces. The algorithm is simple, with a small number of parameters, and the large number of variants proposed in the literature is a sign of the great interest and vigorous research in this field [12,13]. Swarm intelligence algorithms are used in drug design for diverse applications, including gene expression [14], enzyme-inhibitor docking [15], selection of structural descriptors for QSAR models [16], QSAR with support vector machines optimized with PSO [17], and modeling enzyme inhibitors with artificial neural networks trained with PSO [18]. The most important PSO applications in drug discovery are presented in Sect. “[Particle Swarm Optimization](#)”.

The immune system protects an organism against infection by identifying and killing pathogens. Recognition cells known as B-cells and T-cells identify the pathogens that enter into the human body. Receptors situated on the surface of the B-cells and T-cells recognize and bind proteins and protein fragments from pathogens, thus forming high affinity antigen-antibody complexes. The learning and memory capabilities of the biological immune system are used in a novel class of machine learning algorithms, the artificial immune systems (AIS) [19,20,21,22,23,24,25,26,27] (for further de-

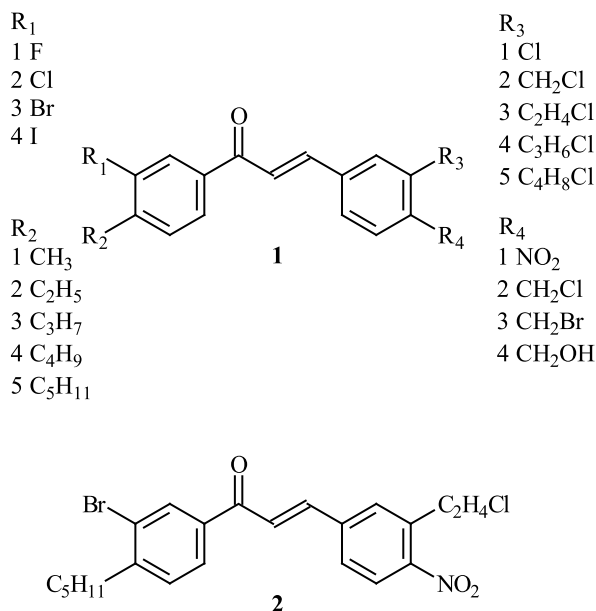
tails on AIS see ► [Immunecomputing](#). The major AIS algorithms and the most important applications are presented in several books and conference proceedings: *Artificial Immune Systems and Their Applications* edited by Dasgupta [28]; *Artificial Immune Systems: A New Computational Intelligence Approach* by de Castro and Timmis [29]; *Immunocomputing: Principles and Applications*, by Tarakanov, Skormin, and Sokolova [30]; *Immunity-Based Systems* by Ishida [31]; *Artificial Immune Systems: ICARIS 2003* edited by Timmis, Bentley and Hart [32]; *Artificial Immune Systems: ICARIS 2004* edited by Nicosia, Cutello, Bentley, and Timmis [33]; *Artificial Immune Systems: ICARIS 2005* edited by Jacob, Pilat, Bentley, and Timmis [34]; *Artificial Immune Systems: ICARIS 2006* edited by Bersini and Carneiro [35].

AIS models were successfully applied to biological and medical problems, such as classification of gene expression data [36,37,38], identification of breast cancer [39], diagnosis of lung cancer [40], recognition of ECG arrhythmia [41], and interpretation of carotid artery Doppler signals [42]. Protein structure prediction starting from the amino acids sequence is a difficult and computationally intensive task, which was investigated with AIS for models based on Dill's hydrophobic-hydrophilic lattice approach [43] and with three-dimensional models [44]. In Sect. "Artificial Immune Systems" we present a review of the AIS applications in drug design and toxicology.

Genetic Algorithms

Compared with other families of artificial intelligence algorithms, evolutionary algorithms are by far the most popular, with the largest number of publications and with the most diverse applications. GA methods are applied with success to solve diverse drug design problems, such as protein-ligand docking [45], structure-based drug design [46], global optimization of QSAR models based on artificial neural networks [47], computer-aided molecular design [48,49], design of combinatorial libraries [50], and feature selection in QSAR models [51,52]. All these problems are difficult to solve thorough a brute force approach due to the huge search space, but GA are very efficient in finding their global optimum with modest computational resources.

Evolutionary algorithms are used with remarkable results in computer-aided molecular design [48,49] to generate novel molecules with prescribed physical, chemical, or biological properties. In pharmaceutical applications, molecular design is focused on discovering chemical structures that can satisfy all requirements of a successful drug, such as affinity and selectivity for the bio-



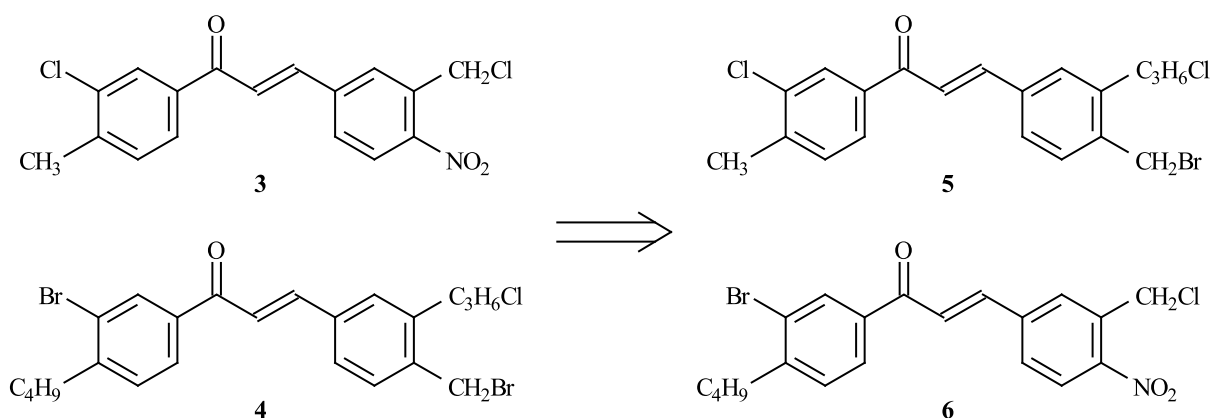
Drug Design with Artificial Intelligence Methods, Figure 1

General formula for a family of chemical compounds (1) that may be encoded with chromosomes having four elements, and an example of molecule (2) from this family

logical target, and good ADME-Tox (absorption, distribution, metabolism, excretion, and toxicity) properties. The most important part of any molecular design application is a proper encoding of the molecular structure into a chromosome. A straightforward translation of chemicals may be achieved if the molecule can be partitioned into a constant skeleton and a series of substituents, such as the family of chemical compounds 1 (Fig. 1) that has four substituents R₁, R₂, R₃, and R₄. Each molecule from this family may be encoded by a chromosome with four elements (/R₁/R₂/R₃/R₄/), with each element recording the index of the respective substituent. Each substitution position has a set of allowed substituents encoded with numbers. For example, compound 2 is represented by the chromosome /3/5/3/1/.

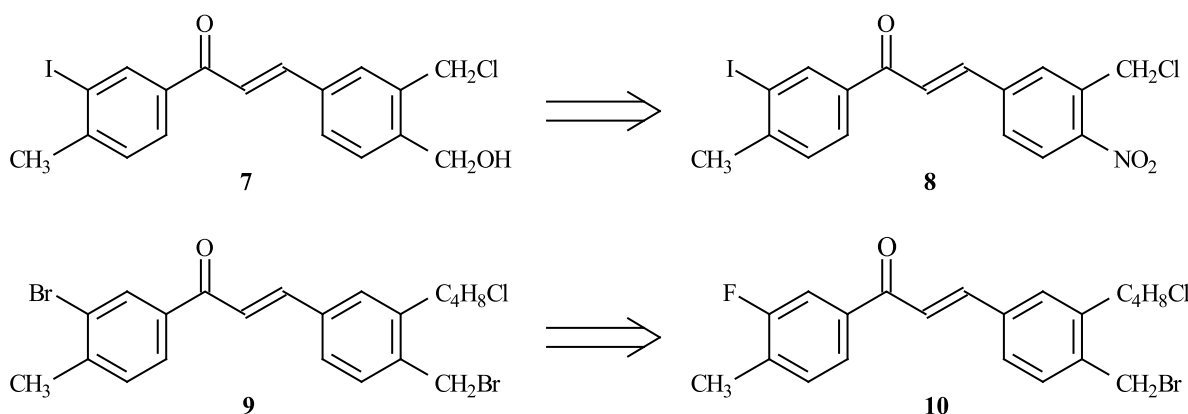
Using this molecular encoding, one can easily define the GA operations of crossover and mutation. The crossover operation involves the exchange of substituents between two parent molecules. For example, parent molecules 3 and 4 generate child molecules 5 and 6 by exchanging substituents R₃ and R₄ (Fig. 2).

The chemical space is also explored with the substituent mutation, as shown in Fig. 3: parent molecule 7 generates child molecule 8 by mutating R₄, and parent molecule 9 generates child molecule 10 by mutating R₁. These examples demonstrate the encoding and evolution of chemical structures in combinatorial libraries of chem-



Drug Design with Artificial Intelligence Methods, Figure 2

Example of molecule crossover: parent molecules 3 and 4 generate child molecules 5 and 6 by exchanging substituents R_3 and R_4 (see molecule 1)



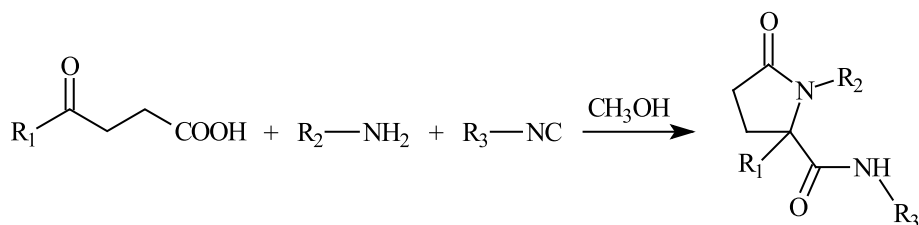
Drug Design with Artificial Intelligence Methods, Figure 3

Examples of molecule mutation: parent molecule 7 generates child molecule 8 by mutating R_4 , and parent molecule 9 generates child molecule 10 by mutating R_1

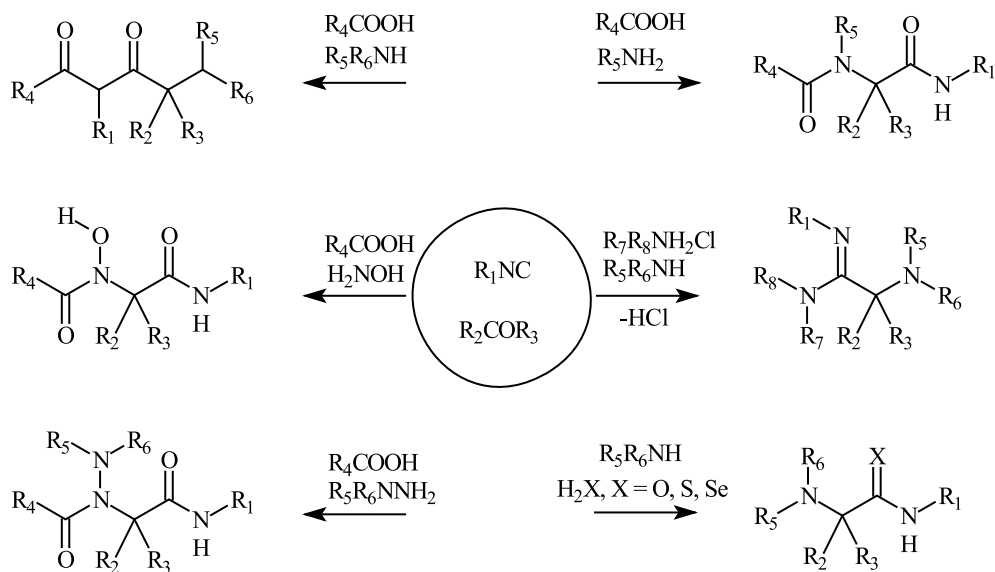
ical compounds [53,54]. The progress in combinatorial chemistry [55,56], virtual screening of chemical libraries, and high throughput techniques dramatically increased the chemical space that can be explored in the quest for molecules with special properties (peptides, nucleic acids, catalysts, pesticides, drugs). Due to the huge chemical space that can be generated through combinatorial chemistry, it is rarely possible to perform an exhaustive synthesis of all possible chemical species. Instead, GA implementations are used to guide the chemical synthesis towards regions containing molecules with target properties [57,58].

An efficient class of reactions that may generate large combinatorial libraries is the Ugi multicomponent reaction (MCR) [59]. Ugi MCRs are one-pot reactions in which three reactants (Fig. 4; U-3CR), four reactants (Fig. 5; U-4CR), or more reactants are converted into the

corresponding product without separation and purification of the intermediates. The diversity of chemical structures generated through MCR reactions comes from the diversity of the groups R from reactants. Using available chemicals, one can design chemical libraries that are too large to synthesize. Instead, a sample of the combinatorial library is synthesized and evaluated in biological assays, followed by an *in silico* exploration based on GA models [60]. If each reactant type in an U-3CR is a set of 1000 different chemical compounds, then the complete library has 10^9 distinct molecular structures. Similarly, an U-4CR library generated from four sets of 1000 chemicals each consists of 10^{12} distinct compounds. It becomes apparent that the vast chemical space available through combinatorial synthesis is too large even for the *in silico* exploration, which explains why evolutionary algorithms are used to guide the chemical synthesis.



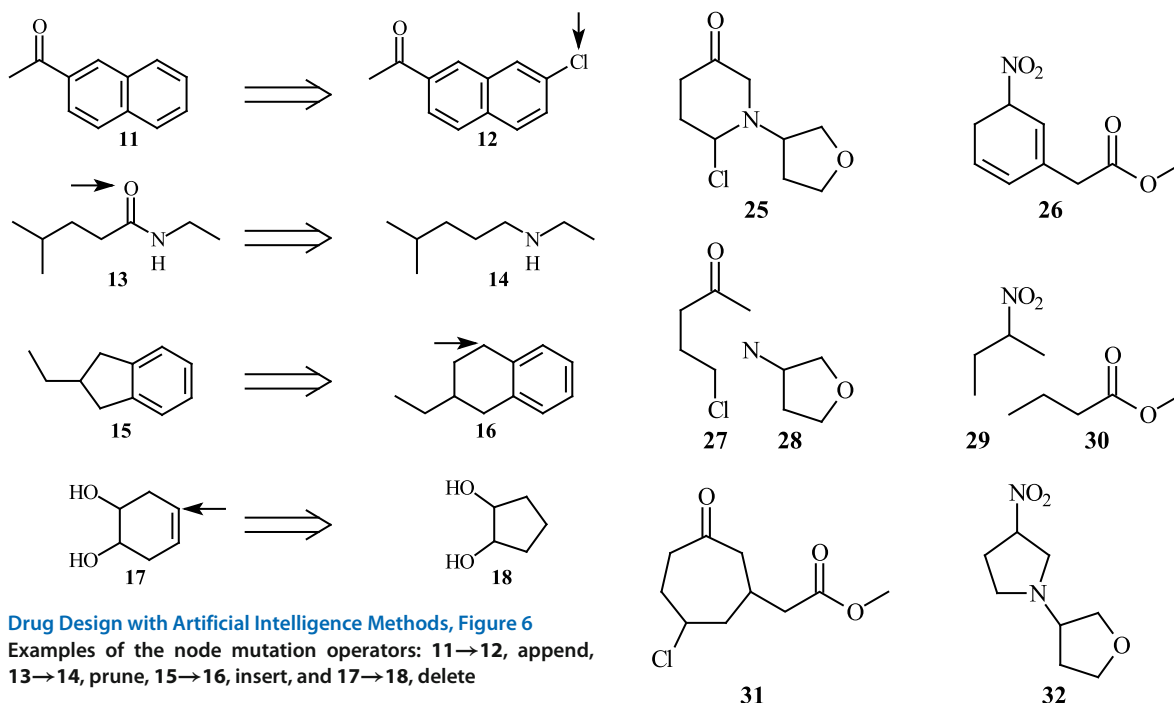
Drug Design with Artificial Intelligence Methods, Figure 4
Example of Ugi 3-component reactions (U-3CR)



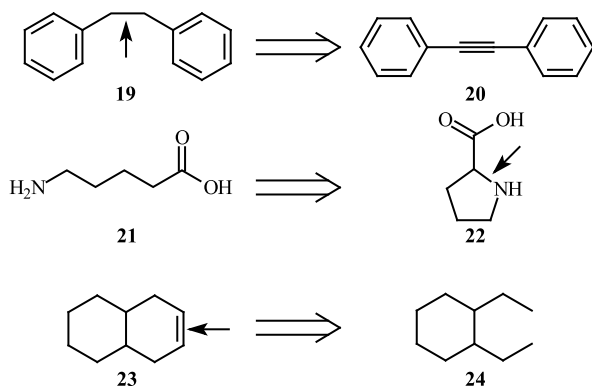
Drug Design with Artificial Intelligence Methods, Figure 5
Examples of Ugi 4-component reactions (U-4CR)

The GA translation of MCR reactions and other combinatorial libraries is straightforward, and in many cases the *in silico* exploration of the chemical space may be performed with standard GA software. This approach has limitations, because the size of the chemical space is fixed by the initial sets of reactants, and the common skeleton remains constant during the simulation. Graph-based GA models solve these limitations by generating molecular structures that are not programmed in the starting building blocks. Such GA systems have crossover and mutation procedures that operate directly on the molecular graph [61], and define a chromosome structure capable to encode a molecular graph [62]. The graph-based GA system proposed by Brown et al. introduces novel crossover and mutation operations for molecular graphs [62] in order to solve the inverse QSAR problem, i.e., to design new chemicals starting from structure-activity models [63]. Four mutations operate on atoms (graph nodes), namely append, prune, insert, and delete (Fig. 6; the site of

the transformation is indicated with an arrow). The append mutation adds an atom and its chemical bond to the molecular graph (11→12). The connecting atom is selected at random from the set of atoms in the molecule that have available valences. The type of the connecting bond is randomly selected from the possible types for the two atoms. The prune mutation removes a terminal atom from the molecular graph (13→14). The insert mutation selects a bond in the molecular graph, cuts it and inserts a molecular fragment between the two disconnected atoms (15→16). The molecular fragment is selected from a library, and may consist of a single atom or a more complex subgraph. Additional tests are performed to ensure that the final chromosome (molecular graph) is a valid chemical structure. The delete mutation selects an atom at random, removes it and then reconnects the molecular graph (17→18). The edge mutations operate on the set of edges in a chromosome (Fig. 7; the site of the transformation is indicated with an arrow). The substi-



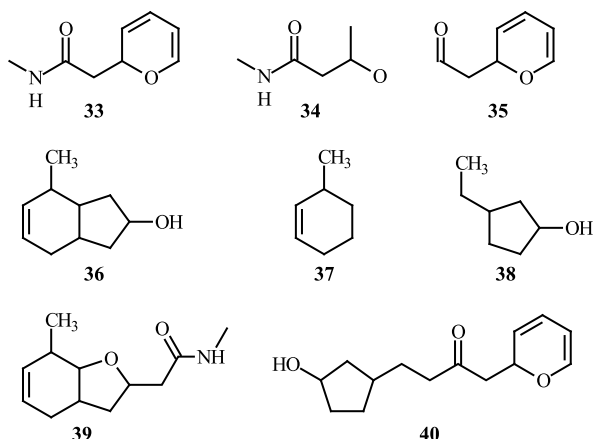
tute mutation selects randomly an edge and then changes its type to another bond type (19→20). The mutation result must correspond to a correct chemical structure. The add mutation adds a new bond between two atoms (21→22), thus making possible the generation of cyclic structures. Finally, the delete mutation deletes a bond that was randomly selected (23→24). The resulting chromosome must represent a connected molecular graph. Two crossover operations are defined for molecular graphs, i. e., multiple crossover and subgraph crossover. The multiple crossover starts from two parent molecules, then each parent molecule is disconnected into two subgraphs, and fi-



Drug Design with Artificial Intelligence Methods, Figure 8

Example of the multiple crossover: 25 and 26, parent molecules; 27 and 28, disconnected subgraphs of 25; 29 and 30, disconnected subgraphs of 26; 31, child molecule generated from subgraphs 27 and 30; 32, child molecule generated from subgraphs 29 and 28

nally, two child molecules are generated by swapping subgraphs from the parent molecules (Fig. 8). Parent molecule 25 generates subgraphs 27 and 28, and parent molecule 26 generates subgraphs 29 and 30. The crossover operation generates child molecule 31 from subgraphs 27 and 30, and then assembles child molecule 32 from subgraphs 29 and 28. In the subgraph crossover, a connected subgraph is selected in each parent molecule, and then the subgraphs are combined to obtain the first child molecule (Fig. 9). The combination of the two fragments tries to retain the topology of the two subgraphs. In the second step, a different subgraph is induced in each parent molecule, and the two subgraphs form the second child molecule. The parent molecule 33 generates the induced connected subgraphs 34 and 35, and the second parent molecule 36 generates the induced connected subgraphs 37 and 38. The first child molecule 39 is obtained by combining subgraphs 34 and 37, and the second child molecule 40 is obtained from subgraphs 35 and 38. The main advantage of the graph-based GA system is its ability to explore chemical structures that are not related to the starting molecules, and to discover novel chemical topologies.



Drug Design with Artificial Intelligence Methods, Figure 9

Example of the subgraph crossover: the first parent molecule **33** and its two induced connected subgraphs **34** and **35**; the second parent molecule **36** and its two induced connected subgraphs **37** and **38**; the first child molecule **39** generated from subgraphs **34** and **37**; the second child molecule **40** generated from subgraphs **35** and **38**

Virtual Screening of Chemical Libraries

QSAR models are very useful tools for the identification of structural features that determine various molecular properties, and may even suggest the mechanism of action for biochemical processes. Thus, QSAR models start from the chemical structure and correlate structural descriptors with molecular properties. Once a QSAR model is established, an inverse process becomes possible, namely setting a target value for a molecular property and then finding all possible chemical structures that might exhibit that property value, within a certain range of variation. This process is called inverse QSAR, and it represents an important step in optimizing the drug-like properties of chemical compounds. Lewis proposed an inverse QSAR strategy that may assist medicinal chemists in deciding how to optimize a library of chemical compounds [64]. The starting point is a dataset of chemical compounds with a molecular property, and a corresponding QSAR model. The inverse QSAR strategy involves an iterative application of several steps, namely generation of new structures, structure filtering based on synthetic feasibility or undesired properties, and QSAR filtering. The first step generates a new chemical library by applying simple chemical transformations to the molecules from the initial dataset. Examples of such transformations are modification of the bond order, adding or removing an atom, adding or removing a fragment, or changing C to N or O. The second step filters molecules that have nonspecific reactivity, such as electrophiles, nucleophiles, acylating agents, or re-

dox systems. Synthetic feasibility rules are used to eliminate compounds that are difficult to synthesize or those that are expensive. Finally, QSAR models are used to select candidates for chemical synthesis. The inverse QSAR strategy developed by Lewis was tested for a combinatorial library of 150 inhibitors of human carbonic anhydrase II, that was used to develop a MLR genetic function approximation QSAR, as implemented in Cerius². The best QSAR model is based on five structural descriptors:

$$\begin{aligned} \text{pIC}_{50} = & 7.5 - 0.6\text{PHI} - 5.7\text{Jurs-RPCG} + 0.2\text{SdsN} \\ & + 1.7\text{NaaS} + 0.001\text{Vm} \\ n = 150 \quad r^2 = 0.81 \quad q_{LOO}^2 = 0.80 \quad F = 127 \end{aligned}$$

where PHI is the molecular flexibility index, Jurs-RPCG is the charge of the most positive atom divided by the total positive charge, SdsN is the E-state index for sp^2 N, NaaS is the electrotopological count for aromatic S, and Vm is the molecular volume inside the contact surface. This QSAR was used as the starting point for performing automated property optimization.

Ant Colony Optimization

The classical ACO algorithm was successfully modified and adapted in numerous variants to solve specific problems from chemistry and drug design. By far the most important ACO application is represented by the feature selection for QSAR models [65,66]. Several ACO implementations were tested in diverse QSAR models, including multi-linear regression, artificial neural networks, and regression trees. Clustering is routinely used to discover novelty in large chemical datasets, based on structural similarities measured by molecular descriptors. Since similar chemicals usually have similar properties, clustering may suggest groups of molecules that interact with the same biological target. Shelokar et al. proposed a clustering algorithm based on ACO assignment of objects in clusters [67]. Many biochemical problems require optimization of continuous variables, whereas the classical ACO implementation optimizes discrete variables. He et al. demonstrated an ACO extension to continuous variables that may be applied to identify optimum parameters for QSAR models [68]. Korb and co-workers introduced a new protein-ligand docking algorithm, PLANTS (Protein-Ligand ANT System), which uses ACO to find a minimum energy conformation for a protein-ligand complex [69]. Compared with docking algorithms based on GA, PLANTS is faster and finds a larger number of good solutions.

Izrailev and Agrafiotis used an ACO approach to identify the best regression tree models in QSAR [65]. Each

ant represents a regression tree, and the pheromone trail is obtained from a reference tree that represents the topological union of all ant trees simulated. The ACO selection of regression trees was evaluated for three QSAR datasets, namely the antifilarial activity of antimycin analogues, the binding affinities of ligands to benzodiazepine/GABA_A receptors, and the inhibition of dihydrofolate reductase by pyrimidines. Each simulation generated 2000 ant trees and then the tree with the best cross-validation predictions was selected as solution. For all three QSAR datasets the ant tree results were significantly better than those obtained with recursive partitioning and with random trees. Using the same three QSAR datasets, Izrailev and Agrafiotis proposed an ACO procedure (ANTSELECT) for feature selection in artificial neural networks QSAR [66]. A number of 100 independent ANTSELECT simulations were performed for each QSAR dataset, with each simulation containing a population of 2000 ants. Structural descriptors are represented as graph vertices, and an ant generates a path by visiting a number of vertices. All vertices on the path represent the selected structural descriptors that are subsequently used as input to an artificial neural network. Features that give good QSAR models receive a larger quantity of pheromones, thus having greater chances to be selected in subsequent iterations. The QSAR results indicate that the ANTSELECT algorithm provides good solutions if the simulations use a sufficient number of ants to evaluate all features in different combinations. A second requirement is to have a pheromone accumulation that distinguishes between good and bad features. Artificial neural networks are sensitive to the input features, and ANTSELECT provides sets of descriptors that result in models with good predictive power.

Nonsteroidal antiinflammatory drugs (NSAID) treat inflammation and pain by inhibiting both cyclooxygenase-1 and cyclooxygenase-2 (COX2). NSAID have serious side effects, such as gastrointestinal ulceration and bleeding, but the observation that acute and chronic inflammation correlates with higher levels of COX2 prompted several drug design studies to identify selective COX2 inhibitors. Shen et al. proposed a novel ACO procedure for feature selection in a QSAR study of 42 COX2 inhibitors [70]. Starting from 85 structural descriptors, the simulation used 100 ants and 200 iterations to select 3 descriptors for the optimum model. The ACO procedure selected a better set of descriptors, compared with a selection made with an evolutionary algorithm.

The drug binding to human serum albumin (HSA) determines its bioavailability, pharmacokinetics, and therapeutic effect. Many drugs are transported by HSA, but only the free drug has pharmacological effect. Gunturi et al.

modeled the HSA binding of 94 diverse drugs starting from a pool of 327 structural descriptors [71]. Since the number of descriptors is too large for a multi-linear regression QSAR, an ACO procedure was implemented to select those features that determine HSA binding. The ACO solutions were cross-validated, and the best QSAR equations with five and six descriptors were selected as final models. The importance of each descriptor was evaluated by the frequency of selection in QSAR models, and it was found that HSA binding depends on hydrophobic interactions, solubility, size, and shape.

Tyrosine kinases are enzymes that transfer a phosphate group from ATP to a tyrosine residue in a protein. These enzymes have important functions in diverse cellular processes, such as metabolism, differentiation, growth, and apoptosis. Shi et al. developed QSAR models for inhibitors of the epidermal growth factor receptor (EGFR), a cell-surface receptor from the tyrosine kinase family [72]. Mutations affecting EGFR expression or activity could result in cancer. The structure of the 61 EGFR inhibitors was encoded with 50 structural descriptors, and ACO was used to select relevant groups of descriptors. The ant population had 100 individuals trained for 200 iterations. The analysis of the descriptors selected with higher frequency by ants reveals the importance of electronic indices, and suggest that electron-donating groups increase the activity of these EGFR inhibitors.

The ability to distinguish between foreign and self proteins is one of the most important characteristics of the immune system. The major histocompatibility complex (MHC) molecules bind short peptides resulting from intracellular processing of foreign and self proteins. The MHC molecule loaded with the peptide migrates to the cell surface where it interacts with T-cell receptors. There are two classes of MHC molecules: (a) MHC class I, which binds peptides derived from endogenously expressed proteins and (b) MHC class II, which binds peptides derived mainly from exogenous or transmembrane proteins. Karpenko et al. devised a novel procedure to predict peptides that bind to MHC II, by using ACO to identify the optimum alignment of a set of variable length peptides [73]. The multiple alignment of all peptides is then utilized to compute a position specific scoring matrix. This matrix assigns different weights to each position and amino acid type, and provides a score for each peptide. Finally, the score is compared with a threshold to determine if the peptide binds or not to MHC II. The predictive power of the scoring matrix was demonstrated on several benchmark datasets, showing that the novel algorithm may be useful to design peptides that bind to MHC II and that may be used in vaccine development.

Major advances in proteomics are a result of significant technological advances in protein purification and mass spectrometry. Another critical component is the automated and reliable protein identification from mass spectrometric data. To improve the protein identification process, Hernandez et al. devised a heuristic algorithm that addresses the difficulties of the current methods, such as poor performance for large databases or for low quality data [74]. The new method based on ACO matches theoretical peptide sequences from a database with a structured representation of the source MS/MS spectrum. Tested with a set of 721 MS/MS spectra, the ACO-based procedure showed success rate of 88.9%, demonstrating that the artificial ants may perform an efficient exploration of the search space.

Particle Swarm Optimization

Particle swarm algorithms are used in diverse biochemistry and drug design applications, to solve problems that require binary or real value optimization. Among the advantages of using PSO in optimization one should count the simple algorithm that translates into small and effective software, fast convergence, small population, and low number of iterations. PSO is applied with success to difficult problems, such as feature selection for gene expression data [14,75], identification of the global minimum geometry of chemical compounds [76], enzyme-inhibitor docking [15], QSAR [16], and protein motif discovery [77].

PSO is an effective replacement of GA for the global optimization of protein-ligand geometry in docking studies. Several PSO modifications of the most popular docking program, AutoDock, were proposed in the literature. The Tribe-PSO algorithm was used in AutoDock to identify the best protein-ligand geometry [78]. In Tribe-PSO the population is divided into several subpopulations or tribes. Each tribe has the same structure and evolution mechanism as the basic PSO model. In the first phase, each tribe evolves independent of the other tribes and converges to an optimum solution. In the second phase, the tribes exchange information regarding the best solution from each tribe, and in the third phase all particles are united into a single population that evolves as a classical PSO model towards the final solution. In a comparative test involving 100 protein-ligand complexes from PDB, over 90% are docked better with Tribe-PSO than with AutoDock. Another PSO modification of AutoDock is SODOCK, which combines the basic PSO model with a local search for the best particle [79]. Compared with four docking methods (GOLD, DOCK, FlexX, and AutoDock) for a set of 37 PDB protein-ligand complexes, SODOCK obtained an aver-

age RMSD of 2.29 Å, whereas all other docking programs had an RMSD higher than 3 Å. In a related implementation, PSO@AUTODOCK, AutoDock is combined with a PSO variant that allows larger movements in the search space [15]. Significant improvement is obtained for 12 out of the 37 test complexes, compared with the SODOCK predictions.

Feature selection is an important step in QSAR and in virtual screening of chemical libraries, because almost all QSAR models are sensitive to the presence of irrelevant descriptors. Another benefit of feature selection is the identification of structural descriptors that may explain the mechanism of a particular structure-activity relationship. Agrafiotis and Cedeño used a binary PSO to select descriptors for a QSAR based on multilayer feed-forward artificial neural networks (MLF ANN) [16]. The real value PSO model may also be used for feature selection, as shown for QSAR models based on *k*-nearest neighbors kernel regression [80]. The target of the PSO model was to find the optimum weight (situated in the range [0, 1]) for each structural descriptor. The features with the largest weights were selected in the QSAR model.

In a comparative study for 42 cyclooxygenase inhibitors, Lü et al. found that binary PSO is superior to GA for feature selection in multi-linear regression (MLR) QSAR [81]. Shen et al. showed that the partial least-squares (PLS) QSAR model could be improved by using structural descriptors selected with a binary PSO [82]. Another approach to feature selection is the optimized blockwise variable combination (OBVC) method that performs a descriptor selection guided by PSO followed by PLS modeling of the data [83,84]. Instead of selecting each descriptor independent of the other descriptors, OBVC operates with groups of descriptors. The size and composition of each group of descriptors is optimized with PSO. OBVC was evaluated in QSAR models for the carcinogenic potency of aromatic amines [83] and for inhibitors of lung carcinoma cells [84]. OBVC was also tested for a QSAR dataset consisting of 37 ligands of the $\alpha 6$ benzodiazepine receptor, and more than 70 structural descriptors (topological, geometric, and quantum indices) [85]. Comparative tests show that OBVC exceeds the predictions obtained with MLR, PLS, and hierarchical PLS. OBVC may suggest several combinations of descriptors with comparable prediction statistics, and can assist the discovery of the most important structural descriptors.

PSO is used also to modify and improve QSAR models, such as the piecewise modeling by particle swarm algorithm (PMPPO) which is a QSAR based on piecewise linear models [86]. PMPPO may be useful for datasets with

high structural diversity, when a single linear model for all compounds might not be the best option. A minimum spanning tree model is used to cluster all compounds, and then PSO is applied to divide the tree into predictive piecewise linear models. PMPSO was applied with good results for angiotensin II antagonists. A variant of this QSAR model is the piecewise hypersphere modeling by particle swarm optimization (PHMPSO) which clusters similar compounds in subsets defined as hyperspheres [87]. The position and size of the hyperspheres are optimized with PSO, and then a QSAR model is fitted for the compounds in each hypersphere. PHMPSO was tested with good results for dihydrofolate reductase inhibitors, epidermal growth factor receptor inhibitors, and benzodiazepine receptor ligands [88]. Another PSO-modified algorithm is the optimized sample-weighted PLS (OSWPLS) which uses PSO to weight each object (chemical compound in QSAR) from the training dataset [89]. The weight determines the importance of each object, and the target of the PSO step is to minimize the error of the calibration model.

Training a neural network QSAR consists of (a) finding the best network topology (number of hidden neurons and distribution of the connections between neurons), and (b) optimization of the connection weights. PSO is very efficient in optimizing the ANN weights, as shown in a QSAR study of inhibitors of platelet-derived growth factor receptor phosphorylation [18]. The versatility of the swarm algorithm is practical in a global optimization of ANN QSAR, namely finding the best topology and set of weights. Shen et al. proposed a hybrid use of PSO in training a MLF ANN, namely a binary PSO to determine the optimum network topology and a continuous PSO to find the optimum connection weights [90]. Extensive tests showed that this combination converges quickly and may avoid the overfitting of the learning dataset of chemicals.

The training process of a radial basis function artificial neural network (RBF ANN) consists of selecting the network topology, finding the centers and widths of the RBF neurons, and computing the connection weights between the hidden and output layers. A hybrid particle swarm optimization (HPSO) was used by Zhou et al. to train an RBF network for drug design studies [91]. In the HPSO algorithm, a discrete PSO is used to optimize the network topology, whereas a continuous PSO is used to optimize the network parameters. The new QSAR approach was tested with a dataset of 40 inhibitors of murine P388 leukemia cells and over 70 Cerius² descriptors. The HPSO network has the highest predictions: PLS, $r = 0.664$; RBF with parameters optimized with PSO, $r = 0.838$; RBF with parameters optimized with K-means, $r = 0.852$; RBF optimized with HPSO, $r = 0.894$. A sim-

ilar trend was found for a second QSAR test, performed with 72 cyclooxygenase-2 inhibitors: RBF optimized with PSO, $r = 0.894$; RBF optimized with K-means, $r = 0.903$; RBF optimized with HPSO, $r = 0.921$. The experimental evidence suggests that the hybrid PSO optimization of RBF ANN has a fast convergence to predictive QSAR models.

Zhou et al. proposed a novel version of nonlinear partial least-square method that is based on structural descriptors transformed by an artificial neural network [92]. The structural descriptors represent the ANN input, whereas the output signals from the neurons in the hidden layer represent the nonlinear input for PLS. The ANN weights are trained with PSO. The novel nonlinear QSAR model was tested with good results for two datasets, namely 53 antitumor agents, and 52 benzodiazepine receptor ligands.

As shown in the QSAR studies reviewed here, PSO is an efficient method to optimize linear and nonlinear structure-activity models. A fast convergence to the global minimum depends on the parameters that control the population size, number of iterations, and weights to update the velocity of each particle. Choosing the best parameters that control a PSO model is a meta-optimization problem that was solved by Meissner et al. with the optimized particle swarm optimization (OPSO) model, in which the control parameters are optimized by a meta-swarm [93]. Although OPSO is more complex than a classical PSO because it contains swarms within a swarm, the system converges fast to good QSAR models. OPSO was tested for the prediction of the blood-brain barrier permeation coefficient with a MLF neural network.

Support vector machines (SVM) represent a class of versatile models that can produce nonlinear classification or regression QSAR equations [94]. PSO can be efficiently applied to select the best structural descriptors for SVM models, as demonstrated in a QSAR for P-glycoprotein substrates [17]. The mathematical formalism of SVM was adapted by Lin et al. for the training of MLF ANN [95]. The parameters of the hybrid method SVM-ANN were optimized with PSO, and the new QSAR model was compared with other two algorithms, namely back-propagation ANN (BP-ANN) and ANN optimized with PSO (PSO-ANN). These methods were compared for a dataset of 111 dihydrofolate reductase inhibitors and for another set of 85 cyclooxygenase-2 inhibitors. The results show that SVM-ANN models have better prediction statistics, and that the PSO procedure converges fast to optimum parameters. A similar QSAR model was developed based on radial basis function ANN [96], by defining a nonlinear SVM model (RBF-SVM) representing a kernel transform

based on RBF ANN optimized with PSO. QSAR models obtained for inhibitors of HIV-1 reverse transcriptase demonstrate that RBF-SVM provides better predictions compared to BP-ANN and SVM.

Artificial Immune Systems

The mechanisms and functions of the biological immune system were used as an inspiration for many AIS algorithms, such as the artificial immune network (aiNet) [97,98], the hierarchical artificial immune network (HaiNet) [37], the artificial immune recognition system (AIRS) [99,100,101,102], the clonal selection algorithm (CLONALG) [103,104], the clonal selection classification system (CSCA) [105], IMMUNOS-81 [106], and IMMUNOS-99 [107]. The pattern recognition capabilities of the artificial immune systems may be applied in modeling structure-activity relationships for drug design or for the computational screening of chemical libraries. In the following sections we review several SAR models obtained with AIRS, CLONALG, CSCA, and IMMUNOS. All AIS models were computed with Weka [108].

AIRS – Artificial Immune Recognition System

The AIRS machine learning algorithm developed by Watkins, Timmis, and Boggess is an efficient and popular pattern recognition adaptation of AIS [99,100,101,102]. Brownlee tested AIRS for a wide range of classification problems [109], confirming its utility as a supervised learning classifier. The main characteristics of AIRS are briefly reviewed below.

An antigen is represented as an n -dimensional vector $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, where each structural descriptor x_i is a real number ($x_i \in R$ for $i = 1, 2, \dots, n$), and an associated class $y = \{+1, -1\}$. An identical encoding is used for antibodies. An artificial recognition ball (ARB) represents a B-cell, and consists of an antibody, a number of resources, and a stimulation value. The similarity between an ARB and an antigen is measured by the stimulation value. The number of resource from an AIRS model is limited, and ARBs compete for their allocation. Resources are allocated to the most stimulated ARBs by removing them from the least stimulated ARBs, and ARBs without resources are eliminated from the cell population. The ARB population is trained during several cycles of competition for limited resources. In each cycle of ARB training, the best ARB classifiers generate mutated clones that enhance the antigen recognition process, whereas the ARBs with insufficient resources are removed from the population. After training, the top ARB classifiers are selected as memory

cells. Finally, the memory cells are used to classify novel antigens (patterns).

The drug design applications reviewed here were obtained with AIRS2, an improved version of AIRS [110]. The AIRS2 algorithm consists of the following steps [109]:

- (1) **Initialization.** In the first phase of the algorithm the system is prepared for the learning process. The training data are normalized between 0 and 1. The Euclidean distance is computed for all pairs of antigens, and then the affinity Af is determined as the ratio between the distance and the maximum distance. The affinity threshold AT is computed as the average affinity for all antigens in the training set. The memory cell pool is populated with randomly selected antigens. At the end of the AIRS algorithm, the memory cell pool represents the recognition ARBs used as classifiers.
- (2) **Train for all antigens.** The AIRS algorithm trains a classifier by passing only once over the entire population of training antigens.
 - (2.1) **Antigen presentation.** Each training antigen is presented to the memory cell pool, and each memory cell receives a stimulation value St , $St = 1 - Af$. The memory cells with the largest stimulation values are selected, and a number of mutated clones are created and added to the ARB pool. The number of clones NC generated is computed with the formula:

$$NC = St \times CR \times HR$$
 where the clonal rate CR and the hypermutation rate HR are user defined parameters.
 - (2.2) **Competition for limited resources.** During this iterative process the algorithm selects those ARBs that have the best recognition capabilities, while optimally allocating the resources to the best ARBs. For each antigen the process trains only those ARBs from the same class with the antigen.
 - (2.2.1) **Perform competition for resources.** The total number of resources is a user defined parameter that limits the number of ARBs.
 - (2.2.1.1) **Stimulation.** The selected antigen is presented to all ARBs and the stimulation is computed for each cell in the ARB pool.
 - (2.2.1.2) **Normalization.** The ARB stimulation values NSt are normalized.
 - (2.2.1.3) **Allocate limited resources.** The amount of resources Rs allocated to each ARB is computed from the normalized stimulation NSt and the clonal rate CR :

$$Rs = NSt \times CR$$

The ARB pool is sorted in the descending order of allocated resources R_s and then resources are removed from the ARB situated at the end of the list until the sum of all allocated resources is lower than the total number of resources.

(2.2.1.4) Remove ARBs with insufficient resources. The ARBs with zero resources are removed from the pool.

(2.2.2) Continue with (2.3) if the stop condition is satisfied. The stop condition for the ARB refinement is met when the average normalized stimulation is higher than a user defined stimulation threshold.

(2.2.3) Generate mutated clones of surviving ARBs.

The number of clones generated for each ARB is:

$$NC = St \times CR$$

where St is the stimulation against the antigen, and CR is the clonal rate. The clones undergo a process of hypermutation, during which the elements of the x vector are randomly modified to increase the antigen recognition.

(2.2.4) Go to (2.2.1)

(2.3) Memory cell selection. In this step, new ARB classifiers are evaluated for inclusion into the memory cell pool. An ARB is inserted into the memory cell pool if its stimulation value is higher than that of the existing best matching memory cell. The existing best matching memory cell is then removed if the affinity between the candidate ARB and the existing memory cell is less than a cut-off value $CutOff$ computed with the formula:

$$CutOff = AT \times ATS$$

where the affinity threshold AT was computed during the initialization phase, and the affinity threshold scalar ATS is a user defined parameter.

(3) Classification. At the end of the training phase, the memory cell pool represents the AIRS classifier. The classification is performed with a k -nearest neighbor method, in which the k best matches to a prediction pattern are identified and the predicted class is determined with a majority vote. The parameter k is user defined, and may be optimized to maximize the prediction performances.

AIRS was applied with success in several drug design structure-activity relationships that are reviewed here. The classification performance of the AIRS algorithm depends on eight user defined parameters: affinity threshold scalar, clonal rate, hypermutation rate, number of nearest neighbors, initial memory cell pool size, number of instances

to compute the affinity threshold, stimulation threshold, and total resources. To illustrate the influence of these parameters, we show the variation of the prediction statistics with the affinity threshold scalar. The statistical indices reported for each AIRS model are: TP_p , true positive in prediction (number of compounds from class +1 classified in class +1); FN_p , false negative in prediction (number of compounds from class +1 classified in class -1); TN_p , true negative in prediction (number of compounds from class -1 classified in class -1); FP_p , false positive in prediction (number of compounds from class -1 classified in class +1); Se_p , prediction selectivity; Sp_p , prediction specificity; Ac_p , prediction accuracy; MCC_p , prediction Matthews correlation coefficient.

Torsade de pointes (TdP) is a polymorphic ventricular arrhythmia that may be caused by drugs that induce the prolongation of the QT interval [111]. QT prolongation and TdP may be caused by a large number of drugs, such as antiarrhythmics, antihistamines, antimicrobials, antidepressants, and antipsychotics. The drug design and development costs may be significantly reduced if, along with other ADME/Tox filters, chemical compounds that have the potential to induce torsade de pointes are eliminated as early as possible. AIRS was applied with success to classify 349 drugs into a subset of 106 drugs that induce torsade de pointes and a subset of 243 drugs that do not induce torsade de pointes [112]. The chemical structure was described with five linear solvation energy relationships (LSER) descriptors, and the prediction of the AIRS models was evaluated with the ten-fold (leave-10%-out) cross-validation. The MCC_p variation with ATS (Table 1) shows that the best predictions are obtained with low values of ATS , in this case $ATS = 0.05$ ($Ac = 0.845$, $MCC = 0.632$). After several steps of optimizations involving the remaining seven parameters, the best AIRS model ($Ac = 0.860$, $MCC = 0.671$) has better predictions than 11 other machine learning algorithms.

Drug Design with Artificial Intelligence Methods, Table 1

AIRS prediction statistics for TdP SAR models based on LSER descriptors and computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
0.01	76	30	213	30	0.7170	0.8765	0.8281	0.5935
0.05	78	28	217	26	0.7358	0.8930	0.8453	0.6323
0.10	78	28	206	37	0.7358	0.8477	0.8138	0.5710
0.30	71	35	210	33	0.6698	0.8642	0.8052	0.5369
0.50	63	43	203	40	0.5943	0.8354	0.7622	0.4333
0.70	55	51	198	45	0.5189	0.8148	0.7249	0.3394
0.90	49	57	198	45	0.4623	0.8148	0.7077	0.2872

Drug Design with Artificial Intelligence Methods, Table 2

AIRS prediction statistics for TdP SAR models based on 2D/3D descriptors and computed for various values of the affinity threshold scalar ATS

ATS	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 159 structural descriptors								
0.01	39	46	213	63	0.4588	0.7717	0.6981	0.2173
0.05	42	43	213	63	0.4941	0.7717	0.7064	0.2484
0.09	43	42	218	58	0.5059	0.7899	0.7230	0.2795
0.15	40	45	204	72	0.4706	0.7391	0.6759	0.1924
0.30	40	45	207	69	0.4706	0.7500	0.6842	0.2039
0.50	36	49	203	73	0.4235	0.7355	0.6620	0.1470
0.70	36	49	201	75	0.4235	0.7283	0.6565	0.1396
0.95	38	47	201	75	0.4471	0.7283	0.6620	0.1604
(b) 13 structural descriptors								
0.01	40	45	236	40	0.4706	0.8551	0.7645	0.3327
0.05	42	43	235	41	0.4941	0.8514	0.7673	0.3484
0.09	40	45	227	49	0.4706	0.8225	0.7396	0.2885
0.15	47	38	226	50	0.5529	0.8188	0.7562	0.3558
0.30	30	55	238	38	0.3529	0.8623	0.7424	0.2336
0.50	24	61	243	33	0.2824	0.8804	0.7396	0.1894
0.70	29	56	246	30	0.3412	0.8913	0.7618	0.2668
0.95	30	55	235	41	0.3529	0.8514	0.7341	0.2182

In a related study, AIRS was applied to the classification of 361 drugs (85 induce torsade de pointes, and 276 do not induce torsade de pointes) based on 159 structural indices computed from the molecular structure [113]. The ATS parameter has a significant influence on the AIRS predictions (Table 2a). A series of fivefold (leave-20%-out) cross-validation tests shows that MCC increases from 0.2173 for ATS = 0.01, peaks at 0.2795 for ATS = 0.09, and then decreases to 0.1604 for ATS = 0.95. To investigate the effect of feature selection on the AIRS prediction quality, Weka was used to reduce the number of features to 13 with the combination SubsetEvaluation and BestFirst. Feature selection significantly improves the TdP predictions (Table 2b), with the best predictions obtained for ATS = 0.15 (MCC = 0.356). These results suggest that feature selection should be explored in order to increase the AIRS prediction power.

A good intestinal absorption is a major requirement for oral drugs [114,115], and various computational models were proposed as fast, reliable, and inexpensive *in silico* methods to assess the intestinal permeability of a chemical compound before synthesis [116,117]. The oral absorption of a drug is influenced by a large number of variables, such as drug formulation and stability, aqueous solubility, contents of the gastrointestinal tract, residence time in the intestine, intestinal metabolism, rate of passive in-

testinal permeability, carrier-mediated influx, and active efflux via transporters. The human intestinal absorption (HIA) of 196 drugs (131 drugs that penetrate the human intestine, and 65 drugs that do not penetrate the intestine) was modeled with the AIRS algorithm [118]. The AIRS classifiers were obtained with 159 structural descriptors from five classes, namely constitutional, topological indices, electrotopological state indices, quantum descriptors, and geometrical indices. The influence of the ATS parameter in L20%O cross-validation was investigated for values between 0.01 and 0.95 (Table 3a). As in previous experiments, MCC increases from 0.3174 for ATS = 0.01 to a maximum of 0.3506 for ATS = 0.09, and then decreases to 0.1997 for ATS = 0.95. After optimizing all eight parameters, the best predictions of the AIRS algorithm (Ac = 0.735, MCC = 0.406) are higher than those obtained with seven other machine learning algorithms, namely Bayesian network, naïve Bayes classifier, updateable naïve Bayes classifier, logistic regression, Gaussian radial basis function network, decision tree with naïve Bayes classifiers at the leaves, and random tree. In a feature selection experiment (SubsetEvaluation and BestFirst) the number of structural descriptors was reduced to 21, which improved considerably the AIRS predictions [119]. The results obtained for the ATS parameter (Table 3b) show a significant increase across the entire range of values, with a maximum of 0.53 for ATS = 0.04.

Drug Design with Artificial Intelligence Methods, Table 3

AIRS prediction statistics for HIA SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 159 structural descriptors								
0.01	105	26	33	32	0.8015	0.5077	0.7041	0.3174
0.04	107	24	33	32	0.8168	0.5077	0.7143	0.3364
0.09	107	24	34	31	0.8168	0.5231	0.7194	0.3506
0.15	107	24	28	37	0.8168	0.4308	0.6888	0.2640
0.30	100	31	30	35	0.7634	0.4615	0.6633	0.2287
0.50	105	26	25	40	0.8015	0.3846	0.6633	0.1997
0.70	105	26	25	40	0.8015	0.3846	0.6633	0.1997
0.95	105	26	25	40	0.8015	0.3846	0.6633	0.1997
(b) 21 structural descriptors								
0.01	113	18	41	24	0.8626	0.6308	0.7857	0.5064
0.04	114	17	42	23	0.8702	0.6462	0.7959	0.5300
0.09	113	18	39	26	0.8626	0.6000	0.7755	0.4796
0.15	111	20	40	25	0.8473	0.6154	0.7704	0.4727
0.30	116	15	30	35	0.8855	0.4615	0.7449	0.3885
0.50	121	10	30	35	0.9237	0.4615	0.7704	0.4500
0.70	120	11	28	37	0.9160	0.4308	0.7551	0.4090
0.95	123	8	28	37	0.9389	0.4308	0.7704	0.4495

P-glycoprotein (Pgp) is responsible for the low cellular accumulation of anticancer drugs, for reduced oral absorption, for low blood-brain barrier penetration, and is involved in hepatic, renal, or intestinal elimination of drugs. Computational methods for the identification of Pgp substrates are useful drug design tools for the early elimination of potential Pgp substrates [120,121]. The immune system classifier AIRS was used to discriminate between 116 Pgp substrates and 85 Pgp nonsubstrates [122]. The SAR models were computed from 159 structural descriptors and the prediction power was estimated with L20%O cross-validation. Low values for the ATS parameter give better predictions, with the highest predictions obtained for $ATS = 0.03$ (Table 4a). The AIRS model optimized for all eight parameters ($Ac = 0.702$, $MCC = 0.380$) is better than five machine learning algorithms (alternating decision tree, Bayesian network, logistic regression with ridge estimator, random tree, and fast decision tree learner), demonstrating that Pgp substrates may be successfully recognized with AIRS. A feature selection step reduces the number of structural descriptors from 159 to 15, and increases the SAR performances over the entire range of ATS values (Table 4b) [119]. The best predictions are obtained with $ATS = 0.01$ ($Ac = 0.736$ and $MCC = 0.467$), but the variation of the prediction statistics is not monotonous with ATS, and no simple rule can be extracted to guide further experiments.

Drug Design with Artificial Intelligence Methods, Table 4

AIRS prediction statistics for Pgp SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 159 structural descriptors								
0.01	85	31	45	40	0.7328	0.5294	0.6468	0.2671
0.03	85	31	48	37	0.7328	0.5647	0.6617	0.3009
0.07	80	36	46	39	0.6897	0.5412	0.6269	0.2320
0.15	78	38	51	34	0.6724	0.6000	0.6418	0.2709
0.30	80	36	42	43	0.6897	0.4941	0.6070	0.1863
0.50	80	36	44	41	0.6897	0.5176	0.6169	0.2092
0.70	80	36	43	42	0.6897	0.5059	0.6119	0.1978
0.95	80	36	43	42	0.6897	0.5059	0.6119	0.1978
(b) 15 structural descriptors								
0.01	86	30	62	23	0.7414	0.7294	0.7363	0.4668
0.03	85	31	59	26	0.7328	0.6941	0.7164	0.4241
0.07	85	31	54	31	0.7328	0.6353	0.6915	0.3681
0.15	88	28	58	27	0.7586	0.6824	0.7264	0.4403
0.30	87	29	57	28	0.7500	0.6706	0.7164	0.4199
0.50	81	35	56	29	0.6983	0.6588	0.6816	0.3544
0.70	78	38	58	27	0.6724	0.6824	0.6766	0.3509
0.95	80	36	56	29	0.6897	0.6588	0.6766	0.3455

Drug Design with Artificial Intelligence Methods, Table 5

AIRS prediction statistics for BZR SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 75 structural descriptors								
0.01	63	19	52	29	0.7683	0.6420	0.7055	0.4137
0.05	64	18	53	28	0.7805	0.6543	0.7178	0.4385
0.10	62	20	52	29	0.7561	0.6420	0.6994	0.4008
0.20	60	22	57	24	0.7317	0.7037	0.7178	0.4356
0.25	65	17	56	25	0.7927	0.6914	0.7423	0.4867
0.30	62	20	56	25	0.7561	0.6914	0.7239	0.4485
0.50	63	19	56	25	0.7683	0.6914	0.7301	0.4611
0.95	63	19	56	25	0.7683	0.6914	0.7301	0.4611
(b) 16 structural descriptors								
0.01	64	18	57	24	0.7805	0.7037	0.7423	0.4857
0.05	65	17	57	24	0.7927	0.7037	0.7485	0.4985
0.10	57	25	55	26	0.6951	0.6790	0.6871	0.3742
0.20	62	20	65	16	0.7561	0.8025	0.7791	0.5591
0.25	60	22	62	19	0.7317	0.7654	0.7485	0.4974
0.30	61	21	61	20	0.7439	0.7531	0.7485	0.4970
0.50	59	23	62	19	0.7195	0.7654	0.7423	0.4854
0.95	59	23	61	20	0.7195	0.7531	0.7362	0.4728

Another successful application of AIRS in drug design was reported for the identification of benzodiazepine receptor (BZR) ligands [123]. The structure of the 163 BZR ligands was encoded with 75 structural descriptors, and AIRS classifiers were trained to discriminate between 82 high affinity ligands (class +1, pIC_{50} between 8.92 and 7.80) and 81 low affinity ligands (class -1, pIC_{50} between 7.77 and 5). A scan of the ATS values (Table 5a) shows that the best predictions are obtained for $ATS = 0.025$ ($Ac = 0.7423$ and $MCC = 0.4867$). The feature selection step further reduces the number of structural descriptors to 16 (Table 5b), and results in better predictions (best $ATS = 0.20$, with $Ac = 0.7791$ and $MCC = 0.5591$).

Numerous organic chemicals are environmental pollutants, and a considerable number of studies are dedicated to the computational prediction of their mechanism of aquatic toxicity (MOA). The reliable prediction of MOA has major applications in selecting the appropriate QSAR model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimens [124,125]. The immune system AIRS was applied for the MOA prediction of 187 chemicals (143 non-polar narcotics, and 44 polar narcotics) [126]. The chemical structure was described with five LSER descriptors, and the AIRS predictions were evaluated with the ten-fold cross-validation. The ATS parameter was modified between 0.01 and 0.95 (Table 6), and the

Drug Design with Artificial Intelligence Methods, Table 6

AIRS prediction statistics for MOA SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
0.01	138	5	40	4	0.9650	0.9091	0.9519	0.8674
0.02	138	5	40	4	0.9650	0.9091	0.9519	0.8674
0.05	138	5	40	4	0.9650	0.9091	0.9519	0.8674
0.15	135	8	39	5	0.9441	0.8864	0.9305	0.8120
0.30	139	4	39	5	0.9720	0.8864	0.9519	0.8653
0.50	138	5	39	5	0.9650	0.8864	0.9465	0.8514
0.70	137	6	39	5	0.9580	0.8864	0.9412	0.8379
0.95	137	6	39	5	0.9580	0.8864	0.9412	0.8379

best predictions were obtained for low ATS values (0.01, 0.02, and 0.05), namely $Ac = 0.9519$ and $MCC = 0.8674$. Based on the high prediction rates obtained with AIRS, such models may be used to identify the aquatic toxicity mechanism and to select the appropriate computational model for new chemical compounds.

CLONALG – Clonal Selection Algorithm

An AIS algorithm that gives a central role to the clonal selection theory is CLONALG, proposed by de Castro and Von Zuben [103,104]. CLONALG implements several mechanisms of the clonal selection: training of a group of memory cells; identification and cloning of the antibodies with the highest recognition power; death of the antibodies with low recognition power; cloning and hypermutation of the antibodies with high recognition power; evaluation and replacement of the clones; generation and preservation of antibody diversity. The CLONALG algorithm, as implemented by Brownlee, consists of the following steps [105]:

- (1) **Initialization.** The CLONALG algorithm starts by generating a pool of N antibodies, which is subsequently partitioned into the memory antibody pool (MAP) and the remaining antibody pool (RAP). MAP contains m antibodies, and at the end of the training process it will represent the solution of the CLONALG classifier. RAP contains the remaining antibodies, $r = N - m$, and it has the role of adding additional diversity during the learning phase.
- (2) **Train antibodies.** The main part of the CLONALG algorithm is an iterative process of exposing the system to all antigens from the training set for a number of G generations (iterations).
- (2.1) **Train for each antigen.** Repeat steps (2.2)–(2.9) for all antigens in the training set. In each generation, an antigen is selected for training once and only once.

(2.2) **Antigen selection.** For each generation, an antigen is randomly selected without replacement from the entire pool of antigens.

(2.3) **Affinity calculation.** The selected antigen interacts with all antibodies, and the affinity is calculated for the interaction between the antigen and every antibody in the system. The affinity measures the similarity between an antigen and an antibody, and is based on the Euclidean distance between the vectors of structural descriptors that characterize the antigen and the antibody.

(2.4) **Select antibodies.** The antibodies are ranked according to their decreasing affinity towards the antigen, and the top n antibodies are selected for further processing.

(2.5) **Clone antibodies.** All n antibodies selected in the previous step are cloned proportionally with their affinity. The number of clones computed for an antibody that is ranked i th according to its affinity, with $i \in [1, n]$, is

$$N_c = \left\lfloor \frac{CF \times N}{i} + 0.5 \right\rfloor$$

where CF is the clonal factor. The total number of clones generated for the entire system of n antibodies is:

$$NC = \sum_{i=1}^n N_c.$$

(2.6) **Affinity maturation.** The clones enter the process of affinity maturation, during which random mutations are performed onto each clone in order to increase its affinity towards the antigen. The degree of affinity maturation is inversely proportional to the initial affinity, namely the lower the initial affinity the greater the mutation rate is.

(2.7) **Evaluate clones.** All clones are exposed to the antigen to compute their affinity.

(2.8) **Select candidates.** The antibodies with the highest affinity are selected to replace antibodies from MAP that have lower affinities.

(2.9) **Replacement.** The RAP group of antibodies is ranked according to the decreasing affinity towards the antigen, and the set of s antibodies with the lowest affinity is replaced with random antibodies.

(3) **Classification.** After training the system for G generations, the MAP group of antigens represents the solution of the CLONALG classifier.

The CLONALG machine learning was tested with success in drug design applications, namely recognition of

glycogen phosphorylase B inhibitors, classification of benzodiazepine receptor ligands, and identification of polar and nonpolar narcotic pollutants. To illustrate the effect of the user defined parameters on the prediction performance of CLONALG, we show the influence of the clonal factor CF on the L20%O cross-validation statistics. The clonal factor is a scaling factor, with values between 0 and 1, that determines the number of clones generated for each selected antibody. Low values for CF result in a local search, whereas for high values the algorithm generates a larger number of clones that may explore a wider region and result in a higher diversity.

CLONALG was applied in drug development for the recognition of glycogen phosphorylase B (GPB) inhibitors, based on a set of 66 compounds and 70 structural descriptors [127]. The subset of active compounds contains 33 chemicals (class +1, pK_i between 6.8 and 2.5), whereas the subset of inactive compounds contains the remaining 33 chemicals (class -1, pK_i between 2.4 and 1.3). The prediction performance depends on the number of clones generated, controlled by the values of CF (Table 7a), with best results for CF = 0.50 (Ac = 0.6212 and MCC = 0.2453), whereas low and high values for CF result in lower predictions. A feature selection step drastically reduces the number of structural descriptors from 70 to 2, while the model prediction increases (Ac = 0.6667 and MCC = 0.3339) for several CF values (Table 7b).

Drug Design with Artificial Intelligence Methods, Table 7

CLONALG prediction statistics for GPB SAR models computed for various values of the clonal factor CF

CF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 70 structural descriptors								
0.01	20	13	15	18	0.6061	0.4545	0.5303	0.0613
0.05	22	11	17	16	0.6667	0.5152	0.5909	0.1839
0.08	22	11	14	19	0.6667	0.4242	0.5455	0.0937
0.15	20	13	17	16	0.6061	0.5152	0.5606	0.1217
0.25	23	10	15	18	0.6970	0.4545	0.5758	0.1562
0.50	23	10	18	15	0.6970	0.5455	0.6212	0.2453
0.65	24	9	16	17	0.7273	0.4848	0.6061	0.2186
0.95	22	11	14	19	0.6667	0.4242	0.5455	0.0937
(b) 2 structural descriptors								
0.01	21	12	22	11	0.6364	0.6667	0.6515	0.3032
0.05	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.08	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.15	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.25	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.50	21	12	22	11	0.6364	0.6667	0.6515	0.3032
0.65	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.95	21	12	22	11	0.6364	0.6667	0.6515	0.3032

Drug Design with Artificial Intelligence Methods, Table 8

CLONALG prediction statistics for BZR SAR models computed for various values of the clonal factor CF

CF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 75 structural descriptors								
0.01	56	26	48	33	0.6829	0.5926	0.6380	0.2767
0.05	57	25	50	31	0.6951	0.6173	0.6564	0.3134
0.10	58	24	51	30	0.7073	0.6296	0.6687	0.3380
0.20	56	26	52	29	0.6829	0.6420	0.6626	0.3252
0.45	56	26	53	28	0.6829	0.6543	0.6687	0.3374
0.60	62	20	52	29	0.7561	0.6420	0.6994	0.4008
0.85	57	25	47	34	0.6951	0.5802	0.6380	0.2773
0.95	63	19	47	34	0.7683	0.5802	0.6748	0.3550
(b) 16 structural descriptors								
0.01	42	40	53	28	0.5122	0.6543	0.5828	0.1682
0.05	57	25	53	28	0.6951	0.6543	0.6748	0.3498
0.10	53	29	55	26	0.6463	0.6790	0.6626	0.3255
0.20	57	25	54	27	0.6951	0.6667	0.6810	0.3620
0.45	64	18	52	29	0.7805	0.6420	0.7117	0.4267
0.60	61	21	52	29	0.7439	0.6420	0.6933	0.3880
0.85	51	31	55	26	0.6220	0.6790	0.6503	0.3014
0.95	57	25	55	26	0.6951	0.6790	0.6871	0.3742

The CLONALG immune system was tested for the classification of 163 benzodiazepine receptor (BZR) ligands (82 high affinity ligands and 81 low affinity ligands) which are encoded with 75 structural descriptors [123]. The clonal factor was modified between 0.01 and 0.95 (Table 8a). The prediction MCC increases from 0.2767 for CF = 0.01, peaks at 0.4008 for CF = 0.60, and then decreases to 0.2888 for CF = 0.90. These results indicate that too few or too many clones are detrimental to the antigen recognition. The number of structural descriptors can be significantly reduced to 16 by feature selection (Table 8b), which also results in a slight increase of the prediction quality (MCC = 0.4267 for CF = 0.45). The optimum CF is situated in the middle of the range of CF values, similarly with the results obtained for the identification of GPB inhibitors.

The mechanism of toxic action of polar and nonpolar narcotic pollutants may be efficiently identified with CLONALG classifiers [128]. The dataset consists of 190 compounds (114 nonpolar pollutants, class +1; 76 polar pollutants, class -1), with each chemical characterized by five structural descriptors, namely the octanol-water partition coefficient, the energy of the highest occupied molecular orbital, the energy of the lowest unoccupied molecular orbital, the most negative partial charge on any non-hydrogen atom in the molecule, and the most positive partial charge on a hydrogen atom. The prediction MCC has no clear-cut variation with CF (Table 9), but the optimum

Drug Design with Artificial Intelligence Methods, Table 9
CLONALG prediction statistics for MOA SAR models computed for various values of the clonal factor CF

CF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
0.01	98	16	73	3	0.8596	0.9605	0.9000	0.8052
0.05	100	14	72	4	0.8772	0.9474	0.9053	0.8116
0.10	103	11	75	1	0.9035	0.9868	0.9368	0.8763
0.15	105	9	68	8	0.9211	0.8947	0.9105	0.8141
0.30	104	10	72	4	0.9123	0.9474	0.9263	0.8503
0.55	110	4	69	7	0.9649	0.9079	0.9421	0.8791
0.70	102	12	73	3	0.8947	0.9605	0.9211	0.8427
0.90	105	9	73	3	0.9211	0.9605	0.9368	0.8720

is still in the middle of the range, as in previous studies, with MCC = 0.8791 for CF = 0.55.

CSCA – Clonal Selection Classification System

The clonal selection classification system, developed by Brownlee, is formulated as a function optimization procedure that maximizes the number of patterns correctly classified and minimizes the number of patterns incorrectly classified [105]. Unlike the AIRS algorithm, in which the system is exposed only once to the set of antigens, CSCA is trained for several generations, and during each generation the entire set of antibodies is exposed to all antigens. The computational steps of the CSCA algorithm are shown in the following procedure:

- (1) **Initialization.** The CSCA algorithm starts by generating a set of N antibodies.
- (2) **Training.** Repeat the training of all antibodies for G generations (iterations).
 - (2.1) **Selection and pruning.** The entire group of antibodies is exposed to the antigen set and a fitness score is computed for each antibody. Then all antibodies are selected and the following three evaluation rules are applied to each antibody:
 - (2.1.1) Remove from the selected set all antibodies with a misclassification score of zero.
 - (2.1.2) Antibodies that have zero correct classification and misclassification higher than zero are reassigned to the class of the majority. Fitness is recalculated.
 - (2.1.3) Remove from the selected set and from the base antibody population all antibodies with a fitness scoring lower than a threshold.
 - (2.2) **Cloning and mutation.** The selected set of antibodies is cloned and mutated.
 - (2.3) **Insert new antibodies.** Insert the clones generated into the main antibody population. A number of n randomly selected antigens from the antigen set are in-

serted into the main antibody population, where n is the number of antibodies selected in step (2.1).

- (3) **Final pruning.** The antibody population is exposed to the entire antigen population, fitness scores are computed for each antibody, and pruning of antibodies is performed as described in step (2.1.3).
- (4) **Select classifier.** The final antibody population represents the CSCA classifier. To classify a new pattern, the classification antibodies are exposed to the pattern, then the k most similar (highest affinity) antibodies are selected and a majority vote assigns the class of the pattern.

The artificial immune system CSCA was applied in several virtual screening studies, namely identification of estrogen receptor ligands, recognition of dihydrofolate reductase inhibitors, classification of angiotensin converting enzyme inhibitors, detection of benzodiazepine receptor ligands, and SAR for thermolysin inhibitors. To demonstrate the influence of the user defined parameters on the CSCA predictions, we present the influence of the clonal scale factor CSF, tested in L20%O cross-validation. CSF is used to increase or decrease the number of clones generated for each antibody, and has a default value of one. Low values for CSF promote a low diversity of solutions, whereas high CSF values increase the diversity of the recognition cells.

CSCA was applied for the classification of 232 chemical compounds into estrogen receptor (ER) ligands (131 chemicals, class +1) and compounds that do not bind to the estrogen receptor (101 chemicals, class -1) [129]. The chemical structure was represented with 312 topological indices computed with Molconn-Z. The clonal scale factor was modified between 0.1 and 4 (Table 10a), with the best predictions obtained for CSF = 2 ($Ac = 0.6207$ and $MCC = 0.2057$), but with no clear trend apparent for the values that give the best predictions. For example, the next best predictions are obtained for CSF = 0.1 ($MCC = 0.1935$), whereas the lowest predictions are obtained with CSF = 0.7 ($MCC = 0.0416$). To investigate the influence of feature selection on the classification abilities of CSCA, 29 structural descriptors were selected with SubsetEvaluation and BestFirst, which results in slightly better predictions for a much lower value of CSF (CSF = 0.1, $Ac = 0.6336$, $MCC = 0.2508$; Table 10b).

Dihydrofolate reductase (DHFR) inhibitors may be efficiently identified with CSCA, as was demonstrated for a dataset of 397 chemicals (198 compounds in class +1, pIC_{50} between 9.81 and 6.08; 199 compounds in class -1, pIC_{50} between 6.06 and 3.30) [130]. CSCA classifiers computed with 70 structural descriptors are used to eval-

Drug Design with Artificial Intelligence Methods, Table 10

CSCA prediction statistics for ER SAR models computed for various values of the clonal scale factor CSF

CSF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 312 structural descriptors								
0.1	98	33	44	57	0.7481	0.4356	0.6121	0.1935
0.3	92	39	35	66	0.7023	0.3465	0.5474	0.0519
0.5	97	34	37	64	0.7405	0.3663	0.5776	0.1149
0.7	92	39	34	67	0.7023	0.3366	0.5431	0.0416
1.0	108	23	29	72	0.8244	0.2871	0.5905	0.1326
2.0	112	19	32	69	0.8550	0.3168	0.6207	0.2057
3.0	99	32	31	70	0.7557	0.3069	0.5603	0.0698
4.0	106	25	30	71	0.8092	0.2970	0.5862	0.1238
(b) 29 structural descriptors								
0.1	91	40	56	45	0.6947	0.5545	0.6336	0.2508
0.3	99	32	45	56	0.7557	0.4455	0.6207	0.2119
0.5	97	34	42	59	0.7405	0.4158	0.5991	0.1651
0.7	94	37	47	54	0.7176	0.4653	0.6078	0.1887
1.0	98	33	47	54	0.7481	0.4653	0.6250	0.2226
2.0	98	33	38	63	0.7481	0.3762	0.5862	0.1338
3.0	96	35	41	60	0.7328	0.4059	0.5905	0.1466
4.0	98	33	42	59	0.7481	0.4158	0.6034	0.1738

uate the effect of the clonal scale factor on the prediction accuracy. Based on the structure of the CSCA algorithm, it should be expected that higher CSF values are useful in identifying better solutions, because more clones are generated, and the system explores a wider diversity of solutions. However, for dihydrofolate reductase inhibitors, the highest predictions are obtained for CSF = 0.2 (Ac = 0.5945 and MCC = 0.1935; Table 11a). Also, for high CSF values, between 0.7 and 3, MCC decreases markedly. A dramatic increase of the CSCA model quality is obtained with a feature selection that reduces the set of structural descriptors to 5 (Table 11b). The best predictions are obtained for a much higher CSF value, namely CSF = 3, with Ac = 0.7834 and MCC = 0.5670. Further tests should be performed with other SAR datasets in order to find the optimum CSF values for various drug screening experiments.

Another set of experiments with CSCA involved the classification of 114 angiotensin converting enzyme (ACE) inhibitors (57 compounds in class +1, pIC₅₀ between 9.94 and 6.41; 57 compounds in class -1, pIC₅₀ between 6.37 and 2.14) [131]. The chemical structure was encoded with 56 structural descriptors, and the CSF influence was evaluated for 16 values between 0.1 and 4. For all but one CSF values the CSCA classifiers give the same prediction indices, with Ac = 0.8684 and MCC = 0.7510. The CSCA insensitivity to the CSF variation is unexpected, and more experiments are necessary to fully un-

Drug Design with Artificial Intelligence Methods, Table 11

CSCA prediction statistics for DHFR SAR models computed for various values of the clonal scale factor CSF

CSF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 70 structural descriptors								
0.1	130	68	92	107	0.6566	0.4623	0.5592	0.1212
0.2	138	60	98	101	0.6970	0.4925	0.5945	0.1935
0.5	129	69	92	107	0.6515	0.4623	0.5567	0.1159
0.7	130	68	87	112	0.6566	0.4372	0.5466	0.0961
1.0	119	79	98	101	0.6010	0.4925	0.5466	0.0940
2.0	142	56	87	112	0.7172	0.4372	0.5768	0.1608
3.0	115	83	99	100	0.5808	0.4975	0.5390	0.0786
4.0	132	66	89	110	0.6667	0.4472	0.5567	0.1167
(b) 5 structural descriptors								
0.1	166	32	144	55	0.8384	0.7236	0.7809	0.5656
0.2	155	43	142	57	0.7828	0.7136	0.7481	0.4975
0.5	151	47	150	49	0.7626	0.7538	0.7582	0.5164
0.7	149	49	148	51	0.7525	0.7437	0.7481	0.4963
1.0	152	46	151	48	0.7677	0.7588	0.7632	0.5265
2.0	154	44	150	49	0.7778	0.7538	0.7657	0.5317
3.0	158	40	153	46	0.7980	0.7688	0.7834	0.5670
4.0	148	50	151	48	0.7475	0.7588	0.7531	0.5063

Drug Design with Artificial Intelligence Methods, Table 12

CSCA prediction statistics for ACE SAR models with 12 structural descriptors computed for various values of the clonal scale factor CSF

CSF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
0.1	45	12	50	7	0.7895	0.8772	0.8333	0.6692
0.3	43	14	49	8	0.7544	0.8596	0.8070	0.6175
0.5	47	10	48	9	0.8246	0.8421	0.8333	0.6668
0.7	44	13	49	8	0.7719	0.8596	0.8158	0.6340
0.9	47	10	50	7	0.8246	0.8772	0.8509	0.7027
2.0	46	11	43	14	0.8070	0.7544	0.7807	0.5622
3.0	46	11	49	8	0.8070	0.8596	0.8333	0.6676
4.0	46	11	48	9	0.8070	0.8421	0.8246	0.6495

derstand this behavior. A feature selection step decreases the pool of structural descriptors to 12 (Table 12), with a slight decrease in the prediction statistics (CSF = 0.9, Ac = 0.8509, MCC = 0.7027). Usually, feature selection provides a smaller set of structural descriptors that increase the predictions of artificial immune systems. The exception encountered for ACE inhibitors should be further investigated to identify possible explanations and better feature selection procedures.

The CSCA immune system was evaluated for the discrimination of 163 benzodiazepine receptor (BZR) ligands (82 high affinity ligands and 81 low affinity ligands) [123]. Starting from a set of 75 structural descriptors, CSF was modified between 0.1 and 4 (Table 13a), with

Drug Design with Artificial Intelligence Methods, Table 13

CSCA prediction statistics for BZR SAR models computed for various values of the clonal scale factor CSF

CSF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 75 structural descriptors								
0.1	55	27	57	24	0.6707	0.7037	0.6871	0.3746
0.3	52	30	58	23	0.6341	0.7160	0.6748	0.3513
0.5	52	30	58	23	0.6341	0.7160	0.6748	0.3513
0.7	57	25	57	24	0.6951	0.7037	0.6994	0.3988
1.0	48	34	65	16	0.5854	0.8025	0.6933	0.3971
2.0	48	34	59	22	0.5854	0.7284	0.6564	0.3169
3.0	54	28	54	27	0.6585	0.6667	0.6626	0.3252
4.0	58	24	52	29	0.7073	0.6420	0.6748	0.3501
(b) 16 structural descriptors								
0.1	56	26	55	26	0.6829	0.6790	0.6810	0.3619
0.3	62	20	51	30	0.7561	0.6296	0.6933	0.3890
0.5	53	29	57	24	0.6463	0.7037	0.6748	0.3506
0.7	60	22	54	27	0.7317	0.6667	0.6994	0.3993
1.0	61	21	45	36	0.7439	0.5556	0.6503	0.3050
2.0	65	17	50	31	0.7927	0.6173	0.7055	0.4166
3.0	58	24	55	26	0.7073	0.6790	0.6933	0.3865
4.0	58	24	51	30	0.7073	0.6296	0.6687	0.3380

the best results obtained for CSF = 0.7 (Ac = 0.6994 and MCC = 0.3988). A small improvement of the CSCA predictions is obtained by reducing the pool of descriptors to 16 by feature selection (Table 13b). Although the model improvement is not big (Ac = 0.7055 and MCC = 0.4166 for CSF = 2), feature selection is still important because the CSCA model can be computed faster, and the selected descriptors may suggest which molecular features influence the biological activity.

CSCA was also tested for a dataset of 76 thermolysin (THER) inhibitors (38 compounds in class +1, pK_i between 10.17 and 5.55; 38 compounds in class -1, pK_i between 5.16 and 0.52) and 64 structural descriptors [132]. For 14 out of 16 CSF values tested in this experiment, the CSCA classifiers have identical predictions, with Ac = 0.6711 and MCC = 0.3162. The best predictions are obtained for CSF = 3.5, with slightly higher prediction statistics, namely MCC = 0.3422 (Table 14a). Feature selection reduces the number of descriptors to 10, which results in a minor improvement (CSF = 2, Ac = 0.6974, MCC = 0.4124, Table 14b).

IMMUNOS

Carter developed the IMMUNOS-81 artificial immune systems as an instance based classifier with some similarity to *k*-nearest neighbor classifiers [106]. Brownlee extended this algorithm by adding elements from other AIS classi-

Drug Design with Artificial Intelligence Methods, Table 14

CSCA prediction statistics for THER SAR models computed for various values of the clonal scale factor CSF

CSF	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 64 structural descriptors								
0.1	24	14	26	12	0.6316	0.6842	0.6579	0.3162
0.5	24	14	26	12	0.6316	0.6842	0.6579	0.3162
1.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
2.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
2.5	19	19	27	11	0.5000	0.7105	0.6053	0.2154
3.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
3.5	25	13	26	12	0.6579	0.6842	0.6711	0.3422
4.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
(b) 10 structural descriptors								
0.1	20	18	27	11	0.5263	0.7105	0.6184	0.2410
0.5	19	19	31	7	0.5000	0.8158	0.6579	0.3328
1.0	20	18	24	14	0.5263	0.6316	0.5789	0.1588
2.0	21	17	32	6	0.5526	0.8421	0.6974	0.4124
2.5	19	19	27	11	0.5000	0.7105	0.6053	0.2154
3.0	21	17	31	7	0.5526	0.8158	0.6842	0.3819
3.5	13	25	32	6	0.3421	0.8421	0.5921	0.2127
4.0	16	22	32	6	0.4211	0.8421	0.6316	0.2901

fiers, such as cloning and hypermutation, to obtain IMMUNOS-99 [107]. A brief description of the IMMUNOS-99 consists of the following steps:

- (1) **Initialization.** The training group of antigens is divided into groups based on class label.
- (2) **Train B-cell groups.** The final IMMUNOS classifier consists of a B-cell population for each class represented in the training set of antigens. Each B-cell population is generated and trained independent of the other B-cell populations. Steps (2.1) and (2.2) are repeated *C* times, where *C* is the number of antigen classes.
 - (2.1) **Create B-cell population.** Generate a B-cell population for the antigen class under training. A fraction of the antigen population from that class is used as seed for the B-cell population.
 - (2.2) **Training.** Train the B-cell class for *G* generations (iterations).
 - (2.2.1) **Expose population.** The B-cell population is exposed to all antigens from all classes, and an affinity value is computed for each B-cell/antigen comparison. A rank-based scoring is established for each B-cell.
 - (2.2.2) **Compute fitness.** A fitness index is computed for each B-cell, based on the rank scores for antigens in the same class and the rank scores for antigens in all other classes. B-cells that recognize better the antigens from the same class have fitness score higher than one,

whereas B-cells that recognize better the antigens from other classes have fitness score lower than one.

(2.2.3) Pruning. A user-defined parameter, between [0, 1], sets the minimum fitness score of a B-cell. All B-cells with fitness scores lower than this threshold are removed from the population.

(2.2.4) Affinity maturation. After pruning, the B-cell population contains only cells that can identify antigens from the same class. To improve the B-cell recognition ability, the system undergoes an affinity maturation process based on cloning and hypermutation.

(2.2.4.1) Order population. The B-cell population is ordered in the descending order of the fitness scores.

(2.2.4.2) Generate clones. Each B-cell is cloned proportional to its fitness rank. The rank ratio for a B-cell is:

$$r_i = \frac{\text{rank}}{S}$$

where r_i is the rank ratio of the i th B-cell, rank is the actual index of the B-cell in the ordered sequence, $\text{rank} \in [1, S]$, and S is the total number of B-cells in the population (class). The number of clones generated for each B-cell is:

$$NC_i = \left\lfloor \frac{r_i}{\sum_{j=1}^S r_j} N + 0.5 \right\rfloor$$

where N is the total number of antigens in the same class.

(2.2.4.3) Mutate clones. The clones are mutated by the inverse of the B-cell rank ratios. As a result of this procedure, clones of B-cells with higher ranks undergo small mutations, whereas clones of B-cells with lower ranks go through large mutations. All clones generated are added to the B-cell population.

(2.2.5) Insert random antigens. In order to increase the diversity of the B-cell population, a random selection of antigens from the same class is added to the B-cell pool. The number of antigens added is equal to the number of B-cells deleted during the pruning process from step (2.2.3). The diversity introduced by the antigen-based B-cells is particularly useful whenever the affinity maturation process converges to a limited number of B-cells.

(3) Final pruning. This step removes B-cells with low fitness after the system finishes the training for each antigen class and for the set number of generations G .

(3.1) Compute fitness. Each B-cell population (class) is exposed to all antigens, one antigen at a time, and only the best matching B-cells receive a score.

(3.2) Pruning. Similarly with the pruning process from step (2.2.3), all B-cells with low fitness scores lower are removed from the population.

(4) Select classifier. The populations of B-cells that survive the final pruning represent the classifier for new, unknown antigens. During the classification process, each B-cell class is exposed to the unknown antigen, and an avidity index is computed. Then the B-cell populations compete for the unknown antigen that takes the class label of the B-cell population with the highest avidity index.

The IMMUNOS-99 system was evaluated in several drug design studies, namely structure-activity relationships for acetylcholinesterase inhibitors, virtual screening of cyclooxygenase-2 inhibitors, recognition of benzodiazepine receptor ligands, and classification of thrombin inhibitors. All examples presented here investigate the influence of the seed population percentage SPP. SPP is a user defined parameter that specifies the percentage of the antigen population from each class that is used as seed for the B-cell population. If SPP = 100% then the initial B-cell population is identical with the antigen population in the same class. The influence of the SPP parameter was investigated in series of L20%O cross-validation experiments. For each drug design dataset, the IMMUNOS-99 classifier was trained for 19 values of the SPP parameter, between 0.05 and 0.95.

IMMUNOS-99 structure-activity relationships were developed for a dataset of 111 acetylcholinesterase (AChE) inhibitors characterized by 63 structural descriptors [133]. The classifiers were trained to discriminate between 55 inhibitors in class +1 (pIC_{50} between 9.52 and 6.87) and 56 inhibitors in class -1 (pIC_{50} between 6.84 and 4.27). The prediction MCC increases from 0.1349 for SPP = 0.05, has a maximum of 0.2847 for SPP = 0.35, and then decreases to 0.2110 for SPP = 0.95 (Table 15a). These results suggest that seeding the B-cell population with less than half of the antigen population improves the prediction statistics. The number of structural descriptors is reduced to 9 by feature selection, which results in a slight decrease in the IMMUNOS-99 predictions (Table 15b).

The virtual screening of cyclooxygenase-2 (COX2) inhibitors may be efficiently done with IMMUNOS-99, as shown for 322 compounds (162 compounds in class +1, pIC_{50} between 9 and 6.60; 160 compounds in class -1, pIC_{50} between 6.59 and 4) [134]. Starting from a set of 74 structural descriptors, several IMMUNOS-99 classifiers were developed to study the influence of the SPP parameter (Table 16a). The results obtained from this series of experiments indicate that the prediction statistics

Drug Design with Artificial Intelligence Methods, Table 15

IMMUNOS-99 prediction statistics for AChE SAR models computed for various values of the seed population percentage SPP

SPP	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 63 structural descriptors								
0.05	30	25	33	23	0.5455	0.5893	0.5676	0.1349
0.10	45	10	22	34	0.8182	0.3929	0.6036	0.2329
0.15	44	11	18	38	0.8000	0.3214	0.5586	0.1382
0.35	45	10	25	31	0.8182	0.4464	0.6306	0.2847
0.50	45	10	22	34	0.8182	0.3929	0.6036	0.2329
0.70	43	12	23	33	0.7818	0.4107	0.5946	0.2072
0.85	44	11	23	33	0.8000	0.4107	0.6036	0.2286
0.95	44	11	23	33	0.8000	0.4107	0.6036	0.2286
(b) 9 structural descriptors								
0.05	35	20	21	35	0.6364	0.3750	0.5045	0.0118
0.10	41	14	18	38	0.7455	0.3214	0.5315	0.0738
0.15	47	8	15	41	0.8545	0.2679	0.5586	0.1510
0.35	48	7	6	50	0.8727	0.1071	0.4865	-0.0313
0.50	53	2	3	53	0.9636	0.0536	0.5045	0.0415
0.70	55	0	3	53	1.0000	0.0536	0.5225	0.1652
0.85	54	1	3	53	0.9818	0.0536	0.5135	0.0949
0.95	54	1	2	54	0.9818	0.0357	0.5045	0.0541

have similar values for a wide range of the SPP parameter, with a small improvement for SPP = 0.45. The number of structural descriptors was reduced by feature selection to 12 important descriptors, thus improving the predictions of the IMMUNOS-99 classifiers (Table 16b). The best results are obtained for SPP = 0.75 (Ac = 0.6429, MCC = 0.3855), but FP is still too large, i.e., too many inactive compounds are predicted as active.

The IMMUNOS-99 immune system was also tested for the dataset of benzodiazepine receptor (BZR) ligands (82 high affinity ligands and 81 low affinity ligands) [123]. The best predictions for the entire pool of 75 structural descriptors were obtained for SPP = 0.75 (Ac = 0.6564, MCC 0.3506; Table 17a). To evaluate the importance of feature selection, the number of structural descriptors was reduced to 16 and the entire analysis was repeated for the full range of SPP values. Although FN decreases (active compounds predicted inactive), FP increases which results in slightly worse predictions (Table 17b). The best predictions are obtained also for SPP = 0.15 (Ac = 0.6503, MCC = 0.3006), but the results suggest that IMMUNOS-99 predictions do not improve with feature selection.

The classification of thrombin (THR) inhibitors with IMMUNOS-99 was investigated for 88 chemicals (44 compounds in class +1, pK_i between 8.48 and 6.70; 44 compounds in class -1, pK_i between 6.68 and 4.36) and 66 structural descriptors [135]. The prediction statistics indi-

Drug Design with Artificial Intelligence Methods, Table 16

IMMUNOS-99 prediction statistics for COX2 SAR models computed for various values of the seed population percentage SPP

SPP	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 74 structural descriptors								
0.05	113	49	72	88	0.6975	0.4500	0.5745	0.1523
0.15	104	58	75	85	0.6420	0.4688	0.5559	0.1124
0.30	93	69	91	69	0.5741	0.5687	0.5714	0.1428
0.45	90	72	96	64	0.5556	0.6000	0.5776	0.1557
0.60	93	69	93	67	0.5741	0.5813	0.5776	0.1553
0.75	90	72	94	66	0.5556	0.5875	0.5714	0.1431
0.85	93	69	89	71	0.5741	0.5563	0.5652	0.1303
0.95	93	69	90	70	0.5741	0.5625	0.5683	0.1366
(b) 12 structural descriptors								
0.05	160	2	25	135	0.9877	0.1562	0.5745	0.2596
0.15	159	3	34	126	0.9815	0.2125	0.5994	0.3041
0.30	158	4	43	117	0.9753	0.2687	0.6242	0.3456
0.45	159	3	41	119	0.9815	0.2562	0.6211	0.3461
0.60	159	3	46	114	0.9815	0.2875	0.6366	0.3744
0.75	159	3	48	112	0.9815	0.3000	0.6429	0.3855
0.85	157	5	50	110	0.9691	0.3125	0.6429	0.3742
0.95	158	4	50	110	0.9753	0.3125	0.6460	0.3852

Drug Design with Artificial Intelligence Methods, Table 17

IMMUNOS-99 prediction statistics for BZR SAR models computed for various values of the seed population percentage SPP

SPP	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 75 structural descriptors								
0.05	48	34	47	34	0.5854	0.5802	0.5828	0.1656
0.15	35	47	68	13	0.4268	0.8395	0.6319	0.2922
0.25	39	43	67	14	0.4756	0.8272	0.6503	0.3232
0.35	35	47	70	11	0.4268	0.8642	0.6442	0.3233
0.50	36	46	69	12	0.4390	0.8519	0.6442	0.3191
0.65	34	48	69	12	0.4146	0.8519	0.6319	0.2960
0.75	36	46	71	10	0.4390	0.8765	0.6564	0.3506
0.95	36	46	70	11	0.4390	0.8642	0.6503	0.3347
(b) 16 structural descriptors								
0.05	57	25	39	42	0.6951	0.4815	0.5890	0.1808
0.15	55	27	51	30	0.6707	0.6296	0.6503	0.3006
0.25	54	28	51	30	0.6585	0.6296	0.6442	0.2883
0.35	55	27	49	32	0.6707	0.6049	0.6380	0.2763
0.50	53	29	49	32	0.6463	0.6049	0.6258	0.2515
0.65	51	31	51	30	0.6220	0.6296	0.6258	0.2516
0.75	50	32	51	30	0.6098	0.6296	0.6196	0.2394
0.95	50	32	51	30	0.6098	0.6296	0.6196	0.2394

cate that the IMMUNOS-99 is not very successful in discriminating thrombin inhibitors from non-inhibitors (Table 18a). In all 19 experiments that explore the influence of the SPP parameter, almost all chemical compounds are

Drug Design with Artificial Intelligence Methods, Table 18
IMMUNOS-99 prediction statistics for THR SAR models computed for various values of the seed population percentage SPP

SPP	TP _p	FN _p	TN _p	FP _p	Se _p	Sp _p	Ac _p	MCC _p
(a) 66 structural descriptors								
0.05	38	6	10	34	0.8636	0.2273	0.5455	0.1179
0.15	43	1	6	38	0.9773	0.1364	0.5568	0.2100
0.25	44	0	2	42	1.0000	0.0455	0.5227	0.1525
0.40	44	0	2	42	1.0000	0.0455	0.5227	0.1525
0.50	43	1	1	43	0.9773	0.0227	0.5000	0.0000
0.65	43	1	2	42	0.9773	0.0455	0.5114	0.0626
0.80	44	0	1	43	1.0000	0.0227	0.5114	0.1072
0.95	43	1	2	42	0.9773	0.0455	0.5114	0.0626
(b) 7 structural descriptors								
0.05	41	3	7	37	0.9318	0.1591	0.5455	0.1432
0.15	44	0	6	38	1.0000	0.1364	0.5682	0.2705
0.25	44	0	3	41	1.0000	0.0682	0.5341	0.1879
0.40	44	0	5	39	1.0000	0.1136	0.5568	0.2454
0.50	44	0	5	39	1.0000	0.1136	0.5568	0.2454
0.65	44	0	5	39	1.0000	0.1136	0.5568	0.2454
0.80	44	0	4	40	1.0000	0.0909	0.5455	0.2182
0.95	44	0	3	41	1.0000	0.0682	0.5341	0.1879

predicted in the class +1 (inhibitors). As a result, FN is small (which is good) but FP is very large (which is bad), and the overall statistics are low. A maximum is identified for SPP = 0.15 (Ac = 0.5568, MCC = 0.2100). Feature selection reduces the pool of descriptors to 7, and results in slightly better models (Table 18b). FP is still too large for the whole range of SPP values, which explains the low values for the statistical indices. Compared with the other three artificial immune systems, IMMUNOS-99 seems to be the most difficult to tune in order to obtain good predictions. Feature selection has no or small effect in improving IMMUNOS-99 models, which suggests that other algorithms should be investigated to reduce the pool of structural descriptors.

Future Directions

Pharmaceutical drug discovery uses computer-assisted molecular design to increase the chances of bringing a drug on the market, and to lower the research and development costs. Computational models are used to simulate the physical, chemical, biological, and toxicological properties of drug candidates, thus replacing expensive and time-consuming large scale experiments. The entire process consists of iterative steps, in which experimental results are used to train computational models, which in turn suggest novel molecules that are synthesized and

tested in the laboratory. We reviewed here the most important artificial intelligence algorithms used in drug design, namely genetic algorithms, ant colony optimization, particle swarm optimization, and artificial immune systems. The main advantage of artificial intelligence algorithms is their ability to explore search spaces of high dimensionality, and to identify the global optimum for complex and difficult problems. Genetic algorithms have a long history of applications in QSAR and drug design, and their operation is thoroughly explored. The other artificial intelligence algorithms were adopted only recently, but they already demonstrated strong results that make them competitors for GA. More important, ACO, PSO and AIS bring new simulation capabilities, thus complementing GA. A promising direction of development is a combined use of these artificial intelligence algorithms that could provide better predictions of molecular properties. Another source of improvement might come from the integration of the molecular graph into the artificial intelligence algorithms, which would complement (or even substitute) the use of structural descriptors.

Bibliography

- Holland J (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor
- Goldberg DE (1989) *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison Wesley, Reading
- Jones G (1998) Genetic and evolutionary algorithms. In: Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) *The Encyclopedia of Computational Chemistry*. Wiley, Chichester, pp 1127–1136
- Terfloth L, Gasteiger J (2001) Neural networks and genetic algorithms in drug design. *Drug Discov Today* 6:S102–S108
- von Homeyer A (2003) Evolutionary algorithms and their applications in chemistry. In: Gasteiger J (ed) *Handbook of Chemoinformatics*, vol 3. Wiley-VCH, Weinheim, pp 1239–1280
- Dorigo M, Maniezzo V, Colomi A (1996) Ant system: Optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern Part B Cybern* 26:29–41
- Dorigo M, Gambardella LM (1997) Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Trans Evol Comput* 1:53–66
- Dorigo M, Di Caro G, Gambardella LM (1999) Ant algorithms for discrete optimization. *Artif Life* 5:137–172
- Dorigo M, Stützle T (2004) *Ant Colony Optimization*. MIT Press, Cambridge
- Dorigo M, Blum C (2005) Ant colony optimization theory: A survey. *Theor Comput Sci* 344:243–278
- Kennedy J, Eberhart R (1995) Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks*, vol 4. pp 1942–1948
- Banks A, Vincent J, Anyakoha C (2007) A review of particle swarm optimization Part I: background and development. *Nat Comput* 6:467–484

13. Banks A, Vincent J, Anyakoha C (2008) A review of particle swarm optimization Part II: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications. *Nat Comput* 7:109–124
14. Chuang LY, Chang HW, Tu CJ, Yang CH (2008) Improved binary PSO for feature selection using gene expression data. *Comput Biol Chem* 32:29–38
15. Namasivayam V, Günther R (2007) PSO@AUTODOCK: A fast flexible molecular docking program based on swarm intelligence. *Chem Biol Drug Des* 70:475–484
16. Agrafiotis DK, Cedeño W (2002) Feature selection for structure-activity correlation using binary particle swarms. *J Med Chem* 45:1098–1107
17. Huang J, Ma G, Muhammad I, Cheng Y (2007) Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J Chem Inf Model* 47:1638–1647
18. Shen Q, Shi WM, Yang XP, Ye BX (2006) Particle swarm algorithm trained neural network for QSAR studies of inhibitors of platelet-derived growth factor receptor phosphorylation. *Eur J Pharm Sci* 28:369–376
19. Hunt JE, Cooke DE (1996) Learning using an artificial immune system. *J Netw Comput Appl* 19:189–212
20. de Castro LN, Von Zuben FJ (1999) Artificial immune systems: Part I Basic theory and applications. FEEC/UNICAMP, Brazil
21. de Castro LN, Von Zuben FJ (2000) Artificial immune systems: Part II A survey of applications. FEEC/UNICAMP, Brazil
22. Timmis J, Neal M, Hunt J (2000) An artificial immune system for data analysis. *Biosystems* 55:143–150
23. Chao DL, Forrest S (2003) Information immune systems. *Genet Programm Evolv Mach* 4:311–331
24. de Castro LN, Timmis JI (2003) Artificial immune systems as a novel soft computing paradigm. *Soft Comput* 7:526–544
25. Musilek P, Lau A, Reformat M, Wyard-Scott L (2006) Immune programming. *Inf Sci* 176:972–1002
26. Timmis J (2007) Artificial immune systems – Today and tomorrow. *Nat Comput* 6:1–18
27. Forrest S, Beauchemin C (2007) Computer immunology. *Immunol Rev* 216:176–197
28. Dasgupta D (1999) Artificial Immune Systems and Their Applications. Springer, Berlin
29. de Castro LN, Timmis J (2002) Artificial Immune Systems: A New Computational Intelligence Approach. Springer, Berlin
30. Tarakanov AO, Skormin VA, Sokolova SP (2003) Immunocomputing: Principles and Applications. Springer, Berlin
31. Ishida Y (2004) Immunity-Based Systems. Springer, Berlin
32. Timmis J, Bentley P, Hart E (2003) Artificial Immune Systems: Second International Conference, ICARIS 2003, Edinburgh, September 1–3. Lecture Notes in Computer Science, vol 2787. Springer, Berlin
33. Nicosia G, Cutello V, Bentley PJ, Timmis JI (2004) Artificial Immune Systems: Third International Conference, ICARIS 2004, Catania, September 13–16. Lecture Notes in Computer Science, vol 3239. Springer, Berlin
34. Jacob C, Pilat ML, Bentley PJ, Timmis J (2005) Artificial Immune Systems: 4th International Conference, ICARIS 2005, Banff, August 14–17. Lecture Notes in Computer Science, vol 3627. Springer, Berlin
35. Bersini H, Carneiro J (2006) Artificial Immune Systems: 5th International Conference, ICARIS 2006, Oeiras, September 4–6. Lecture Notes in Computer Science, vol 4163. Springer, Berlin
36. Ando S, Iba H (2004) Classification of gene expression profile using combinatory method of evolutionary computation and machine learning. *Genet Programm Evolv Mach* 5:145–156
37. Bezerra GB, Cançado GMA, Menossi M, de Castro LN, Von Zuben FJ (2005) Recent advances in gene expression data clustering: A case study with comparative results. *Genet Mol Res* 4:514–524
38. Tsankova D, Georgieva V, Kasabov N (2005) Artificial immune networks as a paradigm for classification and profiling of gene expression data. *J Comput Theor Nanosci* 2:543–550
39. Şahan S, Polat K, Kodaz H, Güneş S (2007) A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Comput Biol Med* 37:415–423
40. Polat K, Güneş S (2008) Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm. *Expert Syst Appl* 34:773–779
41. Polat K, Şahan S, Güneş S (2006) A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia. *Expert Syst Appl* 31:264–269
42. Latifoglu F, Şahan S, Kara S, Güneş S (2007) Diagnosis of atherosclerosis from carotid artery Doppler signals as a real-world medical application of artificial immune systems. *Expert Syst Appl* 33:786–793
43. Cutello V, Nicosia G, Pavone M, Timmis J (2007) An immune algorithm for protein structure prediction on lattice models. *IEEE Trans Evol Comput* 11:101–117
44. Anile AM, Cutello V, Narzisi G, Nicosia G, Spinella S (2007) Determination of protein structure and dynamics combining immune algorithms and pattern search methods. *Nat Comput* 6:55–72
45. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* 19:1639–1662
46. Wang R, Gao Y, Lai LH (2000) LigBuilder: A multi-purpose program for structure-based drug design. *J Mol Model* 6:498–516
47. So SS, Karplus M (1996) Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. *J Med Chem* 39:1521–1530
48. Venkatasubramanian V, Chan K, Caruthers JM (1995) Evolutionary design of molecules with desired properties using the genetic algorithm. *J Chem Inf Comput Sci* 35:188–195
49. Sundaram A, Venkatasubramanian V (1998) Parametric sensitivity and search-space characterization studies of genetic algorithms for computer-aided polymer design. *J Chem Inf Comput Sci* 38:1177–1191
50. Gillet VJ, Willett P, Bradshaw J, Green DVS (1999) Selecting combinatorial libraries to optimize diversity and physical properties. *J Chem Inf Comput Sci* 39:169–177
51. Ivanciuc O, Ivanciuc T, Cabrol-Bass D (2002) QSAR for dihydrofolate reductase inhibitors with molecular graph structural descriptors. *J Mol Struct (Theochem)* 582:39–51
52. Wegner JK, Fröhlich H, Zell A (2004) Feature selection for descriptor based classification models, 2. Human intestinal absorption (HIA). *J Chem Inf Comput Sci* 44:931–939
53. Weber L (1998) Evolutionary combinatorial chemistry: application of genetic algorithms. *Drug Discov Today* 3:379–385
54. Weber L (2005) Current status of virtual combinatorial library design. *QSAR Comb Sci* 24:809–823

55. Gallop MA, Barrett RW, Dower WJ, Fodor SPA, Gordon EM (1994) Applications of combinatorial technologies to drug discovery, 1. Background and peptide combinatorial libraries. *J Med Chem* 37:1233–1251
56. Gordon EM, Barrett RW, Dower WJ, Fodor SPA, Gallop MA (1994) Applications of combinatorial technologies to drug discovery, 2. Combinatorial organic-synthesis, library screening strategies, and future directions. *J Med Chem* 37:1385–1401
57. Weber L (1998) Applications of genetic algorithms in molecular diversity. *Curr Opin Chem Biol* 2:381–385
58. Illgen K, Enderle T, Broger C, Weber L (2000) Simulated molecular evolution in a full combinatorial library. *Chem Biol* 7:433–441
59. Ugi I, Almstetter M, Bock H, Dömling A, Ebert B, Gruber B, Hanusch-Kompa C, Heck S, Kehagia-Drikos K, Lorenz K, Papathoma S, Raditschnig R, Schmid T, Werner B, von Zychlinski A (1998) MCR XVII. Three types of MCRs and the libraries – Their chemistry of natural events and preparative chemistry. *Croat Chem Acta* 71:527–547
60. Weber L (2002) Multi-component reactions and evolutionary chemistry. *Drug Discov Today* 7:143–147
61. Globus A, Lawtonb J, Wipke T (1999) Automatic molecular design using evolutionary techniques. *Nanotechnology* 10:290–299
62. Brown N, McKay B, Gilardoni F, Gasteiger J (2004) A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J Chem Inf Comput Sci* 44:1079–1087
63. Brown N, McKay B, Gasteiger J (2006) A novel workflow for the inverse QSPR problem using multiobjective optimization. *J Comput Aided Mol Des* 20:333–341
64. Lewis RA (2005) A general method for exploiting QSAR models in lead optimization. *J Med Chem* 48:1638–1648
65. Izrailev S, Agrafiotis D (2001) A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J Chem Inf Comput Sci* 41:176–180
66. Izrailev S, Agrafiotis DK (2002) Variable selection for QSAR by artificial ant colony systems. *SAR QSAR Environ Res* 13:417–423
67. Shelokar PS, Jayaraman VK, Kulkarni BD (2004) An ant colony approach for clustering. *Anal Chim Acta* 509:187–195
68. He Y, Chen D, Zhao W (2006) Ensemble classifier system based on ant colony algorithm and its application in chemical pattern classification. *Chemom Intell Lab Syst* 80:39–49
69. Korb O, Stützel T, Exner TE (2006) PLANTS: Application of ant colony optimization to structure-based drug design. *Ant Colony Optimization and Swarm Intelligence. Proceedings. LNCS, vol 4150. Springer, Berlin*, pp 247–258
70. Shen Q, Jiang JH, Tao JC, Shen GL, Yu RQ (2005) Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. *J Chem Inf Model* 45:1024–1029
71. Gunturi SB, Narayanan R, Khandelwal A (2006) In silico ADME modelling 2: Computational models to predict human serum albumin binding affinity using ant colony systems. *Bioorg Med Chem* 14:4118–4129
72. Shi WM, Shen Q, Kong W, Ye BX (2007) QSAR analysis of tyrosine kinase inhibitor using modified ant colony optimization and multiple linear regression. *Eur J Med Chem* 42:81–86
73. Karpenko O, Shi J, Dai Y (2005) Prediction of MHC class II binders using the ant colony search strategy. *Artif Intell Med* 35:147–156
74. Hernandez P, Gras R, Frey J, Appel RD (2003) Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* 3:870–878
75. Shen Q, Shi WM, Kong W, Ye BX (2007) A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. *Talanta* 71:1679–1683
76. Call ST, Zubarev DY, Boldyrev AI (2007) Global minimum structure searches via particle swarm optimization. *J Comput Chem* 28:1177–1186
77. Chang BCH, Ratnaweera A, Halgamuge SK, Watson HC (2004) Particle swarm optimisation for protein motif discovery. *Genet Programm Evolv Mach* 5:203–214
78. Chen K, Li T, Cao T (2006) Tribe-PSO: A novel global optimization algorithm and its application in molecular docking. *Chemom Intell Lab Syst* 82:248–259
79. Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY (2007) SODOCK: Swarm optimization for highly flexible protein-ligand docking. *J Comput Chem* 28:612–623
80. Cedeño W, Agrafiotis DK (2003) Using particle swarms for the development of QSAR models based on K-nearest neighbor and kernel regression. *J Comput Aided Mol Des* 17:255–263
81. Lü JX, Shen Q, Jiang JH, Shen GL, Yu RQ (2004) QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. *J Pharm Biomed Anal* 35:679–687
82. Shen Q, Jiang JH, Jiao CX, Shen GL, Yu RQ (2004) Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *Eur J Pharm Sci* 22:145–152
83. Lin WQ, Jiang JH, Shen Q, Shen GL, Yu RQ (2005) Optimized block-wise variable combination by particle swarm optimization for partial least squares modeling in quantitative structure-activity relationship studies. *J Chem Inf Model* 45:486–493
84. Lin L, Lin WQ, Jiang JH, Shen GL, Yu RQ (2005) QSAR analysis of substituted bis[(acridine-4-carboxamide)propyl] methylamines using optimized block-wise variable combination by particle swarm optimization for partial least squares modeling. *Eur J Pharm Sci* 25:245–254
85. Hu L, Wu H, Lin W, Jiang J, Yu R (2007) Quantitative structure-activity relationship studies for the binding affinities of imidazobenzodiazepines for the $\alpha 6$ benzodiazepine receptor isoform utilizing optimized blockwise variable combination by particle swarm optimization for partial least squares modeling. *QSAR Comb Sci* 26:92–101
86. Shen Q, Jiang JH, Jiao CX, Huan SY, Shen GL, Yu RQ (2004) Optimized partition of minimum spanning tree for piecewise modeling by particle swarm algorithm. QSAR studies of antagonism of angiotensin II antagonists. *J Chem Inf Comput Sci* 44:2027–2031
87. Lin WQ, Jiang JH, Shen Q, Wu HL, Shen GL, Yu RQ (2005) Piecewise hypersphere modeling by particle swarm optimization in QSAR studies of bioactivities of chemical compounds. *J Chem Inf Model* 45:535–541
88. Lin L, Lin WQ, Jiang JH, Zhou YP, Shen GL, Yu RQ (2005) QSAR analysis of a series of 2-aryl(heteroaryl)-2,5-dihydropyrazolo[4,3-c]quinolin-3-(3H)-ones using piecewise

- hyper-sphere modeling by particle swarm optimization. *Anal Chim Acta* 552:42–49
89. Xu L, Jiang JH, Lin WQ, Zhou YP, Wu HL, Shen GL, Yu RQ (2007) Optimized sample-weighted partial least squares. *Talanta* 71:561–566
90. Shen Q, Jiang JH, Jiao CX, Lin WQ, Shen GL, Yu RQ (2004) Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. *J Comput Chem* 25:1726–1735
91. Zhou YP, Jiang JH, Lin WQ, Zou HY, Wu HL, Shen GL, Yu RQ (2006) Adaptive configuring of radial basis function network by hybrid particle swarm algorithm for QSAR studies of organic compounds. *J Chem Inf Model* 46:2494–2501
92. Zhou YP, Jiang JH, Lin WQ, Xu L, Wu HL, Shen GL, Yu RQ (2007) Artificial neural network-based transformation for nonlinear partial least-square regression with application to QSAR studies. *Talanta* 71:848–853
93. Meissner M, Schmuker M, Schneider G (2006) Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training. *BMC Bioinformatics* 7:125
94. Ivanciuc O (2007) Applications of support vector machines in chemistry. In: Lipkowitz KB, Cundari TR (eds) *Reviews in Computational Chemistry*, vol 23. Wiley-VCH, Weinheim, pp 291–400
95. Lin WQ, Jiang JH, Zhou YP, Wu HL, Shen GL, Yu RQ (2007) Support vector machine based training of multilayer feedforward neural networks as optimized by particle swarm algorithm: Application in QSAR studies of bioactivity of organic compounds. *J Comput Chem* 28:519–527
96. Tang LJ, Zhou YP, Jiang JH, Zou HY, Wu HL, Shen GL, Yu RQ (2007) Radial basis function network-based transform for a nonlinear support vector machine as optimized by a particle swarm optimization algorithm with application to QSAR studies. *J Chem Inf Model* 47:1438–1445
97. de Castro LN (2004) Dynamics of an artificial immune network. *J Exp Theor Artif Intell* 16:19–39
98. Bezerra GB, de Castro LN, Von Zuben FJ (2004) A hierarchical immune network applied to gene expression data. In: Nicosia G, Cutello V, Bentley PJ, Timmis JI (eds) *Artificial Immune Systems: Third International Conference, ICARIS 2004*. Catania, September 13–16. LNCS, vol 3239. Springer, Berlin, pp 14–27
99. Watkins A, Timmis J, Boggess L (2004) Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm. *Genet Programm Evol Mach* 5:291–317
100. Meng L, van der Putten P, Wang H (2005) A comprehensive benchmark of the artificial immune recognition system (AIRS). *Advanced Data Mining and Applications, Proceedings Lecture Notes in Artificial Intelligence*, vol 3584. pp 575–582
101. Watkins AB (2001) AIRS: A resource limited artificial immune classifier. Department of Computer Science, vol MS. Mississippi State University, pp 81
102. Watkins AB (2005) Exploiting immunological metaphors in the development of serial, parallel and distributed learning algorithms. Ph D, University of Kent, pp 314
103. de Castro LN, Von Zuben FJ (2000) The clonal selection algorithm with engineering applications. In: Whitley D, Goldberg D, Cantu-Paz E, Spector L, Parmee I, Beyer HG (eds) *GECCO-2000: Proceedings of the Genetic and Evolutionary Computation Conference*, July 10–12. Las Vegas, Morgan Kaufmann, pp 36–37
104. de Castro LN, Von Zuben FJ (2002) Learning and optimization using the clonal selection principle. *IEEE Trans Evol Comput* 6:239–251
105. Brownlee J (2005) Clonal selection theory & CLONAG. The clonal selection classification algorithm (CSCA). Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT). Victoria
106. Carter JH (2000) The immune system as a model for pattern recognition and classification. *J Am Med Inf Assoc* 7:28–41
107. Brownlee J (2005) Immunos-81. The misunderstood artificial immune system. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT). Victoria
108. Witten IH, Frank E (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco, pp 525
109. Brownlee J (2005) Artificial immune recognition system (AIRS). A review and analysis. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT). Victoria
110. Watkins A, Timmis J (2002) Artificial immune recognition system (AIRS): Revisions and refinements. *Artificial Immune Systems: First International Conference, ICARIS 2002*. University of Kent at Canterbury, pp 173–181
111. Fenichel RR, Malik M, Antzelevitch C, Sanguinetti M, Roden DM, Priori SG, Ruskin JN, Lipicky RJ, Cantilena LR (2004) Drug-induced torsades de pointes and implications for drug development. *J Cardiovasc Electrophysiol* 15:475–495
112. Ivanciuc O (2006) Artificial immune system classification of drug-induced torsade de pointes with AIRS (artificial immune recognition system). *Internet Electron J Mol Des* 5:488–502
113. Ivanciuc O (2007) Artificial immune systems in drug design: Structure-activity relationships for torsade de pointes with AIRS (artificial immune recognition system). *Internet Electron J Mol Des* 6:47–62
114. Stenberg P, Luthman K, Artursson P (2000) Virtual screening of intestinal drug permeability. *J Control Release* 65:231–243
115. Ponce YM, Pérez MAC, Zaldivar VR, Sanz MB, Mota DS, Torrens F (2005) Prediction of intestinal epithelial transport of drug in (Caco-2) cell culture from molecular structure using in silico approaches during early drug discovery. *Internet Electron J Mol Des* 4:124–150
116. Linnankoski J, Makela JM, Ranta VP, Urtti A, Yliperttula M (2006) Computational prediction of oral drug absorption based on absorption rate constants in humans. *J Med Chem* 49:3674–3681
117. Iyer M, Tseng YJ, Senese CL, Liu J, Hopfinger AJ (2007) Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. *Mol Pharmaceutics* 4:218–231
118. Ivanciuc O (2006) Artificial immune system prediction of the human intestinal absorption of drugs with AIRS (artificial immune recognition system). *Internet Electron J Mol Des* 5:515–529
119. Ivanciuc O (2007) Feature Selection in AIRS (Artificial Immune Recognition System) Structure-Activity Relationships. *Internet Electron J Mol Des* 6:331–344

120. Crivori P, Reinach B, Pezzetta D, Poggesi I (2006) Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Mol Pharmaceutics* 3:33–44
121. Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, Gasteiger J (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. *J Med Chem* 50:1698–1702
122. Ivanciuc O (2006) Artificial immune systems in drug design: Recognition of P-glycoprotein substrates with AIRS (artificial immune recognition system). *Internet Electron J Mol Des* 5:542–554
123. Ivanciuc O (2006) Structure-activity relationships with artificial immune systems: Classification of benzodiazepine receptor ligands with AIRS, CLONALG, CSCA, and IMMUNOS. *Internet Electron J Mol Des* 5:585–604
124. Verhaar HJM, Solbé J, Speksnijder J, van Leeuwen CJ, Hermens JLM (2000) Classifying environmental pollutants: Part 3. External validation of the classification system. *Chemosphere* 40:875–883
125. Ivanciuc O (2003) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. *Internet Electron J Mol Des* 2:195–208
126. Ivanciuc O (2007) Artificial immune systems in aquatic toxicology: Structure-activity relationships for the mechanism of toxic action with AIRS (artificial immune recognition system). *Internet Electron J Mol Des* 6:13–28
127. Ivanciuc O (2007) Drug design with artificial immune systems: Structure-activity relationships for glycogen phosphorylase B inhibitors with CLONALG (clonal selection algorithm). *Internet Electron J Mol Des* 6:311–319
128. Ivanciuc O (2007) Structure-activity relationships in aquatic toxicology with artificial immune systems: Mechanism of toxic action classification of polar and nonpolar narcotic pollutants with CLONALG (clonal selection algorithm). *Internet Electron J Mol Des* 6:106–114
129. Ivanciuc O (2007) Artificial immune systems structure-activity relationships for estrogen receptor ligands with CSCA (clonal selection classification system). *Internet Electron J Mol Des* 6:81–89
130. Ivanciuc O (2007) Artificial immune systems in the virtual screening of dihydrofolate reductase inhibitors with CSCA (clonal selection classification system). *Internet Electron J Mol Des* 6:253–261
131. Ivanciuc O (2007) Drug design with artificial immune systems: Classification of angiotensin converting enzyme inhibitors with CSCA (clonal selection classification system). *Internet Electron J Mol Des* 6:135–143
132. Ivanciuc O (2007) Artificial immune systems in structure-activity relationships: Classification of thermolysin inhibitors with CSCA (clonal selection classification system). *Internet Electron J Mol Des* 6:209–217
133. Ivanciuc O (2007) Structure-activity relationships for acetylcholinesterase inhibitors with the IMMUNOS artificial immune system. *Internet Electron J Mol Des* 6:167–175
134. Ivanciuc O (2007) Virtual screening of cyclooxygenase-2 inhibitors with the IMMUNOS artificial immune system. *Internet Electron J Mol Des* 6:200–208
135. Ivanciuc O (2007) Structure-activity relationships with the IMMUNOS artificial immune system for thrombin inhibitors. *Internet Electron J Mol Des* 6:262–270

Drug Design with Artificial Neural Networks

OVIDIU IVANCIUC

Department of Biochemistry and Molecular Biology,
University of Texas Medical Branch, Galveston, USA

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Perceptron](#)
[Multilayer Feedforward Artificial Neural Network](#)
[Radial Basis Function Network](#)
[Self-organizing Map](#)
[Counterpropagation Neural Network](#)
[Graph Machine Neural Networks](#)
[Other Neural Networks](#)
[Future Directions](#)
[Bibliography](#)

Glossary

Artificial neuron An artificial neuron is a mathematical function that simulates in a simplified form the functions of biological neurons. Usually, an artificial neuron has four computational functions, namely receives signals through input connections from other neurons or from the environment, sums the input signals, applies a nonlinear functions (transfer function or activation function) to the sum, and sends the result to other neurons or as output from the neural network.

Counterpropagation neural network The counterpropagation neural network is a hybrid network that consists of a self-organizing map as the hidden layer and an output layer that has as output a computed value for the modeled property. The network implements a supervised learning algorithm that converges to a unique solution.

Multilayer feedforward artificial neural network

A multilayer feedforward (MLF) artificial neural network consists of artificial neurons organized in layers. The MLF network has an input layer that receives the structural descriptors for each molecule, an output layer that provides one or more computed properties, and one or more hidden layers situated between the input and the output layers. Each neuron in a hidden layer receives signals from neurons in the preceding layer and sends signals to the neurons in the next layer.

Perceptron A perceptron is a linear classifier that consists of a layer of input neurons and an output neuron. Each connection between an input neuron and the output neuron has a weight. Depending on the sum of the signals received by the output neuron, its output is +1 or -1.

Quantitative structure-activity relationships

Quantitative structure-activity relationships (QSAR) represent regression models that define quantitative correlations between the chemical structure of molecules and their physical properties (boiling point, melting point, aqueous solubility), chemical properties and reactivities (chromatographic retention, reaction rate), or biological activities (cell growth inhibition, enzyme inhibition, lethal dose). The fundamental hypotheses of QSAR are that similar chemicals have similar properties, and that small structural changes result in small changes in property values. The general form of a QSAR equation is $P(i) = f(\mathbf{SD}_i)$, where $P(i)$ is a physical, chemical, or biological property of compound i , \mathbf{SD}_i is a vector of structural descriptors of i , and f is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. A QSAR model for a property P is based on a dataset of chemical compounds with known values for the property P , and a matrix of structural descriptors computed for all chemicals. The learning (training) of the QSAR model is the process of determining the optimum parameters of the regression function f . After the training phase, a QSAR model may be used to predict the property P for novel compounds that are not present in the learning set of molecules.

Radial basis function network The radial basis function (RBF) neural network has three layers, namely an input layer, a hidden layer with a nonlinear RBF activation function and a linear output layer.

Self-organizing map A self-organizing map (SOM) is an artificial neural network that uses an unsupervised learning algorithm to project a high dimensional input space into a two dimensional space called a map. The topology of the input space is preserved in SOM, and points that are close to each other in the SOM grid correspond to input vectors that are close to each other in the input space. A SOM consists of neurons arranged usually in a rectangular or hexagonal grid. Each neuron has a position on the map and a weight vector of the same dimension as the input vectors.

Structural descriptor A structural descriptor (SD) is a numerical value computed from the chemical structure of a molecule, which is invariant to the number-

ing of the atoms in the molecule. Structural descriptors may be classified as constitutional (counts of molecular fragments, such as rings, functional groups, or atom pairs), topological indices (computed from the molecular graph), geometrical (volume, surface, charged-surface), quantum (atomic charges, energies of molecular orbitals), and molecular field (such as those used in CoMFA, CoMSIA, or CoRSA).

Structure-activity relationships Structure-activity relationships (SAR) represent classification models that can discriminate between sets of chemicals that belong to different classes of biological activities, usually active/inactive towards a certain biological receptor. The general form of a SAR equation is $C(i) = f(\mathbf{SD}_i)$, where $C(i)$ is the activity class of compound i (active/inactive, inhibitor/non-inhibitor, ligand/non-ligand), \mathbf{SD}_i is a vector of structural descriptors of i , and f is a classification function such as k -nearest neighbors, linear discriminant analysis, random trees, random forests, Bayesian networks, artificial neural networks, or support vector machines.

Definition of the Subject

The fundamental hypothesis of the structure-property models is that the structural features of molecules determine the physical, chemical and biological properties of chemical compounds. The first studies that use structure-activity relationships to explain the biological properties of sets of compounds were published by Kopp [74], Crum-Brown and Frazer [18], Meyer [88], and Overton [97]. Modern structure-activity relationships (SAR) and quantitative structure-activity relationships (QSAR) models are based on the Hansch model that predicts a biological property as a statistical correlation with steric, electronic, and hydrophobic indices [27,35,36,37]. The Hansch model shaped the general scene of structure-activity correlations, and almost all subsequent SAR and QSAR models are variations that extend the Hansch model with novel classes of descriptors or with more powerful statistical models, such as partial least squares (PLS), artificial neural networks (ANN), support vector machines (SVM), or other machine learning algorithm. A structural descriptor is a numerical representation of some important molecular features, such as empirical indices (Hammett and Taft substituent constants), physical properties (octanol-water partition coefficient, dipole moment, aqueous solubility), counts of substructures or substituents, graph descriptors [13,54,125], topological indices [4,12,55], connectivity indices [65,66], electrotopological indices [67], geometrical descriptors (molecular surface and volume),

quantum indices (atomic charges, HOMO and LUMO energies) [60,124], and molecular fields (steric, electrostatic, and hydrophobic). SAR represent classification models that are used when the experimental property is a class label (+1/−1), such as substrate/non-substrate, inhibitor/non-inhibitor, ligand/non-ligand, toxic/non-toxic, or carcinogen/non-carcinogen. Classification models are used to screen chemical libraries and to identify compounds that have a desired biological activity, such as inhibitor for a particular enzyme. QSAR represent regression models that are used for experimental properties with continuous values, such as hydrophobicity, aqueous solubility, membrane penetration, lethal concentration, or inhibition constant for an enzyme.

The drug discovery and development process is lengthy, expensive, and has a high attrition rate. The duration of the drug development phases, namely pre-clinical (1 to 5 years), clinical (5 to 11 years), and approval (0.5 to 2 years) [9], puts at 18 years the upper limit for bringing a drug to market, without considering the duration of the drug discovery phase. For 168 drugs approved between 1992 and 2002, the median clinical trial and approval periods were 5.1 and 1.2 years, respectively, with no tendency of decreasing despite exceptional technological advances [63]. Cost data for 68 randomly selected drugs from 10 pharmaceutical companies show that better discovery and screening programs can reduce the total cost per approved drug by up to \$US 242 million [20]. A similar analysis shows that the average out-of-pocket cost per new drug is \$US 403 million (2000 dollars), which increases to \$US 802 after adding the opportunity cost (the cost of pursuing one choice instead of another) [21]. Depending on the therapy or the developing firm, the cost per new drug varies between \$US 500 million to \$US 2,000 million [1]. The attrition rate of the chemical compounds is also impressive, considering that the process starts with millions of compounds tested in high throughput screening and eventually ends with one successful drug on the market. From all compounds that enter the clinical trials, 30% fail due to lack of efficacy, 30% fail in toxicological and safety tests, and only 11% finish the trials [73]. A study considering all 548 new chemical entities approved between 1975 and 1999 found that 45 drugs (8.2%) received one or more black box warnings and 16 (2.9%) were withdrawn from the market [76]. Examples of drugs withdrawn from the market due to adverse reactions include troglitazone (Rezulin) in 2000 [25], cerivastatin (Baycol, Lipobay) in 2001 [28], nefazodone (Serzone) in 2003 [17], pemoline (Cylert) in 1999 (Canada) and 2005 (USA) [42], and rofecoxib (Vioxx) in 2004 [22]. For all difficulties, expenses, and failures that are associated with the drug develop-

ment process, the reward for a successful drug can be substantial. The top 2006 best selling drugs (in \$US billions) are lipitor with 14.39 (Pfizer, cholesterol), advair with 6.13 (GlaxoSmithKline, asthma), plavix with 6.06 (Bristol-Myers Squibb, vascular disease), nexium with 5.18 (AstraZeneca, acid reflux), norvasc with 4.87 (Pfizer, hypertension).

Computer-assisted drug design (CADD) uses computational chemistry to increase the chances of finding valuable drug candidates. Among the broad range of computational tools used in CADD, artificial neural networks (ANN) have a special place due to their abilities to model with high accuracies the complex relationships between chemical structures and molecular properties. ANN can be used both for classification (SAR) and regression (QSAR), and they are essential tools in drug discovery cycles to identify active compounds in chemical libraries, and to optimize a wide range of physico-chemical and biological properties, such as enzyme inhibition, target selectivity, or membrane transport. The SAR and QSAR models based on ANN may predict with high accuracy properties for novel chemicals, even before their chemical synthesis. CADD systems based on ANN may increase the chances of bringing a drug to market, when they are integrated into the chemical, biological, pharmacological and clinical procedures used by the pharmaceutical industry.

Introduction

The simulation of brain mechanisms with artificial neural networks was initiated by McCulloch and Pitts [84,101] and by Rosenblatt [106]. Minsky and Papert investigated the mathematical properties of these simple neural networks in their seminal book *Perceptrons* [92]. Although perceptrons are able to solve some problems, they fail to learn the XOR (exclusive-or) operation. As a result, the research in this field was limited until Hopfield demonstrated that by adding nonlinear functions to each artificial neuron the network could simulate complex behavior [43].

Another milestone in ANN research was the discovery of the backpropagation algorithm for the training of multilayer feedforward artificial neural networks [107,108]. Multilayer feedforward (MLF) networks consist of multiple layers of artificial neurons, all connected in a feedforward way, from the input layer to the output layer. MLF networks have at least one hidden layer of neurons, and the neurons in the hidden and output layers have nonlinear activation functions. The hidden neurons and the nonlinear activation functions are essential in modeling nonlinear relationships between input variables and the output

property. MLF networks represent the most widespread ANN type used to model physico-chemical, biological, and toxicological properties of chemical compounds.

In addition to MLF, other networks used in drug design are radial basis function networks, self-organizing maps, counter-propagation networks, graph machine neural networks, and probabilistic neural networks. Several reviews [49,52,144] and books are recommended for a detailed overview of neural networks and their applications: *Neural Computing* by Wasserman [134], *Neural Networks for Pattern Recognition* by Bishop [10], *Pattern Recognition and Neural Networks* by Ripley [104], *Neural Networks for Chemical Engineers* by Bulsari [15], and *Neural Networks in Chemistry and Drug Design* by Zupan and Gasteiger [145].

The artificial neuron and the perceptron are described in Sect. “[Perceptron](#)”. MLF networks are used to model physico-chemical properties, such as melting temperature [75], boiling temperature [33], aqueous solubility [46], and *n*-octanol/water partition coefficients [123]. Drug design applications of MLF networks include models for enzyme inhibition [120], human intestinal absorption [136], virtual screening of chemical libraries [102], and genotoxicity [127]. Another important application of MLF networks is in cancer diagnosis from gene expression data [70,119,135]. Section “[Multilayer Feedforward Artificial Neural Network](#)” is an overview of the most important MLF network models in structure-property prediction and in drug design.

The radial basis function (RBF) neural network belongs to the class of feedforward artificial neural network, and it has three layers, namely an input layer, a hidden layer with a nonlinear RBF activation function and a linear output layer [93]. RBF networks give predictive models for various properties, such as boiling temperature [81], QSAR for enzyme inhibitors [99,100], ligand-target binding affinity [142], and aquatic toxicity [87]. SAR and QSAR applications of the RBF network are presented in Sect. “[Radial Basis Function Network](#)”.

The self-organizing map (SOM) was developed by Kohonen as an unsupervised neural network that transforms a high dimensional input space into a two dimensional space [71,72]. SOM maps the objects by trying to preserve the topological properties of the input space. To predict the class of a new molecule, a vector of structural descriptors is sent as input for a trained SOM and projected onto a neuron from the map. The molecule is classified in the same class with the training molecules that are projected onto the same neuron. SOM networks cluster chemical compounds according to their similarity, and thus can be used to identify groups of molecules with similar proper-

ties [3]. Applications of SOM are found in virtual screening of chemical libraries [105], in modeling the aquatic toxicity of organic chemicals [98], or in structure-activity studies [11]. These and other SOM applications are examined in Sect. “[Self-organizing Map](#)”.

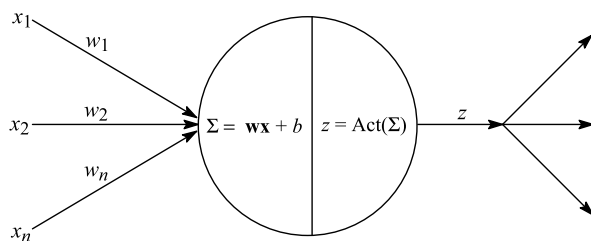
The counterpropagation neural network (CPNN) was developed by Hecht-Nielsen by transforming a self-organizing map into a supervised learning neural network [39,40,41]. The input layer of neurons is connected to the hidden layer that is a self-organizing map, and the excited neuron from SOM sends a signal to the output layer. CPNN may be used for classification or for regression. SAR applications of CPNN include the mutagenicity prediction of aromatic amines [59], and the classification of organic compounds based on their mode of toxic action [117]. Applications of the counterpropagation neural network in SAR and QSAR are presented in Sect. “[Counterpropagation Neural Network](#)”.

Graph machine neural networks learn SAR and QSAR models directly from the molecular graph [32], by translating the chemical structure into the network topology. For each chemical compound presented to the graph machine, the neural network adopts a structure derived from the molecular graph of the chemical. There are several graph machines used with success in SAR and QSAR, such as recursive neural networks [5,89], the Baskin-Palyulin-Zefirov (BPZ) neural device [7], ChemNet [69], and MolNet [47,50]. The BPZ neural device is constructed by analogy with a biological vision system and contains a sensor field, a set of eyes, and a brain. A chemical structure investigated with the BPZ neural device is represented as a molecular matrix that is superimposed over a sensor field that receives the network input. The information received by the sensor field is transmitted to a set of eyes that transforms it into several MLF networks, and sends signals to the next block, the brain. ChemNet encodes a molecular distance matrix between the input and hidden layers, and has as output an atomic property. In MolNet each atom in a molecule has a corresponding input and hidden neuron, and the connections between the input and hidden layers correspond to the weighted interactions between atoms. The output neuron offers the computed value for a molecular property. In Sect. “[Graph Machine Neural Networks](#)” we review structure-activity models computed with graph machine neural networks.

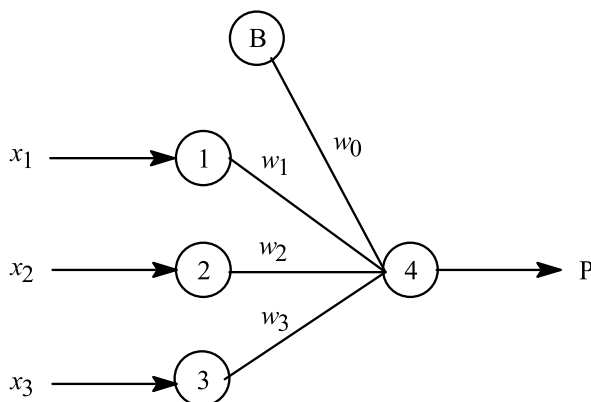
Other ANN used in drug design and molecular property modeling are the Bayesian regularized neural network [31,133], the probabilistic neural network [38,77], and the general regression neural network [141]. Their applications in SAR and QSAR are presented in Sect. “[Other Neural Networks](#)”.

Perceptron

An artificial neuron is a simple model of a biological neuron expressed as a mathematical function. Artificial neurons represent the basic computational units in an artificial neural network. Each artificial neuron has four basic functions (Fig. 1): (a) receives input signals through connections from the neurons in the previous layer or from the environment; (b) makes a summation of the inputs signals; (c) the result of summation is passed to the activation function and transformed with a mathematical function; (d) the result of the activation function is used by a transfer function to produce an output value that is sent to the neurons in the next layer of neurons. The artificial neuron depicted in Fig. 1 receives the input signals x_1, \dots, x_n through connections with the weights w_1, \dots, w_n . The summation of the input signal is computed as the dot product between \mathbf{w} and \mathbf{x} plus a bias parameter b . An activation function Act is applied to the sum Σ to give the output value z . The first artificial neuron was the threshold logic unit proposed by McCulloch and Pitts [84,101]. The activation function was a step function, which has the value zero for a negative argument and one for a positive argument. Other important activation functions are: lin-



Drug Design with Artificial Neural Networks, Figure 1
Structure and functions of an artificial neuron



Drug Design with Artificial Neural Networks, Figure 2
Perceptron structure and signal flow

ear, $\text{Act}(z) = z$; sigmoid, $\text{Act}(z) = 1/(1 + e^{-z})$ [43]; hyperbolic tangent, $\text{Act}(z) = \tanh(z)$; symmetric logarithmoid, $\text{Act}(z) = \text{sign}(z) \ln(1 + |z|)$ [15].

A perceptron is the simplest type of ANN, consisting of a layer of input neurons and an output neuron (Fig. 2) [92,106]. The perceptron shown in Fig. 2 has three input neurons (1, 2, 3), a bias neuron B, and an output neuron 4. An input neuron receives an input signal, scales it and then sends it to the output neuron. A simple perceptron is a linear classifier, which cannot learn simple operations, such as the XOR (exclusive-or) operation. A major development was the introduction of nonlinear activation functions that allows ANN to simulate any nonlinear function [43].

Multilayer Feedforward Artificial Neural Network

A multilayer feedforward artificial neural network, called also a multilayer perceptron (MLP) is a generalization of the simple perceptron obtained by inserting one or more hidden layers of neurons between the input layer and the output layer. Another important characteristic is the use of nonlinear activation functions in the hidden and output layers, although the output neurons can have also a linear activation function. MLF ANN are universal approximators, namely they can approximate any function with arbitrary precision provided that enough neurons are used in the hidden layer.

In SAR and QSAR applications the structural descriptors represent the input data received by the input neurons. As an example, we consider topological descriptors computed from the molecular graph of *t*-butylcyclopropane (Fig. 3a) that are used as input in an MLF ANN with one hidden layer (Fig. 3b). The input layer is formed by neurons 1–4, the hidden layer contains neurons 5–7, and the output layer is represented by neuron 8. The bias neurons B send a signal with the value +1 through weighted connections, and have the role to fine-tune each neuron. The topological descriptors represent the counts of vertex degrees, namely $\text{Deg} = \{3, 2, 1, 1\}$, i. e., there are three atoms with degree 1, two atoms with degree 2, one atom with degree 3, and one atom with degree 4. These numbers are used as input to neurons 1–4, which perform a scaling and then send the values to each neuron from the hidden layer. The output neuron 8 offers the computed value of a property *P* for *t*-butylcyclopropane.

The MLF ANN application to property modeling consists of two phases, namely training and prediction. During the training (learning) phase the weights of all connections are optimized (adjusted) in order to estimate with high precision the investigated molecular property.

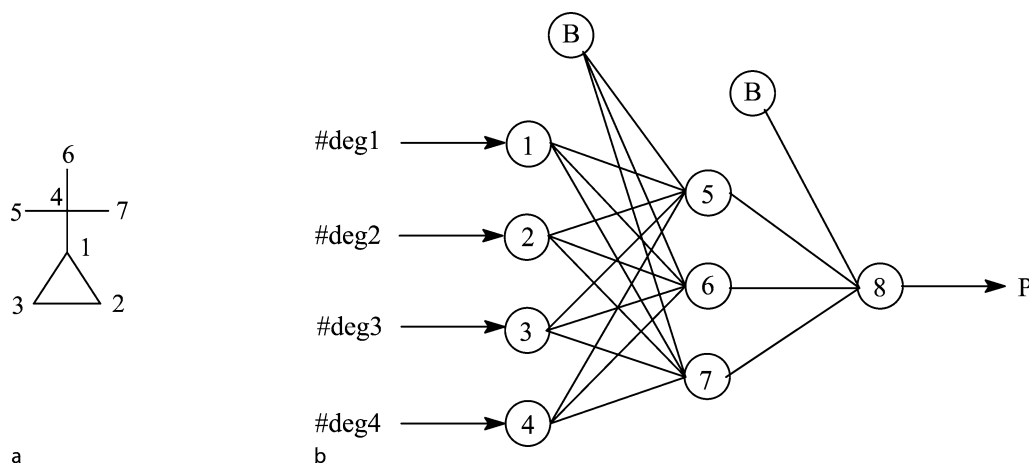
The most popular training method is the backpropagation algorithm [107,108], a variant of the gradient algorithms. More efficient training algorithms are direction-set methods (Powell's method), methods that require the computation of the first derivatives such as conjugate gradient methods (Fletcher-Reeves or Polak-Ribiere), Levenberg-Marquardt, or quasi-Newton (variable metric) methods (Davidon-Fletcher-Powell or Broyden-Fletcher-Goldfarb-Shanno).

MLF ANN can be trained also with artificial intelligence algorithms, such as particle swarm optimization, which was used in a QSAR study of inhibitors of platelet-derived growth factor receptor phosphorylation [113]. Such algorithms can provide a complete optimization of an MLF ANN, namely they can optimize both the network topology and the connection weights. The network topology is defined by the number of hidden layers and by the number of neurons in each hidden layer. The complete optimization of ANN with particle swarm optimization has a fast convergence to the global minimum [85] and avoids the overfitting of the learning dataset of chemicals [112]. The selection of the most important descriptors for the ANN input is an important step in obtaining predictive models. Several artificial intelligence algorithms can be used for an efficient feature selection, such as genetic algorithm [115,116], particle swarm optimization [2], and ant colony optimization [58].

Taskinen and Yliruusi reviewed the MLF ANN applications in predicting several physico-chemical properties of interest in drug development, such as octanol-water partition coefficient, aqueous solubility, boiling temperature, and vapor pressure [121]. The melting temperatures

for sulfur-containing organic compounds were predicted from simple molecular graph indices [75]. The melting temperatures of 717 ionic liquids were modeled with MLF ANN, associative neural networks, support vector machines, *k*-nearest neighbors, and multiple linear regression [126]. The structural descriptors evaluated in these models were electrotopological indices [67] and Dragon descriptors [124]. The moderate accuracy of predictions obtained with all algorithms can be explained by the difficulty to encode the solid state characteristics of ionic liquids. Hall and Story used electrotopological indices and an MLF ANN to model the boiling temperatures of 298 compounds and the critical temperatures of 165 compounds [33]. The best results for the boiling temperatures were obtained with a neural network with five hidden neurons with a mean absolute error of 3.93 K, whereas a network with four hidden neurons gave good predictions for the critical temperatures with a mean absolute error of 4.52 K. In a similar study, atom type electrotopological state indices were used to model the boiling temperatures of alkanes, alcohols and chloroalkanes [34].

Huuskonen et al. developed MLF ANN models for the aqueous solubility of a diverse set of 734 organic compounds [46]. Atom type electrotopological state indices were used as input data in a neural network with five hidden neurons. Good results were obtained both in training ($s = 0.52$) and prediction ($s = 0.75$), showing that the aqueous solubility for a large and diverse set of organic compounds can be predicted from topological indices. Tetko et al. developed a neural network model for the octanol-water partition coefficients of 12,908 organic compounds [123]. The model consists of an ensemble of



Drug Design with Artificial Neural Networks, Figure 3

ANN prediction of molecular properties from the molecular structure: topological descriptors computed from a the molecular graph of *t*-butylcyclopropane are used as input in b a multilayer feedforward neural network with one hidden layer

50 ANN, each ANN having 75 input neurons and 10 neurons in the hidden layer. The leave-one-out cross-validation indicate that the model gives good predictions, with $r^2 = 0.95$. The ^{13}C NMR chemical shift can be predicted from simple topological parameters, as shown for a dataset of acyclic alkenes [56,57].

To compensate for the lack of toxicological data for aquatic organisms, a number of neural networks QSAR models were proposed for the toxicity of organic chemicals against fathead minnow, rainbow trout, bluegill sunfish, *Daphnia magna*, *Tetrahymena pyriformis*, and *Vibrio fischeri* [61]. These ANN models can be used to predict the environmental impact of novel chemicals, and to prioritize their experimental evaluation in aquatic toxicity assays. Huuskonen tested multiple linear regression and artificial neural networks in QSAR models for toxicity of 140 organic chemicals against fathead minnow [45]. A neural network trained with 14 electrotopological state indices provided a predictive model for the experimental toxicity. Basak et al. modeled the toxic modes of action for 283 chemicals with ANN and topological indices [6]. The toxic compounds were classified as narcotics, electrophiles/proelectrophiles, uncouplers of oxidative phosphorylation, acetylcholinesterase inhibitors, and neurotoxicants. The modes of action SAR models provide reliable predictions, with rates of correct classification between 65% and 95%. Lower predictions are obtained for classes with a low number of compounds.

Dopamine antagonists, serotonin antagonists, and serotonin-dopamine dual antagonists are used as antipsychotics. Based on a training dataset consisting of 1135 dopamine antagonists, 1251 serotonin antagonists, and 386 serotonin-dopamine dual antagonists, Kim et al. tested several machine learning algorithms, including ANN and recursive partitioning [68]. The average classification rate in training was 73.6% and in prediction was 69.8%, indicating that these models can be used in virtual screening to identify new active compounds. Siu and Che used multiple linear regression, partial least squares, and ANN to model the α -amino acids affinities for H^+ and for sodium, copper, and silver cations [114]. The cross-validation tests indicate that the best predictions are obtained with the artificial neural network. These QSAR models are useful in elucidating the binding properties of the α -amino acids.

Votano et al. used ANN, k -nearest neighbors, and decision forest to model the mutagenicity of 3363 diverse compounds tested for their Ames genotoxicity [127]. The dataset was split into 2963 training compounds and 400 prediction compounds, and the SAR models were developed with 148 topological indices that included

electrotopological state indices and molecular connectivity indices. All three classifiers gave good predictions, with a slight advantage for the neural network, as indicated by the area under the receiver operator characteristic (AUROC) curve, namely AUROC = 0.93 for ANN, AUROC = 0.92 for k -nearest neighbors, and AUROC = 0.91 for decision forest.

A good intestinal absorption is a desirable property for drugs, but its experimental determination is costly and time-consuming. Wessel et al. used ANN to estimate the percent human intestinal absorption (% HIA) of 86 drugs [136]. A neural network optimized with a genetic algorithm provided the best results, with an error of 9.4% HIA units for training, 19.7% HIA units the cross-validation, and 16.0% HIA units for an external prediction set.

Sutherland et al. evaluated ANN, PLS, genetic function approximation, and genetic PLS in QSAR models for inhibitors of angiotensin converting enzyme, acetylcholinesterase, benzodiazepine receptor, cyclooxygenase-2, dihydrofolate reductase, glycogen phosphorylase B, thermolysin, and thrombin [120]. ANN gave the best predictions for QSAR models obtained with graph indices and geometrical descriptors. Another large scale evaluation of structure-activity models was made for chemicals that target HIV-reverse transcriptase, COX2, dihydrofolate reductase, estrogen receptor, and thrombin [102]. A comparison of several nonlinear modeling algorithms shows that ANN are successful in identifying active compounds from chemical libraries and can be used in virtual screening. Pharmaceutical companies constantly enrich their collections of chemical compounds by acquisitions that increase the structural diversity of molecules available for high throughput screening. Muresan and Sadowski proposed an ANN system to compute an "in-house likeness" score for compound acquisition [94]. The analysis of several datasets shows that a set of atom-type counts used as input to the ANN models represents an efficient way of identifying structural patterns that are missing from an in-house collection.

Votano et al. compared multiple linear regression, ANN, k -nearest neighbors, and support vector machines in QSAR models for the human serum protein binding of 1008 chemicals (808 for training and 200 for prediction) [128]. The best predictions ($r^2 = 0.7$) were obtained with ANN models, indicating that these QSAR can assist the drug design process. Katritzky et al. evaluated ANN and multiple linear regression in QSAR models for 277 inhibitors of glycogen synthase kinase-3 [62]. The descriptors were selected from a pool consisting of subgraph counts, topological indices, geometrical parameters and quantum indices. The QSAR mod-

els highlight the structural factors that influence the inhibition of glycogen synthase kinase-3. Inhibitors of the hERG (human ether-á-go-go-related gene) potassium channel can lead to a prolongation of the QT interval that can trigger torsade de pointes, an atypical ventricular tachycardia. Seierstad and Agrafiotis used ANN to model the nonlinear structure-activity relationship for hERG channel inhibitors [110]. The QSAR models with best predictive power are obtained with a feature selection to prevent over-fitting and by aggregating several ANN into an ensemble model to minimize the instability in predicting novel chemicals. The time- and dose-dependent anti-inflammatory *in vivo* activity of substituted imidazo[1,2-*a*]pyridines was modeled with artificial neural networks [109]. This study shows that ANN have unique properties to study pharmacodynamic and pharmacokinetic properties of drugs and drug-like compounds. Peptides that have a good binding affinity for major histocompatibility complex (MHC) class I molecules are candidates in the development of new vaccines against cancer and viral infections. Filter et al. modeled the peptide binding to MHC class I allele HLA-A*0201 with artificial neural networks [26]. The ANN system can identify peptides that do not correspond to the usual recognition motifs, as shown for several new melanoma-associated peptides.

Artificial neural networks represent the method of choice in cancer prediction and prognosis [19]. Their use improves the accuracy of predicting cancer susceptibility, recurrence and mortality. Molecular data, such as protein biomarkers and microarrays, together with ANN models can improve the understanding of cancer development and progression. The mechanism of action of drugs against 60 malignant cell lines in the National Cancer Institute drug screening program was investigated with ANN [135]. Out of the 141 drugs, the ANN gives incorrect predictions for only 12 of them, showing that the model can be used to guide the screening program. Khan et al. used ANN and gene expression data to identify the class of small, round blue-cell tumors [64]. Although the classification of these tumors is difficult in clinical practice, the neural networks identified the relevant genes and correctly classified all samples. ANN trained with gene microarray data can identify with success several types of cancers, such as colon cancer [111] and esophageal cancer [138].

Radial Basis Function Network

An RBF neural network is a feedforward network with three layers, in which the neurons in the hidden layer have an RBF activation function [93]. RBF neural networks are

universal approximators, can approximate any continuous function with arbitrary precision if enough neurons are allowed in the hidden layer. An early application of the RBF network in structure-property studies was reported by Lohninger for the boiling temperatures of organic compounds [81]. The training process of a radial basis function artificial neural network (RBF ANN) consists of selecting the network topology, finding the centers and widths of the RBF neurons, and computing the connection weights between the hidden and output layers. Zhou et al. showed that the particle swarm optimization is very efficient in training an RBF neural network [143]. The procedure was tested with a dataset of 40 inhibitors of murine P388 leukemia cells and 70 structural descriptors, and the results indicate that the swarm optimization converges fast to a global optimum. Wan and Harrington compared two algorithms for RBF ANN training, namely K-means clustering and linear averaging [131]. The results obtained for two datasets (polychlorobiphenyl mass spectra and Italian olive oil classification) indicate that the linear averaging method gives better predictions.

QSAR models for glycine/NMDA receptor antagonists, which may treat stroke or seizure, were computed for a dataset consisting of 109 compounds [99]. The set of structural descriptors included topological indices, geometric descriptors, quantum indices, and polar surface parameters. The optimum set of descriptors was selected with a genetic algorithm and an RBF network. SAR models for inhibitors of protein tyrosine phosphatase 1B, which are potential agents for the treatment for type 2 diabetes and obesity, were computed for 128 compounds [100]. The classification models were obtained with RBF networks, linear discriminant analysis, and *k*-nearest neighbors. A correct classification rate of 85.7% was obtained for an external prediction set, indicating that the SAR model can be used to find inhibitors in virtual libraries. Neuraminidase is a key target in treating infections with the influenza virus, and many classes of inhibitors are tested for this enzyme. In order to identify the structural features that influence the inhibition of this enzyme, Lü et al. evaluated RBF networks in QSAR models for 46 neuraminidase inhibitors [82]. Five important descriptors were found with a heuristic search among several hundred indices. Prediction tests show that the RBF network results are comparable with those obtained with a multiple linear model, indicating a possible absence of nonlinear structure-activity relationships for these inhibitors.

Zheng et al. compared RBF networks, support vector machines, and partial least square in QSAR models for the Ah receptor binding affinities of polychlorinated, polybrominated, and polychlorinated-brominated

dibenzo-p-dioxins [142]. The best leave-one-out cross-validation predictions are obtained by the RBF network, with $q^2 = 0.88$. Melagraki et al. used RBF networks to model the toxicity of 39 organic compounds against *Vibrio fischeri* [87]. The descriptors used in the simulation include topological indices, electronegativity parameters, and lipophilicity. The good QSAR predictions indicate that the model can be used to evaluate the toxicity of diverse chemicals. QSAR models based on RBF networks were developed for the toxicity of 221 phenols against *Tetrahymena pyriformis* [86]. Prediction tests performed with cross-validation and with an external set show that the RBF network gives more accurate predictions than multiple linear regression.

Self-organizing Map

A self-organizing map is an artificial neural network that projects a high-dimensional input space into a low-dimensional (usually two dimensional) output space [71,72]. A SOM is trained with unsupervised learning, meaning that the training molecules do not need a class label or property value. The input vectors are mapped in a two dimensional space by preserving the topological properties of the input space. The output layer in a SOM is a matrix of neurons, with the property that an input vector is projected into a single output neuron. Several similar input vectors may be projected into the same output neuron, or into adjacent output neurons that form clusters of similar objects. This SOM property is used to discover clusters in collections of chemical compounds. The chemical similarity of large chemical libraries is easily evaluated by projecting the molecules onto a SOM that auto-organizes similar compounds by preserving their similarity relationships in the descriptor space. SOM are used to screen libraries to identify clusters of structurally similar compounds that may have similar biological properties.

SOM may be used also for classification by assigning class labels to a trained network (Fig. 4). This example shows an 8×8 SOM with labels attached to the output neurons. The class labels are attached after training and represent the majority class for objects projected onto a particular output neuron. The objects are organized in four clusters, two for class 1, one for class 2, and one for class 3. The SOM predictions are straightforward, namely a new object is projected onto an output neuron and takes the label of the output neuron as its predicted class. Single molecules may be mapped onto a SOM network by projecting the points situated on the molecular surface [3]. The Cartesian coordinates of each surface point are used as input for the SOM network that provides a two-dimen-

2	2	2		1	1	1	1
2	2				1	1	1
2	2			1	1	1	1
3	3	3					
3	3		1		1		
3	3	3	1	1	1	1	
			1	1	1		

Drug Design with Artificial Neural Networks, Figure 4
A self-organizing map with three classes

sional display of the molecular surface. To add more structural information on the map, output neurons are colored according to the molecular electrostatic potential of the surface points projected onto them.

SOM networks are very efficient in discovering activity clusters in chemical libraries that contain active compounds for several targets. Terfloth and Gasteiger clustered with SOM a number of 299 compounds representing four activity classes, namely 75 5-hydroxytryptamine 5-HT_{1A}-receptor agonists, 75 histamine H₂-receptor antagonists, 74 monoamine oxidase MAO-A inhibitors, and 75 thrombin inhibitors [122]. All classes are generally clustered in distinct regions, but a small number of neurons represent chemicals from different classes. Such degenerate mapping may be solved by increasing the size of the map or by testing other structural descriptors that may discriminate better these activity classes. Drug candidates are routinely screen to determine if they are hERG channel blockers, because such chemicals can cause sudden cardiac death. Roche et al. examined the hERG activity of a chemical library with SOM, PLS, MLF ANN, principal component analysis, and substructure analysis [105]. These classification models offer good predictions for a validation set, with 71% success rate for active compounds and 93% for inactive compounds.

Trypanosoma cruzi is deposited on the skin surface by bugs from the Reduviidae family, and then it infects the human host by penetrating through insect bites, thus causing trypanosomiasis, or Chagas disease. The chronic form of the disease may develop more than 10 years after the infection, and may be fatal. Boiani et al. developed several computational models to evaluate the anti-*Trypanosoma cruzi* activity of *N*-oxide containing heterocycles [11]. A SOM was applied to separate between

activity classes and to select an optimum set of descriptors. Other models tested with success for this dataset are *k*-nearest neighbors and decision tree.

QSAR models have limitations for the domain of chemical structures and biological activities for which they give reliable predictions. For example, interpolation gives better predictions than extrapolation, and extrapolations far away from the training structural space or activity space are not reliable. Also, all compounds used to train a QSAR should have the same mechanism of interaction with the biological target. A SOM can easily identify if a prediction compound is similar or not with the training set of chemicals. If the prediction compound is mapped in the SOM regions occupied by the training compounds, then the QSAR prediction has a high confidence level. SOM is also used to select from a large dataset only those chemicals that act through a common mechanism or form a cluster of compounds with similar structures. Gini et al. applied such an approach to separate molecules into homogeneous groups of chemicals with similar biological and structural properties [30]. These groups were then used to compute local QSAR models for the aquatic toxicity of chemicals against *Pimephales promelas* and *Tetrahymena pyriformis*. A proper validation of a QSAR model gives vital information regarding its prediction error and validity domain. SOM is a good choice to split a dataset into a training set and a validation (prediction) set. Papa et al. applied this method to develop QSAR models for the toxicity of organic chemicals against *Pimephales promelas* [98].

SOM networks are flexible tools to cluster and visualize large chemical libraries. von Korff and Hilpert applied SOM to cluster a library containing 3000 G-protein-coupled receptor (GPCR) ligands for about 130 receptors and 3000 non-GPCR ligands [147]. The chemical structure was described with fingerprints and topological pharmacophores. Cross-validation tests indicate that SOM can separate GPCR ligands from non-GPCR ligands. A possible application of the SAR model is in the design and identification of focused libraries that target a specific GPCR receptor. The supervised self-organizing map (sSOM) is a SOM variant that uses property labels during the training phase. Xiao et al. compared sSOM with *k*-nearest neighbors, PLS and other linear methods in their ability to model a set of 256 inhibitors of dihydrofolate reductase [137]. The predictions tests show that sSOM is insensitive to noise and provides more accurate predictions than linear QSAR methods.

The experimental data obtained in the National Cancer Institute antitumor drug screening program were investigated with SOM to identify clusters of chemicals hav-

ing the same biological action [103]. SOM clusters highlight four classes of mechanism of action of drugs, namely mitosis, nucleic acid synthesis, membrane transport and integrity, and phosphatase- and kinase-mediated cell cycle regulation. SOM and fuzzy C-means clustering were used by Wang et al. to identify marker genes and to classify tumors based on DNA microarray data [132]. The methods were tested with good results for four datasets, namely leukemia, colon cancer, brain tumors, and National Cancer Institute cancer cell lines. For each dataset the algorithm identified sets of marker genes with predictive power for tumor classification.

Counterpropagation Neural Network

Hecht-Nielsen developed the counter-propagation neural network (CPNN) as a supervised learning network that combines a self-organizing map that sends signals to an output layer of neurons [39,40,41]. CPNN applications in chemistry and in structure-activity are reviewed by Zupan et al. [146] and by Vracko [129]. Recent CPNN studies in molecular modeling and structure-activity relationships are reviewed in this section.

Jeziarska et al. used CPNN to select an optimum group of descriptors for mutagenicity QSAR models obtained for 95 aromatic and heteroaromatic amines [59]. Starting from 275 descriptors, CPNN selected six topological indices and the LUMO energy as the best combination of input parameters. The reduced set of descriptors give good predictions, with $R_{cv} = 0.85$ obtained in leave-one-out cross-validation. Numerous organic chemicals are environmental pollutants, and a considerable number of studies are dedicated to the computational prediction of their mechanism of aquatic toxicity (MOA). The reliable prediction of MOA has major applications in selecting the appropriate QSAR model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimens [53]. Spycher et al. developed MOA classifiers for 115 compounds comprising nine MOA classes and 24 descriptors [117]. The four classifiers evaluated, namely CPNN, PLS, logistic regression, and linear discriminant analysis, had an overall correct classification between 52% and 59%, as determined by five-fold cross-validation. Several MOA classes contain a small number of compounds, which might explain the low prediction rate. In a related study, CPNN and logistic regression classifiers were computed for a dataset of 220 phenols with four MOA [118]. The rate of correct predictions in five-fold cross-validation was 92%, much higher than in the previous study that considered nine MOA classes. The higher prediction rate may be a result

of a larger number of compounds in each MOA class and of a more homogeneous set of chemicals.

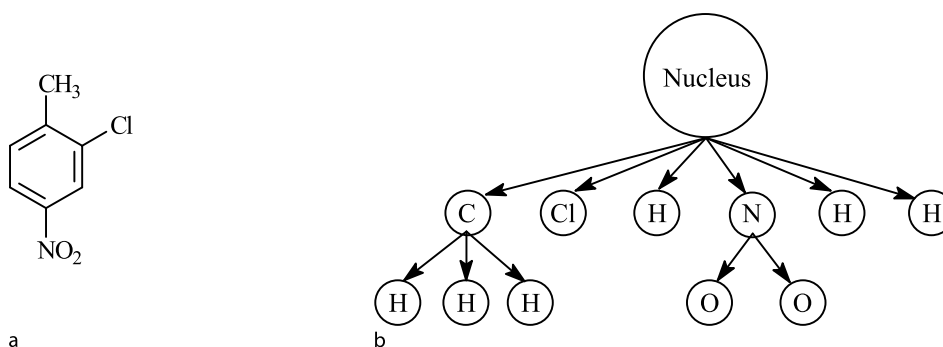
CPNN and support vector machines (SVM) were compared in a QSAR study for the estrogen receptor binding affinity of 131 compounds [29]. Several methods of feature selection were tested, and the predictions obtained with CPNN and SVM vary significantly with the different sets of selected descriptors. Mazurek et al. investigated CPNN structure-selectivity models for a set of artificial metalloenzymes [83]. These catalysts perform the enantioselective hydrogenation of the acetamidoacrylic acid with a high enantiomeric excess (%ee). The structural descriptors are computed from the geometry of the complexes ligand-metalloenzyme, and the CPNN is trained to predict the %ee. The predictions obtained for an external test set ($R = 0.953$ and $RMS = 16.8\%$ ee) provide convincing evidence for the utility of ligand-metalloenzyme descriptors in modeling structure-selectivity relationships. Wagner et al. applied CPNN to structure-activity models for the NF-KB inhibition by 103 sesquiterpene lactones [130]. The activity range for NF-KB inhibition was partitioned in six activity classes, and the chemical structure was expressed with autocorrelation descriptors based on atomic properties. A ten-fold cross-validation shows that the rate of correct classification is 80.6%, and an external validation with 14 new compounds has 78.6% correct predictions. CPNN and k -nearest neighbors classifiers were applied to the classification of plants from the Asteraceae family based on a set of sesquiterpene lactones isolated from individual species [44]. The separation of plants in three tribes and seven subtribes is based on structural descriptors computed from the molecular structure of sesquiterpene lactones. This approach can be applied for the automatic identification of plant species based on chemical analysis and machine learning.

Graph Machine Neural Networks

Usual structure-activity models consider the chemical structure only indirectly, through descriptors computed from various representations of molecules. In these models, the topology of neural networks remains constant for all training and prediction molecules. All training molecules, large or small, are represented by a constant-length vector of descriptors, thus restricting the use of the chemical structure in models. A different approach is taken in artificial neural networks based on graph machines which learn SAR and QSAR models directly from the molecular graph, by translating the chemical structure into the network topology [32]. In this section we review four graph machine neural networks, namely recursive neural networks [5,89,90], the Baskin-Palyulin-Zefirov (BPZ) neural device [7], ChemNet [69], and MolNet [47,50].

Recursive Neural Networks

Recursive neural networks (RecNN) map a directed ordered acyclic graph to a molecular property. All chemical compounds modeled in a RecNN model must be represented as a directed ordered acyclic graph. As an example consider 2-chloro-1-methyl-4-nitrobenzene (Fig. 5a) as a member of series of compounds that have a common skeleton, which in this case is benzene. The molecular structure is transformed into an ordered tree as shown in Fig. 5b. All molecules in the training set are transformed into ordered trees. Using these structured data as input, RecNN reflects the molecular topology as encoded in the ordered tree. The theory and applications of RecNN in SAR and QSAR are presented in several reviews [5,89,90]. RecNN were applied to several structure-property stud-



Drug Design with Artificial Neural Networks, Figure 5

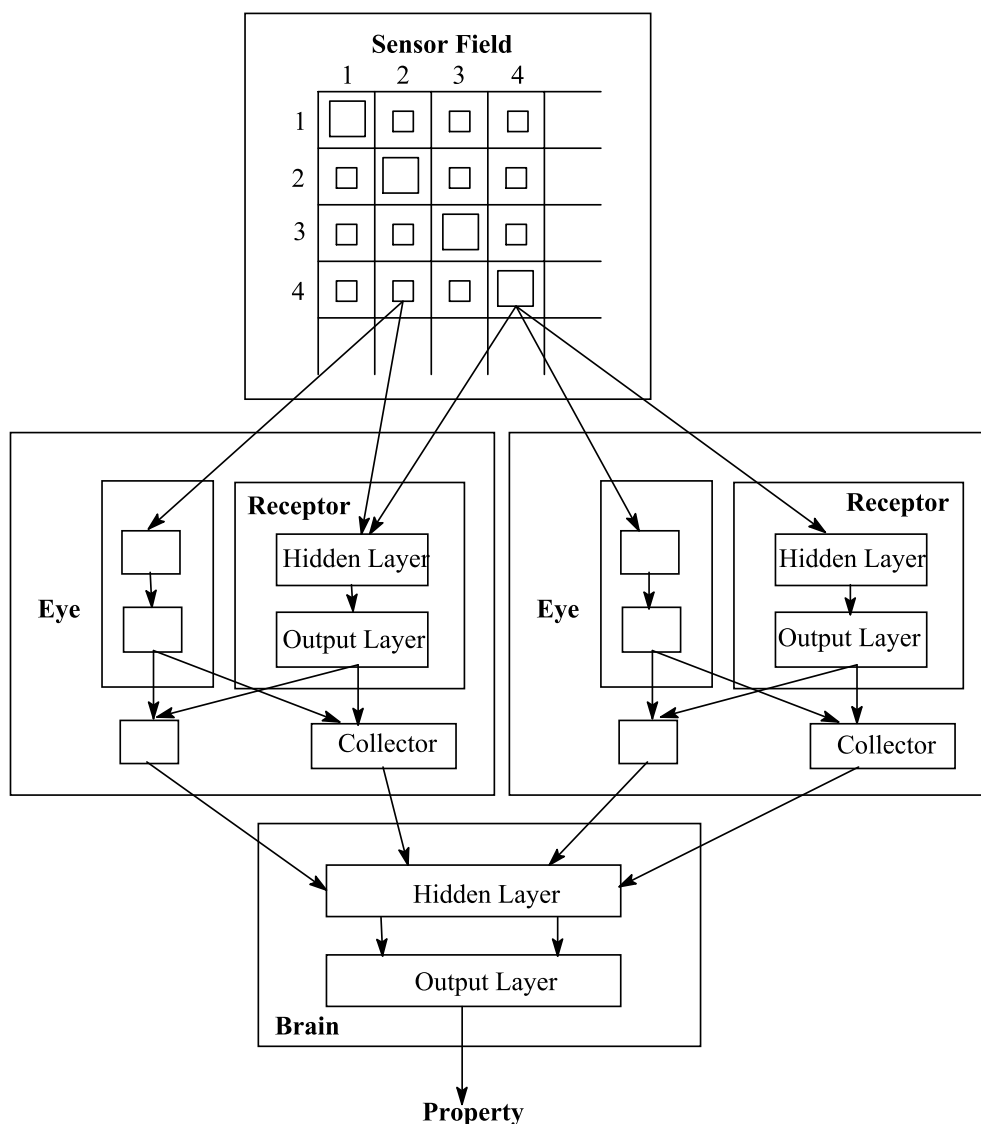
Representation of congeneric chemical compounds in recursive neural networks: a 2-chloro-1-methyl-4-nitrobenzene b represented as a tree

ies, including models for the glass transition temperature of polymers [23,24], QSAR for the binding affinities of ligands to the benzodiazepine receptor [91], and for the Gibbs free energy of solvation in water of organic compounds [8].

Baskin–Palyulin–Zefirov Neural Device

The Baskin–Palyulin–Zefirov (BPZ) neural device (Fig. 6) is a computational implementation of a biological vision system and contains a sensor field, a set of eyes, and a brain [7]. A molecule submitted for evaluation to the

BPZ neural device is represented as a molecular matrix superimposed over the sensor field. The sensor field receives the network input as a matrix representation of a molecule, and then sends this information to a set of eyes. The structural information is processed in the eyes by several multilayer feedforward neural networks, and the output signal is sent to the brain. The signals received from the eyes are processed in the brain by an MLF ANN and the output represents the computed property for the molecule sent as input to the sensor field. A set of rules is used to translate a molecule into the structure of the sensor field and of the eyes.



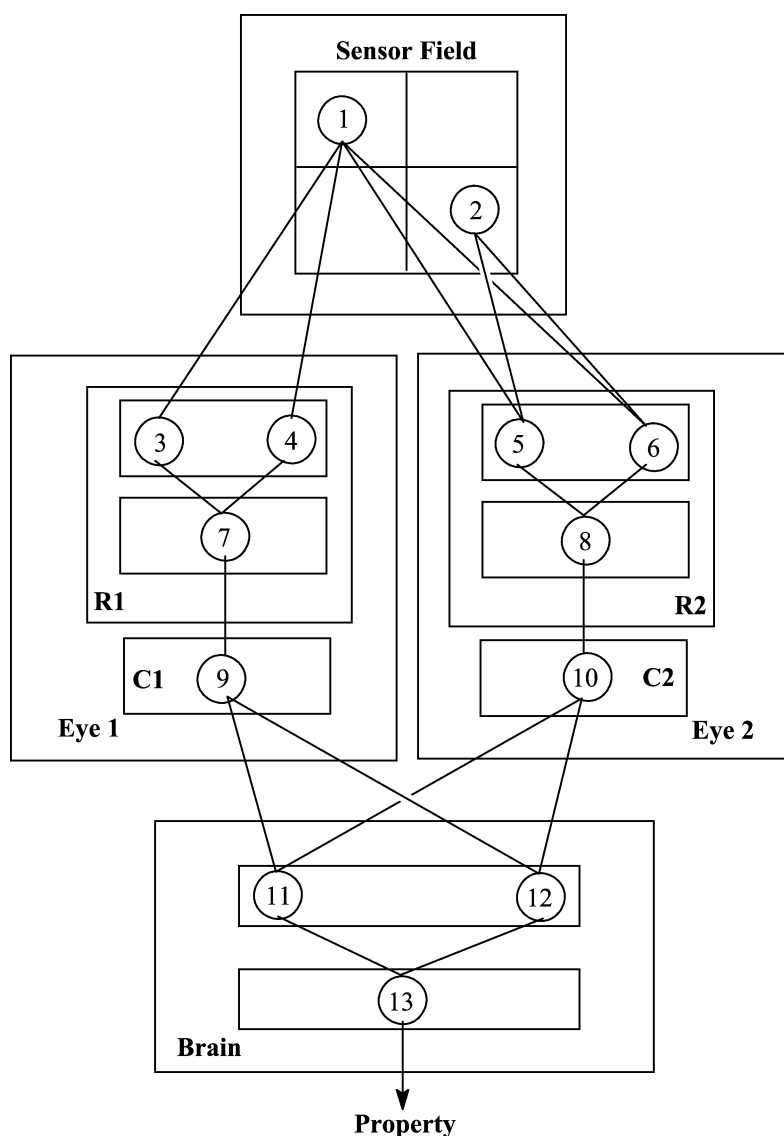
Drug Design with Artificial Neural Networks, Figure 6
The structure of the Baskin–Palyulin–Zefirov neural device

A BPZ neural device with a minimal configuration has one receptor in each eye, and can be used as a template to generate a neural device for any molecule presented to the sensor field (Fig. 7). In the minimal configuration, Eye 1 receives signals from individual atoms, whereas Eye 2 receives from the sensor field signals corresponding to pairs of bonded atoms. The information from the sensor neuron 1 goes to neurons 3 and 4 representing the hidden layer of the receptor R1 from Eye 1. The signal goes then to the output neuron 7 and then to the collector C1 that sends the output to the brain. The information flow is similar in Eye 2, with the only difference that the input signal

represents information regarding bonds between atoms. Neurons 11 and 12 form the hidden layer from the brain, whereas neuron 13 is the output neuron that offers the computed molecular property.

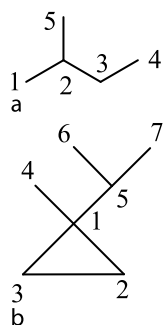
ChemNet

ChemNet is an MLF neural network that translates a molecular structure into the network topology to compute an atomic property [69]. Each molecule is represented by its labeled hydrogen-depleted molecular graph. For each atom i from the molecular graph ChemNet has



Drug Design with Artificial Neural Networks, Figure 7

A minimal configuration of the Baskin-Palyulin-Zefirov neural device



Drug Design with Artificial Neural Networks, Figure 8

Molecular graphs used to generate ChemNet and MolNet neural networks: **a** 2-methylbutane; **b** 1-methyl-1-isopropylcyclopropane

two corresponding neurons with the same label i , namely one in the input layer and the second one in the hidden layer. The output layer has only one neuron, representing the atom from the molecular graph whose property is computed with ChemNet. The connections between the input and hidden layers correspond to the bonding relationships between pairs of atoms, i.e. two pairs of atoms separated by the same number of bonds have identical connection weights. The ChemNet structure is demonstrated for 2-methylbutane (Fig. 8a). The distance matrix for the molecular graph of 2-methylbutane is:

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 3 \\ 2 & 1 & 2 & 3 & 0 \end{bmatrix}$$

Each neuron from the input and hidden layers of ChemNet corresponds to the atom with the same label from the molecular graph of 2-methylbutane (Fig. 9). Input values represent the number of hydrogen atoms attached to each atom in the molecular graph of 2-methylbutane. All pairs of atoms that are separated by the same number of bonds are characterized by input-hidden connections with identical weights. The distance matrix determines the connections that have the same weight. All ChemNet connections corresponding to pairs of atoms separated by one bond have the same weight (Fig. 9a). The connections between atoms separated by two bonds are shown in Fig. 9b, whereas the connections between atoms separated by three bonds are shown in Fig. 9c. All connections from the bias neuron to the 5 hidden neurons have the same weight (Fig. 9d). The ChemNet structure from Fig. 9 computes a property for atom 1. The connections between the hidden layer and the output neuron corre-

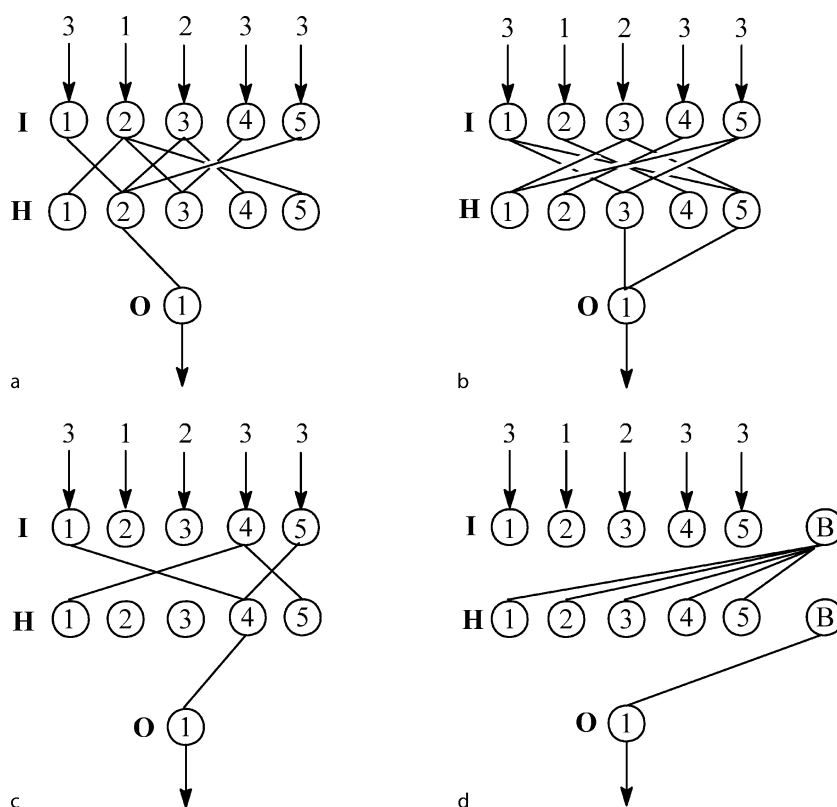
spond to the bonding relationships between atom 1 and all other atoms in the molecule, namely atoms situated at distance 1 (Fig. 9a), distance 2 (Fig. 9b), and distance 3 (Fig. 9c).

MolNet

MolNet is a graph machine that maps the distance matrix in the ANN structure and computes a molecular property using atomic descriptors as input data [47,48,50,51]. The number of neurons in the input and hidden layers is equal to the number of atoms from the molecular graph, and each atom in a molecule has a corresponding neuron in the input and hidden layers. The output layer has only one neuron that provides the computed value for the investigated molecular property. The bonding pattern between two atoms is defined by the ordered sequence of atom types and bond types situated on the shortest path between the two atoms. All identical bonding patterns correspond to network connections with identical weights. An input neuron i is connected to the hidden neuron i with a connection type that depends on the chemical nature of the corresponding atom i . All $i-i$ connections corresponding to the same atomic species have identical weights. The connections from the hidden layer to the output layer (HO connections) are classified according to the type of atoms represented by the hidden neurons. The atoms are partitioned into classes according to their atomic number Z , the hybridization state and the degree. All hidden neurons representing atoms of the same type, either in the same molecule or in different molecules, are connected to the output neuron by connections with identical weights. The bias neuron is connected to each hidden neuron with connections (BH connections) partitioned similarly with the connections between the hidden and output layers, i.e. according to the atom types as defined above. The bias neuron is also connected with the output neuron (BO connection). The MolNet structure of 1-methyl-1-isopropylcyclopropane (Fig. 8b) is based the distance matrix computed from the molecular graph:

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 2 & 2 \\ 1 & 0 & 1 & 2 & 2 & 3 & 3 \\ 1 & 1 & 0 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 0 & 1 & 1 \\ 2 & 3 & 3 & 3 & 1 & 0 & 2 \\ 2 & 3 & 3 & 3 & 1 & 2 & 0 \end{bmatrix}$$

The MolNet connections between the input and hidden layers corresponding to 1-methyl-1-isopropylcyclopropane are shown in Fig. 10. The neural network is sep-



Drug Design with Artificial Neural Networks, Figure 9

ChemNet topology for 2-methylbutane (Fig. 8a) that computes a property of atom 1. Each neuron in the input (I) and hidden (H) layers corresponds to the atom with the same label from the molecular graph of 2-methylbutane. The connections between atoms situated at topological distances 1, 2, and 3 are presented in a, b, and c, respectively, whereas the connections from the bias neuron B are presented in d. Input values represent the number of hydrogen atoms attached to each atom in the molecular graph of 2-methylbutane

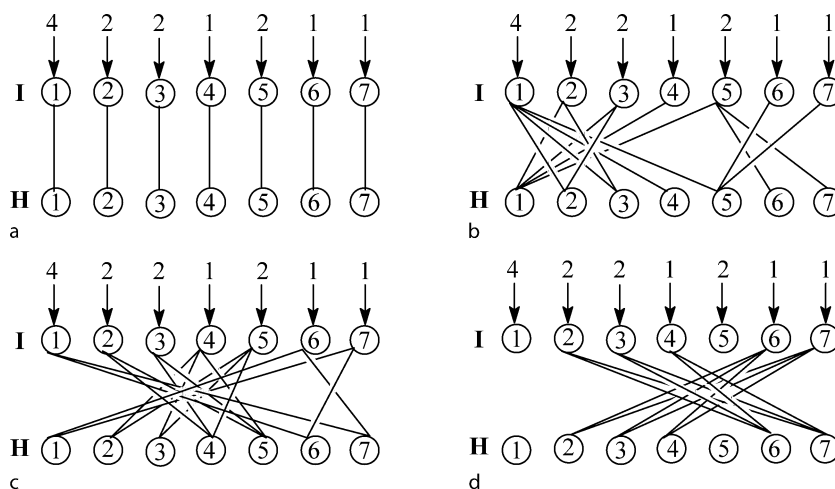
arated into four sections, each corresponding to a bonding relationship. Because the molecule considered here has only carbon atoms and single bonds, all connections depicted in a section from Fig. 10 have identical weights. Input values represent vertex degrees, but any atomic index may be used as input to the network. The network consists of connections between atoms with the same label (Fig. 10a), and between atoms situated at topological distances 1 (Fig. 10b), 2 (Fig. 10c), and 3 (Fig. 10d), respectively.

The connections between the hidden and output layers (HO connections) are determined by the structure of the molecule presented to MolNet. HO connections in alkanes are classified according to the degree of the carbon atoms: hidden neurons representing atoms with identical degrees are linked to the output neuron with connections having identical weights. The connections between the bias neuron and the neurons in the hidden layer (BH connections) are classified according to the same rules used for

HO connections. MolNet connections between the hidden and output layers for 1-methyl-1-isopropylcyclopropane are presented in Fig. 11.

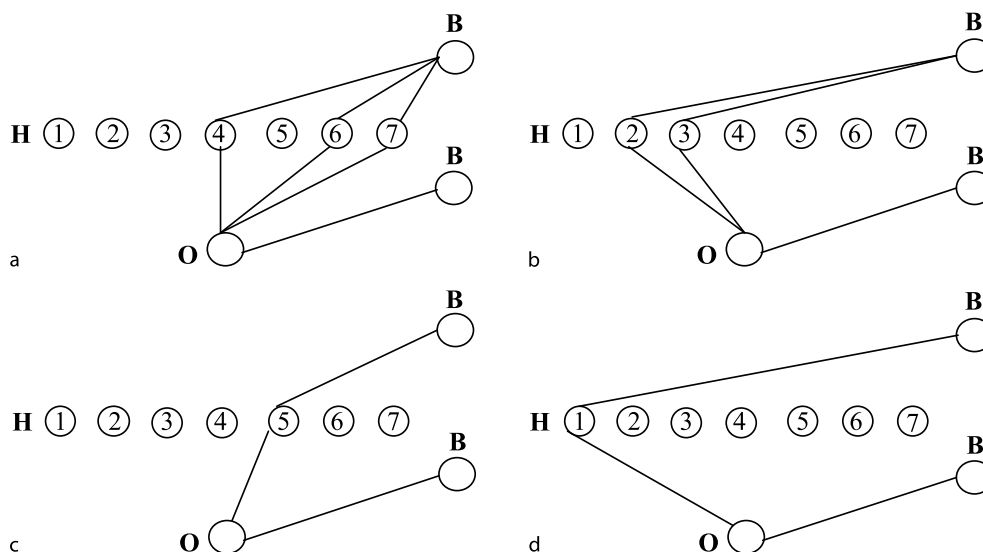
Other Neural Networks

In this section we present an overview of other artificial neural networks applied in structure-activity models and in drug design. A Bayesian regularized neural network (BRNN) was evaluated in QSAR for K_m values of cytochrome P450 3A4 substrates [133]. Correlations with electrotopological state indices give predictive and robust models. Bruneau and McElroy applied BRNN to models for the distribution coefficient $\log D_{7.4}$ computed for a set of 5000 compounds [14]. The neural model obtained is stable and can be applied for compounds in different ionization states. The steady-state volume of distribution V_{ss} for drugs and drug-like chemicals was computed with BRNN for a set of 199 compounds (human V_{ss}) and a sec-



Drug Design with Artificial Neural Networks, Figure 10

MolNet connections between the input (I) and hidden (H) layers for 1-methyl-1-isopropylcyclopropane (Fig. 8b); each neuron corresponds to the carbon atom with the same label from the molecular graph of 1-methyl-1-isopropylcyclopropane. The connections between atoms with the same label are presented in a, whereas the connections between atoms situated at topological distances 1, 2, and 3 are presented in b, c, and d, respectively. Input values represent vertex degrees



Drug Design with Artificial Neural Networks, Figure 11

MolNet connections between the hidden (H) and output (O) layers for 1-methyl-1-isopropylcyclopropane (Fig. 8b); the bias neuron is labeled with B. The bias-hidden and hidden-output connections to/from hidden neurons representing atoms with the degree 1, 2, 3, and 4 are presented in a, b, c, and d, respectively

ond set of 2086 compounds (rat V_{ss}) [31]. Both models have good predictions statistics, and can be used to filter chemicals in the early stages of the drug discovery process. Caballero et al. computed a BRNN model for the inhibition constant of 46 inhibitors of the cysteine protease cruzain [16]. The neural network was also used to rank the importance of the structural descriptors.

Probabilistic neural networks (PNN) and general regression neural network (GRNN) have similar architectures, with the difference that PNN perform classification whereas GRNN perform regression. Niwa applied PNN to identify the biological target from the chemical structure of their ligands [96]. The neural network was trained with a set of 799 compounds having activities against seven

biological targets. On average, 90% of the compounds were correctly classified. PNN classification models were developed for the genotoxic potential of 85 quinolones and 115 quinolines [38]. The quinolone dataset had 23 genotoxic and 62 nongenotoxic compounds, whereas the quinoline dataset had 44 genotoxic and 71 nongenotoxic chemicals. An ensemble of nine PNN models was developed for each classification model, and the final class attribution (genotoxic/nongenotoxic) was decided by a majority vote of the trained classifiers. Simulated annealing was used to select between three and ten structural descriptors for each PNN classifier. The ensemble PNN model for quinolone derivatives was able to predict correctly 16 of the 23 genotoxic chemicals and 60 of the 62 nongenotoxic compounds, with an overall accuracy of 89.4%, an overall accuracy for genotoxic class of 69.6%, and an overall accuracy for nongenotoxic class of 96.8%. Other structure-activity and drug design applications of PNN include the prediction of estrogen receptor agonists [78], identification of drugs that penetrate the blood-brain barrier [79], estimation of the genotoxicity [77], prediction of P-glycoprotein substrates [140], classification of chemicals that are toxic for *Tetrahymena pyriformis* [139], and identification of factor Xa inhibitors [80]. The general regression neural network was applied for the computation of the total clearance CL_{tot} of 503 compounds [141]. The CL_{tot} of a drug characterizes its bioavailability and elimination, and thus may be used to determine its dose and steady-state concentration. Based on the statistics obtained for a validation dataset of 105 compounds, the best predictions are obtained with the support vector regression followed by GRNN. The percent human intestinal absorption (%HIA) of 86 drug and drug-like chemicals was computed with GRNN and PNN [95]. For an external prediction set, GRNN has a root mean square error of 22.8%HIA and PNN has 80% rate of correct classification. The models for human intestinal absorption are based on topological indices, thus making the GRNN and PNN models strong candidates for screening in large virtual libraries.

Future Directions

Artificial neural networks have a unique appeal in structure-activity predictions because they promise to deliver superior modeling power inspired by the functions and mechanisms of the human brain. Although the initial exaggerated expectations never materialized, and artificial intelligence is still a faraway dream, the drug design applications reviewed here provide compelling evidence for the exceptional modeling abilities of artificial neural net-

works. Current applications span all major SAR and QSAR properties relevant to drug design and discovery, and new applications are reported in the literature. Future developments should consider an integrated development of optimally predictive models, in which feature selection, network topology optimization, network training, and validation form a single optimization problem. We anticipate significant improvement in the optimization algorithms used in ANN training, inspired by the remarkable results obtained with particle swarm optimization and ant colony optimization. Novel ANN algorithms and procedures should be adopted from the disciplines of computer science and machine learning. It is expected that the continuous increase in computer power will make feasible applications of ANN ensembles for datasets relevant to the pharmaceutical industry. Significant improvement in SAR and QSAR models may come from graph machine applications that incorporate structured data into the ANN topology.

Bibliography

1. Adams CP, Brantner VV (2006) Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff* 25:420–428
2. Agrafiotis DK, Cedeño W (2002) Feature selection for structure-activity correlation using binary particle swarms. *J Med Chem* 45:1098–1107
3. Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Sadowski J, Teckentrup A, Wagener M (1998) The use of self-organizing neural networks in drug design. *Perspect Drug Discov Design* 9–11:273–299
4. Balaban AT, Ivanciuc O (1999) Historical development of topological indices. In: Devillers J, Balaban AT (eds) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers, Amsterdam, pp 21–57
5. Baldi P, Pollastri G (2004) The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. *J Mach Learn Res* 4:575–602
6. Basak SC, Grunwald GD, Host GE, Niemi GJ, Bradbury SP (1998) A comparative study of molecular similarity, statistical, and neural methods for predicting toxic modes of action. *Environ Toxicol Chem* 17:1056–1064
7. Baskin II, Palyulin VA, Zefirov NS (1997) A neural device for searching direct correlations between structures and properties of chemical compounds. *J Chem Inf Comput Sci* 37:715–721
8. Bernazzani L, Duce C, Micheli A, Mollica V, Sperduti A, Starita A, Tiné MR (2006) Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. *J Chem Inf Model* 46:2030–2042
9. Berndt ER, Gottschalk AHB, Strobeck MW (2005) Opportunities for improving the drug development process: Results from a survey of industry and the FDA. National Bureau of Economic Research Workshop on Innovation Policy and the Economy, NBER Working Paper No 11425, Washington, DC

10. Bishop CM (1996) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 504 pp
11. Boiani M, Cerecetto H, M González, Gasteiger J (2008) Modeling anti-Trypanosoma cruzi activity of *N*-oxide containing heterocycles. *J Chem Inf Model* 48:213–219
12. Bonchev D (1983) *Information Theoretic Indices for Characterization of Chemical Structure*. Research Studies Press, Chichester
13. Bonchev D, Rouvray DH (eds) (1991) *Chemical Graph Theory. Introduction and Fundamentals*. Abacus Press/Gordon & Breach Science Publishers, New York
14. Bruneau P, McElroy NR (2006) $\log D_{7.4}$ Modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. *J Chem Inf Model* 46:1379–1387
15. Bulsari AB (1995) *Neural Networks for Chemical Engineers*. Elsevier, Amsterdam, 609 pp
16. Caballero J, Tundidor-Camba A, Fernández M (2007) Modeling of the inhibition constant (K_i) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized genetic neural networks. *QSAR Comb Sci* 26:27–40
17. Choi S (2003) Nefazodone (Serzone) withdrawn because of hepatotoxicity. *Can Med Assoc J* 169:1187–1187
18. Crum-Brown A, Frazer T (1868–1869) On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans Royal Soc Edinburgh* 25:257–274
19. Cruz JA, Wishart DS (2006) Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2:59–78
20. DiMasi JA (2002) The value of improving the productivity of the drug development process: faster times and better decisions. *Pharmacoeconomics* 20(S3):1–10
21. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. *J Health Econ* 22:151–185
22. Dieppe PA, Ebrahim S, Martin RM, Jüni P (2004) Lessons from the withdrawal of rofecoxib. *Br Med J* 329:867–868
23. Duce C, Micheli A, Solaro R, Starita A, Tiné MR (2006) Prediction of chemical-physical properties by neural networks for structures. *Macromol Symp* 234:13–19
24. Duce C, Micheli A, Starita A, Tiné MR, Solaro R (2006) Prediction of polymer properties from their structure by recursive neural networks. *Macromol Rapid Commun* 27:711–715
25. Faich GA, Moseley RH (2001) Troglitazone (Rezulin) and hepatic injury. *Pharmacoepidemiol Drug Saf* 10:537–547
26. Filter M, Eichler-Mertens M, Bredenbeck A, Losch FO, Sharav T, Givchchi A, Walden P, Wrede P (2006) A strategy for the identification of canonical and non-canonical MHC I-binding epitopes using an ANN-based epitope prediction algorithm. *QSAR Comb Sci* 25:350–358
27. Fujita T, Iwasa J, Hansch C (1964) A new substituent constant, π , derived from partition coefficients. *J Am Chem Soc* 86:5175–5180
28. Furberg CD, Pitt B (2001) Withdrawal of cerivastatin from the world market. *Curr Control Trials Cardivasc Med* 2:205–207
29. Ghafourian T, Cronin MTD (2006) The effect of variable selection on the non-linear modelling of oestrogen receptor binding. *QSAR Comb Sci* 25:824–835
30. Gini G, Craciun MV, König C, Benfenati E (2004) Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. *J Chem Inf Comput Sci* 44:1897–1902
31. Gleeson MP, Waters NJ, Paine SW, Davis AM (2006) In silico human and rat V_{ss} quantitative structure-activity relationship models. *J Med Chem* 49:1953–1963
32. Goulon-Sigwalt-Abram A, Duprat A, Dreyfus G (2005) From Hopfield nets to recursive networks to graph machines: Numerical machine learning for structured data. *Theor Comput Sci* 344:298–334
33. Hall LH, Story CT (1996) Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *J Chem Inf Comput Sci* 36:1004–1014
34. Hall LH, Story CT (1997) Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological state indices using artificial neural networks. *SAR QSAR Environ Res* 6:139–161
35. Hansch C (1969) A quantitative approach to biochemical structure-activity relationships. *Acc Chem Res* 2:232–239
36. Hansch C, Fujita T (1964) $\rho - \sigma - \pi$ analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc* 86:1616–1626
37. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178–180
38. He L, Jurs PC, Kretsoulas C, Custer LL, Durham SK, Pearl GM (2005) Probabilistic neural network multiple classifier system for predicting the neurotoxicity of quinolone and quinoline derivatives. *Chem Res Toxicol* 18:428–440
39. Hecht-Nielsen R (1987) Counterpropagation networks. *Appl Optics* 26:4979–4984
40. Hecht-Nielsen R (1988) Applications of counterpropagation networks. *Neural Netw* 1:131–139
41. Hecht-Nielsen R (1990) *Neurocomputing*, Addison-Wesley, Reading
42. Hogan V (2000) Pemoline (Cylert): market withdrawal. *Can Med Assoc J* 162:106–106
43. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558
44. Hristozov D, Da Costa FB, Gasteiger J (2007) Sesquiterpene lactones-based classification of the family Asteraceae using neural networks and *k*-nearest neighbors. *J Chem Inf Model* 47:9–19
45. Huuskonen J (2003) QSAR modeling with the electrotopological state indices: Predicting the toxicity of organic chemicals. *Chemosphere* 50:949–953
46. Huuskonen J, Rantanen J, Livingstone D (2000) Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur J Med Chem* 35:1081–1088
47. Ivanciuc O (1998) Artificial neural networks applications, Part 9. MolNet prediction of alkane boiling points. *Rev Roum Chim* 43:885–894
48. Ivanciuc O (1999) Artificial neural networks applications. Part 11. MolNet prediction of alkane densities. *Rev Roum Chim* 44:619–631
49. Ivanciuc O (1999) Molecular graph descriptors used in neural network models. In: Devillers J, Balaban AT (eds) *Topological*

- Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Amsterdam, pp 697–777
50. Ivanciuc O (1999) The neural network MolNet prediction of alkane enthalpies. *Anal Chim Acta* 384:271–284
51. Ivanciuc O (2000) Artificial neural networks applications. Part 12. The prediction of alkane heat capacities with the MolNet neural network. *Rev Roum Chim* 45:391–403
52. Ivanciuc O (2001) New neural networks for structure-property models. In: Diudea MV (ed) *QSPR/QSAR Studies by Molecular Descriptors*. Nova Science Publishers, Huntington, pp 213–231
53. Ivanciuc O (2003) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. *Internet Electron J Mol Des* 2:195–208
54. Ivanciuc O (2003) Graph theory in chemistry. In: Gasteiger J (ed) *Handbook of Chemoinformatics*, vol 1. Wiley-VCH, Weinheim, pp 103–138
55. Ivanciuc O (2003) Topological indices. In: Gasteiger J (ed) *Handbook of Chemoinformatics*, vol 3. Wiley-VCH, Weinheim, pp 981–1003
56. Ivanciuc O, Rabine J-P, Cabrol DB, Panaye A, Doucet JP (1996) ^{13}C NMR chemical shift prediction of sp^2 carbon atoms in acyclic alkenes using neural networks. *J Chem Inf Comput Sci* 36:644–653
57. Ivanciuc O, Rabine J-P, Cabrol DB, Panaye A, Doucet JP (1997) ^{13}C NMR chemical shift prediction of the sp^3 carbon atoms in the α position relative to the double bond in acyclic alkenes. *J Chem Inf Comput Sci* 37:587–598
58. Izrailev S, Agrafiotis DK (2002) Variable selection for QSAR by artificial ant colony systems. *SAR QSAR Environ Res* 13:417–423
59. Jezierska A, Vračko M, Basak SC (2004) Counter-propagation artificial neural networks as a tool for the independent variable selection: Structure-mutagenicity study on aromatic amines. *Mol Divers* 8:371–377
60. Jurs P (2003) Quantitative structure-property relationships. In: Gasteiger J (ed) *Handbook of Chemoinformatics*, vol 3. Wiley-VCH, Weinheim, pp 1314–1335
61. Kaiser KLE (2003) The use of neural networks in QSARs for acute aquatic toxicological endpoints. *J Mol Struct (Theochem)* 622:85–95
62. Katritzky AR, Pacureanu LM, Dobchev DA, Fara DC, Duchowicz PR, Karelson M (2006) QSAR modeling of the inhibition of glycogen synthase kinase-3. *Bioorg Med Chem* 14:4987–5002
63. Keyhani S, Diener-West M, Powe N (2006) Are development times for pharmaceuticals increasing or decreasing? *Health Aff* 25:461–468
64. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med* 7:673–679
65. Kier LB, Hall LH (1976) *Molecular Connectivity in Chemistry and Drug Research*. Academic Press, New York
66. Kier LB, Hall LH (1986) *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth
67. Kier LB, Hall LH (1999) *Molecular Structure Description. The Electrotopological State*. Academic Press, San Diego
68. Kim H-J, Choo H, Cho YS, Koh HY, No KT, Pae AN (2006) Classification of dopamine, serotonin, and dual antagonists by decision trees. *Bioorg Med Chem* 14:2763–2770
69. Kireev DB (1995) ChemNet: A novel neural network based method for graph/property mapping. *J Chem Inf Comput Sci* 35:175–180
70. Ko D, Xu W, Windle B (2005) Gene function classification using NCI-60 cell line gene expression profiles. *Comput Biol Chem* 29:412–419
71. Kohonen T (1989) *Self-Organization and Associative Memory*, 3 edn. Springer, Berlin
72. Kohonen T (1995) *Self-Organizing Maps*. Springer, Berlin
73. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov* 3:711–715
74. Kopp H (1844) Ueber den Zusammenhang zwischen der chemischen Constitution und einigen physikalischen Eigenschaften bei flüssigen Verbindungen. *Ann Chem Pharm* 50:71–144
75. Koziol J (2002) Neural network modeling of melting temperatures for sulfur-containing organic compounds. *Internet Electron J Mol Des* 1:80–93
76. Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SN, Bor DH (2002) Timing of new black box warnings and withdrawals for prescription medications. *J Am Med Assoc* 287:2215–2220
77. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Cao ZW, Chen YZ (2005) Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem Res Toxicol* 18:1071–1080
78. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Chen YZ (2006) Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *J Mol Graph Modell* 25:313–323
79. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J Chem Inf Model* 45:1376–1384
80. Lin HH, Han LY, Yap CW, Xue Y, Liu XH, Zhu F, Chen YZ (2007) Prediction of factor Xa inhibitors by machine learning methods. *J Mol Graph Modell* 26:505–518
81. Lohninger H (1993) Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters. *J Chem Inf Comput Sci* 33:736–744
82. Lü WJ, Chen YL, Ma WP, Zhang XY, Luan F, Liu MC, Chen XG, Hu ZD (2008) QSAR study of neuraminidase inhibitors based on heuristic method and radial basis function network. *Eur J Med Chem* 43:569–576
83. Mazurek S, Ward TR, Novič M (2007) Counter propagation artificial neural networks modeling of an enantioselectivity of artificial metalloenzymes. *Mol Divers* 11:141–152
84. McCulloch WS, Pitts WH (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 7:115–133
85. Meissner M, Schmuker M, Schneider G (2006) Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training. *BMC Bioinformatics* 7:125
86. Melagraki G, Afantitis A, Makridima K, Sarimveis H, Igglessi-Markopoulou O (2006) Prediction of toxicity using a novel RBF neural network training methodology. *J Mol Model* 12:297–305
87. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Alexandridis A (2006) A novel RBF neural network training methodology to predict toxicity to *Vibrio fischeri*. *Mol Divers* 10:213–221
88. Meyer H (1899) *Zur Theorie der Alkoholnarkose, welche*

- Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung. *Arch Exp Pathol Pharmacol* 42:109–118
89. Micheli A, Portera F, Sperduti A (2005) A preliminary empirical comparison of recursive neural networks and tree kernel methods on regression tasks for tree structured domains. *Neurocomputing* 64:73–92
 90. Micheli A, Sperduti A, Starita A (2007) An introduction to recursive neural networks and kernel methods for cheminformatics. *Curr Pharm Design* 13:1469–1495
 91. Micheli A, Sperduti A, Starita A, Bianucci AM (2001) Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. *J Chem Inf Comput Sci* 41:202–218
 92. Minsky ML, Papert SA (1969) *Perceptrons*. MIT Press, Cambridge
 93. Moody J, Darken CJ (1989) Fast learning in networks of locally-tuned processing units. *Neural Comput* 1:281–294
 94. Muresan S, Sadowski J (2005) "In-house likeness": Comparison of large compound collections using artificial neural networks. *J Chem Inf Model* 45:888–893
 95. Niwa T (2003) Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J Chem Inf Comput Sci* 43:113–119
 96. Niwa T (2004) Prediction of biological targets using probabilistic neural networks and atom-type descriptors. *J Med Chem* 47:2645–2650
 97. Overton CE (1901) *Studien über die Narkose. Zugleich ein Beitrag zur Allgemeinen Pharmakologie*. Gustav Fisher Verlag, Jena
 98. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). *J Chem Inf Model* 45:1256–1266
 99. Patankar SJ, Jurs PC (2002) Prediction of glycine/NMDA receptor antagonist inhibition from molecular structure. *J Chem Inf Comput Sci* 42:1053–1068
 100. Patankar SJ, Jurs PC (2003) Classification of inhibitors of protein tyrosine phosphatase 1B using molecular structure based descriptors. *J Chem Inf Comput Sci* 43:885–899
 101. Pitts WH, McCulloch WS (1947) How we know universals: The perception of auditory and visual forms. *Bull Math Biophys* 9:127–147
 102. Plewczynski D, Spieser SAH, Koch U (2006) Assessing different classification methods for virtual screening. *J Chem Inf Model* 46:1098–1106
 103. Rabow AA, Shoemaker RH, Sausville EA, Covell DG (2002) Mining the National Cancer Institute's tumor-screening database: Identification of compounds with similar cellular activities. *J Med Chem* 45:818–840
 104. Ripley BD (2008) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 416 pp
 105. Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G (2002) A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *Chem Bio Chem* 3:455–459
 106. Rosenblatt F (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington
 107. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323:533–536
 108. Rumelhart DE, McClelland JL (eds) (1986) *Parallel Distributed Processing*. MIT Press, Cambridge, 344 pp
 109. Saxena AK, Schaper K-J (2006) QSAR analysis of the time- and dose-dependent anti-inflammatory in vivo activity of substituted imidazo[1:2-*a*]pyridines using artificial neural networks. *QSAR Comb Sci* 25:590–597
 110. Seierstad M, Agrafiotis DK (2006) A QSAR model of hERG binding using a large, diverse, and internally consistent training set. *Chem Biol Drug Des* 67:284–296
 111. Selaru FM, Xu Y, Yin J, Zou T, Liu TC, Mori Y, Abraham JM, Sato F, Wang S, Twigg C, Olaru A, Shustova V, Leytin A, Hytiroglou P, Shibata D, Harpaz N, Meltzer SJ (2002) Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology* 122:606–613
 112. Shen Q, Jiang J-H, Jiao C-X, Lin W-Q, Shen G-L, Yu R-Q (2004) Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. *J Comput Chem* 25:1726–1735
 113. Shen Q, Shi W-M, Yang X-P, Ye B-X (2006) Particle swarm algorithm trained neural network for QSAR studies of inhibitors of platelet-derived growth factor receptor phosphorylation. *Eur J Pharm Sci* 28:369–376
 114. Siu F-M, Che C-M (2006) Quantitative structure-activity (affinity) relationship (QSAR) study on protonation and cationization of α -amino acids. *J Phys Chem A* 110:12348–12354
 115. So S-S, Karplus M (1996) Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. *J Med Chem* 39:1521–1530
 116. So S-S, Karplus M (1996) Genetic neural networks for quantitative structure-activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/GABA_A receptors. *J Med Chem* 39:5246–5256
 117. Spycher S, Nendza M, Gasteiger J (2004) Comparison of different classification methods applied to a mode of toxic action data set. *QSAR Comb Sci* 23:779–791
 118. Spycher S, Pellegrini E, Gasteiger J (2005) Use of structure descriptors to discriminate between modes of toxic action of phenols. *J Chem Inf Model* 45:200–208
 119. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21:631–643
 120. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure-activity relationships. *J Med Chem* 47:5541–5554
 121. Taskinen J, Yliruusi J (2003) Prediction of physicochemical properties based on neural network modelling. *Adv Drug Deliv Rev* 55:1163–1183
 122. Terfloth L, Gasteiger J (2001) Neural networks and genetic algorithms in drug design. *Drug Discov Today* 6:S102–S108
 123. Tetko IV, Tanchuk VY, Villa AEP (2001) Prediction of *n*-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J Chem Inf Comput Sci* 41:1407–1421
 124. Todeschini R, Consonni V (2003) Descriptors from molecular geometry. In: Gasteiger J (ed) *Handbook of Chemoinformatics*, vol 3. Wiley-VCH, Weinheim, pp 1004–1033

125. Trinajstić N (1992) Chemical Graph Theory. CRC Press, Boca Raton
126. Varnek A, Kireeva N, Tetko IV, Baskin II, Solov'ev VP (2007) Exhaustive QSPR studies of a large diverse set of ionic liquids: How accurately can we predict melting points? *J Chem Inf Model* 47:1111–1122
127. Votano JR, Parham M, Hall LH, Kier LB, Oloff S, Tropsha A, Xie Q, Tong W (2004) Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* 19:365–377
128. Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, Tropsha A (2006) QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J Med Chem* 49:7169–7181
129. Vracko M (2005) Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. *Curr Comput-Aided Drug Des* 1:73–78
130. Wagner S, Hofmann A, Siedle B, Terfloth L, Merfort I, Gasteiger J (2006) Development of a structural model for NF-KB inhibition of sesquiterpene lactones using self-organizing neural networks. *J Med Chem* 49:2241–2252
131. Wan C, Harrington PB (1999) Self-configuring radial basis function neural networks for chemical pattern recognition. *J Chem Inf Comput Sci* 39:1049–1056
132. Wang J, Johannsen TH, Myklebost O, Hovig E (2003) Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinformatics* 4:60
133. Wang Y-H, Li Y, Li Y-H, Yang S-L, Yang L (2005) Modeling K_m values using electrotopological state: Substrates for cytochrome P450 3A4-mediated metabolism. *Bioorg Med Chem Lett* 15:4076–4084
134. Wasserman PD (1989) Neural Computing. Van Nostrand Reinhold, New York, 230 pp
135. Weinstein JN, Kohn KW, Grever MR, Viswanadhan VN, Rubinstein LV, Monks AP, Scudiero DA, Welch L, Koutsoukos AD, Chiousa AJ, Paull KD (1992) Neural computing in cancer drug development: Predicting mechanism of action. *Science* 258:447–451
136. Wessel MD, Jurs PC, Tolan JW, Muskall SM (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. *J Chem Inf Comput Sci* 38:726–735
137. Xiao Y-D, Clauset A, Harris R, Bayram E, Santago P, Schmitt JD (2005) Supervised self-organizing maps in drug discovery, 1. Robust behavior with overdetermined data sets. *J Chem Inf Model* 45:1749–1758
138. Xu Y, Selaru FM, Yin J, Zou TT, Shustova V, Mori Y, Sato F, Liu TC, Olaru A, Wang S, Kimos MC, Perry K, Desai K, Greenwald BD, Krasna MJ, Shibata D, Abraham JM, Meltzer SJ (2002) Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Res* 62:3493–3497
139. Xue Y, Li H, Ung CY, Yap CW, Chen YZ (2006) Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chem Res Toxicol* 19:1030–1039
140. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci* 44:1497–1505
141. Yap CW, Li ZR, Chen YZ (2006) Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *J Mol Graph Modell* 24:383–395
142. Zheng G, Xiao M, Lu XH (2005) QSAR study on the Ah receptor-binding affinities of polyhalogenated dibenzo-*p*-dioxins using net atomic-charge descriptors and a radial basis neural network. *Anal Bioanal Chem* 383:810–816
143. Zhou Y-P, Jiang J-H, Lin W-Q, Zou H-Y, Wu H-L, Shen G-L, Yu R-Q (2006) Adaptive configuring of radial basis function network by hybrid particle swarm algorithm for QSAR studies of organic compounds. *J Chem Inf Model* 46:2494–2501
144. Zupan J (2003) Neural networks. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 3. Wiley-VCH, Weinheim, pp 1167–1215
145. Zupan J, Gasteiger J (1999) Neural Networks in Chemistry and Drug Design. Wiley-VCH, Weinheim
146. Zupan J, Novič M, Gasteiger J (1995) Neural networks with counter-propagation learning-strategy used for modeling. *Chemometrics Intell Lab Syst* 27:175–187
147. von Korff M, Hilpert K (2006) Assessing the predictive power of unsupervised visualization techniques to improve the identification of GPCR-focused compound libraries. *J Chem Inf Model* 46:1580–1587

Drug Design with Machine Learning

OVIDIU IVANCIUC

Department of Biochemistry and Molecular Biology,
University of Texas, Medical Branch, Galveston, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Decision Trees](#)

[Lazy Learning and *k*-Nearest Neighbors](#)

[Bayesian Methods](#)

[Support Vector Machines](#)

[Comparative Studies](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Bayesian classifier Bayes' theorem of conditional probability is a method of statistical inference that represents the basis of several classification machine learning models used in drug design and chemoinformatics to classify libraries of compounds into active and inactive chemicals. A Bayesian classifier considers each structural feature or descriptor independent of the other descriptors, and the probability that a compound is active is proportional to the ratio of active to inactive compounds that have the same structural feature or have the same value for that descriptor. The final probability that a compound is active is computed as the prod-

uct of all descriptor-based probabilities. Structural descriptors that are real numbers are usually binned prior to their evaluation with a Bayesian classifier.

Decision tree A decision tree is a sequence of rules applied to selected structural descriptors. The training phase comprises the selection of the structural descriptors that are evaluated, the order in which the rules are applied, and the decision taken at each leaf. Usually, each rule evaluates a descriptor (\geq or $<$ than a threshold) and splits the objects into two or more populations. Then each population is selected and the splitting procedure is performed with a new rule, until a stopping condition is met (for example, when all objects in the population belong to the same class). The prediction phase starts from the root node and evaluates each rule on a pathway determined by the outcome (true or false) of the previous rule. When a leaf is reached the algorithm predicts the class of the object (classification trees) or the numerical value of a property (regression trees).

***k*-nearest neighbors** *k*-nearest neighbors (*k*-NN) is a supervised learning algorithm that predicts the property of an object based on a local interpolation model. In classification, the class of a new object is predicted based on the majority vote of its *k* nearest neighbors. In regression, the property value for a new object is predicted as an average value of the property values for its *k* nearest neighbors.

Lazy learning Lazy learning is a memory based local learning that defers the computation until a prediction is requested for an object. The first step is to insert the query object into the space of the training objects, and to identify the training objects located in a set neighborhood. The predicted property of the query object is based on an interpolation of the properties of the objects situated in the neighborhood.

Machine learning Machine learning is an important field of artificial intelligence, and includes a diversity of methods and algorithms that extract rules and functions from large datasets, such as decision trees, lazy learning, *k*-nearest neighbors, Bayesian methods, Gaussian processes, support vector machines, and kernel algorithms. Machine learning algorithms extract information from experimental data by computational and statistical methods and generate a set of rules, functions or procedures that allow them to predict the properties of novel objects that are not included in the learning set.

Quantitative structure-activity relationships

Quantitative structure-activity relationships (QSAR) represent regression models that define quantita-

tive correlations between the chemical structure of molecules and their physical properties (boiling point, melting point, aqueous solubility), chemical properties and reactivities (chromatographic retention, reaction rate), or biological activities (cell growth inhibition, enzyme inhibition, lethal dose). The fundamental hypotheses of QSAR are that similar chemicals have similar properties, and that small structural changes result in small changes in property values. The general form of a QSAR equation is $P(i) = f(\mathbf{SD}_i)$, where $P(i)$ is a physical, chemical, or biological property of compound *i*, \mathbf{SD}_i is a vector of structural descriptors of *i*, and *f* is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. A QSAR model for a property *P* is based on a dataset of chemical compounds with known values for the property *P*, and a matrix of structural descriptors computed for all chemicals. The learning (training) of the QSAR model is the process of determining the optimum parameters of the regression function *f*. After the training phase, a QSAR model may be used to predict the property *P* for novel compounds that are not present in the learning set of molecules.

Support vector machines Support vector machines (SVM) are a class of supervised machine learning methods based on the structural risk minimization and the statistical learning theory of Vapnik. SVM may be applied to data classification and regression, using selected objects (support vectors) to generate the SVM model. Nonlinear classification problems are transformed into linear classification problems by using kernel functions that combine the input space into a higher-dimensional feature space in which a hyperplane may discriminate the classes. An SVM classification model computes a maximum margin hyperplane that separates the classes in the feature space. The maximal margin hyperplane maximizes the distance to the hyperplane of the closest patterns from the two classes. An SVM regression model builds a regression tube with the property that all objects inside the tube do not contribute to the overall error of the model. The shape of the regression tube is determined by selected objects (support vectors) situated outside the tube.

Structural descriptor A structural descriptor (SD) is a numerical value computed from the chemical structure of a molecule, which is invariant to the numbering of the atoms in the molecule. Structural descriptors may be classified as constitutional (counts of molecular fragments, such as rings, functional groups, or atom pairs), topological indices (computed from the molec-

ular graph), geometrical (volume, surface, charged-surface), quantum (atomic charges, energies of molecular orbitals), and molecular field (such as those used in CoMFA, CoMSIA, or CoRSA).

Structure-activity relationships Structure-activity relationships (SAR) represent classification models that can discriminate between sets of chemicals that belong to different classes of biological activities, usually active/inactive towards a certain biological receptor. The general form of a SAR equation is $C(i) = f(\mathbf{SD}_i)$, where $C(i)$ is the activity class of compound i (active/inactive, inhibitor/non-inhibitor, ligand/non-ligand), \mathbf{SD}_i is a vector of structural descriptors of i , and f is a classification function such as k -nearest neighbors, linear discriminant analysis, random trees, random forests, Bayesian networks, artificial neural networks, or support vector machines.

Definition of the Subject

The process of drug discovery has the goal to identify lead chemicals that have a significant activity against a selected biological target. A disease state may be the result of changes in the structure and function of cell-signaling receptors, enzymes, hormone receptors, or other functional proteins. The drug target is a protein whose activity is modulated by its interaction with a chemical compound, and thus may control a disease. The lead compounds identified in the drug discovery step are optimized in the drug development phase that results in a small number of chemicals that are evaluated in human clinical trials. The first priority in drug development is to increase the biological activity of a lead compound while preserving its drug-like properties. The lead compound is expanded into a chemical library that conserves the structure responsible for the biological activity (pharmacophore) and adds chemical groups that might improve its activity. Then the chemicals are synthesized and tested in biological assays against the target, which may result in the identification of more active compounds. The cycle “library design – chemical synthesis – biological assay” is repeated several times until a number of good candidates are identified for clinical trials. The design of effective drug candidates is a multi-objective optimization problem, because simultaneous with a good biological activity, the chemicals must pass several other important tests, including pharmacokinetics, pharmacodynamics, toxicity, mutagenicity, metabolism, and excretion.

Computer-assisted drug design (CADD) uses computational chemistry to increase the chances of finding valuable drug candidates. CADD methods include machine

learning, structure-activity relationships (SAR), quantitative structure-activity relationships (QSAR), molecular mechanics, quantum mechanics, molecular dynamics, and drug-protein docking. SAR and QSAR are based on the theory that chemical structure determines all physical, chemical and biological properties of a molecule. To obtain structure-activity models, the chemical structure is characterized with structural descriptors and then ML models are used to identify a statistical relationship between the descriptor space and a molecular property. Other considerations for the application of SAR and QSAR in drug design are that similar molecules have similar properties, and that a small modification in a chemical structure result in a small modification in its properties.

Machine learning (ML) procedures are applied in drug discovery in an iterative way, in which experimental activity data are used to train ML models which in turn offer predictions for novel molecules with improved biological properties. Then the molecules designed with ML are tested in biological assays, and the best candidates are selected for further cycles of optimization. ML models may greatly improve the chances of finding lead molecules by screening a larger diversity of chemical topologies. In the classical approach, a number of chemicals in the range 10^3 – 10^4 is synthesized and assayed to identify lead compounds, whereas the introduction of combinatorial chemistry and high-throughput screening (HTS) increased these numbers to 10^5 – 10^6 . Although the robots offer great speed and reliability in HTS, the chemical compounds still have to be synthesized before screening, thus limiting the diversity of structures evaluated in HTS. This is why ML models are used to enhance the lead identification process with a virtual HTS (vHTS) screening of 10^7 – 10^8 molecules. Based on experimental results from a HTS campaign, ML models are computed and then used in vHTS experiments. These ML are mainly classification models that predict molecules that have a high probability of interacting with the selected target. vHTS has several obvious advantages compared to HTS, mainly coming from the fact that the chemical compounds are generated only in a computer, which opens the possibility to explore a larger diversity of chemical skeletons and a much higher number of molecules.

Classification (SAR) and regression (QSAR) ML models are applied during the drug development cycles to optimize the biological activity, target selectivity, and other physico-chemical and biological properties of selected chemicals. ML models are essential tools in increasing the chances of bringing a drug to market, but only when integrated with the usual chemical, biological, pharmacological and clinical procedures used by the pharmaceuti-

cal industry. Boid enumerated several successful CADD applications in which computational methods had a decisive contribution to the discovery of a drug, namely norfloxacin (Merck, antibacterial), losartan (Merck, antihypertensive), dorzolamide (Merck, antiglaucoma), ritonavir (Abbott, antiviral), indinavir (Merck, antiviral), donepezil (Esai, anti-Alzheimer's disease), zolmitriptan (AstraZeneca, antimigraine), nelfinavir (Pfizer, antiviral), amprenavir (GlaxoSmithKline, antiviral), zanamivir (GlaxoSmithKline, antiviral), oseltamivir (Roche, antiviral), lopinavir (Abbott, antiviral), imatinib (Novartis, antineoplastic), erlotinib (OSI, antineoplastic) [17]. This is an impressive list that adds convincing evidence to advocate the integration of machine learning and other chemoinformatics methods in drug discovery and development.

Introduction

Machine learning is an important field of artificial intelligence, and includes a diversity of methods and algorithms that extract rules and functions from large datasets, such as linear discriminant analysis (LDA), artificial neural networks (ANN), decision trees, lazy learning, k -nearest neighbors, Bayesian methods, Gaussian processes, support vector machines (SVM), and kernel algorithms. Several influential books are recommended for a detailed overview of ML algorithms: *Pattern Recognition and Neural Networks* by Ripley [110], *Machine Learning and Data Mining* by Kononenko and Kukar [83], *Pattern Recognition and Machine Learning* by Bishop [15], *Data Mining: Practical Machine Learning Tools and Techniques* by Witten and Frank [151], *Introduction to Machine Learning* by Alpaydin [4], *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman [52], *Pattern Classification* by Duda, Hart, and Stork [35], *Machine Learning* by Mitchell [96], and *Neural Networks for Pattern Recognition* by Bishop [16].

ML algorithms extract information from experimental data by computational and statistical methods and generate a set of rules, functions or procedures that allow them to predict the properties of novel objects that are not included in the learning set. In drug design the task is to learn how chemical structure determines some important drug property, such as physico-chemical properties (aqueous solubility, skin penetration, hydrophobicity, intestinal absorption, blood-brain barrier penetration), biological activity (enzyme inhibition), metabolism, toxicity, mutagenicity, or excretion. Each molecule is associated with a set of structural descriptors and an experimental property. The structural descriptors are used as input for an ML algorithm and the property is the

target (output) of the model. A structural descriptor is a structural feature of the chemical structure, and can be a list of substructures or substituents, graph descriptors [19,64,136], topological indices [8,18,65], connectivity indices [74,75] electrotopological indices [76], or indices derived from the molecular geometry and quantum calculations [73,133]. The experimental property that is modeled may be a class label (+1/−1), such as soluble/non-soluble, inhibitor/non-inhibitor, ligand/non-ligand, toxic/non-toxic, mutagen/non-mutagen, carcinogen/non-carcinogen. For this type of data one obtains a classification model or SAR. Classification models are mainly used to filter large chemical libraries and reduce them to a small number of chemicals with a high probability of having a desired property (ligand for a biological target, inhibitor for an enzyme, non-toxic, non-mutagen, or non-carcinogen). The experimental property may be also a continuous value, such as inhibition constant to an enzyme, binding constant to a target, hydrophobicity, aqueous solubility, skin penetration, or intestinal absorption, in which case one computes a regression model or QSAR. Regression models, which may be linear or nonlinear, are used to optimize the drug-like properties of a chemical compound. A drug-related property may be considered categorical variable or a continuous variable, depending on the drug development phase in which is applied. For example, in developing CNS (central nervous system) drugs, the blood-brain barrier penetration is used as a categorical variable in the early stages of drug discovery, to identify and eliminate compounds that do not pass the blood-brain barrier. In later stages of drug development this property is used as a continuous variable in QSAR models to optimize the brain concentration of a small number of selected chemicals that show promising pharmacological properties. Both SAR and QSAR models belong to the class of supervised learning algorithms, because the target (output) value is provided together with the input chemical structures and descriptors. Supervised learning algorithms have two distinct phases. The first one is learning or training, in which a ML method is used to learn the relationships between input (structural descriptors) and output (an experimental property provided for each molecule). The result is a model that can be a statistical function or a set of rules. The second phase of a ML algorithm is the prediction, when the trained model is used to predict the property for novel chemicals that were not present in the training set or that are not even synthesized. The predictions obtained are then used as a guide in synthesizing and testing novel chemicals.

In unsupervised learning the dataset contains only chemical structures and structural descriptors, without

output values. The ML objective is to identify how molecules form clusters based on their structural similarity. Obviously, starting from a particular set of structural descriptors, different similarity indices and different clustering algorithms will result in different clusters. Clustering algorithms are applied in drug design to identify similar molecules in chemical libraries. For example, starting from a lead compound one can find similar compounds in catalogs and databases, and thus purchase chemicals that might be useful in the drug discovery process. Another application is to identify which chemical structures are under-represented or missing from a company collection, and to guide compound synthesis and acquisition.

To illustrate two basic approaches in data classification we consider the objects from Fig. 1a. The two classes of objects may represent two populations of molecules, for example enzyme inhibitors form class +1, whereas non-inhibitors form class -1. From the distribution of the two classes one can identify a cluster for class +1 and a distinct cluster for class -1. Such a clear-cut situation is not the norm in drug design application, because classes of molecules may overlap which makes more difficult the property prediction for novel chemicals. The problem considered in Fig. 1a is to use the information provided by the two classes of objects to predict the class of the unknown object marked with “?”. The first ML tested is *k*-nearest neighbors (*k*-NN), that belongs to the class of lazy learning algorithms, and performs a local approximation of the model that is computed only in the moment of the prediction. An object is classified in the most populated class among its *k* nearest neighbors, and for two-class problems *k* must be odd to avoid undecided situations. We test here the simplest variant, 1-nearest neighbor, which considers only the closest neighbor. An inspection of the distances between the prediction object • and all other objects shows that the closest object belongs to class -1, and thus the prediction object is assigned to class -1 (Fig. 1b). The predicted class may change if one considers a larger number of nearest neighbors, and the optimum value for *k* is usually determined through cross-validation.

An efficient procedure to classify linearly separable classes is represented by a hyperplane *H* (Fig. 1c) that separates the descriptor space into a region for class +1 and another region for class -1. A new object that is located in the space region that belongs to class +1 will be assigned to class +1 irrespective of its distance to the separating hyperplane. Similarly, if the new object is situated in the space region that belongs to class -1, then the object is predicted in class -1. The unknown object from Fig. 1a is assigned to class +1 by the hyperplane *H* (Fig. 1c). For the same classification problem we obtained two differ-

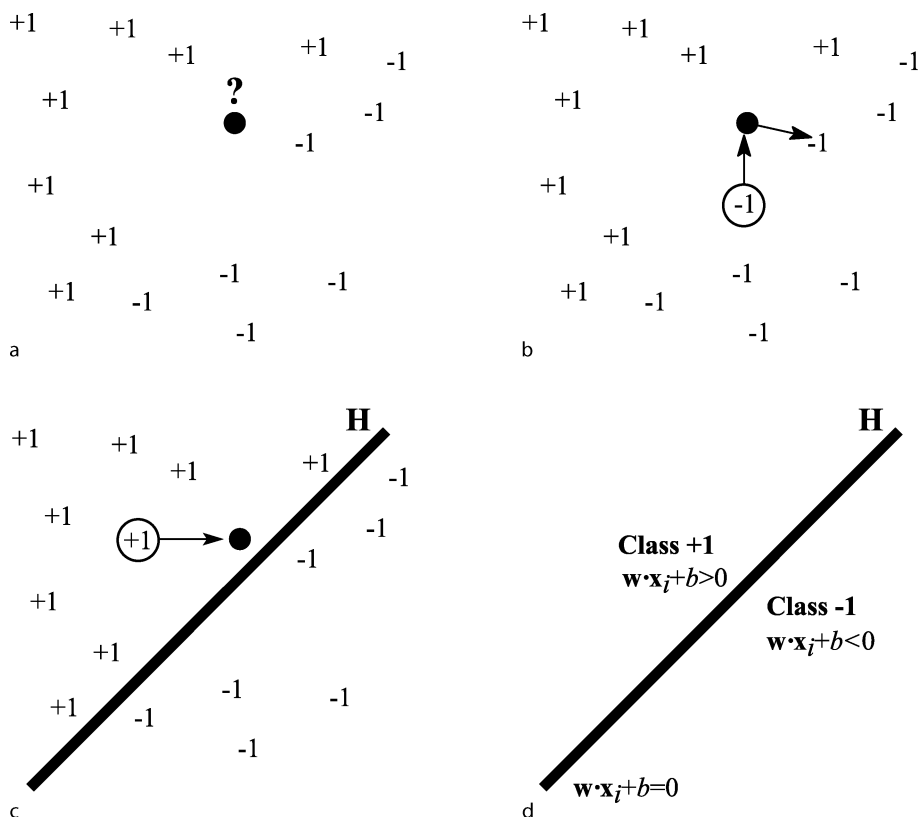
ent predictions for the unknown object, namely class -1 with 1-NN and class +1 with a hyperplane. Such divergent predictions obtained with different ML methods are quite common, and it is not unusual to obtain different predictions for the same object just by changing the ML parameters. These practical aspects of ML should receive proper consideration, and it is a good practice to evaluate a large diversity of ML methods to obtain consensus predictions.

The two ML demonstrated here are very different in their formulation and their properties. *k*-NN can be used for multi-class discrimination and it is a nonlinear classifier that is useful for classes that cannot be separated with a linear hypersurface. Also, *k*-NN does not produce a statistical function to replace the data, and the learning phase is missing. The second classifier replaces the data with the equation of a hyperplane (Fig. 1d). This function may be used only for two-class problems in which the objects may be separated by linear hypersurface. A hyperplane *H* has the formula $\mathbf{w} \cdot \mathbf{x} + b = 0$, where \mathbf{w} is the normal vector to *H*. A molecule *i* characterized by a vector of structural descriptors \mathbf{x}_i belongs to class +1 if $\mathbf{w} \cdot \mathbf{x}_i + b > 0$, or to class -1 if $\mathbf{w} \cdot \mathbf{x}_i + b < 0$. The rules for predicting the class for an unknown object *k* are:

$$\text{class}(k) = \begin{cases} +1 & \text{if } \mathbf{w} \cdot \mathbf{x}_k + b > 0 \\ -1 & \text{if } \mathbf{w} \cdot \mathbf{x}_k + b < 0. \end{cases} \quad (1)$$

The hyperplane classifier may be extended with kernel functions to nonlinear classification, as it will be shown in Sect. “Support Vector Machines” in which we give an overview of support vector machines and their applications in structure-activity models.

A decision tree performs a mapping of the structural descriptors to a conclusion about a property of the object. The learning process establishes a series of successive rules that are based on the numerical values of a subset of descriptors. Rules are applied in a set sequence and at each branching point a rule is evaluated and its outcome (true or false) determines which branch is selected in the next step. Starting from the root node, the rules are evaluated and depending on the actual numerical values of the descriptors the pathway ends on a leaf that assigns a class label or a value to the investigated property. The class of decision tree algorithms includes the C4.5 decision tree [108], the NBTree decision tree with naïve Bayes classifiers at the leaves [82], the alternating decision tree ADTree [42], random trees and random forests [22]. Decision trees are computationally very efficient, and their predictions are usually better than those obtained with more complicated ML algorithms. Drug design applications of decision trees include diverse



Drug Design with Machine Learning, Figure 1

Classification with machine learning: a a class prediction (+1/-1) is sought for the object •; b a k -nearest neighbor classifier ($k = 1$) predicts the object • in class -1; c a linear classifier defined by the hyperplane H predicts the object • in class +1; d the classification hyperplane H separates the space in region for class +1 and another region for class -1

applications, such as prediction of drugs that cross the blood-brain barrier [30], structure-activity relationships for estrogen receptor ligands [135], prediction of *P*-glycoprotein substrates [91], evaluation of the cytochrome P450 metabolism [154], identification of drug-like chemicals [114], design of target chemical libraries [31], prediction of protein-protein interaction [98], and modeling the oral absorption of drugs [57]. A review of these studies is presented in Sect. “Decision Trees”.

Lazy learning is a memory-based local learning method that stores the training objects and performs all computations only when a prediction request is received by the system [5,6]. The prediction is performed by inserting the unknown object into the prediction space and then identifying all objects situated in its neighborhood. Each prediction is made by computing a local model from the neighboring objects. Lazy learning may be used both for classification and regression, with predictions based on local interpolations of selected objects according to a distance measure. Local learning may perform better

than global learning, as shown in a number of SAR and QSAR studies that use various techniques of lazy learning [47,85,159].

The k -nearest neighbors (k -NN) algorithm is a lazy learning method that can be used for classification and regression [1]. The training set of objects is stored in the memory, and then each prediction is made by computing a local interpolation model. The prediction object is added to the descriptor space in which the training objects are placed and then its k nearest neighbors are identified. For classification, the class of the prediction object is assigned to class of the majority of its neighbors. Ties are avoided by selected an odd value for k , and the best value for k is usually determined by cross-validation. For regression, the predicted property for the query object is an average value of the property values for its k nearest neighbors. A distance weighting may be added in such a way to increase the contribution of closer objects and to decrease the influence of more distant objects. Another lazy learning algorithm is K^* which uses an entropy-

based distance function [26]. Basak was an early proponent of the k -NN regression for physico-chemical and biological properties [10,11,49]. The method was further refined by Tropsha by optimizing the descriptor selection, k , and the distance weighing function [60,160,161]. Other k -NN applications were proposed for three-dimensional QSAR [2] and for ligand-based virtual screening of chemical libraries [54]. In Sect. “[Lazy Learning and \$k\$ -Nearest Neighbors](#)” we review SAR and QSAR models computed with lazy learning and k -nearest neighbors.

Bayes’ theorem provides mathematical tools that explain how the probability that a theory is true is affected by a new piece of evidence [12]. Several Bayesian classifiers [71,151] were proposed to estimate the probability that a chemical is active based on a set of descriptors and a chemical library with known class attributes (active/inactive). Bayesian classifiers consider that each descriptor is statistically independent of all other descriptors. Binary descriptors, structural counts, fingerprints and other descriptors that have integer values are directly evaluated with Bayesian classifiers, whereas real number descriptors are discretized and transformed into an array of bins. The probability that a compound is active is proportional to the ratio of active to inactive compounds that have the same structural feature or have the same value for that descriptor. A final probability is computed as the product of all descriptor-based probabilities. Bayesian classifiers have numerous applications in chemoinformatics and structure-activity models, such as for the prediction of multidrug resistance reversal [125], to estimate the phospholipidosis inducing potential [105], to identify chemical compounds similar to natural products [39], to improve high-throughput docking [78,79], and to screen virtual chemical libraries [77]. Drug design applications of Bayesian methods are presented in Sect. “[Bayesian Methods](#)”.

Support vector machines belong to a heterogeneous group of machine learning methods that use kernels to solve efficiently high-dimensional nonlinear problems. SVM extend the generalized portrait algorithm developed by Vapnik [144] by using elements of statistical learning theory [143] that describe the ML properties that guarantee dependable predictions. Vapnik elaborated further the statistical learning theory in three more recent books, *Estimation of Dependencies Based on Empirical Data* [138], *The Nature of Statistical Learning Theory* [139], and *Statistical Learning Theory* [140]. Vapnik and co-workers developed the current formulation of the SVM algorithm at AT&T Bell Laboratories [20,25,28,33,50,116,141,142].

SVM models have several interesting and appealing properties, namely the maximum margin classification,

the kernel transformation of the input space into a feature space where a hyperplane may separate the classes, and a unique solution. The SVM introduction generated an enormous interest, comparable only with that produced by the development of artificial neural networks. Many applications were investigated in a short period of time, simultaneous with developments of novel SVM algorithms, such as least squares SVM [126,127] and other kernel algorithms [21,58,119]. The theory and applications of SVM are presented in a number of books, including *Learning with Kernels* by Schölkopf and Smola [115], *Learning Kernel Classifiers* by Herbrich [53], *An Introduction to Support Vector Machines* by Cristianini and Shawe-Taylor [29], and *Advances in Kernel Methods: Support Vector Learning* by Schölkopf, Burges, and Smola [117]. SVM have numerous applications in drug design, from screening of chemical libraries to QSAR [70], with accurate predictions that frequently surpass those obtained with more established algorithms. In Sect. “[Support Vector Machines](#)” we review the most important aspects of SVM models and we present several applications to drug design, SAR and QSAR.

With an ever-increasing number of machine learning algorithms available, it is difficult to select those methods that have better chances of solving a particular SAR or QSAR problem. To address this topic, we review in Sect. “[Comparative Studies](#)” several papers that compare diverse ML models for drug design and structure-activity relationships. A comparison of 21 machine learning algorithms is presented for data selected from the National Cancer Institute 60-cell line screening panel (NCI-60) [69]. The structure-anticancer activity models are developed for three cell lines, namely for lung large cell carcinoma NCI-H460, glioma SF-268, and melanoma SK-MEL-5. Finally, we present an assessment of the results obtained in the machine learning competition CoEPrA 2006 (Comparative Evaluation of Prediction Algorithms, <http://www.coepra.org/>).

Decision Trees

A decision tree represents a series of rules that perform a mapping of the structural descriptors to a prediction for a molecular property. The class of decision tree algorithms includes the C4.5 decision tree [108], the NBTree decision tree with naïve Bayes classifiers at the leaves [82], the alternating decision tree ADTree [42], random trees and random forests [22]. A typical decision tree algorithm, such as C4.5 and related methods, generates a sequence of rules based on splitting the objects into two subgroups based on a selected structural descriptor. A threshold of the descriptor separates the objects into a subgroup of objects

that have a descriptor value lower than the threshold, and another subgroup of objects that have a descriptor value higher than the threshold. The descriptor and the threshold are selected such as to maximize the difference in entropy, which maximizes the separation of objects based on their class, i. e., one subgroup contains mainly object from one class whereas the second subgroup contains mainly objects from the other class. The process is repeated until a proper separation in classes is obtained for all objects.

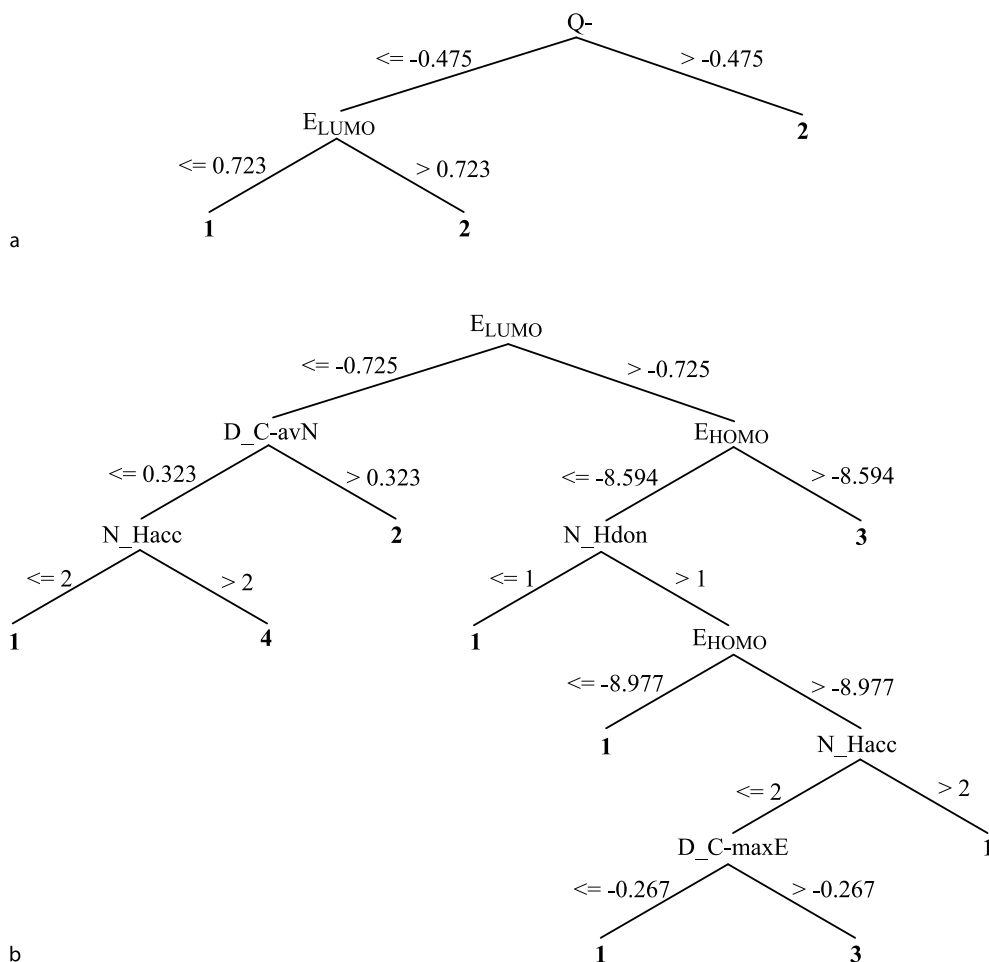
The structure of a C4.5 decision tree is demonstrated for the classification of narcotic pollutants into polar chemicals (class 1) and nonpolar chemicals (class 2) [66,109,137]. The dataset consists of 190 compounds, from which 76 are polar and 114 are nonpolar. The structural descriptors are the octanol-water partition coefficient $\log K_{ow}$, the energy of the highest occupied molecular orbital E_{HOMO} , the energy of the lowest unoccupied molecular orbital E_{LUMO} , the most negative partial charge on any non-hydrogen atom in the molecule Q^- , and the most positive partial charge on a hydrogen atom Q^+ . The J48 (C4.5) decision trees computed with Weka [41,151] shows that only two descriptors, namely Q^- and E_{LUMO} , can separate the leaning dataset (Fig. 2a). The first rule of this decision tree separates molecules based on their Q^- value, namely those with $Q^- \leq -0.475$ are sent to the second rule whereas those with $Q^- > -0.475$ are classified in class 2. The second rule separates chemicals according to their E_{LUMO} value, namely those with $E_{LUMO} \leq 0.723$ are classified in class 1 whereas those with $E_{LUMO} > 0.723$ are classified in class 2.

The second example considers a J48 (C4.5) decision trees with four classes (Fig. 2b) representing four modes of toxic action of phenols in the *Tetrahymena pyriformis* assay, namely polar narcotics (class 1), oxidative uncouplers (class 2), proelectrophiles (class 3), and soft electrophiles (class 4). The dataset consists of 220 chemicals and seven structural descriptors, namely $\log K_{ow}$, E_{HOMO} , E_{LUMO} , the maximum donor (electrophilic) delocalizability for C atoms D_{C-maxE} , the average acceptor (nucleophilic) delocalizability for C atoms D_{C-avN} , the hydrogen bond donor count N_{Hdon} , and the hydrogen bond acceptor count N_{Hacc} . Out of the seven descriptors, only $\log K_{ow}$ is not selected to form a splitting rule, whereas E_{HOMO} and N_{Hacc} are selected in two rules each. Decision trees are efficient ML algorithms that provide accurate and fast predictions, with the added benefit of performing also a selection of the most useful descriptors.

Combinatorial chemistry and high-throughput screening accelerate the drug discovery process by producing vast quantities of biological assay results. The data

mining of all these chemical and biological experiments requires faster structure-activity algorithms, as shown by Rusinko et al for a library of monoamine oxidase inhibitors that was evaluated with recursive partitioning [112]. The advantage of RP is that it scales linearly with the number of descriptors, and it can be used for datasets containing more than 10^5 chemicals and 2×10^6 structural descriptors. To generate chemical libraries focused on a specific target, Deng et al developed the structural interaction fingerprints, a class of descriptors that encode ligand-target binding interactions [31]. These fingerprints were used in a decision tree model to identify ligands for MAP kinase p38, showing that the method can be used with success for the structure-based screening of combinatorial chemical libraries. Statistical studies of FDA approved drugs show that there are certain physico-chemical properties that define drug-like structural requirements, independent of the drug targets. Drug-like filters are used to eliminate at early stages those compounds that do not have the general structural features of a drug. Schneider et al applied decision trees to the classification of drug-like compounds based on drug-like indices, hydrophobicity, and molar refractivity [114]. This simple filter correctly identifies 83% of drugs and 39% of the non-drugs. The classification trees suggest several properties that separate drugs from non-drugs, such as a molecular weight higher than 230, a molar refractivity higher than 40, as well as the presence of structural elements such as rings and functional groups. Natural products are frequently screened to identify active compounds for selected targets. A database of 240 Chinese herbs containing 8264 compounds was screened with a random forest algorithm to identify inhibitors for several targets, including cyclooxygenases, lipoxygenases, aldose reductase, and three HIV targets [36]. The screening results show that random forests give dependable predictions even for unbalanced libraries in which inactive compounds are more numerous than active compounds. A literature search confirmed the computational predictions for 83 inhibitors identified in Chinese herbs.

Torsade de pointes (TdP) is a polymorphic ventricular arrhythmia that may be caused by drugs that induce the prolongation of the QT interval by inhibiting the heart potassium channel hERG (human ether-á-go-go). Gepp and Hutter investigated the TdP potential of 339 drugs, and found that a decision tree has a success rate of up to 80% in predicting the correct class [46]. Ekins et al combined a recursive partitioning (RP) tree with Kohonen and Sammon maps in order to identify hERG inhibitors [38]. RP is used as a fast and accurate filter to screen and prioritize drug databases for an in-depth assessment with Ko-



Drug Design with Machine Learning, Figure 2

Classification with J48 (C4.5) decision trees: **a** polar (class 1) and nonpolar (class 2) narcotic pollutants; **b** toxicity modes of phenols against the *Tetrahymena pyriformis*, namely polar narcotics (class 1), oxidative uncouplers (class 2), proelectrophiles (class 3), and soft electrophiles (class 4)

honen and Sammon maps. The human intestinal absorption of potential drugs may be investigated with in silico methods that assess the intestinal permeability of a chemical compound before synthesis. Hou et al used recursive partitioning to classify the intestinal absorption (poor or good) of chemical compounds [57]. The SAR model based on decision tree classification has good results for a training set of 481 compounds (95.9% success rate for the poor absorption class and 96.1% for the good absorption class) and a prediction set of 98 chemicals (100% success rate for the poor absorption class and 96.8% for the good absorption class).

All drugs that have a CNS (central nervous system) action must penetrate the blood-brain barrier (BBB), and BBB permeability models are used to filter compounds that cannot pass the barrier. Andres and Hutter found that

a decision tree provides fast and accurate results, with an accuracy of 96% for 186 training compounds and 84% for 38 test chemicals [3]. The drugs that pass the BBB are predicted with a higher accuracy (94%) than those that do not pass the BBB (89%). Deconinck et al predicted BBB permeability with classification and regression trees (CART), alone and aggregated in a boosting approach [30]. The training was performed for 147 drugs, and the structural descriptors were computed with Dragon [133]. The percentage of correctly classified drugs in cross-validation is 83.6% for a single tree, and increases to 94% for a boosting model that contains 150 trees. The single tree model may be used for a fast screening of large libraries, but for more reliable predictions a boosting model should be preferred.

The drug affinity for the cytochrome P450 2C9 was predicted with a consensus method based on four ML al-

gorithms [59]. Two of these algorithms are variants of recursive partitioning, the third is a local interpolation method, and the fourth is a subgraph search method. These four models were trained with a set of 276 compounds, and then the SAR models were tested for a prediction set of 50 chemicals. The consensus predictions have an accuracy of 94%, demonstrating that the computational methods may predict which drugs are metabolized by P450 2C9. The metabolic stability of 161 drugs against six isoforms of the cytochrome P450 (1A2, 2C9, 2C19, 2D6, 2E1, and 3A4) was evaluated with a classification tree [154]. The SAR models identified several structural characteristics that determine the P450 specificity, namely 3A4 substrates are larger compounds, 2E1 are smaller molecules, 2C9 substrates are anionic, and 2D6 substrates are cationic.

P-glycoprotein (Pgp) mediated drug efflux is responsible for the low cellular accumulation of anticancer drugs, for reduced oral absorption, and for low blood-brain barrier penetration. Pgp affects also the hepatic, renal, or intestinal elimination of drugs. Li et al. proposed a decision tree model to differentiate Pgp substrates from non-substrates [91]. Four-point 3D pharmacophores obtained from the three-dimensional structure of 163 chemical compounds were subsequently used as descriptors for the classification tree. Nine pharmacophores were selected in the decision tree, giving an accuracy of 87.7% for the training set and 87.6% for the prediction set. These pharmacophores highlight the most important structural features of Pgp substrates, and can be used as a fast filter for chemical libraries. The renal clearance of 130 drugs was investigated with Molconn-Z topological indices [76] and recursive partitioning [32]. RP results for the separation of high-clearance compounds from low-clearance compounds show 88% correct predictions for the training set of 130 compounds and 75% correct predictions for a prediction set of 20 compounds.

The prediction power of classification and regression algorithms can be significantly improved by using an ensemble of models that are combined into a final prediction by an averaging or a voting procedure. Such ML algorithms are called ensemble, committee, or jury methods. Tong et al proposed an ensemble method, decision forest (DF), which is based on classification and regression trees [134]. A DF is obtained in four steps: (a) generate a CART; (b) generate a second CART based only on structural descriptors that were not used in the first CART; (c) repeat the first two steps until no more trees can be generated; (d) predict a property of a chemical compound based on all trees generated. DF was used to predict the estrogen receptor binding activity of organic com-

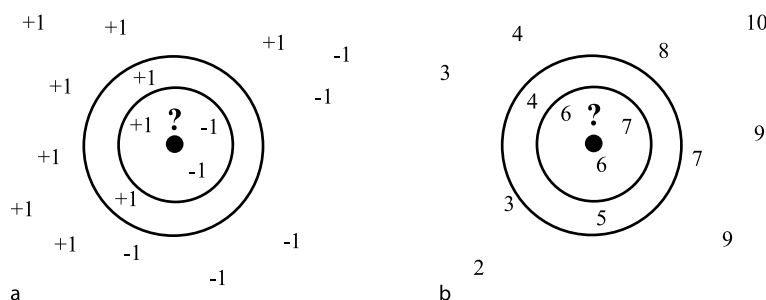
pounds from Molconn-Z topological indices [135]. After 2000 simulations of a 10-fold cross-validation, the predicted accuracy for a dataset of 232 compounds was 81.9%, and the accuracy for another dataset of 1092 chemicals was 79.7%. The DF models can be used to identify potential endocrine disruptors, which are chemicals that affect the endocrine functions. A similar algorithm is the random forest (RF) [22], which was extensively tested for several structure-activity studies [128]. RF has three properties that are important in SAR, namely the algorithm evaluates the importance of descriptors, identifies similar compounds, and measures the prediction performance. The skin sensitization activity of 131 chemicals was modeled with RF, using Dragon and Molconn-Z indices [90]. Compared with single tree models, RF improve significantly the ability to identify chemical allergens.

Lazy Learning and *k*-Nearest Neighbors

Lazy learning is a local learning method that keeps in the memory the entire set of training objects, and computed an interpolation model based on the neighborhood on the prediction object Atkeson, [56]. The prevalent lazy learning algorithm in SAR and QSAR is the *k*-nearest neighbors [1]. A *k*-NN prediction is based on a descriptor space, which may be the entire set of structural descriptors evaluated in a study or a subset that gives better predictions, a distance function, and a neighborhood *k* of the prediction object. The *k*-NN classification may be applied to two or more classes (Fig. 3a). The prediction object, shown here as a black circle with an unknown class “?”, is inserted into the descriptor space and then its *k* nearest neighbors are identified. The class of the prediction object is the class of the majority of its neighbors. In the example from Fig. 3a, the predicted class is -1 if $k = 3$, and $+1$ if $k = 5$. The prediction statistics of the *k*-NN classifier depend on the value of *k*, and its optimum value is usually determined by cross-validation.

A similar procedure is used for *k*-NN regression (Fig. 3b). For regression, each training object has a real number as property value, and the predicted property for the query object is an average value of the property values for its *k* nearest neighbors. For the situation depicted in Fig. 3b, the predicted value for $k = 3$ is $(6+6+7)/3 = 6.3$, whereas the prediction for $k = 5$ is 5.6. A distance weight is usually added to increase the contribution of closer objects and to decrease the influence of more distant objects. For a distance function *d*, a possible weighting of the property values is the inverse distance $1/d$.

Lazy learning applications in drug design are still rare, mainly because the community of practitioners only re-



Drug Design with Machine Learning, Figure 3

Applications of the k -NN algorithm to a classification and b regression

cently discovered the rich field of machine learning algorithms. Kumar et al used locally linear embedding for non-linear dimensionality reduction coupled with lazy learning for two QSAR benchmarks, the Selwood dataset and the steroids dataset originally analyzed with CoMFA [85]. Guha et al applied local lazy regression to three QSAR models, namely to 179 artemisinin analogues with antimalarial activity, to 79 platelet-derived growth factor inhibitors, and to 756 inhibitors of dihydrofolate reductase [47]. In all cases, the local models give better predictions compared to global regression models. An automated lazy learning quantitative structure-activity relationship (ALL-QSAR) approach was developed based on locally weighted linear regression models computed from the compounds in training set that are chemically most similar to a test compound [159]. ALL-QSAR was applied with good results for three datasets, namely 48 anticonvulsant compounds, 48 dopamine receptor antagonists, and to model the toxicity of 250 phenols against *Tetrahymena pyriformis*. Sommer and Kramer reported several applications of lazy learning to chemical datasets with a high structural diversity [123]. Collections of noncongeneric compounds are usually difficult to model, but their approach obtained good results for several datasets, namely 739 mutagenic chemicals from the Carcinogenic Potency Database, 342 chemicals tested for biodegradability, and 1481 from the AIDS Antiviral Screen of the NCI Developmental Therapeutics Program.

The fundamental assumptions in k -NN applications to SAR and QSAR are that molecules with similar structures are close in the descriptor space, and that similar molecules have similar properties. Basak advocated the use of k -NN in QSAR as an alternative to the multiple linear regression [10,11]. In this approach, the chemical compounds are represented with a collection of topological indices, and the Euclidean distance is used to identify the structures most similar with the test molecule. This k -NN method was applied with success to model the hy-

drophobicity of 4067 compounds [10], the mutagenicity of a diverse set of 520 chemicals [11], the boiling temperature of 139 hydrocarbons and the mutagenicity of 95 aromatic and heteroaromatic amines [9]. Good models are obtained with optimized k values over a range between 1 and 25, as shown in a QSAR model for vapor pressure [48]. The descriptors selected in the structural space are essential in establishing a suitable similarity measure. General collections of descriptors usually give good predictions, but property tailored structural spaces can increase significantly the predictivity of the model. Tailored structural spaces include only those descriptors that contribute significantly to the model, as demonstrated for hydrophobicity and boiling temperature [49]. A structural space consisting of electrotopological state indices gives reliable predictions for hydrophobicity, Henry's law constant, vapor pressure, and OH-radical bimolecular rate constant [24].

Tropsha applied k -NN to numerous QSAR models, demonstrating its value in comparisons with more established algorithms. For a dataset of 29 dopamine antagonists, k -NN predictions are much better than those obtained with CoMFA, which represents the standard 3D QSAR model [55]. k -NN may be used also as a robust method to select those structural descriptors that are important in modeling a particular biological property, as shown for 58 estrogen receptor ligands [161] and for 48 amino acid derivatives with anticonvulsant activity [120]. In a more elaborate implementation, k -NN is used to select descriptors, to find the best value for k , and to optimize the function that weights the contribution of its neighbors [60]. The anticancer activity of 157 epipodophylotoxin derivatives was modeled with molecular connectivity indices and k -NN [152]. In a comparison with a CoMFA model obtained for the same dataset it was found that k -NN gives better predictions in a cross-validation test. The metabolic stability of 631 drug candidates was evaluated with k -NN and a collection of structural descriptors that included molecular connectivity indices and atom

pairs [121]. The prediction in a test set of compounds had accuracy higher than 85%, whereas an external validation set of 107 chemicals was predicted with a 83% success rate.

k-NN is also used as a component in more elaborate drug design algorithms. Ajmani et al combined molecular field analysis and *k*-NN to obtain a robust QSAR model that was tested for three datasets, namely the CoMFA steroids, cyclooxygenase-2 inhibitors, and anticancer compounds [2]. Another QSAR model that incorporates *k*-NN is Complimentary Ligands Based on Receptor Information (CoLiBRI) [103]. CoLiBRI represents both ligands and receptor binding sites in the same space of universal chemical descriptors. In a test consisting of 800 X-ray protein-ligand receptors from PDB, CoLiBRI ranked the correct ligands in the top 1% chemicals selected, and was able to quickly eliminate improbable ligands. The skin permeability coefficients of 110 compounds were predicted with an ensemble consisting of a collection of *k*-NN and ridge regression models [99].

The melting temperature of organic compounds is particularly difficult to model, due mainly to a lack of good descriptors for the interactions in the solid and liquid states. Nigsch et al used several types of molecular descriptors to model with *k*-NN the melting temperature of 4119 diverse organic molecules with [101]. The contribution of selected neighbors was evaluated with four weighting functions (arithmetic and geometric average, inverse distance weighting, and exponential weighting), with better results obtained by using the exponential weighting scheme. The optimized model has an average error of 42.2 °C. In a comparison between global and local (*k*-NN) regression models for the blood-brain distribution of 328 chemicals, it was found that *k*-NN consistently gives better predictions as measured in cross-validation [84].

Bayesian Methods

Although Bayesian classifiers were only recently applied to drug design and structure-activity studies, they have several characteristics that make them particularly useful in screening large chemical libraries. Bayesian classifiers give dependable predictions, can tolerate noise and errors in experimental data, and can be computed fast compared to other ML algorithms, which is particularly important in screening large collections of molecules. Bayes' theorem describes the relationship between the prior and conditional probabilities of two events *A* and *B*, when *B* has a non-zero probability [12]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where the terms have the following meaning: *P*(*A*) is the prior or marginal probability of *A*, and does not consider any information regarding *B*; *P*(*B*) is the prior (marginal) probability of *B*; *P*(*A*|*B*) is the conditional (posterior) probability of *A* given the occurrence of event *B*; *P*(*B*|*A*) is the conditional (posterior) probability of *B* given the occurrence of event *A*. Bayesian classifiers [71,151] evaluate a set of descriptors to predict the activity of a chemical based on the known descriptors and activities of a training set of chemicals. Each descriptor is considered to be statistically independent of all other descriptors. Based on a selected descriptor, the probability that a compound is active is proportional to the ratio of active to inactive compounds that have the same value for that descriptor. The final prediction is obtained by multiplying the descriptor-based probabilities.

A variety of drug discovery and development problems can be solved efficiently with Bayesian classifiers. The virtual screening of large chemical libraries represents a major application of Bayesian classifiers, mainly due to their tolerance to noise in experimental data [146]. Klon and Diller proposed a method to prioritize chemicals for synthesis and screening based on physico-chemical properties [77]. It was found that the Bayesian classifier is superior to the usual pairwise Tanimoto similarity filter. High-throughput docking is an essential tool for structure-based drug design, but the current scoring functions do not provide an accurate ranking of the molecules. Various improvements to the scoring functions were proposed, including Bayesian classifiers [78,79]. Simulations with two biological targets showed that a Bayesian classifier increases the enrichment in active compounds. Finding the potential targets of chemicals is increasingly relevant in drug discovery. Nidhi et al developed a multi-class Bayesian model in which chemicals represented with extended-connectivity fingerprints are evaluated against 964 target classes [100]. The practical utility of the approach is demonstrated by its high success rate, with 77% correct target identification. Adverse drug reactions are investigated during early phases of drug discovery in preclinical safety pharmacology tests for various targets. Multi-class Bayesian classifiers developed for 70 such targets can identify 93% of ligands with a 94% correct classification rate [14]. In a test using chemicals from World Drug Index, the classifier identifies 90% of the adverse drug reactions with a 92% correct classification rate. Adverse reactions of several drugs withdrawn from the market were also predicted with good accuracy.

Watson used 2D pharmacophore feature triplet vectors and Bayesian classifiers to find active compounds in library screening [150]. The good enrichment in active

compounds obtained for several chemogenomic databases demonstrates the practical utility of this approach. Natural products represent an excellent source of starting structures for the design of novel drugs. Ertl trained a Bayesian classifier to provide the natural product-likeness score for a chemical [39]. This score measures the similarity between a molecule and the structural space covered by natural products, and can separate between natural products and synthetic molecules. Usual similarity search in chemical libraries is based on a set of structural descriptors and a similarity (or distance) measure. Bender et al showed that the retrieval rate of active compounds increases considerably by using Bayes affinity fingerprints that describe the ligand bioactivity space [13]. In this space, Bayes scores for ligands from about 1000 activity classes describe the chemical structures. This approach was able to improve with 24% the results from a standard similarity screening.

Bayesian methods are also used in various structure-activity models. Klon et al demonstrated several applications of modified Bayesian classifiers that model continuous numerical data with a Gaussian distribution [80]. Several models developed for absorption, distribution, metabolism and excretion property prediction show that the new classifiers have better performance compared to classical Bayesian classifiers. Multidrug resistance represents the ability of cancer cells to become simultaneously resistant to several drugs. A possible solution to this problem is the use of multidrug resistance reversal (MDDR) agents. Sun found that a Bayesian classifier based on atom types could be used to identify MDDR agents in a set of 424 training compounds [125]. The prediction of MDDR agents in a group of 185 test compounds had a success rate of 82.2%. Phospholipidosis is an adverse effect that is investigated during preclinical studies in animals, and may stop the development of an otherwise promising drug candidate. Pelletier et al proposed to identify phospholipidosis with an *in silico* Bayesian classifier as a viable alternative to more expensive and time-consuming *in vivo* tests [105]. Based on two simple descriptors, pK_a and ClogP, the model has a success rate of 83% for a dataset of 201 compounds. UGT is an enzyme family that catalyzes the reaction between glucuronic acid and a chemical that has a nucleophilic group. Drug glucuronidation represents an important mechanism for their elimination, but the selectivity of different UGT isoforms is difficult to predict. The Bayesian classifier suggested by Sorich et al predicts the glucuronidation site for eight UGT isoforms based on the partial charge and Fukui function of the nucleophilic atom and the presence of an aromatic ring attached to the nucleophilic atom [124]. The models have a good predic-

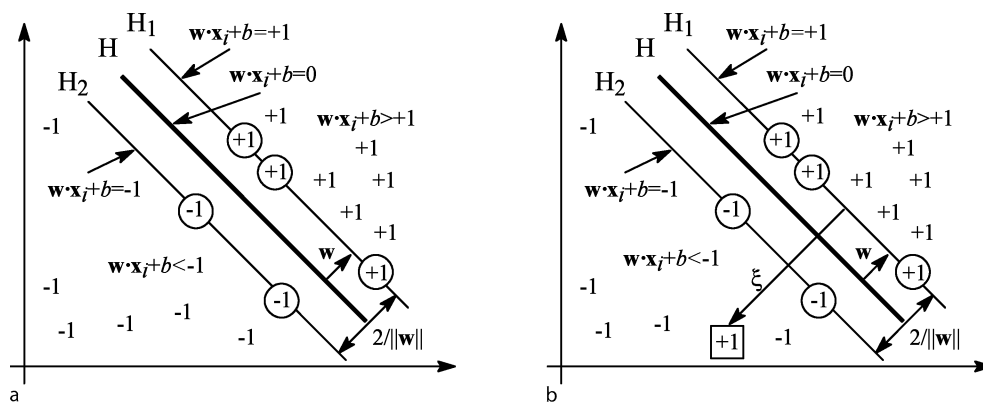
tive ability, with a sensitivity and specificity in the range 75–80%. The screening test to identify bitter compounds presented by Rodgers et al uses selected substructural features identified from 649 bitter chemicals and 13,530 random drug-like compounds [111]. A Bayesian classifier trained with these data can predict correctly 72.1% of the bitter compounds. The bitter taste is determined mainly by substructures representing sugar moieties and highly branched carbon skeletons.

Support Vector Machines

Support vector machines represent a major development for SAR and QSAR models, as suggested by the large number of publications that apply SVM and related kernel methods to drug design. To better understand the mathematical basis of SVM and the parameters that influence their results, we start this section with a brief theoretical presentation of SVM for classification and regression. Several kernel functions are presented together with SVM plots that demonstrate the complex shapes that may be simulated with these functions. The influence of various kernels on QSAR predictions is shown for a dataset of benzodiazepine receptor ligands, and several applications in drug design are reviewed.

Hard Margin Linear SVM Classification

The most simple case of SVM is the classification of two classes of objects that may be separated by a hyperplane (Fig. 4a). It is obvious from this figure that there are many hyperplanes that can separate the two classes of objects, but not all of them give good predictions for novel objects. Intuitively, a hyperplane that maximizes the separation of the two classes, like hyperplane H, is expected to offer the best predictions. The SVM algorithm determines a unique hyperplane H that has the maximum margin, i. e., the maximum distance between hyperplanes H_1 and H_2 . Hyperplane H_1 , which is determined by three objects from class +1 that are shown inside circles, defines the border with class +1. Hyperplane H_2 , which is determined by two objects from class -1 that are shown also inside circles, defines the border with class -1. The objects represented inside circles are called support vectors and they determine the unique solution for the SVM model. By removing all other objects and keeping only the support vectors one obtains the same SVM model. The hyperplane H is defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$, where \mathbf{w} is the normal vector to H, the hyperplane H_1 that borders class +1 is defined by $\mathbf{w} \cdot \mathbf{x} + b = +1$, and the hyperplane H_2 that borders class -1 is defined by $\mathbf{w} \cdot \mathbf{x} + b = -1$. The distance between the origin and the hyperplane H is $|b|/\|\mathbf{w}\|$, and the SVM



Drug Design with Machine Learning, Figure 4

SVM classification models: a linearly separable data; b linearly non-separable data

margin (the distance between hyperplanes H_1 and H_2) is $2/\|\mathbf{w}\|$. A molecule i characterized by a vector of structural descriptors \mathbf{x}_i belongs to class +1 if $\mathbf{w} \cdot \mathbf{x}_i + b \geq +1$, or to class -1 if $\mathbf{w} \cdot \mathbf{x}_i + b \leq -1$. Obviously, support vectors from class +1 are determined by $\mathbf{w} \cdot \mathbf{x}_i + b = +1$, whereas support vectors from class -1 are determined by $\mathbf{w} \cdot \mathbf{x}_i + b = -1$.

Based on the formula of the SVM margin, it follows that the maximum separation hyperplane is obtained by maximizing $2/\|\mathbf{w}\|$, which is equivalent to minimizing $\|\mathbf{w}\|^2/2$. The linear SVM model (Fig. 4a) is formulated as:

$$\begin{aligned} \text{minimize } f(\mathbf{x}) &= \frac{\|\mathbf{w}\|^2}{2} \\ \text{with the constraints } g_i(\mathbf{x}) &= y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \\ & i = 1, \dots, m \end{aligned} \quad (2)$$

where y_i is the class (+1 or -1) of the object i . This optimization equation is a quadratic programming that is further transformed with a Lagrangian function into its dual formulation. All SVM formulations are solved for the dual formulation, because the solution obtained for this simple case is easily extended to more complex situations. The minimization problem from Eq. (2) is expressed with a Lagrangian function:

$$\begin{aligned} L_P(\mathbf{w}, b, \mathbf{A}) &= f(\mathbf{x}) + \sum_{i=1}^m \alpha_i g_i(\mathbf{x}) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i (y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + \sum_{i=1}^m \alpha_i \end{aligned}$$

$$\begin{aligned} &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i y_i \mathbf{w} \cdot \mathbf{x}_i \\ &\quad - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i, \end{aligned} \quad (3)$$

where $\mathbf{A} = (\alpha_1, \alpha_2, \dots, \alpha_m)$ is the set of Lagrange multipliers for the objects that have $\alpha_i \geq 0$, and P in L_P indicates the primal formulation of the problem. The Lagrangian function L_P must be minimized with respect to \mathbf{w} and b , and maximized with respect to α_i , subject to the constraints $\alpha_i \geq 0$. The dual optimization problem L_D is:

$$\begin{aligned} \text{maximize} \\ L_D(\mathbf{w}, b, \mathbf{A}) &= \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{subject to } \alpha_i &\geq 0, \quad i = 1, \dots, m \\ \text{and } \sum_{i=1}^m \alpha_i y_i &= 0. \end{aligned} \quad (4)$$

The optimization problem is usually solved with the sequential minimal optimization (SMO) proposed by Platt [106]. For a complete formulation of the SVM solution see the books and reviews mentioned in Introduction [70]. The training objects with $\alpha > 0$ constitute the support vectors, whereas the objects with $\alpha = 0$ may be removed from the learning set without affecting the SVM model. The vector \mathbf{w} is obtained as:

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \quad (5)$$

and b is computed as the average value for all support vectors. The SVM model is the optimum separation hyperplane (\mathbf{w}, b) that can now be used to predict the class mem-

bership for new objects. The class of a molecule k with the structural descriptors \mathbf{x}_k is:

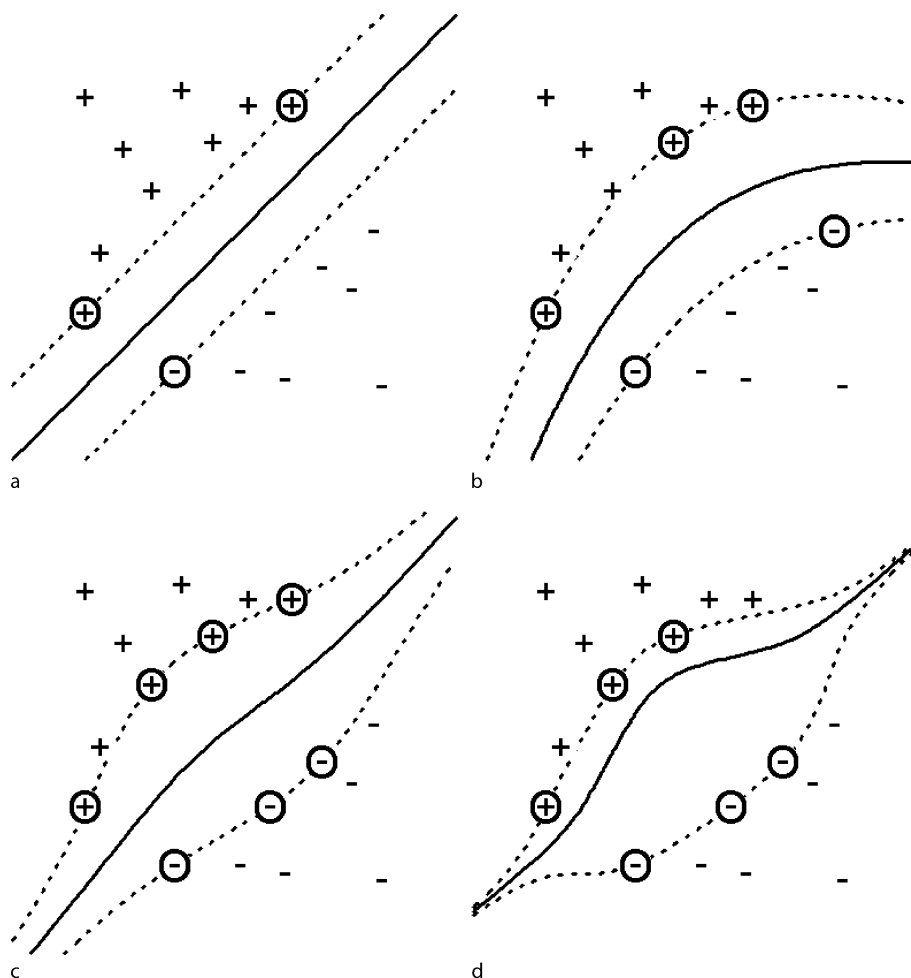
$$\text{class}(k) = \begin{cases} +1 & \text{if } \mathbf{w} \cdot \mathbf{x}_k + b > 0 \\ -1 & \text{if } \mathbf{w} \cdot \mathbf{x}_k + b < 0 \end{cases} \quad (6)$$

The classifier is further simplified by substituting \mathbf{w} with its expression (5):

$$\text{class}(k) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x}_k + b \right), \quad (7)$$

where the summation goes only over the support vectors. The SVM described in the above equations is called a “hard margin” SVM because it does not allow for classification errors.

The class separation with a linear SVM is simple and intuitive. However, SVM are usually used with nonlinear kernels that have a much more complex shape, which might result in too complicated separation surfaces. To demonstrate the shape of the separation surface for linearly separable data we compare results obtained with a linear kernel and several nonlinear kernels that will be introduced later. All calculations were performed with R (<http://www.r-project.org/>) and the kernlab package. In all figures, class +1 patterns are represented by “+” and class -1 patterns are represented by “-”. The SVM hyperplane is depicted with a continuous line, whereas the margins of the SVM hyperplane are shown with dotted lines. Support vectors from class +1 are represented as “+” inside a circle, and support vectors from the class -1 are depicted as “-” inside a circle.



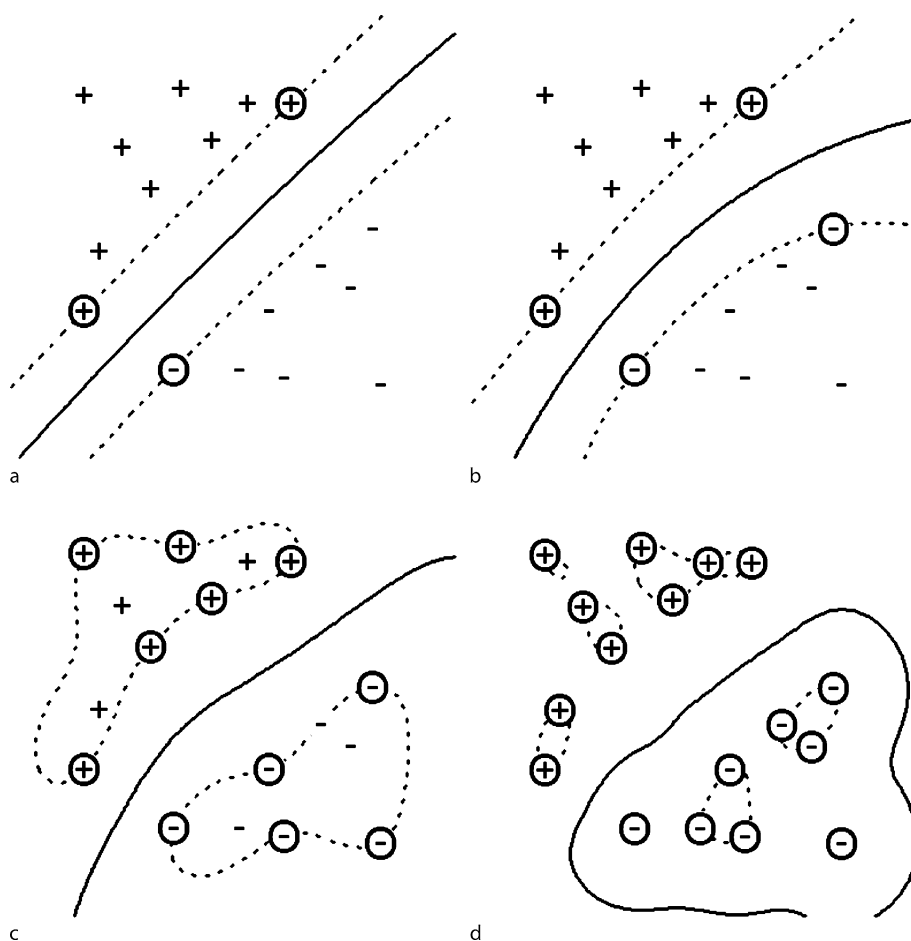
Drug Design with Machine Learning, Figure 5

SVM classification models for linearly separable data: a dot kernel (linear); b polynomial kernel, degree 2; c polynomial kernel, degree 3; d polynomial kernel, degree 10

The linear kernel gives an SVM model that has an optimum separation of the two classes (Fig. 5a) with three support vectors, namely two from class +1 and one from class -1. The hyperplane has the maximum width, and no object is situated inside the margins (represented with dotted lines). The classification of new objects is made by applying Eq. (7) to the three support vectors. The same dataset is modeled with a degree 2 polynomial kernel (Fig. 5b) in which case the SVM model has five support vectors, namely three for class +1 and two for class -1. The margin width varies, being larger in the middle and smaller towards the extremes. The hyperplane topology is different from that obtained with a linear kernel, and obviously for some objects the two SVM models will give different predictions. The SVM classifier obtained with a degree 3 polynomial kernel (Fig. 5c) has four support vectors from class +1 and three support vectors from class -1, and the margin becomes smaller towards the ends. Higher order poly-

nomial kernels produce a complex separation hyperplane, as shown also for a degree 10 polynomial kernel (Fig. 5d), in which case the margin almost vanished towards the extremes.

The Gaussian radial basis function (RBF) kernel is the most common option for SVM, and in Fig. 6 we present four SVM models obtained for different σ values. For low σ values the RBF kernel (Fig. 6a, $\sigma = 0.01$) approximates a linear kernel (Fig. 5a) and the two hyperplanes are very similar. As σ increases the nonlinear behavior of the kernel becomes apparent (Fig. 6b, $\sigma = 0.1$), but the SVM model still has a good separation of the data. As σ increases to 1 (Fig. 6c) and then to 5 (Fig. 6d) the nonlinearity of the SVM model becomes more apparent, and the number of support vectors increases. In the last case all data points are support vectors, and the SVM model has a very complex topology. The examples presented here show that the RBF kernel may generate too complex hy-



Drug Design with Machine Learning, Figure 6

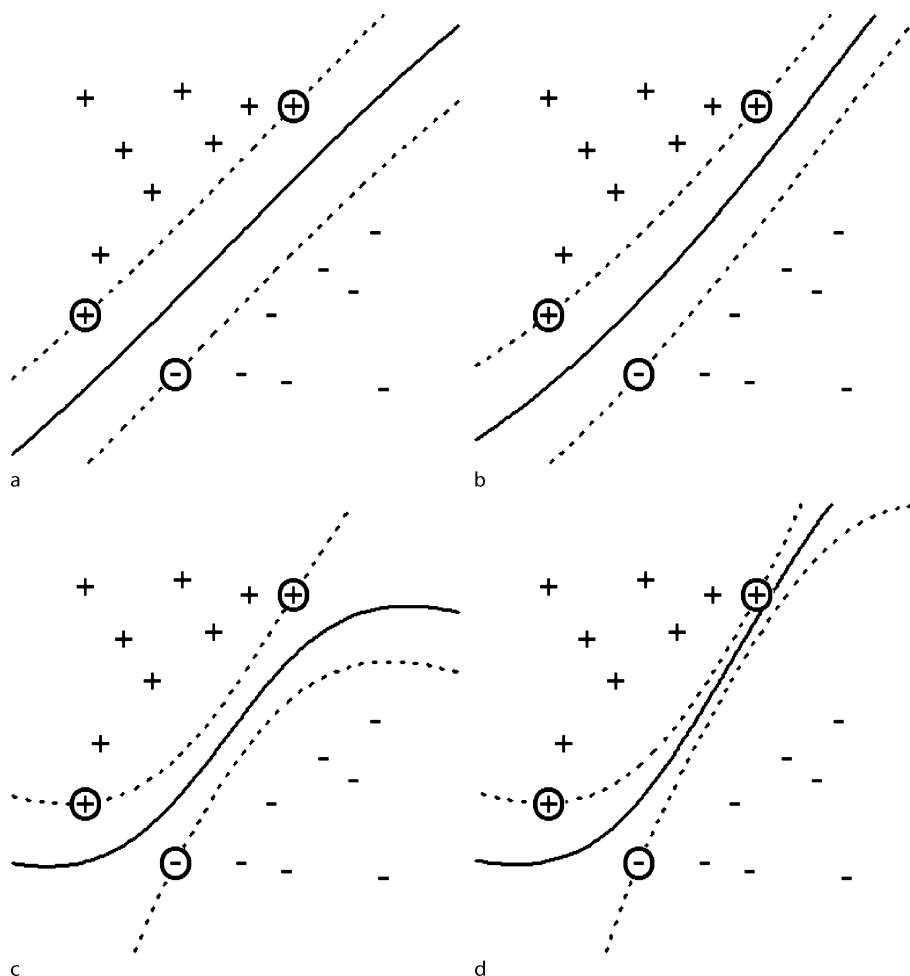
SVM classification models for linearly separable data with the Gaussian RBF kernel: a $\sigma = 0.01$; b $\sigma = 0.1$; c $\sigma = 1$; d $\sigma = 5$

perplanes that do not properly reflect the structure of the data.

Another important kernel is the hyperbolic tangent (\tanh), which is also used in artificial neural networks. Some combinations of parameters result in SVM models almost identical with those obtained with the linear kernel (Fig. 7a) or similar (Fig. 7b). Higher values for the two parameters result in SVM models with smaller margins and nonlinear separation of the two classes (Fig. 7c,d). The SVM models demonstrated here for linearly separable data show that nonlinear kernels, such as RBF and \tanh , may give SVM models similar to those obtained with linear kernels, for certain values of their parameters. However, nonlinear kernels may also give SVM models that are too complex and that do not follow the structure of the data. As a general rule, the SVM computed with nonlinear kernels should be compared with a linear kernel SVM.

Soft Margin Linear SVM Classification

The classification problem solved in the previous section considered a two-class situation with classes separated with a linear classifier. However, such cases are rare in drug design applications, and most often the classes of compounds have regions where they superpose. There are several reasons why classes of chemicals cannot be separated with a linear classifier. The most common problem is the identification of the structural descriptors that determine the property. There are thousands of structural descriptors proposed in the literature, but it is difficult to try them all in a SAR or QSAR study. It is always a possibility that the proper type of descriptors is not even discovered. The second problem is to identify the subset of descriptors that give the best predictions, which is a time-consuming task that is usually solved with heuristic al-



Drug Design with Machine Learning, Figure 7

SVM classification models for linearly separable data with the hyperbolic tangent (\tanh) kernel: a $a = 0.1, b = 0$; b $a = 0.1, b = 0.5$; c $a = 0.5, b = 0$; d $a = 0.5, b = 0.5$

gorithms that give a sub-optimal solution. The mapping between the descriptor space and the property might be nonlinear, and obviously a linear classifier fails to model the data. As a final point we have to mention that experimental data may be affected by noise, measurement errors, interference with other biological targets, or unaccounted factors that determine the outcome of the experiments. Although the hard margin linear SVM is limited in its abilities to model real data, the formalism presented in the previous section is the basis for developing more complex SVM models. The linear SVM with soft margin is such an extension that allows classification errors.

The classification problem shown in Fig. 4b is very similar with the linear separable case from Fig. 4a, because all but one object can be separated with a linear hyperplane H . The exception is the object from class +1 that is situated in the -1 space region (shown inside a square). A penalty term is introduced for misclassified data, with the property that it is zero for objects correctly classified, and it has a positive value for objects that are not situated on the correct side of the classifier. The value of the penalty increases when the distance to the corresponding hyperplane increases. The penalty is called a slack variable and it is denoted with ξ . For the situation depicted in Fig. 4b the penalty for the misclassified object is measured starting from hyperplane H_1 because this hyperplane border the +1 region. For +1 molecules situated in the buffer zone between H and H_1 , and for -1 molecules situated in the buffer zone between H and H_2 , the slack variable takes values between 0 and 1. Such patterns are not considered to be misclassified, but have a penalty added to the objective function. The constraints of the objective function for soft margin linear SVM are:

$$\begin{cases} \mathbf{w} \cdot \mathbf{x}_i + b \geq +1 - \xi_i & \text{if } y_i = +1 \\ \mathbf{w} \cdot \mathbf{x}_i + b \leq -1 + \xi_i & \text{if } y_i = -1 \\ \xi_i > 0, \quad \forall i. \end{cases} \quad (8)$$

The soft margin SVM should balance two opposite conditions, namely a maximum margin and a minimum of errors that is equivalent with minimizing a sum of slack variables. The optimization problem for a linear SVM with classification errors is;

$$\begin{aligned} &\text{minimize} \quad \frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^m \xi_i \\ &\text{with the constraints} \quad (9) \\ &\quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq +1 - \xi_i, \quad i = 1, \dots, m \\ &\quad \xi_i \geq 0, \quad i = 1, \dots, m \end{aligned}$$

where C is called capacity and is a parameter that weights the penalty for classification errors. A large value for the capacity parameter C represents a high penalty for classification errors, which forces the SVM solution to minimize the number of misclassified objects by reducing the margin of the classifier. A small capacity C lowers the value of the penalty term and maximizes the margin, which makes the solution less sensitive to classification errors. The optimization problem is solved with Lagrange multipliers, similar with the procedure outlined for the hard margin linear SVM, when the primal Lagrangian expression is:

$$\begin{aligned} L_P(\mathbf{w}, b, \mathbf{A}, \mathbf{M}) = & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] \\ & - \sum_{i=1}^m \mu_i \xi_i \end{aligned} \quad (10)$$

where the Lagrange multipliers $\mathbf{A} = (\alpha_1, \alpha_2, \dots, \alpha_m)$ are assigned to each constraint $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq +1 - \xi_i$, and the Lagrange multipliers $\mathbf{M} = (\mu_1, \mu_2, \dots, \mu_m)$ are assigned to each constraint $\xi_i \geq 0, \forall i = 1, \dots, m$. The dual optimization problem is:

$$\begin{aligned} &\text{maximize} \\ L_D(\mathbf{w}, b, \mathbf{A}, \mathbf{M}) &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ &\text{subject to } 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m \\ &\text{and } \sum_{i=1}^m \alpha_i y_i = 0. \end{aligned} \quad (11)$$

The expression for the vector \mathbf{w} is identical with that obtained for hard margin SVM:

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i. \quad (12)$$

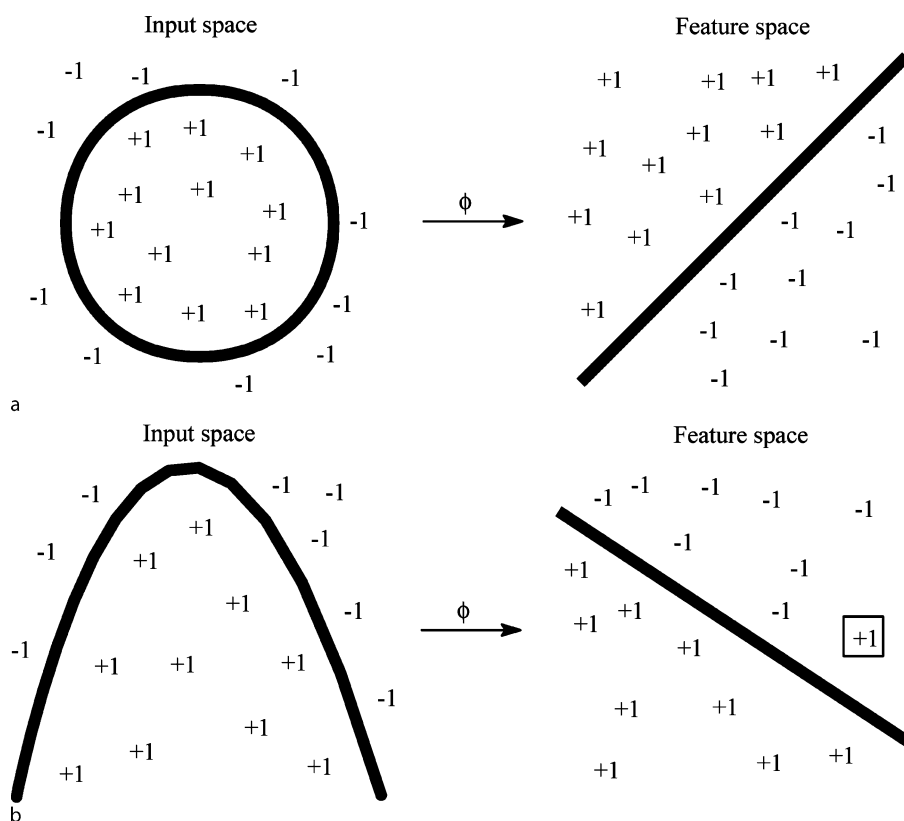
The SVM classifier, which is determined by the pair (\mathbf{w}, b) , may be used for the classification of new molecules. The class of a molecule k with the structural descriptors \mathbf{x}_k is:

$$\text{class}(k) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x}_k + b \right) \quad (13)$$

where the summation goes only over the support vectors.

Nonlinear Classification with Kernels

Nonlinear relationships between structural descriptors and drug properties may be modeled with SVM equations that replace the linear classifier with nonlinear kernel functions. The SVM extension to nonlinear relationships is



Drug Design with Machine Learning, Figure 8

Patterns that are nonlinear separable in the input space are linear separable in the feature space: **a** the class separation is complete; **b** a pattern from class +1 (depicted in a square) is situated in the -1 region of the hyperplane

based on the property of feature functions $\phi(\mathbf{x})$ that transform the input spaces into a feature space of higher dimensionality, in which a linear separation of the classes is possible. This transformation is depicted in Fig. 8a. In the input space, which is represented by the structural descriptors, a circle can discriminate the two classes of objects. Using a proper feature function ϕ the input space is transformed into a feature space in which the two classes can be separated with a linear classifier. Figure. 8b illustrates another situation in which another nonlinear function can discriminate the classes in input space, which is then transformed by the function ϕ into a linear classifier in feature space. The challenge is to find that feature function ϕ that maps the objects into a feature space in which the classes are linearly separable. In real-life applications it is not even desirable to achieve a perfect separation, because the function may map noise or errors in data (see Fig. 8b, where a +1 object, depicted in a square, is situated in the -1 region of the hyperplane). Due to the fact that in feature space the objects may be separated with a linear classifier, it is possible to use the results for linear SVM

presented in the previous sections, which is a fundamental property of SVM.

The mapping from the input space to the feature space may be expressed as:

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \\ \rightarrow \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_h(\mathbf{x})) \quad (14)$$

A soft margin SVM in the feature space is obtained by simply substituting the input vectors \mathbf{x} with the feature vectors $\phi(\mathbf{x})$. We have to point here that a good linear classifier in feature space depends on a proper form of the function ϕ . The class of a molecule k with the structural descriptors \mathbf{x}_k is:

$$\text{class}(k) = \text{sign}[\mathbf{w} \cdot \phi(\mathbf{x}_k) + b] \\ = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k) + b \right) \quad (15)$$

The prediction for a molecule k requires the computation of the dot product $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k)$ for all support vectors \mathbf{x}_i .

This is an important observation showing that we do not have to know the actual expression of the feature function ϕ . Furthermore, nonlinear SVM use kernels, which are a special class of functions that compute the dot product $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k)$ in the input space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j). \quad (16)$$

Several popular kernel functions are presented below. The inner product of two vectors defines the linear (dot) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j. \quad (17)$$

The dot kernel is a linear classifier, and should be used as a reference to demonstrate an eventual improvement of the classification with nonlinear kernels. The polynomial kernel is useful mainly for lower degrees, because for higher degrees it has the tendency to overfit the data:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d. \quad (18)$$

The Gaussian radial basis functions (RBF) kernel is very flexible, and depending on the values of the parameter σ it can display a wide range of nonlinearity:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma \|\mathbf{x}_i - \mathbf{x}_j\|^2). \quad (19)$$

The hyperbolic tangent (tanh) function has a sigmoid shape and it is the most used transfer function for artificial neural networks. The corresponding kernel has the formula:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(a\mathbf{x}_i \cdot \mathbf{x}_j + b). \quad (20)$$

The nonlinearity of the anova kernel is controlled by the parameters γ and d :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_i \exp(-\gamma(\mathbf{x}_i - \mathbf{x}_j)) \right)^d. \quad (21)$$

Hard Margin Nonlinear SVM Classification

The formulation of the hard margin nonlinear SVM classification is obtained from the linear SVM by substituting input vectors \mathbf{x} with feature functions $\phi(\mathbf{x})$. The dual problem is:

maximize

$$L_D(\mathbf{w}, b, \mathbf{A}) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$$

subject to $\alpha_i \geq 0, \quad i = 1, \dots, m$

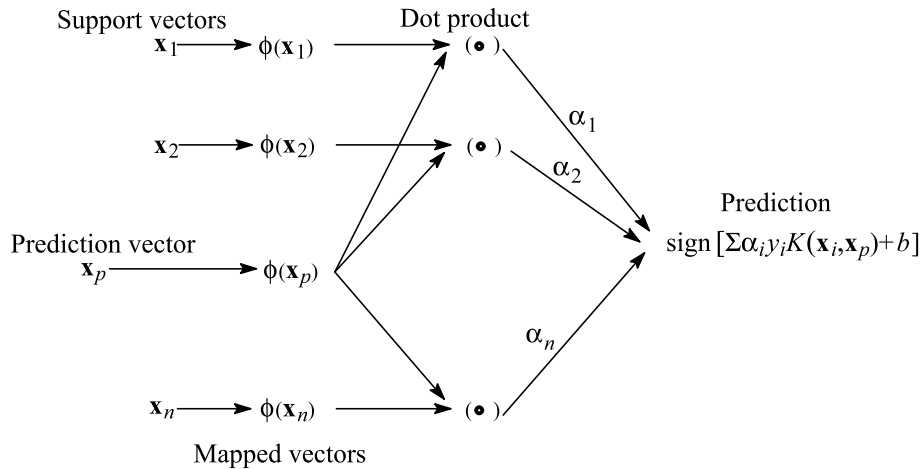
$$\text{and } \sum_{i=1}^m \alpha_i y_i = 0. \quad (22)$$

The optimum separation hyperplane is determined by the vector \mathbf{w} :

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \phi(\mathbf{x}_i). \quad (23)$$

In the final expression of the classifier, the dot product for two feature functions $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k)$ is replaced with a kernel function $K(\mathbf{x}_i, \mathbf{x}_k)$:

$$\text{class}(k) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_k) + b \right). \quad (24)$$



Drug Design with Machine Learning, Figure 9
Network representation of support vector machines

The structure of a nonlinear SVM may be represented in a network format (Fig. 9). The network input is represented by a set of support vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ and a prediction object \mathbf{x}_p . The input space is transformed with a feature function ϕ into the feature space, and the dot product of feature functions is computed and multiplied with the Lagrangian multipliers α . The output is Eq. (24) in which the dot product of feature functions is substituted with a kernel function K .

Soft Margin Nonlinear SVM Classification

The expression of the soft margin nonlinear SVM classification is obtained from the corresponding linear SVM formula by substituting input vectors \mathbf{x} with feature functions

$\phi(\mathbf{x})$. The dual problem is:

maximize

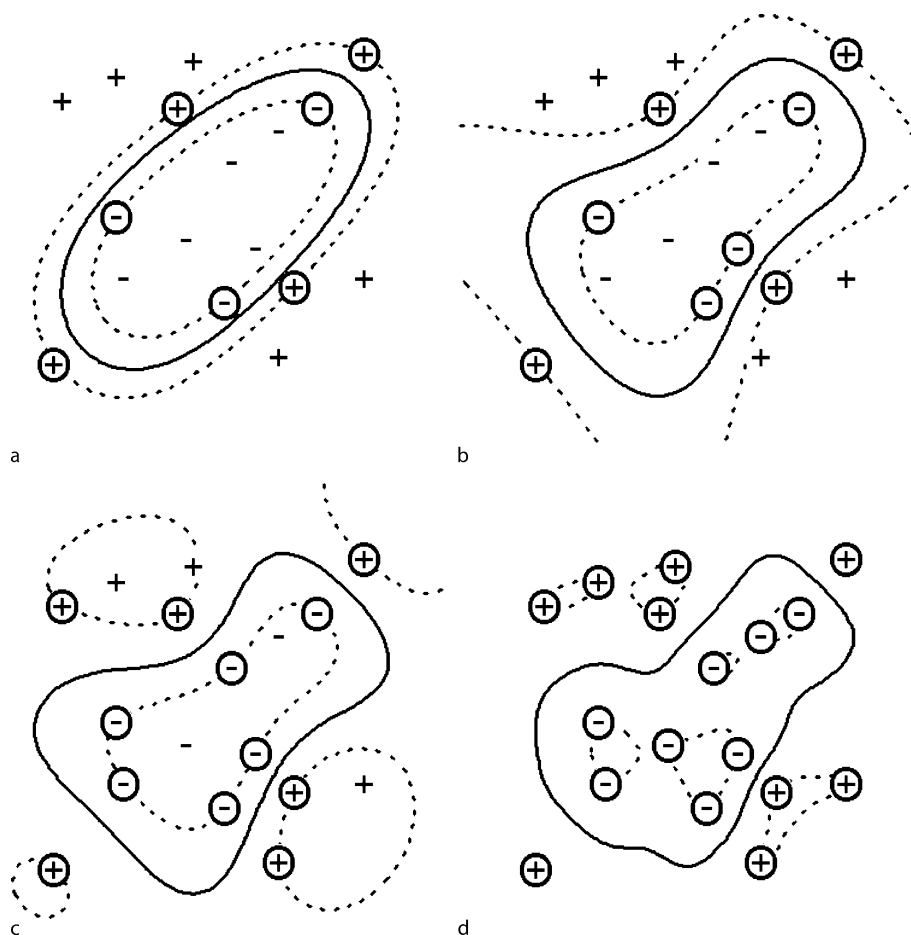
$$L_D(\mathbf{w}, b, \mathbf{A}) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$$

subject to $0 \leq \alpha_i \leq C, \quad i = 1, \dots, m$

$$\text{and } \sum_{i=1}^m \alpha_i y_i = 0 \quad (25)$$

with a solution identical with that for hard margin nonlinear SVM classification from Eq. (24).

In Fig. 10 we demonstrate the nonlinear SVM classification for the Gaussian RBF kernel. The best separation



Drug Design with Machine Learning, Figure 10

SVM classification models for linearly non-separable data with the Gaussian RBF kernel: a $\sigma = 0.1$; b $\sigma = 0.5$; c $\sigma = 1$; d $\sigma = 5$

of the two classes is obtained for a low σ value (Fig. 10a, $\sigma = 0.1$). As the value of σ increases the kernel nonlinearity increases and the separation surface becomes more complex. It is clear from this example that the kernel nonlinearity must match the structure of the data. Also, highly nonlinear functions do not translate in better SVM models. In practical applications the parameters that determine the kernel shape must be optimized to provide the best predictions in a cross-validation test.

SVM Regression

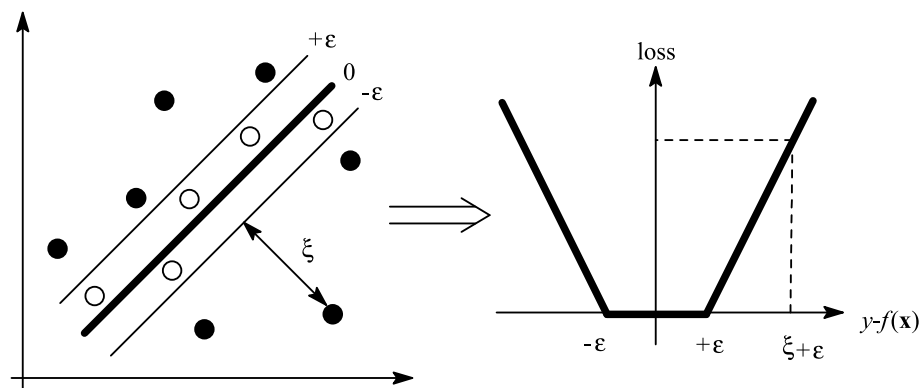
Vapnik extended the SVM algorithm for regression [139] by using an ε -insensitive loss function (Fig. 11). The SVM regression (SVMR) determines a function $f(\mathbf{x})$ with the property that for all learning objects \mathbf{x} it has a maximum deviation ε from the target (experimental) values y , and it has a maximum margin. Starting from the training objects, SVMR computes a model representing a tube with radius ε fitted to the data. All objects situated inside the regression tube have an error equal to zero. For the hard margin SVMR no object is allowed outside the tube, whereas for the soft margin SVMR uses positive slack variables to penalize the objects situated outside the tube [44,45,93,94,122]. If Fig. 11 the objects situated inside the regression tube are shown as white circles, whereas the objects outside the regression tube are depicted as black circles. The slack variable ξ introduces a penalty proportional with the distance between the object and the margin of the regression tube. Support vectors are selected only from objects situated outside the tube.

To demonstrate the kernel influence on the SVM regression we model the inhibition constant K_i for bovine milk xanthine oxidase [51]. The structural descriptor used is ClogP, the computed hydrophobicity. To model the

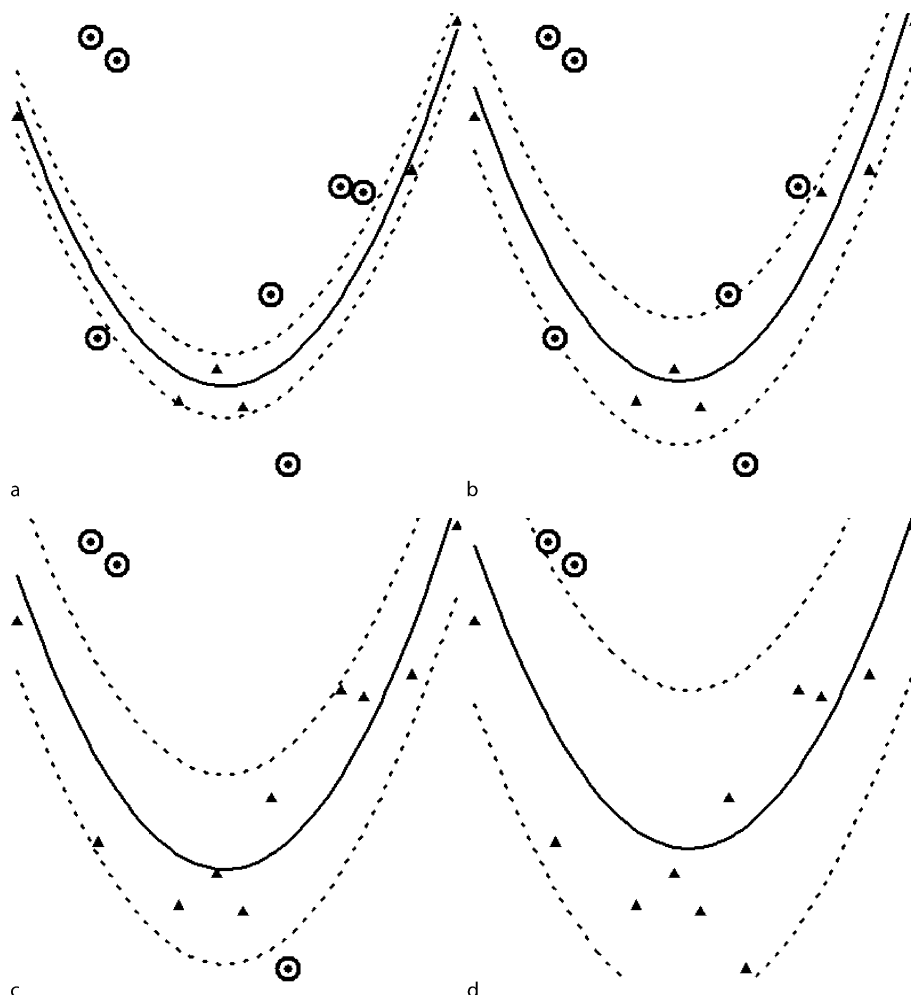
nonlinear dependence between ClogP and K_i we apply a degree 2 polynomial kernel (Fig. 12) with different values for the ε -insensitive parameter. The molecules situated inside the regression tube are indicated with black triangles, whereas those located outside the regression tube are represented as black dots (support vectors are shown as black dots inside a circle). A low value for ε produces a SVMR model with a larger number of support vectors, but an overall good shape for the relationship between ClogP and K_i (Fig. 12a, $\varepsilon = 0.1$). As ε increases the diameter of the regression tube includes almost all molecules, and thus the model depends on a small number of support vectors because objects situated inside the tube do not influence the regression model.

We explore further two higher order polynomials, namely a degree 5 polynomial kernel (Fig. 13a) and a degree 8 polynomial kernel (Fig. 13b). Both kernels overfit the data, and have shapes that are too complex. The Gaussian RBF kernel with $\sigma = 0.2$ is a good fit for the K_i data (Fig. 13c), but an increase in σ results in a regression that start to fit the errors in the experimental data (Fig. 13d).

The next set of SVM regression QSAR evaluates the tanh kernel (Fig. 14). The first combination of parameters is insensitive to the relationship between ClogP and K_i (Fig. 14a). Although the K_i data are best represented by a downward parabola, the second tanh QSAR has the shape of an upward parabola, and misses the most apparent correlation between ClogP and K_i (Fig. 14b). The third tanh QSAR is overfitted, mainly in its right side (Fig. 14c). The last combination of parameters for the tanh QSAR shows a very good fit of the data and obviously it is a good candidate for the prediction of new molecules (Fig. 14d). Considering the large variation in shape of the tanh QSAR it is apparent that its parameters should be varied over a large range to find the best regression model.



Drug Design with Machine Learning, Figure 11
SVM regression with ε -insensitive loss function



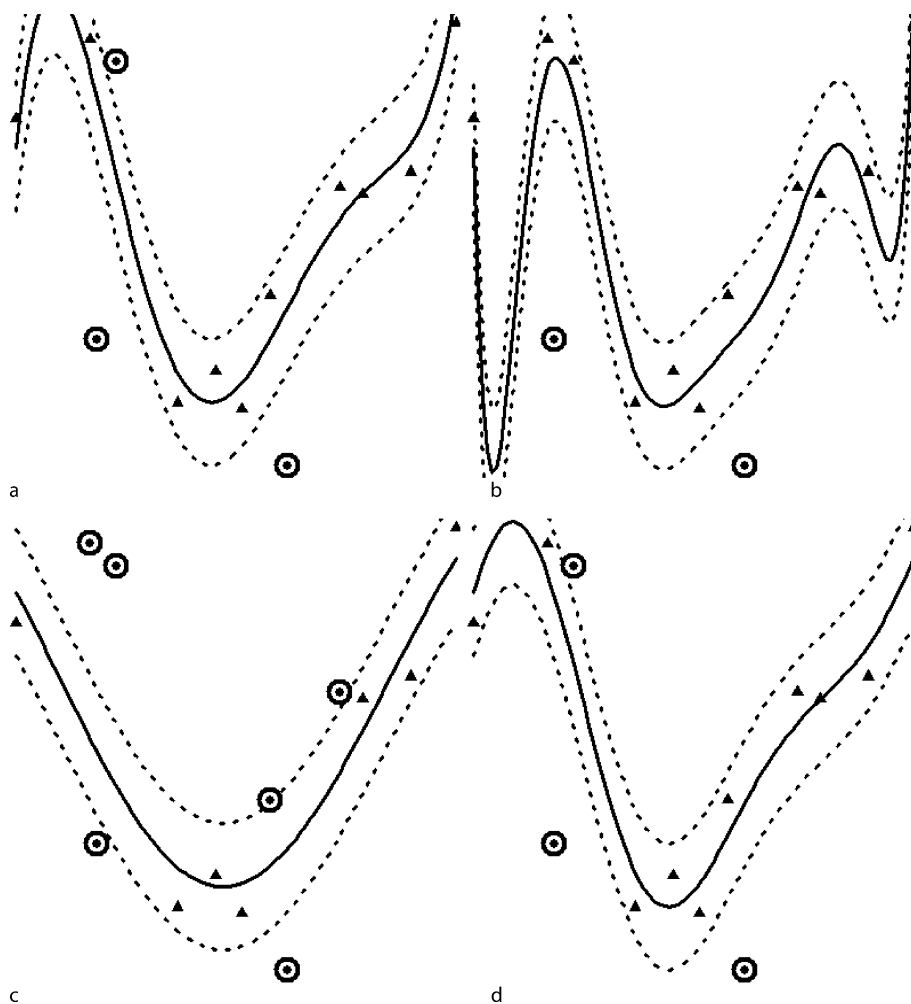
Drug Design with Machine Learning, Figure 12

SVM regression models with a degree 2 polynomial kernel ($C = 1000$): a $\epsilon = 0.1$; b $\epsilon = 0.2$; c $\epsilon = 0.3$; d $\epsilon = 0.5$

Since the dataset considered is a real QSAR problem, it is interesting to examine the behavior of other, less common, kernels. The anova RBF kernel shows an overfit of the data, which is more prominent in its left side (Fig. 15a). The spline kernel has a too low nonlinearity, and the model has no value in predicting new chemicals (Fig. 15b). The degree 1 Bessel kernel is a very good fit for the correlation between ClogP and K_i (Fig. 15c), whereas a degree 5 Bessel kernel overfits the data in the left side (Fig. 15d). The results presented here are only to illustrate the shape of various kernels, and to demonstrate their ability to model nonlinear relationships. Other combinations of parameters and capacity C may give better or worse correlations, and the most predictive QSAR may be identified only by optimizing the kernel parameters.

Kernel Comparison in QSAR

Several kernels have the property to give SVM models that are universal approximators to arbitrary functions, namely the SVM model can approximate any function to any level of precision. However, the theory cannot guarantee that such a model will be also optimally predictive. The essential requirement for a SAR or QSAR model is its ability to provide reliable predictions for new molecules. The simple data fitting ability with a low error is not enough for an SVM model, because such models are usually overfitted and give bad predictions. The experimental data used in SAR and QSAR are affected by experimental errors and sometimes by the fact that good structural descriptors are missing from the model.



Drug Design with Machine Learning, Figure 13

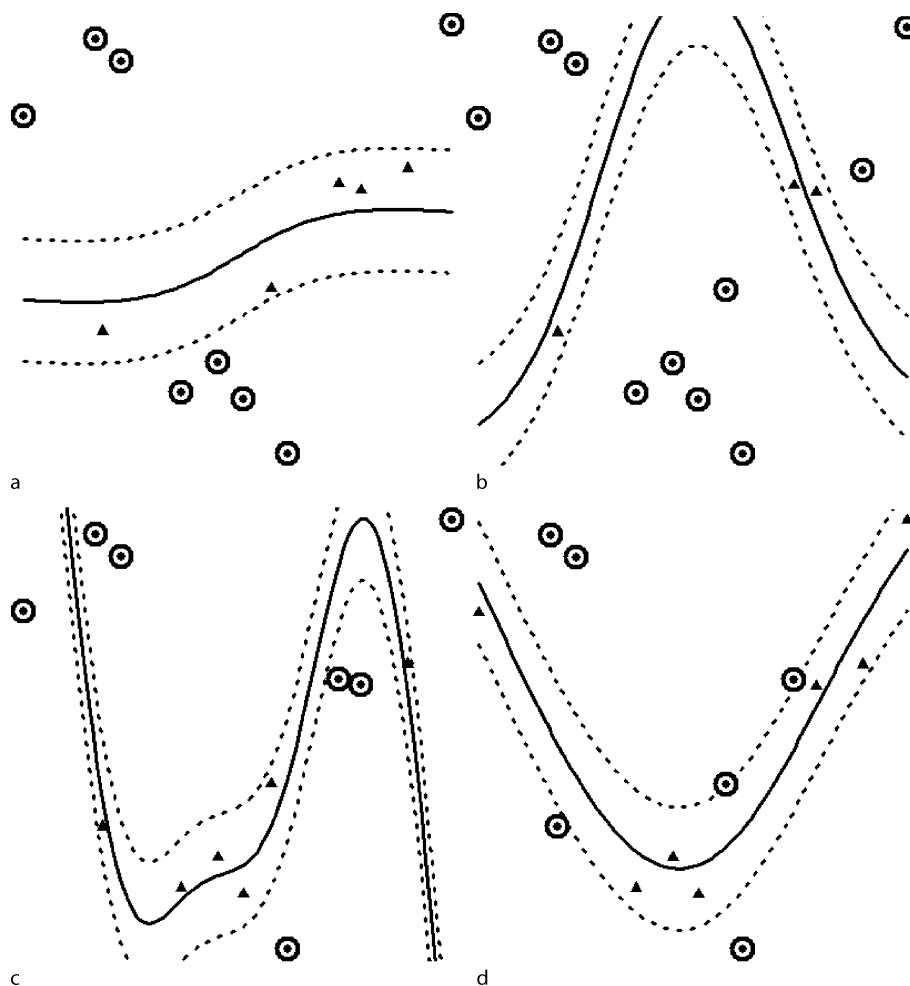
SVM regression models ($C = 1000$, $\epsilon = 0.2$): a degree 5 polynomial kernel; b degree 8 polynomial kernel; c Gaussian RBF kernel, $\sigma = 0.1$; d Gaussian RBF kernel, $\sigma = 0.5$

Therefore it is important to evaluate a number of kernels in order to find the most predictive SVM model and to avoid fitting the noise or the errors from the data. To offer a comparative evaluation of several kernels we review here a number of SVM classification and regression models in which five kernels are compared, namely linear, polynomial, Gaussian RBF, tanh, and anova. Each kernel was evaluated for a large range of parameters, and all results are compared for cross-validation. The SVM models were computed with mySVM, an efficient SVM program authored by Rüping (<http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html>).

The carcinogenic activity of a group of 46 methylated and 32 non-methylated polycyclic aromatic hydrocarbons was modeled with semiempirical quantum indices [61]. From the set of 78 chemicals, 34 are carcinogenic and

44 are non-carcinogenic. The accuracy for leave-10%-out (L10%O) cross-validation tests shows that the best predictions are obtained with the Gaussian RBF kernel: RBF, $\sigma = 0.5$, $AC = 0.86$; anova, $\gamma = 0.5$, $d = 1$, $AC = 0.84$; degree 2 polynomial, $AC = 0.82$; linear, $AC = 0.76$; tanh, $a = 2$, $b = 0$, $AC = 0.66$. All SVM classification models were obtained with $C = 10$. The increase in prediction accuracy for RBF and anova kernel, compared with the linear SVM, indicates that there is a nonlinear relationship between the quantum indices and hydrocarbon carcinogenicity.

A structure-odor classification was performed for 98 tetra-substituted pyrazines representing three odor classes, namely 32 green, 23 nutty, and 43 bell-pepper [62]. L10%O predictions were obtained for three classification tests, in which one odor type is selected as class +1 and



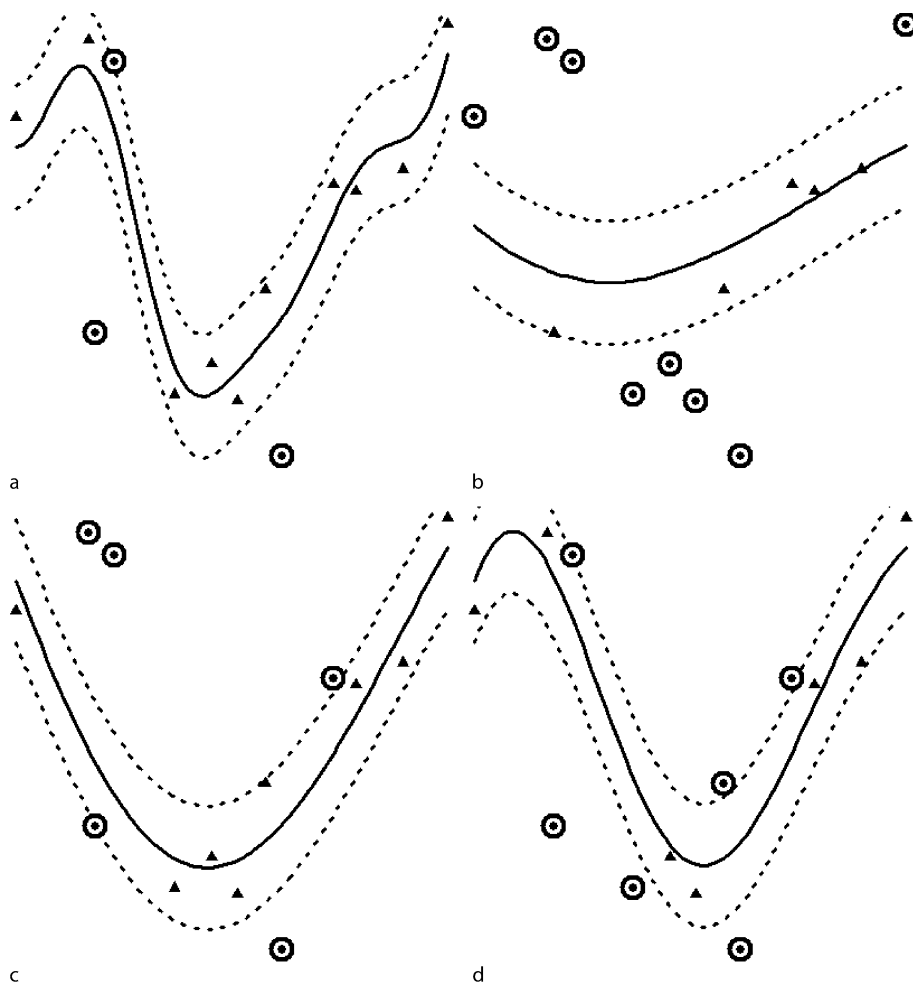
Drug Design with Machine Learning, Figure 14

SVM regression models with a hyperbolic tangent (tanh) kernel ($\epsilon = 0.2$): **a** $a = 1, b = 0, C = 1$; **b** $a = 1, b = 1, C = 1$; **c** $a = 0.9, b = -1, C = 10$; **d** $a = 0.5, b = -0.7, C = 10$

the remaining two odor types form class -1. For green aroma compounds the best predictions are obtained with a polynomial kernel: degree 2 polynomial, $C = 1000$, $AC = 0.86$; anova, $C = 100$, $\gamma = 0.5$, $d = 1$, $AC = 0.84$; linear, $C = 10$, $AC = 0.80$; RBF, $C = 10$, $\sigma = 0.5$, $AC = 0.79$; tanh, $C = 10$, $a = 0.5$, $b = 0$, $AC = 0.73$. For nutty aroma compounds the anova kernel has slightly better predictions: anova, $C = 100$, $\gamma = 0.5$, $d = 1$, $AC = 0.92$; linear, $C = 10$, $AC = 0.89$; degree 2 polynomial, $C = 10$, $AC = 0.89$; RBF, $C = 10$, $\sigma = 0.5$, $AC = 0.89$; tanh, $C = 100$, $a = 0.5$, $b = 0$, $AC = 0.79$. The SVM models for bell-pepper aroma compounds show that three kernels (RBF, polynomial, and anova) give good predictions: RBF, $C = 10$, $\sigma = 0.5$, $AC = 0.89$; degree 2 polynomial, $C = 10$, $AC = 0.88$; anova, $C = 10$, $\gamma = 0.5$, $d = 1$, $AC = 0.87$. linear, $C = 10$, $AC = 0.74$; tanh,

$C = 100$, $a = 2$, $b = 1$, $AC = 0.68$. In all three SVM experiments the best model outperformed the linear kernel, showing that nonlinear relationships are useful in structure-aroma studies.

The relationships between the structure of chemical compounds and their aqueous toxicity are highly relevant because many chemicals are environmental pollutants. The mechanism of toxic action (MOA) of organic chemicals may be predicted from hydrophobicity and experimental toxicity against *Pimephales promelas* and *Tetrahymena pyriformis* [67]. SVM classification was used to discriminate between 126 nonpolar narcotics and 211 chemicals that have other MOA. The RBF, anova, linear and polynomial kernels have similar statistics: RBF, $\sigma = 1$, $AC = 0.80$; anova, $\gamma = 0.5$, $d = 2$, $AC = 0.80$; linear, $AC = 0.79$; degree 2 polynomial, $AC = 0.79$; tanh,



Drug Design with Machine Learning, Figure 15

SVM regression models ($\epsilon = 0.2$): **a** anova RBF kernel, $\sigma = 0.2$, degree = 5, $C = 1000$; **b** spline kernel, $C = 0.05$; **c** Bessel kernel, $\sigma = 1$, order = 1, degree = 1, $C = 1000$; **d** Bessel kernel, $\sigma = 0.6$, order = 1, degree = 5, $C = 1000$

$a = 0.5$, $b = 0$, $AC = 0.53$. These results indicate that there is an almost linear relationship between descriptors and MOA. A similar conclusion was obtained for the MOA classification of polar and nonpolar narcotic compounds [66]. In a related study for the classification of MOA for narcotic and reactive compounds [63], the degree 2 polynomial offered the best predictions: degree 2 polynomial, $C = 10$, $AC = 0.92$; anova, $C = 10$, $\gamma = 0.5$, $d = 1$, $AC = 0.87$; linear, $C = 1000$, $AC = 0.86$; RBF, $C = 100$, $\gamma = 0.5$, $AC = 0.83$; tanh, $C = 10$, $a = 0.5$, $b = 0$, $AC = 0.78$.

SVM regression is demonstrated for QSAR models obtained for 52 benzodiazepine receptor ligands [68]. Five structural descriptors were used as SVM input, and the prediction was evaluated with leave-5%-out (L5%O) and L10%O cross-validation. A number of 34 experiments

were conducted with five kernels (Tab. 1). The predictions obtained with the linear kernel L are used as baseline to compare the results provided by the other kernels. All four polynomial kernels P give very bad predictions, as indicated by the statistical indices. The RBF kernel R has better predictions than L only for $\sigma = 0.25$ (experiment 6), whereas all the other RBF models are worse than L. Overall, the best results are obtained with the tanh kernel T for $a = 0.5$ and $b = 0$ (experiment 11) which is a clear improvement compared to the linear kernel. Finally, the anova kernel A has uneven predictions. Some are good (experiments 22 and 20) whereas experiments 25–34 have very bad predictions. The SVM regression experiments presented in Tab. 1 show that kernel parameters are of paramount importance in obtaining a predictive QSAR.

Drug Design with Machine Learning, Table 1

Support vector regression statistics for benzodiazepine receptor ligands

Exp	Kernel	p_1	p_2	$q_{L5\%O}^2$	RMSE _{L5%O}	$q_{L10\%O}^2$	RMSE _{L10%O}
1	L			0.261	0.98	0.273	0.97
2	P	2		< -100	> 10	< -100	> 10
3	P	3		< -100	> 10	< -100	> 10
4	P	4		< -100	> 10	< -100	> 10
5	P	5		< -100	> 10	< -100	> 10
6	R	0.25		0.368	0.91	0.370	0.91
7	R	0.5		0.226	1.00	0.324	0.94
8	R	1.0		0.221	1.01	0.310	0.95
9	R	1.5		0.236	1.00	0.258	0.98
10	R	2.0		0.208	1.01	0.205	1.02
11	T	0.5	0.0	0.453	0.84	0.498	0.81
12	T	1.0	0.0	0.416	0.87	0.411	0.87
13	T	2.0	0.0	0.396	0.89	0.394	0.89
14	T	0.5	1.0	0.070	1.10	0.120	1.07
15	T	1.0	1.0	0.389	0.89	< -10	5.36
16	T	2.0	1.0	0.297	0.96	0.348	0.92
17	T	0.5	2.0	-0.365	1.33	-0.405	1.35
18	T	1.0	2.0	0.106	1.08	0.243	0.99
19	T	2.0	2.0	0.345	0.92	0.376	0.90
20	A	0.25	1	0.397	0.89	0.377	0.90
21	A	0.5	1	0.324	0.94	0.331	0.93
22	A	1.0	1	0.398	0.88	0.412	0.87
23	A	1.5	1	0.299	0.95	0.374	0.90
24	A	2.0	1	0.293	0.96	0.339	0.93
25	A	0.25	2	-1.289	1.73	-0.921	1.58
26	A	0.5	2	-1.375	1.76	-0.873	1.56
27	A	1.0	2	-0.606	1.44	-0.465	1.38
28	A	1.5	2	-0.241	1.27	-0.195	1.25
29	A	2.0	2	-0.071	1.18	-0.060	1.17
30	A	0.25	3	-2.998	2.28	-1.934	1.95
31	A	0.5	3	-1.282	1.72	-0.983	1.61
32	A	1.0	3	-0.060	1.17	-0.094	1.19
33	A	1.5	3	0.060	1.11	-0.097	1.19
34	A	2.0	3	-0.062	1.18	-0.073	1.18

In this section we presented comparative SAR and QSAR models obtained with five kernels. In developing SVM models it is important to include the linear kernel as a baseline comparison. If nonlinear kernels do not improve the predictions of the linear kernel, than the structure-activity relationship is linear and a linear SVM should be used. The parameters of nonlinear kernels should be optimized in order to obtain the best prediction. The most predictive SAR and QSAR model may be obtained with anyone of the five kernels investigated, and there is no apparent rule that can predict which kernel is the best. Therefore, it is important to experiment with a large diversity of kernels in order to find a predictive SVM model.

Applications in Drug Design

Torsade de pointes (TdP) is an important adverse drug reaction that is responsible for almost one-third of all drug failures during drug development, and resulted in several drugs being withdrawn from the market. Yap et al compared several ML for their ability to separate TdP compounds from those that do not induce TdP [156]. The prediction accuracy shows that SVM is the most reliable ML: SVM with Gaussian RBF kernel 91.0%, k -NN 88.5%, probabilistic neural network 78.2%, and C4.5 decision tree 65.4%. Inhibitors of the hERG potassium channel can lead to a prolongation of the QT interval that can trigger TdP,

an atypical ventricular tachycardia. A dataset of 39 hERG inhibitors (26 for learning and 13 for test) was evaluated by a pharmacophore ensemble coupled with SVM regression [87]. The pharmacophore ensemble encodes the protein conformational flexibility, and the SVM brings the capability of modeling nonlinear relationships. The QSAR model has very good statistics, with $q^2 = 0.89$ and $r^2 = 0.94$ for the test set. A structure-activity SVM study for hERG inhibitors was performed by Tobita et al for 73 drugs [132]. Using a combination of structural descriptors and molecular fragments, the SVM model had a prediction accuracy higher than 90%. Selecting an optimum group of descriptors is an important phase in computing a predictive structure-activity model. The SVM recursive feature elimination algorithm was evaluated by Xue et al for three SAR datasets, namely *P*-glycoprotein substrates, human intestinal absorption, and compounds that cause TdP [153]. The prediction statistics show that a proper descriptor selection may significantly increase the SVM performance.

The drug metabolism by human cytochrome P450 isoforms 3A4, 2D6, and 2C9 was studied by Terfloth et al for 379 drugs and drug analogs [131]. The best SVM model has a cross-validation accuracy of 89% and a prediction accuracy of 83% for 233 validation compounds. Unbalanced data sets, with a majority of inactive compounds and only a few active chemicals, represent the usual situation in library screening. To overcome the ML bias toward the larger class, Eitrich et al used SVM with oversampling to identify cytochrome P450 2D6 inhibitors [37]. An ensemble of SVM models was presented by Yap and Chen as suitable classifiers for inhibitors and substrates of cytochromes P450 3A4, 2D6, and 2C9 [155]. Descriptors computed with Dragon were used to develop predictive models, with MCC (Matthews correlation coefficient) statistics, namely 0.899 for 3A4, 0.884 for 2D6, and 0.872 for 2C9. The classification of cytochrome P450 3A4 inhibitors and non-inhibitors was investigated by Arimoto et al with SVM, *k*-NN, recursive partitioning, logistic regression, and Bayesian classifier [7]. The chemical structure of the 4470 compounds was encoded with fingerprints and MolconnZ topological indices, and the best predictions were obtained with SVM trained with fingerprints. A comparison of single and ensemble classifiers was made by Merkwirth et al for SVM, *k*-NN, and ridge regression [95]. The tests conducted for two libraries of chemicals, one for unspecific protein inhibition and another one for inhibitors of cytochrome P450 3A4, showed that single and ensemble SVM models give similar results, with better predictions than *k*-NN and ridge regression.

Du et al predicted the genotoxicity of 140 thiophene derivatives with SVM and linear discriminant analysis (LDA) [34]. Seven structural descriptors were used as input data for the classification models, and the SVM parameters were identified by grid search. The accuracy of the SVM model (92.9% in training and 92.6% in prediction) is significantly higher than the LDA accuracy (81.4% in training and 85.2% in prediction). Recursive partitioning and SVM were compared in a test for mutagenicity prediction from substructure descriptors [92]. The prediction accuracy for 2199 mutagens is similar for SVM (81.4%) and recursive partitioning (80.2%).

SVM classification was compared with linear discriminant analysis in predicting drug bioavailability [149]. The learning set comprises 167 compounds characterized by five structural descriptors. The SVM parameters were optimized with grid search. The classification results indicate that SVM (accuracy 85.6%) provides better models for bioavailability compared to LDA (accuracy 72.4%). Sakiyama et al studied SAR models for the metabolic stability of drug candidates [113]. The structure-activity models were obtained with random forest, support vector machine, logistic regression, and recursive partitioning. Classification models were obtained for a collection of 1952 compounds characterized by 193 descriptors. All ML models had accuracy > 0.8, with slightly higher results for random forest and SVM.

The identification of drug-like compounds based on atom type counts was investigated with SVM, linear programming machines, linear discriminant analysis, bagged *k*-nearest neighbors, and bagged decision trees C4.5 [97]. The classification dataset consisted of drug-like compounds from World Drug Index and non-drug compounds from Available Chemicals Directory. The smallest errors were obtained with SVM trained with polynomial and Gaussian RBF kernels. A computational filter for chemical compounds with antidepressant activity was developed by Lepp et al based on 21 biological targets related to depression [88]. An SVM model obtained with atom type descriptors gave high enrichment levels, and showed that a compound selected as active interacts on average with 2.3 targets. SVM, artificial neural networks, and multiple linear regression were compared in a model for apoptosis induction by 4-aryl-4-H-chromenes [40]. A leave-one-out cross-validation test showed that SVM gives the best predictions. Briem and Günther used a dataset of 565 kinase inhibitors and 7194 inactive compounds to compare four machine learning algorithms, namely SVM with a Gaussian RBF kernel, artificial neural networks, *k*-nearest neighbors, and recursive partitioning [23]. The best predictions were obtained with SVM, followed closely by

k -NN. The SVM ability to identify active compounds from a chemical library was investigated by Jorissen and Gilson for sets of molecules that interact with the α_{1A} adrenoceptor, and cyclin-dependent kinase, cyclooxygenase-2, factor Xa, and phosphodiesterase-5 [72]. A comparison over ten QSAR datasets was performed to compare the stochastic gradient boosting method (SGB, an ensemble of classification and regression trees) with decision tree, random forest, partial least squares, k -NN, Bayes classifier, and SVM [129]. The best results predictions were obtained with random forest, followed by SVM and SGB.

The enzyme-substrate interaction is very specific, which points to the existence of recognition elements located on the enzyme interface. A surface patch ranking (SPR) method was proposed for the identification of those protein residues that determine the substrate specificity [157]. SPR incorporates an SVM module that can highlight the residues important in ligand binding. The method was applied with good results for several homologous enzymes, namely guanylyl/adenylyl cyclases, lactate/malate dehydrogenases, and trypsin/chymotrypsin. Potential applications are residue selection for mutagenesis experiments and functional annotation of proteins. Single base DNA mutations that result in an amino acid substitution are a common cause of monogenic disease. These substitutions reduce the protein stability, as measured by reduction in hydrophobic area, overpacking, backbone strain, and loss of electrostatic interactions [158]. An SVM classifier was trained with a set of mutations causative of disease and another set of non-disease causing mutations. The cross-validation tests show that the SVM identifies 74% of disease mutations, and confirms that loss of protein stability is a determinant factor in monogenic disease.

Comparative Studies

SAR and QSAR models have two major components, namely a set of structural descriptors and a machine learning or ensemble of ML algorithms. The studies reviewed in previous sections show that there is no universal family of descriptors that gives the best predictions for any molecular property, which justifies the continuous development of novel structural descriptors. Similarly, no ML algorithm performs optimally for any property and any set of descriptors. Extensive computational experiments show that for a given property, the best model can be identified only by an empirical comparison of a large number of ML methods. In this section we review several studies that sought to compare the predictive abilities of ML models.

The aqueous solubility of drugs and drug-related compounds was modeled by Schroeter et al with Gaussian process, random forest, SVM, and ridge regression [118]. The domain of applicability for each model was estimate by computing error levels. The aqueous solubility of 988 organic chemicals was modeled with four regression algorithms, namely partial least squares, random forest, SVM, and MLF artificial neural networks [104]. The best predictions were obtained with random forest, which also gave good results for an external validation set of 330 chemicals. The melting temperatures of ionic liquids were modeled with SVM, associative neural networks, k -NN, partial least squares, multilayer feedforward (MLF) artificial neural networks, and multiple linear regression [145]. Slightly better results were obtained with SVM, associative neural networks, and MLF artificial neural networks.

The human serum protein binding of 808 compounds was simulated with multiple linear regression, MLF artificial neural networks, k -NN, and SVM [148]. An ensemble of MLF artificial neural networks provided the best predictions for an external validation set of 200 compounds. The blood-brain barrier permeability of 415 chemicals was modeled by Li et al with logistic regression, linear discriminate analysis, k -NN, C4.5 decision tree, probabilistic neural network, and SVM [89]. It was found that all ML models are improved by using a descriptor set selected with recursive feature elimination. Plewczynski et al compared SAR models for five biological targets obtained with SVM, random forest, artificial neural networks, k -NN, trend vectors, Bayesian classifier, and decision tree [107]. There are significant differences in performance depending on the target and objective, i. e., high enrichment or maximum number of actives. The toxicity risk assessment of organic chemicals was evaluated with SVM, k -NN, Bayesian classifier, and self-organizing maps (SOM) [147]. Both SOM and SVM can separate toxic from nontoxic compounds based on structural fragments. A new ML algorithm that shows promising results is kScore, which in several tests performed better than SVM, k -NN, recursive partitioning, artificial neural networks, Gaussian process, and Bayesian classifier [102].

A comparison of 21 machine learning algorithms is presented for data selected from the National Cancer Institute 60-cell line screening panel (NCI-60) [69]. The SMILES codes of the chemical compounds [130] were used to compute fingerprints with Open Babel (<http://openbabel.org/>). The descriptor selection (based on Cfs-SubsetEval and BestFirst) and all machine learning models were computed with Weka [41,151], and all prediction results are for 10-fold (leave-10%-out) cross-validation. The machine learning algorithms used are briefly listed

Drug Design with Machine Learning, Table 2

Machine learning anticancer SAR for lung large cell carcinoma NCI-H460

Exp	Model	TP _p	FN _p	TN _p	FP _p	Ac _p	MCC _p
1	IBk $k = 9W(1 - d)$	1386	663	1035	515	0.6727	0.3414
2	IBk $k = 5W(1 - d)$	1410	639	1011	539	0.6727	0.3383
3	IBk $k = 9$ NoW	1404	645	1015	535	0.6721	0.3378
4	IBk $k = 7W(1 - d)$	1386	663	1027	523	0.6705	0.3364
5	IBk $k = 9W(1/d)$	1399	650	1012	538	0.6699	0.3334
6	IBk $k = 5$ NoW	1433	616	983	567	0.6713	0.3324
7	KStar	1364	685	1034	516	0.6663	0.3299
8	IBk $k = 7$ NoW	1404	645	1001	549	0.6682	0.3290
9	SVM RBF $\sigma = 0.01$	1494	555	926	624	0.6724	0.3286
10	IBk $k = 7W(1/d)$	1396	653	1003	547	0.6666	0.3263
11	IBk $k = 3W(1 - d)$	1420	629	982	568	0.6674	0.3252
12	RandomForest $T = 40$	1509	540	902	648	0.6699	0.3217
13	IBk $k = 3$ NoW	1446	603	954	596	0.6669	0.3210
14	IBk $k = 5W(1/d)$	1408	641	985	565	0.6649	0.3210
15	RandomForest $T = 50$	1499	550	903	647	0.6674	0.3171
16	RandomForest $T = 30$	1499	550	902	648	0.6671	0.3164
17	IBk $k = 3W(1/d)$	1428	621	955	595	0.6621	0.3125
18	RandomForest $T = 20$	1487	562	889	661	0.6602	0.3021
19	IBk $k = 1$ NoW	1457	592	911	639	0.6580	0.3000
20	IBk $k = 1W(1/d)$	1457	592	911	639	0.6580	0.3000
21	IBk $k = 1W(1 - d)$	1457	592	911	639	0.6580	0.3000
22	RandomForest $T = 10$	1494	555	878	672	0.6591	0.2990
23	JRip	1568	481	810	740	0.6607	0.2972
24	J48	1481	568	884	666	0.6571	0.2959
25	ADTree Ehp	1536	513	836	714	0.6591	0.2956
26	ADTree Ezp	1536	513	836	714	0.6591	0.2956
27	ADTree Erp	1580	469	791	759	0.6588	0.2922
28	NaiveBayes	1130	919	1147	403	0.6327	0.2919
29	NaiveBayesUpdateable	1130	919	1147	403	0.6327	0.2919
30	ADTree Eap	1563	486	806	744	0.6582	0.2919
31	Logistic	1514	535	848	702	0.6563	0.2911
32	BayesNet	1256	793	1054	496	0.6418	0.2903
33	NBTree	1406	643	930	620	0.6491	0.2857
34	RBFNetwork	1578	471	769	781	0.6521	0.2774
35	PART	1404	645	917	633	0.6449	0.2766
36	REPTree	1551	498	785	765	0.6491	0.2723
37	DecisionTable	1549	500	752	798	0.6393	0.2507
38	RandomTree	1384	665	885	665	0.6305	0.2464
39	Ridor	1617	432	642	908	0.6277	0.2201
40	OneR	1745	304	427	1123	0.6035	0.1565
41	ConjunctiveRule	1899	150	163	1387	0.5729	0.0562
42	DecisionStump	2049	0	0	1550	0.5693	0.0000

here, using their notation in Weka: BayesNet, Bayesian network; NaiveBayes, naïve Bayesian classifier [71]; NaiveBayesUpdateable, naïve Bayesian classifier with estimator classes [71]; Logistic, logistic regression with a ridge estimator [86]; RBFNetwork, Gaussian radial basis func-

tion network; IBk, k -NN classifier with distance weight (NoW – no distance weight, $W(1/d)$ – weighted with $1/d$, $W(1 - d)$ – weighted with $(1 - d)$) and k an odd number between 1 and 9 [1]; KStar, K^* lazy learner with entropy-based distance function [26]; ADTree, alternating

Drug Design with Machine Learning, Table 3
Machine learning anticancer SAR for glioma SF-268

Exp	Model	TP _p	FN _p	TN _p	FP _p	Ac _p	MCC _p
1	IBk $k = 9W(1/d)$	1338	682	1177	524	0.6759	0.3530
2	SVM RBF $\sigma = 0.01$	1403	617	1118	583	0.6775	0.3513
3	IBk $k = 7W(1/d)$	1344	676	1165	536	0.6743	0.3490
4	IBk $k = 5W(1/d)$	1351	669	1152	549	0.6727	0.3449
5	IBk $k = 7W(1 - d)$	1325	695	1173	528	0.6713	0.3443
6	IBk $k = 9W(1 - d)$	1301	719	1185	516	0.6681	0.3395
7	IBk $k = 5W(1 - d)$	1330	690	1160	541	0.6692	0.3391
8	IBk $k = 5$ NoW	1372	648	1120	581	0.6697	0.3368
9	IBk $k = 7$ NoW	1353	667	1135	566	0.6686	0.3360
10	KStar	1307	713	1168	533	0.6651	0.3325
11	IBk $k = 3$ NoW	1396	624	1089	612	0.6678	0.3311
12	IBk $k = 3W(1 - d)$	1358	662	1121	580	0.6662	0.3304
13	IBk $k = 9$ NoW	1323	697	1150	551	0.6646	0.3298
14	RandomForest $T = 40$	1408	612	1074	627	0.6670	0.3287
15	RandomForest $T = 50$	1408	612	1072	629	0.6665	0.3275
16	RandomForest $T = 30$	1418	602	1062	639	0.6665	0.3269
17	IBk $k = 3W(1/d)$	1363	657	1108	593	0.6641	0.3254
18	J48	1376	644	1082	619	0.6606	0.3169
19	RandomForest $T = 20$	1403	617	1045	656	0.6579	0.3095
20	Logistic	1404	616	1028	673	0.6536	0.3003
21	NBTree	1240	780	1164	537	0.6461	0.2974
22	RandomForest $T = 10$	1411	609	1015	686	0.6520	0.2965
23	BayesNet	1220	800	1176	525	0.6439	0.2948
24	IBk $k = 1$ NoW	1396	624	1020	681	0.6493	0.2916
25	IBk $k = 1W(1/d)$	1396	624	1020	681	0.6493	0.2916
26	IBk $k = 1W(1 - d)$	1396	624	1020	681	0.6493	0.2916
27	ADTree Erp	1407	613	993	708	0.6450	0.2819
28	RBFNetwork	1519	501	884	817	0.6458	0.2800
29	NaiveBayes	1064	956	1272	429	0.6278	0.2790
30	NaiveBayesUpdateable	1064	956	1272	429	0.6278	0.2790
31	ADTree Ehp	1406	614	986	715	0.6428	0.2774
32	ADTree Ezp	1406	614	986	715	0.6428	0.2774
33	REPTree	1380	640	1009	692	0.6420	0.2771
34	JRip	1418	602	960	741	0.6391	0.2689
35	ADTree Eap	1443	577	936	765	0.6393	0.2684
36	PART	1328	692	1024	677	0.6321	0.2593
37	DecisionTable	1438	582	915	786	0.6324	0.2538
38	RandomTree	1357	663	984	717	0.6291	0.2510
39	Ridor	1705	315	630	1071	0.6275	0.2454
40	DecisionStump	674	1346	1414	287	0.5611	0.1877
41	ConjunctiveRule	665	1355	1419	282	0.5601	0.1869
42	OneR	1257	763	932	769	0.5883	0.1702

decision tree (search type: Eap, expand all paths; Ehp, expand the heaviest path; Ezp, expand the best z -pure path; Erp, expand a random path) [42]; DecisionStump, one-level binary decision tree with categorical or numerical class label; J48, C4.5 decision tree [108]; NBTree, decision

tree with naïve Bayes classifiers at the leaves [82]; RandomForest, random forest (the number of random trees is between 10 and 50) [22]; RandomTree, a tree that considers k randomly chosen attributes at each node; REPTree, fast decision tree learner; ConjunctiveRule, conjunc-

Drug Design with Machine Learning, Table 4

Machine learning anticancer SAR for melanoma SK-MEL-5

Exp	Model	TP _p	FN _p	TN _p	FP _p	Ac _p	MCC _p
1	IBk $k = 9W(1 - d)$	1379	655	1119	532	0.6779	0.3541
2	IBk $k = 9$ NoW	1401	633	1099	552	0.6784	0.3532
3	IBk $k = 7$ NoW	1422	612	1081	570	0.6792	0.3531
4	IBk $k = 5W(1 - d)$	1401	633	1095	556	0.6773	0.3508
5	IBk $k = 7W(1 - d)$	1390	644	1104	547	0.6768	0.3506
6	RandomForest $T = 50$	1487	547	1019	632	0.6801	0.3504
7	IBk $k = 7W(1/d)$	1414	620	1079	572	0.6765	0.3479
8	IBk $k = 5W(1/d)$	1417	617	1074	577	0.6760	0.3465
9	IBk $k = 9W(1/d)$	1404	630	1084	567	0.6752	0.3458
10	RandomForest $T = 30$	1478	556	1019	632	0.6776	0.3456
11	IBk $k = 5$ NoW	1427	607	1063	588	0.6757	0.3451
12	RandomForest $T = 20$	1482	552	1013	638	0.6771	0.3443
13	RandomForest $T = 40$	1481	553	1013	638	0.6768	0.3437
14	SVM RBF $\sigma = 0.03$	1414	620	1061	590	0.6716	0.3373
15	KStar	1326	708	1119	532	0.6635	0.3279
16	RandomForest $T = 10$	1481	553	976	675	0.6668	0.3222
17	IBk $k = 3W(1 - d)$	1390	644	1053	598	0.6630	0.3204
18	IBk $k = 3$ NoW	1416	618	1025	626	0.6624	0.3171
19	IBk $k = 3W(1/d)$	1393	641	1044	607	0.6613	0.3166
20	IBk $k = 1$ NoW	1427	607	987	664	0.6551	0.3005
21	IBk $k = 1W(1/d)$	1427	607	987	664	0.6551	0.3005
22	IBk $k = 1W(1 - d)$	1427	607	987	664	0.6551	0.3005
23	BayesNet	1232	802	1132	519	0.6415	0.2901
24	PART	1348	686	1033	618	0.6461	0.2875
25	RBFNetwork	1541	493	857	794	0.6507	0.2856
26	NBTree	1313	721	1055	596	0.6426	0.2832
27	J48	1378	656	1000	651	0.6453	0.2831
28	Logistic	1426	608	958	693	0.6469	0.2830
29	JRip	1483	551	899	752	0.6464	0.2785
30	RandomTree	1401	633	963	688	0.6415	0.2731
31	NaiveBayes	1068	966	1224	427	0.6220	0.2698
32	NaiveBayesUpdateable	1068	966	1224	427	0.6220	0.2698
33	DecisionTable	1494	540	863	788	0.6396	0.2634
34	REPTree	1431	603	918	733	0.6374	0.2622
35	ADTree Erp	1436	598	889	762	0.6309	0.2478
36	ADTree Ehp	1429	605	892	759	0.6299	0.2459
37	ADTree Ezp	1429	605	892	759	0.6299	0.2459
38	ADTree Eap	1323	711	981	670	0.6252	0.2441
39	Ridor	1469	565	817	834	0.6204	0.2230
40	DecisionStump	1279	755	932	719	0.6000	0.1930
41	OneR	1279	755	932	719	0.6000	0.1930
42	ConjunctiveRule	1404	630	770	881	0.5900	0.1605

tive rule learner; DecisionTable, decision table majority classifier [81]; JRip, a propositional rule learner based on RIPPER [27]; OneR, rule classifier that uses the minimum-error attribute for prediction [56]; PART, a PART decision list that builds a partial C4.5 decision tree in each iteration

and transforms the best leaf into a rule ; Ridor, a Ripple-Down Rule learner [43]; SVM, LibSVM. More details for each ML may be found in Weka.

The first series of SAR models for anticancer compounds considers the cell line NCI-H460 (lung large cell

Drug Design with Machine Learning, Table 5

Machine learning algorithms with best results in the CoEPra 2006 (comparative evaluation of prediction algorithms, <http://www.coepra.org/>) competition

Task	Rank	Machine Learning	Task	Rank	Machine Learning
C1	1	naïve Bayes	R1	1	LS-SVM linear kernel
	2	LS-SVM RBF kernel		2	random forest
	3	CART		3	Gaussian process
	4	C4.5		4	kernel PLS
	5	SVM string kernel		5	SVM RBF kernel
C2	1	SVM linear kernel	R2	1	kernel PLS
	2	kScore		2	kNN
	3	SVM RBF kernel		3	PLS
	4	ClassificationViaRegression		4	SVM string kernel
	5	random forest		5	kScore
C3	1	SVM linear kernel	R3	1	random forest
	2	LS-SVM RBF kernel		2	kernel PLS
	3	SVM		3	SVM string kernel
	4	C4.5		4	PLS
	5	random forest		5	Gaussian process
C4	1	LS-SVM quadratic kernel	R3D2	1	kernel PLS
	2	SVM with AA sequence		2	SVM string kernel
	3	SVM binary encoding of AA		3	random forest
	4	kScore		4	Gaussian process
	5	random forest		5	SVM binary encoding of AA

carcinoma) with 3599 chemicals (2049 in class +1, 1550 in class −1) and 30 descriptors (Tab. 2). For an easier comparison of the ranking of all ML methods, the results are ordered after the Matthews correlation coefficient MCC. The prediction statistical indices show that *k*-NN, in its several variants evaluated here, constantly gives the best predictions. Three other ML with good predictions are KStar, SVM, and RandomForest.

The second group of structure-activity models is obtained for the cell line SF-268 (glioma) with 3721 chemicals (2020 in class +1, 1701 in class −1) and 36 descriptors (Tab. 3). A *k*-NN classifier is in the first position, followed by SVM and other *k*-NN variants. In decreasing order of the predictions, the list continues with another lazy learning method, KStar, followed by RandomForest and J48.

The third set of experiments considers the cell line SK-MEL-5 (melanoma) with 3685 chemicals (2034 in class +1, 1651 in class −1) and 30 descriptors (Tab. 4). All *k*-NN give good predictions, including the top five results. RandomForest ranks better compared with the previous two cell lines, and it is followed by SVM and KStar. As a conclusion of these three sets of experiments, we find that a small group of ML algorithms gives constantly the best predictions. The top results are obtained with *k*-NN, which has good statistics for the whole range of parameters tested here. This is an important result, because *k*-NN

is a robust method that is much easier to compute than SVM. Three other ML algorithms have good predictions, namely SVM, KStar, and RandomForest. Although KStar is not used in drug design, the results reported here indicate that this ML algorithm should be added to the toolbox of structure-activity methods. We have to note the absence of the Bayesian classifiers, although they are the preferred method in some library screening experiments.

The studies reviewed in this section show that it is difficult to perform comprehensive comparisons for a large number of machine learning algorithms. To address this issue, we organized a competition for the ML evaluation in blind predictions, CoEPra 2006 (Comparative Evaluation of Prediction Algorithms, <http://www.coepra.org/>). For each prediction task the participants received a training dataset (structure, descriptors, and class attribution or activity) and a prediction dataset (structure and descriptors). All predictions received before deadline were compared with the experimental results, and then the ML models were ranked after different statistical indices (Tab. 5). The competition had four classification (C in Tab. 5) tasks and four regression (R in Tab. 5) tasks. All datasets consisted of peptides evaluated for their binding affinity to the major histocompatibility complex (MHC) proteins. To evaluate the general performance of the algorithms tested in competition, for each task we present the ML from the top

five positions. In general, each task has a different winner and order for the top ML. The most frequent ML is SVM, in several variants. The least squares SVM (LS-SVM) is particularly robust, insensitive to noise, and with good performance. Currently, LS-SVM is not used in virtual screening, but the results from the competition show that it should be included in the set of preferred methods. Other predictive ML methods are random forest, Gaussian process, and kScore. The CoEPrA competition offered interesting conclusions with direct implications in drug design, SAR, and QSAR. Similar experiments are necessary to evaluate datasets consisting of drugs and drug-like chemicals.

Future Directions

For a long period of time the machine learning selection available to the computational chemist was restricted to a small number of algorithms. However, recent publications explore a much larger diversity of statistical models, with the intention to help the drug discovery process with reliable predictions. As an example of the current trend, one should consider the fast transition of the support vector machines from small-scale experiments to the preferred method in many industrial applications. Following the SVM success, other kernel methods were adopted in chemoinformatics and drug design. The results reviewed here show that it is not possible to predict which combination of structural descriptors and machine learning will give the most reliable predictions for novel chemicals. Therefore, in order to identify good models, it is always a good idea to compare several types of descriptors and as many ML algorithms as computationally possible. Equally important is a good plan for the experiments, that includes a proper validation and whenever possible tests with external data that were not used in training or cross-validation. Of great importance is an objective and comprehensive evaluation of ML algorithms, which can be performed in settings similar with the CoEPrA competition. Such blind predictions may offer an unbiased ranking of the machine learning algorithms.

Bibliography

1. Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. *Mach Learn* 6:37–66
2. Ajmani S, Jadhav K, Kulkarni SA (2006) Three-dimensional QSAR using the *k*-nearest neighbor method and its interpretation. *J Chem Inf Model* 46:24–31
3. Andres C, Hutter MC (2006) CNS permeability of drugs predicted by a decision tree. *QSAR Comb Sci* 25:305–309
4. Alpaydin E (2004) Introduction to machine learning. MIT Press, Cambridge, p 445
5. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. *Artif Intell Rev* 11:11–73
6. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning for control. *Artif Intell Rev* 11:75–113
7. Arimoto R, Prasad MA, Gifford EM (2005) Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. *J Biomol Screen* 10:197–205
8. Balaban AT, Ivanciuc O (1999) Historical development of topological indices. In: Devillers J, Balaban AT (eds) *Topological indices and related descriptors in QSAR and QSPR*. Gordon & Breach Science Publishers, Amsterdam, pp 21–57
9. Basak SC, Grunwald GD (1995) Molecular similarity and estimation of molecular properties. *J Chem Inf Comput Sci* 35:366–372
10. Basak SC, Bertelsen S, Grunwald GD (1994) Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. *J Chem Inf Comput Sci* 34:270–276
11. Basak SC, Bertelsen S, Grunwald GD (1995) Use of graph theoretic parameters in risk assessment of chemicals. *Toxicol Lett* 79:239–250
12. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Philos Trans Roy Soc London* 53:370–418
13. Bender A, Jenkins JL, Glick M, Deng Z, Nettles JH, Davies JW (2006) “Bayes affinity fingerprints” improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J Chem Inf Model* 46:2445–2456
14. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL (2007) Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *Chem Med Chem* 2:861–873
15. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin, p 740
16. Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press, Oxford, p 504
17. Boid DB (2007) How computational chemistry became important in the pharmaceutical industry. In: Lipkowitz KB, Cundari TR (eds) *Reviews in computational chemistry*, vol 23. Wiley, Weinheim, pp 401–451
18. Bonchev D (1983) Information theoretic indices for characterization of chemical structure. Research Studies Press, Chichester
19. Bonchev D, Rouvray DH (eds) (1991) Chemical graph theory. Introduction and fundamentals. Abacus Press/Gordon & Breach Science Publishers, New York
20. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) *Proc of the 5th annual ACM workshop on computational learning theory*. ACM Press, Pittsburgh, pp 144–152
21. Bottou L, Chapelle O, DeCoste D, Weston J (2007) Large-scale kernel machines. MIT Press, Cambridge, p 416
22. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
23. Briem H, Günther J (2005) Classifying “kinase inhibitor-likeness” by using machine-learning methods. *Chem Bio Chem* 6:558–566
24. Cash GG (1999) Prediction of physicochemical properties from Euclidean distance methods based on electrotopological state indices. *Chemosphere* 39:2583–2591

25. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. *IEEE Trans Neural Netw* 10:1055–1064
26. Cleary JG, Trigg LE (1995) K^* : an instance-based learner using and entropic distance measure. In: Prieditis A, Russell SJ (eds) *Proc of the 12th international conference on machine learning*. Morgan Kaufmann, Tahoe City, pp 108–114
27. Cohen WW (1995) Fast effective rule induction. In: Prieditis A, Russell SJ (eds) *Proc of the 12th international conference on machine learning*. Morgan Kaufmann, Tahoe City, pp 115–123
28. Cortes C, Vapnik V (1995) Support vector networks. *Mach Learn* 20:273–297
29. Cristianini N, Shawe-Taylor J (2000) *An introduction to support vector machines*. Cambridge University Press, Cambridge
30. Deconinck E, Zhang MH, Coomans D, Vander Heyden Y (2006) Classification tree models for the prediction of blood-brain barrier passage of drugs. *J Chem Inf Model* 46:1410–1419
31. Deng Z, Chuaqui C, Singh J (2006) Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J Med Chem* 49:490–500
32. Doddareddy MR, Cho YS, Koh HY, Kim DH, Pae AN (2006) In silico renal clearance model using classical Volsurf approach. *J Chem Inf Model* 46:1312–1320
33. Drucker H, Wu DH, Vapnik VN (1999) Support vector machines for spam categorization. *IEEE Trans Neural Netw* 10:1048–1054
34. Du H, Wang J, Watzl J, Zhang X, Hu Z (2008) Classification structure-activity relationship (CSAR) studies for prediction of genotoxicity of thiophene derivatives. *Toxicol Lett* 177: 10–19
35. Duda RO, Hart PE, Stork DG (2000) *Pattern classification*. 2nd edn. Wiley, New York
36. Ehrman TM, Barlow DJ, Hylands PJ (2007) Virtual screening of chinese herbs with random forest. *J Chem Inf Model* 47:264–278
37. Eitrich T, Kless A, Druska C, Meyer W, Grotendorst J (2007) Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J Chem Inf Model* 47:92–103
38. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y (2006) Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *J Med Chem* 49:5059–5071
39. Ertl P, Roggo S, Schuffenhauer A (2008) Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* 48:68–74
40. Fatemi MH, Gharaghani S (2007) A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine. *Bioorg Med Chem* 15:7746–7754
41. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. *Bioinformatics* 20:2479–2481
42. Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In: Bratko I, Dzeroski S (eds) *Proc of the 16th international conference on machine learning (ICML (1999))*. Morgan Kaufmann, Bled, pp 124–133
43. Gaines BR, Compton P (1995) Induction of ripple-down rules applied to modeling large databases. *Intell J Inf Syst* 5:211–228
44. Gao JB, Gunn SR, Harris CJ (2003) SVM regression through variational methods and its sequential implementation. *Neurocomputing* 55:151–167
45. Gao JB, Gunn SR, Harris CJ (2003) Mean field method for the support vector machine regression. *Neurocomputing* 50:391–405
46. Gepp MM, Hutter MC (2006) Determination of hERG channel blockers using a decision tree. *Bioorg Med Chem* 14:5325–5332
47. Guha R, Dutta D, Jurs PC, Chen T (2006) Local lazy regression: making use of the neighborhood to improve QSAR predictions. *J Chem Inf Model* 46:1836–1847
48. Gute BD, Basak SC (2001) Molecular similarity-based estimation of properties: a comparison of three structure spaces. *J Mol Graph Modell* 20:95–109
49. Gute BD, Basak SC, Mills D, Hawkins DM (2002) Tailored similarity spaces for the prediction of physicochemical properties. *Internet Electron J Mol Des* 1:374–387
50. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46:389–422
51. Hansch C, Garg R, Kurup A, Mekapati SB (2003) Allosteric interactions and QSAR: on the role of ligand hydrophobicity. *Bioorg Med Chem* 11:2075–2084
52. Hastie T, Tibshirani R, Friedman JH (2003) *The elements of statistical learning*. Springer, Berlin, p 552
53. Herbrich R (2002) *Learning kernel classifiers*. MIT Press, Cambridge
54. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model* 46:462–470
55. Hoffman B, Cho SJ, Zheng W, Wyrick S, Nichols DE, Mailman RB, Tropsha A (1999) Quantitative structure-activity relationship modeling of dopamine D_1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K -nearest neighbor methods. *J Med Chem* 42:3217–3226
56. Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. *Mach Learn* 11:63–90
57. Hou T, Wang J, Zhang W, Xu X (2007) ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J Chem Inf Model* 47:208–218
58. Huang T-M, Kecman V, Kopriva I (2006) *Kernel based algorithms for mining huge data sets*. Springer, Berlin, p 260
59. Hudelson MG, Ketkar NS, Holder LB, Carlson TJ, Peng C-C, Waldher BJ, Jones JP (2008) High confidence predictions of drug-drug interactions: predicting affinities for cytochrome P450 2C9 with multiple computational methods. *J Med Chem* 51:648–654
60. Itskowitz P, Tropsha A (2005) k -nearest neighbors QSAR modeling as a variational problem: theory and applications. *J Chem Inf Model* 45:777–785
61. Ivanciuc O (2002) Support vector machine classification of the carcinogenic activity of polycyclic aromatic hydrocarbons. *Internet Electron J Mol Des* 1:203–218
62. Ivanciuc O (2002) Structure-odor relationships for pyrazines

- with support vector machines. *Internet Electron J Mol Des* 1:269–284
63. Ivanciuc O (2002) Support vector machine identification of the aquatic toxicity mechanism of organic compounds. *Internet Electron J Mol Des* 1:157–172
 64. Ivanciuc O (2003) Graph theory in chemistry. In: Gasteiger J (ed) *Handbook of chemoinformatics*, vol 1. Wiley, Weinheim, pp 103–138
 65. Ivanciuc O (2003) Topological indices. In: Gasteiger J (ed) *Handbook of chemoinformatics*, vol 3. Wiley, Weinheim, pp 981–1003
 66. Ivanciuc O (2003) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. *Internet Electron J Mol Des* 2:195–208
 67. Ivanciuc O (2004) Support vector machines prediction of the mechanism of toxic action from hydrophobicity and experimental toxicity against pimephales promelas and tetrahymena pyriformis. *Internet Electron J Mol Des* 3:802–821
 68. Ivanciuc O (2005) Support vector regression quantitative structure-activity relationships (QSAR) for benzodiazepine receptor ligands. *Internet Electron J Mol Des* 4:181–193
 69. Ivanciuc O (2005) Machine learning applied to anticancer structure-activity relationships for NCI human tumor cell lines. *Internet Electron J Mol Des* 4:948–958
 70. Ivanciuc O (2007) Applications of support vector machines in chemistry. In: Lipkowitz KB, Cundari TR (eds) *Reviews in computational chemistry*, vol 23. Wiley, Weinheim, pp 291–400
 71. John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Besnard P, Hanks S (eds) *UAI '95: Proc of the 11th annual conference on uncertainty in artificial intelligence*. Morgan Kaufmann, Montreal, pp 338–345
 72. Jorissen RN, Gilson MK (2005) Virtual screening of molecular databases using a support vector machine. *J Chem Inf Model* 45:549–561
 73. Jurs P (2003) Quantitative structure-property relationships. In: Gasteiger J (ed) *Handbook of chemoinformatics*, vol 3. Wiley, Weinheim, pp 1314–1335
 74. Kier LB, Hall LH (1976) *Molecular connectivity in chemistry and drug research*. Academic Press, New York
 75. Kier LB, Hall LH (1986) *Molecular connectivity in structure-activity analysis*. Research Studies Press, Letchworth
 76. Kier LB, Hall LH (1999) *Molecular structure description. The electrotopological state*. Academic Press, San Diego
 77. Klon AE, Diller DJ (2007) Library fingerprints: a novel approach to the screening of virtual libraries. *J Chem Inf Model* 47:1354–1365
 78. Klon AE, Glick M, Davies JW (2004) Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J Med Chem* 47:4356–4359
 79. Klon AE, Glick M, Thoma M, Acklin P, Davies JW (2004) Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. *J Med Chem* 47:2743–2749
 80. Klon AE, Lowrie JF, Diller DJ (2006) Improved naïve Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J Chem Inf Model* 46:1945–1956
 81. Kohavi R (1995) The power of decision tables. In: Lavrac N, Wrobel S (eds) *ECML-95 8th european conference on machine learning*. Lecture Notes in Computer Science, vol 912. Springer, Heracleion, pp 174–189
 82. Kohavi R (1996) Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In: Simoudis E, Han J, Fayyad UM (eds) *Proc of the 2nd international conference on knowledge discovery and data mining (KDD-96)*. AAAI Press, Menlo Park, pp 202–207
 83. Kononenko I, Kukar M (2007) *Machine learning and data mining: introduction to principles and algorithms*. Horwood, Westergate, p 454
 84. Kononov DA, Coomans D, Deconinck E, Vander Heyden Y (2007) Benchmarking of QSAR models for blood-brain barrier permeation. *J Chem Inf Model* 47:1648–1656
 85. Kumar R, Kulkarni A, Jayaraman VK, Kulkarni BD (2004) Structure-activity relationships using locally linear embedding assisted by support vector and lazy learning regressors. *Internet Electron J Mol Des* 3:118–133
 86. le Cessie S, van Houwelingen JC (1992) Ridge estimators in logistic regression. *Appl Statist* 41:191–201
 87. Leong MK (2007) A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem Res Toxicol* 20:217–226
 88. Lepp Z, Kinoshita T, Chuman H (2006) Screening for new antidepressant leads of multiple activities by support vector machines. *J Chem Inf Model* 46:158–167
 89. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J Chem Inf Model* 45:1376–1384
 90. Li S, Fedorowicz A, Singh H, Soderholm SC (2005) Application of the random forest method in studies of local lymph node assay based skin sensitization data. *J Chem Inf Model* 45:952–964
 91. Li W-X, Li L, Eksterowicz J, Ling XB, Cardozo M (2007) Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. *J Chem Inf Model* 47:2429–2438
 92. Liao Q, Yao J, Yuan S (2007) Prediction of mutagenic toxicity by combination of recursive partitioning and support vector machines. *Mol Divers* 11:59–72
 93. Mangasarian OL, Musicant DR (2000) Robust linear and support vector regression. *IEEE Trans Pattern Anal Mach Intell* 22:950–955
 94. Mangasarian OL, Musicant DR (2002) Large scale kernel regression via linear programming. *Mach Learn* 46:255–269
 95. Merkwirth C, Mauser HA, Schulz-Gasch T, Roche O, Stahl M, Lengauer T (2004) Ensemble methods for classification in cheminformatics. *J Chem Inf Comput Sci* 44:1971–1978
 96. Mitchell TM (1997) *Machine learning*. McGraw-Hill, Maidenhead, p 432
 97. Müller K-R, Rätsch G, Sonnenburg S, Mika S, Grimm M, Heinrich N (2005) Classifying 'drug-likeness' with kernel-based learning methods. *J Chem Inf Model* 45:249–253
 98. Neugebauer A, Hartmann RW, Klein CD (2007) Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. *J Med Chem* 50:4665–4668
 99. Neumann D, Kohlbacher O, Merkwirth C, Lengauer T (2006) A fully computational model for predicting percutaneous drug absorption. *J Chem Inf Model* 46:424–429
 100. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction

- of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46:1124–1133
101. Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JBO (2006) Melting point prediction employing *k*-nearest neighbor algorithms and genetic parameter optimization. *J Chem Inf Model* 46:2412–2422
102. Oloff S, Muegge I (2007) kScore: a novel machine learning approach that is not dependent on the data structure of the training set. *J Comput-Aided Mol Des* 21:87–95
103. Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A (2006) Chemometric analysis of ligand receptor complementarity: Identifying complementary ligands based on receptor information (CoLiBRI). *J Chem Inf Model* 46:844–851
104. Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO (2007) Random forest models to predict aqueous solubility. *J Chem Inf Model* 47:150–158
105. Pelletier DJ, Gehlhaar D, Tilloy-Ellul A, Johnson TO, Greene N (2007) Evaluation of a published in silico model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. *J Chem Inf Model* 47:1196–1205
106. Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ (eds) *Advances in kernel methods – support vector learning*. MIT Press, Cambridge, pp 185–208
107. Plewczynski D, Spieser SAH, Koch U (2006) Assessing different classification methods for virtual screening. *J Chem Inf Model* 46:1098–1106
108. Quinlan R (1993) *C4.5: programs for machine learning*. Morgan Kaufmann, San Mateo
109. Ren S (2002) Classifying class I and class II compounds by hydrophobicity and hydrogen bonding descriptors. *Environ Toxicol* 17:415–423
110. Ripley BD (2008) *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, p 416
111. Rodgers S, Glen RC, Bender A (2006) Characterizing bitterness: identification of key structural features and development of a classification model. *J Chem Inf Model* 46:569–576
112. Rusinko A, Farnen MW, Lambert CG, Brown PL, Young SS (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J Chem Inf Comput Sci* 39:1017–1026
113. Sakiyama Y, Yuki H, Moriya T, Hattori K, Suzuki M, Shimada K, Honma T (2008) Predicting human liver microsomal stability with machine learning techniques. *J Mol Graph Modell* 26:907–915
114. Schneider N, Jäckels C, Andres C, Hutter MC (2008) Gradual in silico filtering for druglike substances. *J Chem Inf Model* 48:613–628
115. Schölkopf B, Smola AJ (2002) *Learning with kernels*. MIT Press, Cambridge
116. Schölkopf B, Sung KK, Burges CJC, Girosi F, Niyogi P, Poggio T, Vapnik V (1997) Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans Signal Process* 45:2758–2765
117. Schölkopf B, Burges CJC, Smola AJ (1999) *Advances in kernel methods: support vector learning*. MIT Press, Cambridge
118. Schroeter TS, Schwaighofer A, Mika S, ter Laak A, Suelzle D, Ganzer U, Heinrich N, Müller K-R (2007) Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J Comput-Aided Mol Des* 21:485–498
119. Shawe-Taylor J, Cristianini N (2004) *Kernel methods for pattern analysis*. Cambridge University Press, Cambridge
120. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A (2002) Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using *k*-nearest neighbor and simulated annealing PLS methods. *J Med Chem* 45:2811–2823
121. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A (2003) Development and validation of *k*-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J Med Chem* 46:3013–3020
122. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14:199–222
123. Sommer S, Kramer S (2007) Three data mining techniques to improve lazy structure-activity relationships for noncongeneric compounds. *J Chem Inf Model* 47:2035–2043
124. Sorch MJ, McKinnon RA, Miners JO, Smith PA (2006) The importance of local chemical structure for chemical metabolism by human uridine 5'-diphosphate-glucuronosyltransferase. *J Chem Inf Model* 46:2692–2697
125. Sun H (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J Med Chem* 48:4031–4039
126. Suykens JAK (2001) Support vector machines: a nonlinear modelling and control perspective. *Eur J Control* 7:311–327
127. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) *Least squares support vector machines*. World Scientific, Singapore
128. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 43:1947–1958
129. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q (2005) Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J Chem Inf Model* 45:786–799
130. Swamidass SJ, Chen J, Phung P, Ralaivola L, Baldi P (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* 21[S1]:i359–i368
131. Terfloth L, Bienfait B, Gasteiger J (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J Chem Inf Model* 47:1688–1701
132. Tobita M, Nishikawa T, Nagashima R (2005) A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. *Bioorg Med Chem Lett* 15:2886–2890
133. Todeschini R, Consonni V (2003) Descriptors from molecular geometry. In: Gasteiger J (ed) *Handbook of chemoinformatics*, vol 3. Wiley, Weinheim, pp 1004–1033
134. Tong W, Hong H, Fang H, Xie Q, Perkins R (2003) Decision forest: Combining the predictions of multiple independent decision tree models. *J Chem Inf Comput Sci* 43:525–531
135. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Env Health Perspect* 112:1249–1254
136. Trinajstić N (1992) *Chemical graph theory*. CRC Press, Boca Raton

137. Urrestarazu Ramos E, Vaes WHJ, Verhaar HJM, Hermens JLM (1998) Quantitative structure-activity relationships for the aquatic toxicity of polar and nonpolar narcotic pollutants. *J Chem Inf Comput Sci* 38:845–852
138. Vapnik VN (1979) Estimation of dependencies based on empirical data. Nauka, Moscow
139. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
140. Vapnik VN (1998) Statistical learning theory. Wiley, New York
141. Vapnik VN (1999) An overview of statistical learning theory. *IEEE Trans Neural Netw* 10:988–999
142. Vapnik V, Chapelle O (2000) Bounds on error expectation for support vector machines. *Neural Comput* 12:2013–2036
143. Vapnik VN, Chervonenkis AY (1974) Theory of pattern recognition. Nauka, Moscow
144. Vapnik V, Lerner A (1963) Pattern recognition using generalized portrait method. *Automat Remote Control* 24:774–780
145. Varnek A, Kireeva N, Tetko IV, Baskin II, Solov'ev VP (2007) Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *J Chem Inf Model* 47:1111–1122
146. Vogt M, Bajorath J (2008) Bayesian similarity searching in high-dimensional descriptor spaces combined with Kullback–Leibler descriptor divergence analysis. *J Chem Inf Model* 48:247–255
147. von Korff M, Sander T (2006) Toxicity-indicating structural patterns. *J Chem Inf Model* 46:536–544
148. Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, Tropsha A (2006) QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J Med Chem* 49:7169–7181
149. Wang J, Du H, Yao X, Hu Z (2007) Using classification structure pharmacokinetic relationship (SCPR) method to predict drug bioavailability based on grid-search support vector machine. *Anal Chim Acta* 601:156–163
150. Watson P (2008) Naïve Bayes classification using 2D pharmacophore feature triplet vectors. *J Chem Inf Model* 48:166–178
151. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco, p 525
152. Xiao Z, Xiao Y-D, Feng J, Golbraikh A, Tropsha A, Lee K-H (2002) Antitumor agents. 213. Modeling of epipodophyllotoxin derivatives using variable selection *k*-nearest neighbor QSAR method. *J Med Chem* 45:2294–2309
153. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (2004) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J Chem Inf Comput Sci* 44:1630–1638
154. Yamashita F, Hara H, Ito T, Hashida M (2008) Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. *J Chem Inf Model* 48:364–369
155. Yap CW, Chen YZ (2005) Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J Chem Inf Model* 45:982–992
156. Yap CW, Cai CZ, Xue Y, Chen YZ (2004) Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol Sci* 79:170–177
157. Yu G-X, Park B-H, Chandramohan P, Munavalli R, Geist A, Samatova NF (2005) In silico discovery of enzyme-substrate specificity-determining residue clusters. *J Mol Biol* 352:1105–1117
158. Yue P, Li Z, Moulton J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 353:459–473
159. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha A (2006) A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J Chem Inf Model* 46:1984–1995
160. Zhang S, Golbraikh A, Tropsha A (2006) Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J Med Chem* 49:2713–2724
161. Zheng WF, Tropsha A (2000) Novel variable selection quantitative structure-property relationship approach based on the *k*-nearest-neighbor principle. *J Chem Inf Comput Sci* 40:185–194

Drug Design, Molecular Descriptors in

ALEXANDRU T. BALABAN
Texas A&M University, Galveston, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Molecular Properties and Biological Activities
 Linear Free Energy Relationships, QSAR and QSPR
 Molecular Descriptors
 Graph Theoretical Intermezzo
 Topological Indices (TIs)
 A Real-World Example of Using Molecular Descriptors Including TIs
 Three-Dimensional Molecular Descriptors
 Molecular Similarity, Dissimilarity, and Clustering of Molecular Descriptors
 Hierarchical Ordering of Structures and of Molecular Descriptors
 General Problems Connected with 2D-QSAR for Drug Design
 Interpretation of Topological Indices
 Reverse Engineering of Molecular Descriptors
 Future Directions
 Bibliography

Glossary

ADMET Absorption, distribution, metabolism, excretion, and toxicity, five factors which determine whether a compound with favorable features resulting from

a computer-assisted drug design will also be a useful medicinal drug.

Local vertex invariant (LOVI) A number associated with a vertex of a molecular graph. It does not depend on the arbitrary numerical labeling of vertices.

Molecular descriptor A mathematical entity (either a number, or a set of numbers characterizing substituents or molecular fragments) associated with a chemical structure, allowing quantitative manipulations of such structures (correlations with properties, clustering, ordering or partial ordering, determination of similarity or dissimilarity). Constitutional isomerism is associated with topological indices or other two-dimensional descriptors, whereas stereoisomerism requires three-dimensional descriptors.

Molecular graph Constitutional chemical formula representing atoms by points (vertices) and covalent bonds by lines (edges). Hydrogen atoms are usually omitted (in hydrogen-depleted graphs) and carbon atoms are not shown explicitly, but other atoms are assigned weights according to the nature of the heteroatom. Multiple bonds are shown explicitly in multigraphs.

Pharmacophore A set of chemical features that determine the biological activity.

Quantitative structure-activity relationship (QSAR)

Mathematical correlation (e.g. a mono- or multi-parametric equation) between a physical or a chemical property and molecular descriptor(s).

Receptor Physiological macromolecule (usually a protein, but sometimes DNA or RNA), also called target, to which a drug (ligand) binds.

Quantitative structure-property relationship (QSPR)

Mathematical correlation (e.g. a mono- or multi-parametric equation) between a biological property and molecular descriptor(s).

Quantitative structure-toxicity relationship (QSTR)

Mathematical correlation (e.g. a mono- or multi-parametric equation) between toxicologic properties and molecular descriptor(s).

Topological index (TI) A number characterizing a molecular constitutional graph, resulting from mathematical operations on the LOVIs, or edges/vertices of the graph.

Definition of the Subject

Empirical observations of association between chemical structures and their physical, chemical, or biological properties have a long history [1], but the real development of mathematical correlations started in the 20th century. For physico-chemical properties, topologi-

cal indices and their applications were first published by Wiener in 1947 followed soon by Platt and Hosoya, as will be shown later. Quantitative structure-activity relationships (QSARs) have 1962 as official birthdates with the connection between Hammett substituent constants, lipophilicity, and biological activities, as will be discussed below.

The huge development in analytical and synthetic methods, leading to combinatorial chemistry, high-throughput screening, and virtual screening (ligand- or structure-based) on one hand, and on the other hand to the progress in computational hardware and software that allows the processing of millions of virtual structure libraries, have made it possible to harness and combine the power of these workhorses for exploring the infinity of “druglike” structures in order to extract information for drug design. Whereas high-throughput screening detects biologically active compounds from a mixture of existing molecules produced by combinatorial synthesis, virtual screening uses computer programs simulating imagined structures. Virtual screening is of two types: ligand-based when the receptor’s data are not known, or structure-based when the target structure is available, especially at high atomic resolution.

Introduction

There are four main reasons why new medicinal drugs have to be continuously explored, tested, and brought to the market: (i) bacterial resistance, exemplified nowadays by the far lower success of older antibiotics as compared to the situation when they were first introduced; (ii) after completion of the Human Genome Project (with the corollaries of genomics, proteomics, and metabolomics) it has become possible to design molecules that interact with specific proteins, alleviating or curing ailments due either to the lack or to the abundance of these proteins; (iii) the increased life span brought to prominence many diseases characteristic of old age, that need new treatments and new types of drugs; and (iv) it may be possible to replace defective proteins by gene therapy, which needs either modified viral vectors or non-viral vectors (cationic lipids) – in both cases by using molecular design [2,3].

The traditional manner in which a new drug was developed consisted in uncovering a lead compound whose molecular structure was then systematically varied by synthetic chemists, followed in each case by sequential testing in vitro (i.e. in cell cultures), in vivo (on lower and then on higher organisms), and finally by clinical testing on humans. An exponential decrease in the number of drug candidates follows each of these steps, but the final cost

still attains staggering proportions – nowadays for each new drug reaching the market, the “big pharma” companies invest hundreds of millions of dollars. It is interesting to note that only half a dozen countries (USA, UK, France, Germany, Switzerland, and Sweden) account for most of the new medicinal drugs that are launched on the market. For lowering the cost and time needed for new drugs, these companies must resort to such strategies as mergers, outsourcing, high-throughput screening of thousands of compounds synthesized in one batch by robots using combinatorial chemistry approaches, and especially by preceding the sequence “chemical laboratory – in vitro – in vivo” with a novel strategy, aptly called “in silico”. Taking into account that patents expire after 20 years and that after this interval generic drug manufacturers (located mainly in Asia) are able to sell such drugs at a much lower price, which make them more convenient for insurance companies and for the public at large, it is understandable why new drugs cost so much: the “big pharma” companies, which are practically the only ones investing in drug discovery and testing, must recover their investment in about ten years (this is the lifetime of a new drug from the date of FDA approval till generic drugs appear on the market). A medicinal drug that is withdrawn from the market causes enormous financial losses, also due to court litigations. Two relevant papers are mentioned in this regard [4,5].

Molecular Properties and Biological Activities

A first classification of molecular properties is whether such properties depend on individual molecules in vacuum (e. g. heats of formation, molecular weights, molecular dimensions) or on molecules interacting among themselves in bulk or condensed phases (e. g. normal boiling points, critical pressure and temperature, solubility and partition coefficients) [6]. A second classification is according to the type of properties (physical, chemical, or biological), and of course subdivisions for each of these properties (for instance toxicity, bioactivity in interaction with certain enzymes for certain cell cultures or animals).

All properties are measured and expressed on numerical scales, achieving a metric and ordering of properties. In drug design there is not much incentive in establishing correlations between various properties because one wishes to explore structures that have not yet been prepared. On the other hand, chemical structures are discrete entities and in order to establish a structure-property correlation one has to associate numbers with structural formulas. A simple way to do this is to translate struc-

tures into molecular invariants, which can be topological indices [7,8,9,10,11,12,13,14] or assemblies of molecular fragments (substructures). *The present review will concentrate on the former approach.*

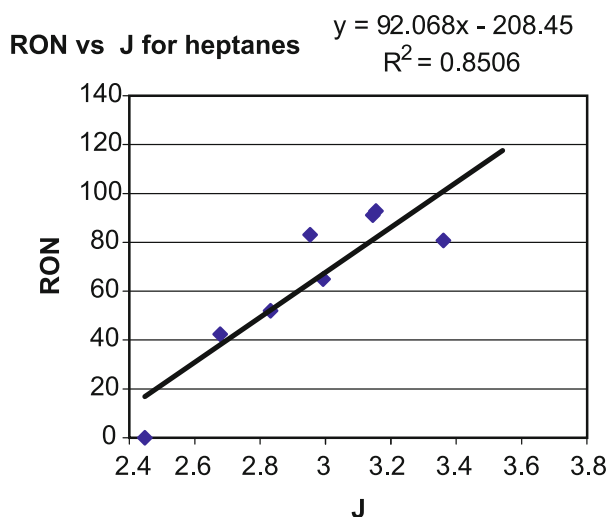
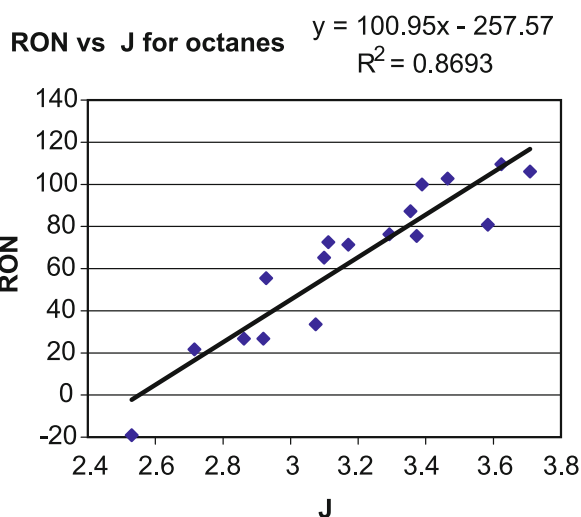
All elementary textbooks of Organic Chemistry mention that with increasing *branching*, the normal boiling points (NBPs) of isomeric alkanes decrease whereas the Research Octane Numbers (RONs) increase. The former (NBP) is a molecular physical property, and the latter (RON) is a molecular chemical property. Both are bulk properties that can be measured precisely and accurately (unlike biological activities which are usually measured with poorer reproducibility). On examining the data presented in Table 1 for all isomers of heptane and octane one sees, however, that the intercorrelation factor for RON vs. NBP is $r^2 = 0.6454$ for heptanes and only 0.2068 for octanes; for the combined set it is even lower. On sorting each of these two sets of isomeric alkanes according to decreasing NBP and increasing RON, one obtains different orderings. Therefore “*branching*” means something different for the NBP and for the RON properties, as it was stressed in the first sentence of this paragraph.

On the other hand, a shape parameter such as index J (to be discussed later) presents better correlation factors, as seen in Fig. 1. One cannot deny a certain arbitrariness in the experimentally defined number RON as the amount k (%) of a binary mixture of k parts of 2,2,4-trimethylpentane or isooctane and $100 - k$ parts of n-heptane that develops the same traction in a standard spark-initiated engine as the gasoline being tested. This number then defines a certain operational type of branching, just as there is another type of branching also defined operationally by dividing into 100 parts (called Kelvin or Celsius degrees) the temperature interval between the phase transitions of water under normal pressure, and comparing the boiling temperatures of various substances at normal pressure on these temperature scales. The conclusion is that for molecular branching one needs many molecular descriptors that can account for the corresponding types of branching defined by operational experiments. And this is also valid for other fuzzy properties such as *cyclicality*, *centricity*, or *complexity*. Yet these notions are deeply ingrained and familiar to all of us, and they serve for allowing us to ascribe to objects a certain metric. If we can also establish a numerical scale for such notions, then we can find quantitative structure-property relationships (QSPR for physical-chemical properties) or quantitative structure-activity relationships (QSAR for biological activities). Of course, in some cases the ordering induced by such a numerical scale is only a partial order.

Drug Design, Molecular Descriptors in, Table 1

Six molecular descriptors and two physico-chemical properties (research octane numbers, and normal boiling points, in °C) of all C_7H_{16} and C_8H_{18} alkanes

Alkanes sorted by <i>J</i>	Path code	<i>n</i>	<i>W</i>	<i>Z</i>	<i>J</i>	<i>Q</i>	$^1\chi$	<i>P</i>	RON	NBP
Heptane	6,5,4,3,2,1	7	56	2 1	2.447 5	91	3.4142	7.09205	0	98.4
2-Methylhexane	6,6,4,3,2	7	52	1 8	2.678 3	101	3.2701	6.83472	42.4	90.1
3-Methylhexane	6,6,5,3,1	7	50	1 9	2.831 8	107	3.3081	6.78577	52.0	91.9
2,4-Dimethylpentane	6,7,4,4	7	48	1 5	2.953 2	117	3.1259	6.48502	83.1	79.2
3-Ethylpentane	6,6,6,3	7	48	2 0	2.992 3	117	3.3461	6.47501	65.0	89.8
2,3-Dimethylpentane	6,7,6,2	7	46	1 7	3.144 2	125	3.1807	6.47021	91.1	80.5
2,2-Dimethylpentane	6,8,4,3	7	46	1 4	3.154 5	125	3.0607	6.46178	92.8	86.0
3,3-Dimethylpentane	6,8,6,1	7	44	1 6	3.360 4	137	3.1213	6.44163	80.8	93.5
2,2,3-Trimethylbutane	6,9,6	7	42	1 3	3.541 2	153	2.9434	5.98502	112.1	80.9
Octane	7,6,5,4,3,2,1	8	84	3 4	2.530 1	140	2.5301	8.3987	−19.0	125.7
2-Methylheptane	7,7,5,4,3,2	8	79	2 9	2.715 8	152	2.7158	8.15952	21.7	117.7
3-Methylheptane	7,7,6,4,3,1	8	76	3 1	2.862 1	160	2.8621	8.11363	26.8	117.0
4-Methylheptane	7,7,6,5,2,1	8	75	3 0	2.919 6	164	2.9196	8.08953	26.7	117.7
2,5-Dimethylhexane	7,8,5,4,4	8	74	2 5	2.927 8	170	2.9278	7.83266	55.5	106.8
3-Ethylhexane	7,7,7,5,2	8	72	3 2	3.074 4	176	3.0744	7.83117	33.5	114.0
2,4-Dimethylhexane	7,8,6,5,2	8	71	2 6	3.098 8	178	3.0988	7.81045	65.2	109.0
2,2-Dimethylhexane	7,9,5,4,3	8	71	2 3	3.111 8	180	3.1118	7.80573	72.5	109.0
2,3-Dimethylhexane	7,8,7,4,2	8	70	2 7	3.170 8	182	3.1708	7.79459	71.3	112.0
3,4-Dimethylhexane	7,8,8,4,1	8	68	2 9	3.292 5	194	3.2925	7.74181	76.3	117.7
3,3-Dimethylhexane	7,9,7,4,1	8	67	2 5	3.354 9	196	3.3734	7.72595	87.3	109.8
3-Ethyl-2-methylpentane	7,8,8,5	8	67	2 8	3.373 4	202	3.3549	7.40006	75.5	118.0
2,2,4-Trimethylpentane	7,10,5,6	8	66	1 9	3.388 9	210	3.3889	7.39752	100.0	99.2
2,3,4-Trimethylpentane	7,9,8,4	8	65	2 4	3.464 2	210	3.4642	7.39677	102.7	114.8
3-Ethyl-3-methylpentane	7,9,9,3	8	64	2 8	3.583 2	220	3.5832	7.38083	80.8	113.5
2,2,3-Trimethylpentane	7,10,8,3	8	63	2 2	3.623 3	222	3.6233	7.36514	109.6	115.6
2,3,3-Trimethylpentane	7,10,9,2	8	62	2 3	3.708 3	234	3.7083	7.32097	106.1	118.3
2,2,3,3-Tetramethylbutane	7,12,9	8	58	1 7	4.020 4	274	4.0240	6.82729	> 100	106.5



Drug Design, Molecular Descriptors in, Figure 1

Plot of Research Octane Numbers (RON) vs. index *J* for C_7H_{16} and C_8H_{18} alkanes

Linear Free Energy Relationships, QSAR and QSPR

Louis P. Hammett published his influential book *Physical Organic Chemistry* in 1940, but he wrote several other reviews earlier on correlations based on electronic effects of substituents in benzene derivatives, showing how chemical structure determined reaction rates [15,16]. He proved that chemical acid-base equilibria (ionization constants of *meta*- or *para*-substituted benzoic acids) can be used to correlate and predict rates of reactions involving benzylic side-chains influenced by the same substituents in *meta*- or *para*-positions. The Hammett equation (1) links specific rates (k , also called rate constants) with equilibrium constants K in terms of Hammett electronic constants σ and reaction constants ρ for unsubstituted (k_H , K_H) and substituted benzene derivatives (k_X , K_X). The σ_X constants are determined for *meta*- or *para*-substituents from pK_a values of X-substituted benzoic acids; electron-donor substituents have $\sigma < 0$, and electron-acceptor substituents have $\sigma > 0$. Steric effects in *ortho*-positions complicate the situation, but Taft introduced steric parameters that can remedy this problem.

$$\rho \times \sigma = \log(k_X/k_H) \quad \text{with} \quad \sigma = \log(K_X/K_H) \quad (1)$$

in water at 25 °C.

Not only kinetic data, but also various physical, chemical, and biological activity data obey such linear free energy relationships (LFER). In the early 60s, Corwin Hansch, Albert Leo and T. Fujita proposed a similar LFER Eq. (3) to account for the partition coefficient between water and amphiphilic solvents (hydrophobicity, denoted by π , a structure-dependent physical property) [17,18,19,20].

$$\pi = [C_{\text{org}}]/[C_{\text{aq}}] \quad (2)$$

where $[C_{\text{org}}]$ and $[C_{\text{aq}}]$ are molar concentrations of the solute at equilibrium in the organic and aqueous phase, respectively. This was the first approach to modern Quantitative Structure-Property Relationships (QSPR) and it continued with Quantitative Structure-Activity Relationships (QSAR) involving biological activities. By choosing as solvent 1-octanol, the lipophilicity π becomes $P = K_{\text{ow}}$ and is determined experimentally as

$$\log \pi = \log P = \log[C_o]/[C_w] = \log[C_{\text{org}}] - \log[C_{\text{aq}}] \quad (3)$$

Compounds with $\log P > 0$ are lipophilic and those with $\log P < 0$ are lipophobic. Several reviews have been published by Hansch in Chem. Rev. drawing attention to the database of about 10^4 organic reactions with known Hammett ρ data, and a similar number of QSAR data for various biological activities; combining these two treasure

troves may offer insight on the mechanism of each biological activity [22,23]. By analogy with the Hammett equation, lipophilicity values can be calculated using a variety of approaches, among which only a few will be mentioned [23,24,25,26].

- (i) The Hansch–Fujita hydrophobic substituent constants

$$\pi_X = (1/\rho) \log(P_X/P_H), \quad (4)$$

where P_X and P_H are the partition coefficients for an X-substituted and unsubstituted compound, respectively. For the 1-octanol/water system, ρ is assumed to be 1. By summing up for all i substituents in the molecule, one obtains the lipophilicity of any compound A:

$$\log P_A = \log P_H + \sum_i \pi_{X,i} \quad (5)$$

- (ii) The Leo–Hansch hydrophobic fragmental constants (calculated $\log P$, CLOGP) can be obtained by means of a computer program or from the Hansch and Leo's books [24,25].
- (iii) The Bodor hydrophobic model (BLOGP) is a non-linear 18-parameter model based on 10 molecular descriptors such as molecular weight, number of carbon atoms, ovality index based on optimized geometry, charges on oxygen and nitrogen atoms.
- (iv) The Gombar hydrophobic model (VLOGP) uses 363 molecular descriptors based on molecular topology (vide infra) such as Kier shape indices and electro-topological indices.
- (v) The topological index of hydrophobicity is based on the Randić molecular connectivity (v.i.).

In many cases, the biological activity is a linear or non-linear function of the lipophilicity, meaning that the determining step is the penetration of the substance through the cell bilayer membrane.

Other molecular descriptors that are being used in drug design involve indicators for: hydrogen bond donors; hydrogen bond acceptors; electrical charges for atoms; conformational flexibility for rotatable chemical bonds; topological surface area; and various quantum-chemical indices, which will be discussed in other chapters of this book.

Molecular Descriptors

One can use two kinds of molecular descriptors, namely computed and experimentally determined ones. Of course, for the latter, one needs to have the substances

whose descriptors one wishes to measure, which makes such descriptors undesirable for drug design, therefore such descriptors will not be discussed here.

The simplest computed molecular descriptors are numerical data on the various types of atoms, chemical bonds, or molecular fragments in the structure. For example, one can enumerate how many carbon, halogen, oxygen, or nitrogen atoms there are in the molecule; how many sp^3 - (primary, secondary, tertiary, quaternary), sp^2 -, or sp -hybridized atoms of each kind; how many hydrogen-bond donor and hydrogen-bond acceptor groups; how many functional groups such as alcohol, phenol, aldehyde, ketone, ester, amine (primary, secondary, tertiary, quaternary), amide, nitrile, oxime, etc. Using force-field calculations or quantum-chemical programs one can compute and optimize molecular geometries and molecular orbital energies, so that the results of these calculations will provide numerical data that can be used as molecular descriptors. Global dimensions of molecules can be expressed as ovality indices. Molecular weights and calculated log P values are also important indicators for drug design [27,28,29,30,31].

In the following, we will describe a class of molecular descriptors known as topological indices (TIs) because they are obtained from the constitutional formula of a compound, and therefore they include (in one number) information on the topology of the molecule, i. e. the way atoms are interconnected. Stereochemical information is not provided at this stage, but may be added via special upgrading of the descriptors. Other limitations of TIs are that they are more or less degenerate (i. e. more than one structure corresponds to the same value of the TI), that their interpretation is often unclear, and that with few exceptions one cannot retrieve the molecular structure from the numerical value of the TI; however, the last two drawbacks of TIs will be discussed again later.

Graph Theoretical Intermezzo

A *graph* is a set V of elements (v_i) called vertices and a set E of elements (e_i) called edges, that couple two distinct vertices [32,33,34,35,36]. The number of edges meeting at a vertex is called the *degree* of that vertex. The sum of all vertex degrees of a graph is twice the number of edges. A *walk* in a graph is an alternating sequence of vertices and edges, beginning and ending with vertices, in which each edge is incident with the two vertices immediately preceding and following it; an example of a walk is: $v_1 e_1 v_2 e_2 v_3 e_3 v_4 e_4 v_3$, and it is evident that edges of a walk need not be distinct but can be repeated. A *path* is a walk whose vertices (and thus necessarily all edges) are distinct.

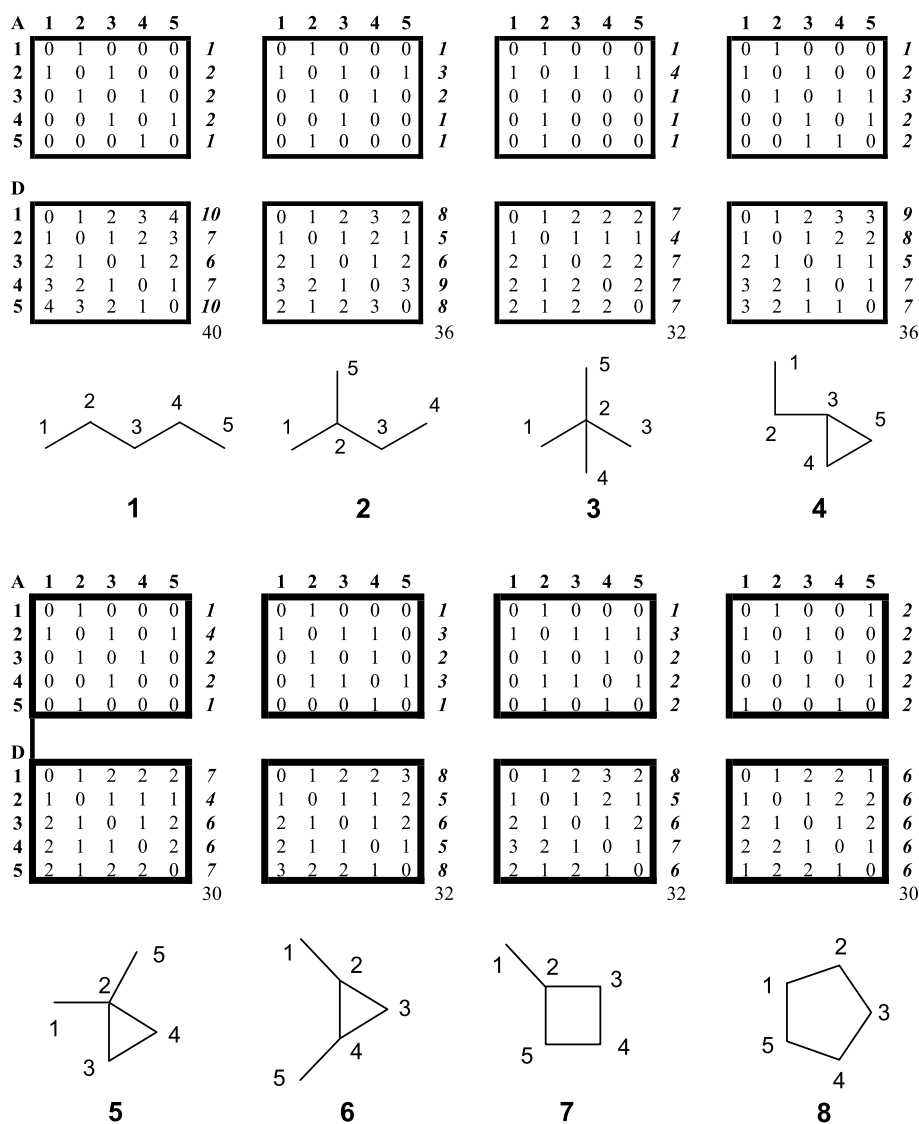
In a *connected graph* any two vertices are connected by one or more paths. All graphs discussed in this paper will be connected graphs. A *cycle* (or a closed walk) is a walk with ≥ 3 distinct edges where the initial and final vertices are identical. A *tree* is a connected graph devoid of circuits; if n is its number of vertices (also called the order of the graph) the number of edges is $e = n - 1$. The *cyclomatic number* μ of a graph is the smallest number of edges that have to be deleted in order to convert it into a tree. In a tree every two vertices are joined by a unique path. The numbers of paths in a graph or a tree are limited, but the number of walks is not. *Chemical graphs* have vertices of degree at most 4. In addition, they are graph-theoretically planar, i. e. they can be embedded on a plane without crossing edges. Molecular formulas for covalently-bonded organic compounds (which represent the vast majority of medicinal drugs) can be translated mathematically into molecular graphs, by assimilating atoms with vertices (points) and covalent bonds with edges (lines). Normally, a vertex is not connected to itself, but multiple bonds are a common occurrence in chemical formulas, and the corresponding type of graph is called a multigraph; edges do not have directions, i. e. we will not use directed graphs (digraphs). However, we need to distinguish various kinds of atoms and for this purpose vertex-weighted graphs will be employed. Unless otherwise stated, hydrogen atoms will not be included so that we will have only *hydrogen-depleted-graphs* or *-multigraphs*. Measured physico-chemical or biological properties of molecules are expressed in numerical form with corresponding units, and therefore they have metric properties, i. e. one can arrange these properties in increasing or decreasing order. On the other hand, chemical formulas or the corresponding molecular graphs are not metric, but for QSPR and QSAR studies we need to find ways to assign metric properties to chemical formulas. One of the simplest and computationally cheapest ways to achieve this performance is to associate numerical invariants to graphs and multigraphs. In graph-theoretical parlance, two vertices are called adjacent if they are connected by an edge, and two edges are called adjacent if they share one vertex. A vertex is incident to an edge if it is an endpoint of that edge. The number n of vertices of a graph is called the *order* of the graph. A simple graph has no loops or multiple edges. For the purpose of chemical information, one can associate graphs with various kinds of matrices, resulting in a one-to-one correspondence between vertices and "local vertex invariants" (LOVIs). Two of the most frequently used matrices are the adjacency matrix A , and the distance matrix D , which are symmetrical matrices with entries a_{ij} and d_{ij} , respectively. For adjacent vertices (i. e. points connected by an edge), $a_{ij} = 1$, and

it is 0 otherwise. It will be noted that **A** is identical to the Hückel matrix for conjugated systems. The row and column sums (written in Fig. 2 at the right of each matrix in **bold italics**) of **A** are called vertex degrees, v_i , a notion familiar to organic chemists who call vertices of degree 1 “primary carbon atoms”, vertices of degree 2 “secondary carbon atoms”, etc.

$$v_i = \sum_j a_{ij} . \quad (6)$$

Figure 2 presents as examples the **A** and **D** matrices for eight graphs with five carbon atoms: 1–3 are the three

possible isomers of alkanes C_5H_{12} and 4–8 are monocyclic hydrocarbons C_5H_{10} . It is evident from Fig. 2 that vertex degrees as LOVIs are not highly discriminating because the primary atoms of **2** have all the same value ($v_i = 1$) although vertex 4 is not equivalent to vertices 1 and 5; similarly, the three secondary vertices of **1** and **4** have all the same value ($v_i = 2$) although vertices 2 and 4 in **1**, or 4 and 5 in **2** are not equivalent to the third vertex. It will be seen from Fig. 2 that both **A** and **D** have the same entries 1, and on their main diagonal the same 0 values; however, all other zero values of **A** are replaced in **D** by integer values higher than 1.



Drug Design, Molecular Descriptors in, Figure 2

Adjacency and distance matrices for eight 5-vertex graphs 1–8

Entries d_{ij} of the distance matrix are the shortest topological distances (i.e. the smallest number of connected edges, or shortest path) between vertices i and j . The row or column sums of \mathbf{D} are called distance degrees, distance sums, or briefly distasums, d_i :

$$d_i = \sum_j d_{ij} . \quad (7)$$

Examination of Fig. 2, where at the right-hand side of each \mathbf{D} matrix one sees the distasums, shows that now the primary carbon atoms in **2** have different d_i values (8 for vertices 1 and 5, and 9 for vertex 4); also the secondary carbon atoms in **1** (7 for vertices 2 and 4, and 6 for vertex 3) as well as in **4** (7 for vertices 4 and 5, and 8 for vertex 2). The same is true for graphs **5–8** where non-equivalent vertices have different d_i values.

Three other symmetrical matrices that were used for devising LOVIs are: (i) the resistance distance matrix whose entries represent the electrical resistance between two vertices if each edge has a resistance of one ohm; for acyclic graphs, there is no difference between the topological distance and the resistance distance; (ii) the reciprocal distance matrix, having as entries reciprocal values of topological distances; Harary indices were devised on the basis of this matrix; (iii) the detour matrix whose entries are the numbers of edges in the longest paths between two vertices. Whereas the previous matrices characterize graphs uniquely, nonisomorphic graphs may have the same detour matrix. For more information of symmetrical and non-symmetrical matrices (distance-valency matrix, Szeged matrix, Cluj matrix, Wiener matrix, hyper-Wiener matrix) the reader should consult three recent reviews [37,38,39].

Topological Indices (TIs)

A topological index (TI) results by applying various mathematical operations to graph constituents, such as combining LOVIs of all graph vertices into a single number. The combination may be a simple operation such as summation, or a more elaborate one. If one would add all vertex degrees of a graph, the result is an integer which has the same value for all constitutional isomers (different substances with the same molecular formula, translated into hydrogen-depleted graphs with the same number n of vertices but with different connectivities):

$$\sum_i v_i = 2n - 2 + 2\mu = 2e \quad (8)$$

where the cyclomatic number μ is obtained from the number n of points and e of edges:

$$\mu = e - n + 1 . \quad (9)$$

Thus, all alkanes have $\mu = 0$, and all monocycloalkanes have $\mu = 1$.

When multiple bonds are present and each edge is counted separately, in the corresponding hydrogen-depleted *multigraphs* double bonds are equivalent to rings, and a triple bond counts as two double bonds. For most of the simple organic compounds with the following numbers of C, H, monovalent Hal, and trivalent N atoms, the cyclomatic number can also be found as follows: $\mu = C + 1 - (H + \text{Hal} - N)/2$. Any divalent sulfur and oxygen atoms do not affect μ .

Integer TIs based on integer LOVIs are considered to belong to the 1st generation, real-number TIs based on integer LOVIs to the 2nd generation, and real-number TIs based on real-number LOVIs to the 3rd generation. We will proceed in an approximate chronological order according to these generations but this review will not be exhaustive.

The Wiener Index W A young medical student, Harold Wiener, published in 1947 in the Journal of the American Chemical Society two papers on QSPR for normal boiling points and for heats of isomerization or vaporization of alkanes, using a simple molecular descriptor which is now known as the Wiener index (W). Although the initial definition was valid only for alkanes [40], this index was soon afterwards shown by Hosoya [41] to be equal to the half-sum of all entries in the distance matrix \mathbf{D} of the hydrogen-depleted graph (values of $2W$ can be seen for all constitutional isomers of heptane and octane in Table 1, and for **1–8** in Fig. 2):

$$W = \frac{1}{2} \sum_i d_i = \frac{1}{2} \sum_i \sum_j d_{ij} . \quad (10)$$

It can be easily deduced that in the alkane series, W increases with the number n of carbon atoms, but one can observe in Fig. 2 that with increasing branching among isomers, W decreases. It will be seen later that this may be a drawback for W because increasing complexity requires both these features (size and branching) to increase.

Also, it can be observed that W does not have a high discriminating ability among isomers (in other words, W has a high degeneracy). Thus, in Table 1 the degeneracy of W is indicated by boldface numbers; in Fig. 2, two pairs of non-isomorphic graphs, namely **5** and **8** with $2W = 30$, as well as **3**, **6** and **7**, with $2W = 32$ share the same value

for W . In conclusion, W is the first TI to have been invented, is easily computed, does not provide for encoding the presence of heteroatoms or multiple bonding, yet it correlates fairly well with many properties. For the initial correlations, Harold Wiener combined index W with a second descriptor, the number of distances of length 3.

The Platt Index and the Gordon–Scantlebury Index

Almost simultaneously with Wiener, J.R. Platt published in J. Chem. Phys. (1947) and J. Phys. Chem. (1952) two QSPR studies for additive properties of saturated hydrocarbons (paraffins) using as a molecular descriptor the sum (F) of all edge degrees (number of edges adjacent to each edge) [42,43]. It was shown later by M. Gordon and G.R. Scantlebury [44] that $F/2$ is the number of ways a path of length 2 (i. e. propane subgraph) can be superimposed on the given graph.

The Hosoya Index In 1971 Haruo Hosoya published the first in a series of papers, proposing another molecular descriptor (Z) that he called *topological index* [41]. This name is now used as a class denomination. For a hydrogen-deleted graph G ,

$$Z = \sum p(G, k) \quad (11)$$

where $p(G, k)$ is the number of ways in which k edges of the graph may be chosen so that no two of them are adjacent. The summation starts with $p(G, 0) = 1$ and $p(G, 1) = e$ (the number of edges in a simple graph). This index Z shows remarkable correlation ability for many alkane properties.

The Zagreb Indices Looking back at Eq. (8) where summing all vertex degrees affords a trivial result (twice the number of edges), it is easy to understand that there are ways [45] to improve on this by considering the following two new indices, named “Zagreb indices”:

$$M_1 = \sum_i v_i^2 \quad (12)$$

$$M_2 = \sum_{(\text{edge } ij)} v_i \times v_j. \quad (13)$$

The Schultz Indices were introduced in 1989–1990 [46]. The first of them is the molecular topological index MTI:

$$\text{MTI} = \sum_i E_i \quad (14)$$

where the “intricacy elements” E_i are the elements in the i th row of the E matrix built from the product of the sum

of adjacency and distance matrices with \mathbf{v} the $1 \times n$ vertex degree matrix:

$$\mathbf{E} = \mathbf{v}(\mathbf{A} + \mathbf{D}). \quad (15)$$

The determinant of the adjacency-plus-distance matrix was proposed as an index called TI for alkanes:

$$\text{TI} = \det |\mathbf{A} + \mathbf{D}|. \quad (16)$$

Interestingly, from such indices Schultz went on to define three-dimensional molecular descriptors encoding chirality.

Centric Indices for acyclic graphs were proposed by Balaban in 1979 by summing squares of numbers of vertices of degree 1 pruned stepwise from the periphery towards the graph center, which is either one vertex or a pair of adjacent vertices [47]. Later, generalizations for graph centers of cyclic graphs or of substituents were proposed [48,49,50,51].

Bonchev’s Overall Connectivity The sum of all adjacencies (vertex degrees) in a graph is called the total adjacency A :

$$A = \sum_i v_i = \sum_i \sum_j a_{ij}. \quad (17)$$

It was shown earlier (Eq. 8) that A is twice the number e of edges. If, according to Danail Bonchev, one sums A values for all connected subgraphs of a given graph (up to a selected threshold for the number of vertices in the subgraph) one obtains the total (or overall) connectivity, or topological complexity, TC [52]. The subgraphs in acyclic graphs (trees) are paths of various lengths or clusters of various sizes. Because in practice the total number of subgraphs increases very rapidly with the graph size, a pragmatic solution is to take into account only subgraphs up to a certain number e of edges as the *partial topological complexity* eTC . An advantage of TC and eTC is the fact that such indices increase both with size and with branching, whereas W and the Randić index χ (vide infra) increase with size and decrease with branching.

We start now with 2nd generation TIs.

The Randić Connectivity Index The following desirable qualities of TIs have been enumerated by Randić [53]:

1. Direct structural interpretation
2. Good correlation with at least one property
3. Good discrimination of isomers
4. Locally defined
5. Generalizable to “higher” analogues
6. Linearly independent

7. Simplicity
8. Not based on physico-chemical properties
9. Not trivially related to other TIs
10. Efficiency of construction
11. Based on familiar structural concepts
12. Showing a correct size dependence
13. Gradual change in value with gradual change in structure.

To these qualities, one may add:

14. Low requirement in computer (CPU) time.

The Randić connectivity index χ or ${}^1\chi$ [54] was proposed in 1975 as the sum over all edges $i-j$ of the reciprocal square root of products of vertex degrees:

$$\chi = \sum_{(i-j)} (v_i \times v_j)^{-1/2}. \quad (18)$$

This TI is the most popular and most frequently cited in the literature. It provides good correlations with many physico-chemical properties and biological activities, and fulfills most of the above requirements. It increases with the number of vertices and decreases with increasing branching, like Hosoya's index Z or the NBPs of alkanes.

The Kier–Hall Higher-Order Indices and Related Indices An edge is equivalent to a path of length one and this relation between edges and paths prompted Kier and Hall in 1976 to introduce higher-order indices for various paths of longer lengths (${}^2\chi$, ${}^3\chi$, etc.) or other subgraphs such as clusters with 3 or 4 vertices, or cycles. They then suggested taking into account the presence of heteroatoms in the molecular scaffold proposing “valence connectivity indices” ${}^v\chi$, where weights (defined in accordance to positions in the Periodic System) replace vertex degrees [55]. By their two initial and influential books they “converted” many pharmaceutical chemists to look at TIs as a convenient tool for drug design [11,12].

Kier introduced also a “flexibility index” [56]. By taking into account additional molecular features such as electronegativity, Kier and Hall described “electrotopological state indices” [13], and the corresponding software MOLCONN is distributed by Hall Associates. All these indices were reviewed recently as book chapters [57].

Information Theoretic Indices were first described in 1977 by Bonchev and Trinajstić [58]. On applying Shannon's formula for summands belonging to different equivalence classes before the summation of LOVIs, one can reduce the degeneracy of many TIs. Thus, by sorting topological distances d_{ij} into groups of g_i summands having

the same d value and only then applying the global summation, one can convert the 1st-generation Wiener index $W = 1/2 \sum_i i g_i$ into the 2nd-generation corresponding total information-based index on the distribution of distances, with much lower degeneracy than W :

$$I_D^W = W \log_2 W - \sum_i i g_i \log_2 i. \quad (19)$$

One can also obtain the *mean* information on the distance degree distribution. A similar approach affords the total and the mean information on other descriptors such as the Hosoya index or the vertex degree distribution [14]. Moreover, the 3D-analog of the Wiener number (WG) introduced by Bonchev and Mekenyan in 1985, and its information-based counterpart, like other information-based TIs, is now part of software packages CODESSA, DRAGON, and OASIS (the last one was developed in Bourgas).

In 1979, Basak and coworkers devised a different type of information-theoretic indices from constitutional formulas that include hydrogen atoms [59,60]. Atoms are partitioned into disjoint subsets, and probabilities p_i are assigned to sets of equivalence classes based on chromatic coloring or to orbits of the automorphism group of the graph. Then, information content (IC), structural information content (SIC) and complementary information content (CIC) are defined as:

$$IC = - \sum_i p_i \log_2 p_i \quad (20)$$

$$SIC = IC / \log_2 n \quad (21)$$

$$CIC = \log_2 n - IC. \quad (22)$$

Analogous but different information-theoretic formulas for four indices (U, V, X, Y) were proposed by Balaban and Balaban [61].

Mean Square Distances in Graphs A very simple descriptor is derived from the observation that compact graphs have lower average distances than elongated graphs [62]. By investigating various ways of averaging distances, Balaban proposed the following formula for the mean square distance index $D^{(2)}$:

$$D^{(2)} = \left(\sum_i i^2 g_i / \sum_i g_i \right)^{1/2}. \quad (23)$$

For acyclic graphs, it was shown that even better correlations with octane numbers or heptanes and octanes could be obtained (instead of summing up all distances) by considering only distances between primary carbon atoms (endpoints).

Average Distance-Based Connectivity J (Balaban Index)

With the aim of obtaining a TI with low degeneracy that does not increase appreciably with size but depends mainly on branching, Balaban replaced vertex degrees in Eq. (18) by distasums (s_i) and introduced a term that averaged the result with respect to the number of edges and rings, keeping the summation over all edges $i-j$ and the inverse square root operation [63]:

$$J = [e/(\mu + 1)] \sum_{(i-j)} (d_i \times d_j)^{1/2}. \quad (24)$$

Interestingly, it was proved that for an infinite vinylic polymer $\text{H}_2\text{C}=\text{CH}-\text{R}$ the J index is a rational multiple of the number $\pi = 3.14159\dots$, and for polyethylene with $\text{R}=\text{H}$, it is exactly π [64]. For accounting for the presence of heteroatoms or multiple bonds, parameters based on electronegativities or covalent radii relatively to carbon seem to work well [65,66]. Unlike previously described topological indices, the J index has a low degeneracy, so that for alkanes, the first isomers with the same J values have 12 carbon atoms [67]. Because of its special nature, index J should not be used in monoparametric correlations for properties that depend on the molecular size but it can be used in such correlations for isomeric sets such as those of Fig. 1.

Eigenvalues of Matrices as TIs It was shown by Lovasz and Pelikan that the largest eigenvalue of the adjacency matrix increases with graph branching [68]. Medeleuanu and Balaban investigated several types of eigenvalues for various matrices [69], and similar explorations were reported by Randić and coworkers [70].

The *Harary index* was proposed independently by chemists in Zagreb [71] and Bucharest [72] and is based on the matrix containing reciprocals of topological distances. Such LOVIs are then combined into an index by a Randić-type formula (see operation 4 on the next page). It places larger weights on atom pairs close to each other, unlike TIs based on the distance matrix.

Gutman's Szeged Index Gutman introduced the edge-Szeged matrix where entries ${}^e\text{SZ}_{ij}$ are products of numbers of vertices on lying closer to i or j after edge ij is removed from the graph (vertices equidistant to i and j are not counted). For trees, the Szeged and Wiener matrices and indices are identical [73].

Diudea Indices Diudea and Balaban devised two different approaches for the same purpose by using "regressive distances" or "regressive vertex degrees" by means of nonsymmetrical matrices whose entries in row i repre-

sent sums of topological distances or vertex degrees, respectively, for vertices in shells around vertex i ; an exponentially decreasing factor multiplies each successive shell [74,75]. Further work by Diudea produced new matrices (Cluj matrix) and modified Szeged matrices from which other descriptors were obtained [76,77,78].

Walk Counts (Rücker Indices) By raising adjacency matrices to higher powers one can obtain easily total walk counts of graphs. They can be used in correlations with NBP's of alkanes and cycloalkanes [79,80,81].

Indices Based on Path Counts The numbers of p_L paths with various lengths L have been seen to determine Kier-Hall indices ${}^L\chi$. On examining NBPs of alkanes, Randić observed that there are regular variations when looking at NBPs of octane isomers arranged according to p_2 and p_3 and this led to the idea of orthogonalized descriptors to be mentioned further below. The enumeration of all paths with lengths from 1 to the maximum value in a graph (called the *path code*) provides distinct sequences for all alkanes up to octane isomers. Among possible indices obtained from path codes, index Q modeled after the Zagreb index as sum of squares of all path numbers (but normalized for the number of rings) has a fairly high degeneracy, but index P appears to provide the best correlations with physical properties [82]. In Table 1 one can see for all constitutional isomers of heptanes and octanes the path code and both indices Q and P along with other four TIs.

$$P = \sum_i \left\{ p_i^{1/2} / [i^{1/2}(\mu + 1)] \right\}. \quad (25)$$

Triplet LOVIs and Derived Topological Indices

Triplet indices were introduced for converting 1st or 2nd generation TIs into 3rd generation TIs, by combining symmetrical or non-symmetrical matrices with two column vectors (one for the main diagonal and the other for the free term) and then solving the resulting system of linear equations [83]. As a result, one can have a wide variety of TIs that can be adapted to the structures under investigation. For instance, the **A** or **D** matrices of neopentane (2,2-dimethylpropane, structure **3** in Fig. 2) may be combined with column vectors $[a]$, $[b]$ that are arbitrary constants, or that convey chemical (e.g. atomic numbers, electronegativities, bond multiplicity, etc.) or topological information (e.g. vertex degrees, distance sums, etc.).

Taking into account that the hydrogen-depleted graph of neopentane has only two kinds of vertices (primary and quaternary carbon atoms, with LOVIs $x_1 = x_3 = x_4 = x_5$, and x_2 , respectively), for the simplest case when the two column vectors are constants a and b ,

	x_1	x_2	x_3	x_4	x_5	
x_1	0	1	0	0	0	1
x_2	1	0	1	1	1	4
x_3	0	1	0	0	0	1
x_4	0	1	0	0	0	1
x_5	0	1	0	0	0	1

	x_1	x_2	x_3	x_4	x_5	
x_1	a	1	0	0	0	$=b$
x_2	1	a	1	1	1	$=b$
x_3	0	1	a	0	0	$=b$
x_4	0	1	0	a	0	$=b$
x_5	0	1	0	0	a	$=b$

	x_1	x_2	x_3	x_4	x_5	
x_1	0	1	2	2	2	7
x_2	1	0	1	1	1	4
x_3	2	1	0	2	2	7
x_4	2	1	2	0	2	7
x_5	2	1	2	2	0	7

	x_1	x_2	x_3	x_4	x_5	
x_1	a	1	2	2	2	$=b$
x_2	1	a	1	1	1	$=b$
x_3	2	1	a	2	2	$=b$
x_4	2	1	2	a	2	$=b$
x_5	2	1	2	2	a	$=b$

Drug Design, Molecular Descriptors in, Figure 3

Upper left and right sides: adjacency and distance matrix for 2,2-dimethylpropane (structure 3 in Fig. 2) with row sums (vertex degrees and distance sums, respectively). Lower left and right sides: addition of column vectors a on the main diagonal and b as the free term, converts matrices A and D into systems of linear equations whose solutions provide real-number LOVIs, x_1 through x_5 , for the five vertices

one obtains for the triplet system $[A, a, b]$ the following LOVIs, as seen from Fig. 3:

$x_1 = x_3 = x_4 = x_5 = b(a-1)/(a^2-4)$ and $x_2 = b(a-4)/(a^2-4)$. Thus, the ratio x_2/x_1 (which for vertex degrees as LOVIs was 4/1) becomes $x_2/x_1 = (a-4)/(a-1)$. For instance, with $a = 5$ and $b = 7$, one obtains $x_1 = 4/3$ and $x_2 = 1/3$ (in this case, a and b are arbitrary constants without any chemical or graph-theoretical significance).

Likewise for the triplet system $[D, a, b]$ one obtains the following LOVIs:

$x_1 = x_3 = x_4 = x_5 = b(a-1)/(a^2+6a-4)$ and $x_2 = b(a+2)/(a^2+6a-4)$. Thus, the ratio x_2/x_1 (which for distance sums as LOVIs was 7/4) becomes $x_2/x_1 = (a+2)/(a-1)$. For instance, with $a = 5$ and $b = 7$, one obtains $x_1 = 24/51$ and $x_2 = 49/51$.

After obtaining real-number LOVIs one combines them into TIs by various operations such as:

1. Sum of weights $\sum_i x_i$;
2. Sum of squared weights $\sum_i x_i^2$;
3. Sum of square root of weights $\sum_i x_i^{1/2}$;
4. Sum of cross-products for edge endpoints $\sum_{ij} x_i \times x_j$;

and one denotes the TI thus obtained by the initial of the matrix and of the two vectors followed by the number of the operation, e.g. AZV1 (adjacency matrix A, atomic number Z on the main diagonal, vertex degree V as the free term, and summation of the LOVIs, i.e. operation 1). Interestingly, Basak et al. used the TRIPLET and POLLY programs for a variety of QSAR studies [84].

Estrada's Generalized Topological Index It was found by Ernesto Estrada that many 1st and 2nd-generation TIs (such as Zagreb indices, Randić index, Balaban's J index, Wiener index, Harary index, Schultz MTI index, Gutman index, and hyper-Wiener index) can be derived from a unique generalized formula for generalized topological indices (GTIs) [85,86,87,88,89,90]. Again, a matrix and two vectors are involved, but the result is an all-encompassing powerful tool, because it also optimizes the variable parameters for the database. These GTIs represent points in a 6-or 8-dimensional space of topological parameters, which can be optimized for describing a specific property. From the infinity of possible GTIs defined by the vector-matrix-vector multiplication, by a systematic change of 6 parameters from 0.0 to 0.9 with a step of 0.1, one reduces the set to only 10^6 ! As an example of the power of this generalization, Estrada and Gutierrez optimized the J index according to 16 octane numbers from the set of isomeric octanes resulting in a correlation factor $R = 0.992$ and standard deviation $s = 3.5$ with a cubic equation [91].

Estrada's Topological-Substructure Molecular Design QSAR is based on the computation of the spectral moments of the bond matrix and is able to identify compounds with a certain biological activity among a database with compounds with many other types of biological activities. Thus, sedatives/hypnotics were found in the Merck database by using a TOSS-MODE discrimination

model [92,93]. By computing fragment contribution models, it was possible to predict that among a virtual library of 327 antioxidants (flavonoid and cinnamic acid derivatives), 70 (from which only about 20 have been described in the literature) should be more active than the most powerful antioxidants in the Brazilian propolis [94].

Orthogonal Descriptors It was mentioned earlier that even when molecular descriptors are intercorrelated, yet they cover different aspects of the multidimensional space determining the variation of properties/activities, a procedure of taking the difference between such descriptors (i. e. a projection of one vector on the other) leads to surprisingly high correlations, as shown by Randić [95,96].

Variable Connectivity Indices are derived from Randić's connectivity indices by optimizing correlations for a given set of data under the assumption that entries in the connectivity matrix may vary. Thus for a compound with a heteroatom, instead of having an entry 1 when adjacencies with this heteroatom are involved, one allows the corresponding entry to be optimized with respect to the final factor of correlation, leading to a weighted entry. Even atoms of the same kind but in different environments (e. g. primary, secondary, tertiary or quaternary carbon atoms) can be allowed to have different weights [97,98]. Pompe observed that in certain cases ("anticonnectivity") the weight can be negative [99]. The field was reviewed recently by Randic and Pompe [100].

Optimal Descriptors Based on SMILES Notation Weininger introduced a simple yet powerful linear notation for encoding chemical structures, called Simplified Molecular Input Line Entry System (SMILES) that can be manipulated by computer programs for interconversion between structural formula and adjacency matrix. Inspired by Randić's optimal (flexible) descriptors for a set of compounds with a certain property/activity expressed numerically, Toropov, Gutman and coworkers assigned optimized numerical values to each individual symbol the SMILES notation for these compounds, including lower and upper case letters, brackets, or numbers, and producing thus satisfactory correlations with the given property/activity [101]. Other studies involved toxicity [102,103] or various biological activities [104].

Gálvez's Charge Indices for Discriminant Analysis for drug design combines the use of molecular connectivity, charge indices, and in-house molecular descriptors accounting for molecular shape [105]. Starting from two data sets, each comprising active and inactive compounds,

and using linear discriminant analysis, one can obtain separation between activity classes. The predicted active compounds are tested for the predicted activity or (in case a virtual library was analyzed) synthesized and then tested. It was thus possible to discover novel activities for known compounds and to synthesize new compounds with certain biological activities [106].

A Real-World Example of Using Molecular Descriptors Including TIs

As an example of how these simple molecular descriptors (which are easy and fast to compute for any imaginable structure) can be applied for biomedical purposes, we present a study by Lahana and coworkers [107] who described a rational design of immunosuppressive decapeptides using 13 molecular descriptors for each amino acid including four topological and shape indices (Balaban's *J* and three Kier–Hall indices), along with log *P*, ellipsoidal and molar volumes, molar refractivity, dipole moment, plus four simple counts of O and N atoms, ethyl and hydroxyl groups. Starting from 24 known bioactivities of active and inactive decapeptides, each descriptor had intervals of favorable and unfavorable ranges of values. The virtual library of decapeptides with structure RXXXXXXXXY (where R = arginine, Y = tyrosine, and X allowing a combinatorial choice involving six natural and one non-natural amino acids), amounted to about a quarter million. This library was reduced 10,000-fold to 26 structures by rapidly computing all their descriptors and selecting those that fitted into the favorable domains. Then in a much slower and elaborate computational effort involving molecular dynamics simulations, a set of five optimal structures was obtained. All five were synthesized and tested for mouse heart allograft model. One decapeptide displayed an immunosuppressive activity approximately 100 times higher than the lead compound [107,108,109].

One should mention that usually the reduction in candidate structures is not as large as in this case. Moreover, often multiparametric correlations involving simple molecular descriptors are followed by more elaborate 3D modeling which is moderately more demanding for CPU time. Only then can one achieve the final stage of molecular docking and rational drug design that is described by other monographs of this Encyclopedia.

Three-Dimensional Molecular Descriptors

The TIs discussed so far provide information on constitution (connectivity), but not on stereochemistry. For many biologically active compounds, this is a serious limitation

because enantiomers with the same constitution usually have different biological activities, as known from the different odors of enantiomers or from the teratogenic consequences of using the racemic mixture of the pharmaceutical Thalidomide, one of the enantiomers of which had tragic results and produced limbless babies. Several ideas have been published on how to proceed from two-dimensional topology to three-dimensional geometry (only reviews are cited) [110,111,112,113]. Unlike the calculations for TIs, such 3D calculations usually require longer computer-time (CPU-intensive).

Cramer defined for his *comparative molecular field analysis* (CoMFA) a 3-D grid so as either to (i) include the most active or the most rigid pharmacophore (ligand), or (ii) model the receptor cavity defined by its molecular docking with the potential bonding sites [114]. At each intersection point of this imaginary grid one computes the molecular geometry using force-field (molecular mechanics) methods or Gaussian approximations. Cross-validation is achieved by the leave-one-out procedure from the studied series, and the error is measured by a parameter named PRESS (predicted residual sum squares). Tripos Co. owns the patent and computer programs. The most difficult problems in CoMFA studies are connected with conformational aspects of flexible molecules, and with molecular alignment. Various remedies have been proposed, and the literature grows continuously both for CoMFA and for CoMSIA (*comparative molecular similarity analysis*) [115,116,117,118].

Todeschini proposed the Weighted Holistic Invariant Molecular (WHIM) approach and applied it to molecular surfaces (MS-WHIM) to derive new 3D theoretical descriptors. A 3D QSAR study performed on a series of steroids, comparing the WHIM description to CoMFA fields, showed that MS-WHIM descriptors provide meaningful quantitative structure–activity correlations. The concise number of indices, the ease of their calculation, and their invariance to the coordinate system make MS-WHIM an attractive tool for 3D-QSAR studies [119,120,121,122,123,124,125]. Other descriptors have been developed, which encompass size, shape, symmetry, and atom distribution, such as the EVA descriptor (normal coordinate EigenValues) [126] derived from the vibrational frequencies; all of these are independent of structural alignment [127].

Molecular Similarity, Dissimilarity, and Clustering of Molecular Descriptors

On the basis of the postulate that similar structures will have similar biological activities, a proven strategy in find-

ing lead compounds for drug design is to look at structures with descriptors that are close neighbors in large databases [128]. With fuzzy bipolar pharmacophore autocorrelograms, the neighborhood behavior (as distinct from the clustering behavior) provide alternative and complementary criteria for evaluating performances of molecular similarity metrics [129]. The approach initiated by Mezey (“molecular holograms” at various resolution levels) has allowed shape simulations to be made for various biologically important molecules such as amino acids or even macromolecular compounds [130]. The rapidly increasing performance of computers now makes it possible to perform elaborate calculations for quantum-chemical parameters [131,132]. With ab-initio calculations of molecular volumes and areas of projection, it was possible to classify molecular shapes (e.g. spheroids, disk-like, rodlike, starlike, crosslike, egglike) by means of descriptors for ovality, roughness, size-corrected parameters and also (based on higher central momenta) skewness and kurtosis [133]. The functional diversity of compound libraries is discussed in detail by Gorse and Lohana [108].

Molecular similarity is important when examining potential candidates for enzyme inhibitors (a classical example is that of sulfonamides that mimic *para*-aminobenzoic acid and thus act as bactericides). On the other hand, molecular dissimilarity (diversity) is important when looking for lead compounds in large databases such as the Cambridge Structural Database (for X-ray crystallographic data) or the National Cancer Institute Databases. Burden [134] suggested using molecular identification numbers obtained from the two lowest EVAs of the connectivity matrix obtained from the hydrogen-depleted graph. A further step was undertaken by Pearlman who proposed the BCUT (Burden–Chemical Abstracts–University of Texas) approach, which consists of taking the lowest and highest EVAs, but using modified matrices instead of atomic numbers. Relevant data for intermolecular interactions are inserted on the main diagonal, namely atomic charges, polarizabilities, and H-bond donor and acceptor abilities. By means of the program CONCORD for generating 3-D structures from 2-D structures, it is possible to introduce scaled information on the 3-D structure as the off-diagonal entries. With the “auto-choose” algorithm, Pearlman obtained a program for comparing the diversities of large databases at a relatively low CPU cost.

Clustering of TIs has been investigated for large and diverse databases [135,136] and the interesting result was that one can group together most TIs into a few classes, so that one should be able to select molecular descriptors that

correlate with different aspects of the molecular structure. Thus it was found that the largest cluster contains TIs that depend mainly of the size of the graph such as W ; a separate cluster contains path connectivities of intermediate order; such indices of higher order appear as a separate cluster; information-based indices are clustered separately; shape descriptors also form a separate cluster; triplet indices of various kinds are distributed among many of these clusters, proving the diversity of such TIs when they are properly selected.

Hierarchical Ordering of Structures and of Molecular Descriptors

On the basis of mathematical arguments, Bertz advocated the idea that line graphs of successive orders provide the most logical ordering of chemical structures, and published a list of ordered alkanes [137]. It was a surprise for him to observe that among all TIs, index J provided the same ordering of isomeric alkanes up to heptanes, and that even for higher alkanes the differences were minimal. In Table 1 heptanes and octane isomers were sorted according to increasing J values [138,139,140]. It is interesting to observe that apart from degeneracy, W values are ordered similarly. Among other criteria for accepting new TIs, a logical ordering of isomers is useful, and the selection of index P (Table 1) took also this criterion into consideration [82].

Basak and coworkers published several papers proving that molecular descriptors obey a hierarchy from the simplest (*topostructural indicators* of the graph constitution) through *topochemical indicators* that add chemical information on heteroatoms and bond multiplicity, then *geometrical indicators*, which contain additional information on the stereochemistry of molecules, to *quantum-chemical descriptors* that are CPU-intensive. In most QSAR studies, it was found that correlation factors increase appreciably from topostructural to topochemical descriptors, but then the improvement was minimal when investing more computational efforts for including geometrical or quantum-chemical information [141]. A thorough review of quantum-chemical descriptors contains details about their applications for QSPR and QSAR [142].

General Problems Connected with 2D-QSAR for Drug Design

With several older and newer journals covering the area of computer-aided drug design (CADD) and with the success of in-silico molecular drug discovery, it is small wonder that all “big pharma” companies sustain groups of chemists, biochemists and computer specialists. A few spe-

cific aspects will be discussed briefly here, but the area is too vast for in-depth coverage, and even the bibliography can be only sketchy [143].

Several computer programs are available for QSAR using hundreds of molecular descriptors and selecting a prescribed number of those that lead to the highest correlation factors: Katritzky and Karelson's CODESSA and CODESSAPRO [144], Todeschini's DRAGON, Hall and Kier's MOLCONN, Basak's POLLY and TRIPLET, Bonchev and Mekenyan's OASIS, and Diudea's TOPOCLUJ, among other ones [145]. Some of the newer statistical validation procedures are included such as leave-one-out, or division into training and test sets. For similarity evaluation, one may select k nearest neighbors (KNN), artificial neural networks [146,147], support vector machines [148], genetic (evolutionary) algorithms [149,150] and combinations thereof. Statistical methods include principal component analysis (PCA), partial least squares (PLS), nonlinear iterative partial least-squares (NIPALS), decision trees, etc.

In order for a molecule to be drug-like (i. e. a drug candidate) it must fulfill several criteria, as discussed by Lipinsky's rule-of-five, which states that candidate compounds are likely to have unfavorable absorption, permeation and bioavailability characteristics if they contain more than 5 H-bond donors, more than 10 H-bond acceptors, a log P greater than 5, and/or a molecular mass of more than 500 Da. Only then may a compound comply with ADMET requirements (absorption, distribution, metabolism, excretion, and toxicity) [151]. Modeling structures for predicting ADMET properties may require non-linear techniques such as support vector machines (SVMs) and Bayesian neural networks (BNNs) [152]. Special cases are compounds that need to penetrate the blood-brain-barrier (BBB) [153,154].

Often in developing combinatorial virtual databases one needs to have descriptors for subgraphs (fingerprints, molecular fragments, or simply substituents). Topological descriptors must depend on the position of the “root” vertex and one should also consider the interaction between fragments and the main part of the molecule [155,156].

Interpretation of Topological Indices

In the introductory section about molecular descriptors it was mentioned that TIs do not appear to have a clear physical interpretation. Moreover, Charton argued that TIs cannot express any “fundamental property” and therefore whenever they show good correlation ability with bulk physical properties (e. g. phase change properties such as normal boiling points, melting points, critical tempera-

tures, heats of melting or vaporization, solubilities of noble gases in liquid compounds, or ratios of Van der Waals constants) they must contain “hidden” correlations to the real causes of such phenomena, namely dipole moment, polarizability, hydrogen bonding, and steric effects. The same should be true, according to him, for QSAR of TIs with biological effects [157].

However, the great advantages of TIs are their ease of calculation with low CPU requirements for any chemical constitution, making the search of virtual combinatorial libraries a useful first step in any drug design effort. Subsequent steps for the selected structures may then involve more elaborate computational methods (such as 3D descriptors, CoMFA and related procedures, BCUT, WHIM, etc.), docking into receptor cavities of proteins with known steric features, quantum-chemical descriptors with flexible geometries and variable ranges of acceptable energies [158]. One should note that on going from 1D or 2D topology to 3D geometry [111], several molecular descriptors (“topographic indices” including “chirality topological indices”), modeled after topological indices, have been described. One of the first is the 3D Schultz index [159], followed by Pyka’s index [160,161,162]. To follow the example of Lahana et al.’s strategy described above, the major reduction of candidate structures will occur at the first step, so that the CPI-intensive second step becomes feasible. Then the final step will be the chemical synthesis and biological/clinical testing of a restricted number of potential candidates.

Nevertheless, one should mention that there exist several proposals about possible interpretations of TIs. Soon after Wiener published his ideas about what we call now Wiener’s TI W [40], Platt provided the hypothesis that $W^{-1/3}$ approximates the mean molecular diameter [43]. Randić et al. [163,164] argued that just as in quantum chemistry the HOMO-LUMO concepts have no interpretation in VB model, or conversely Kekulé valence structures have no place in MO models, TIs need not be interpretable in physical chemistry concepts. Surprisingly, the Pauling bond orders for 40 smaller benzenoids show a better correlation with the connectivity index $^1\chi$ than with the Coulson bond orders. This should not mean that $^1\chi$ is a kind of bond order. However, recalling that edges in hydrogen-depleted constitutional graphs are paths of length one, this indicates that *paths and walks that lie at the basis of many TIs do define fundamental properties of chemical constitution*. Then row sums of entries in distance matrices (distance sums) must also have a meaning, assigning higher weights to longer paths. Vice-versa, the same sums for reversed distance matrices and recip-

rocal distance matrices, from which the reversed Wiener index and the Harary index, respectively, are obtained, should also justify the assignment of lower weights to longer paths, where intramolecular interactions, such as inductive electronic effects, are weaker. Weighting edges in molecular graphs shows how chemical bonds have different contributions to the TIs [164]. Additional support for the fact that TIs may have an intrinsic meaning is provided by Estrada’s finding of a unique formula unifying all the main TIs [85,86,87,88,89,90]. Other discussions on the interpretation and intercorrelation of TIs may be found in [165,166,167,168,169,170].

Reverse Engineering of Molecular Descriptors

Topological indices have been proved to be useful in drug design for both lead discovery and lead optimization, as attested by numerous published reports. The result of a QSAR study is an equation in terms of several descriptors, obtained with a training and a testing set (the latter can be replaced by the leave-one-out procedure), or with a ridge regression validation. The problem that arises often is then to find related structures that can improve or optimize the result, i.e. to find other structures that have the same set of descriptors. This is equivalent to finding non-isomorphic graphs that have the same TI. For highly or moderately degenerate TIs such as the Wiener, Randić, Hosoya, and centric Balaban indices, this “reverse engineering” problem is mathematically relatively simple as shown by Zefirov and coworkers [171,172,173] or Luque–Ruiz et al. [174], but will result in large numbers of graphs with given TI values or value ranges. For less degenerate TIs such as J , this would be a “harder nut to crack”.

In 2005, at an American Chemical Society Symposium on reverse engineering of molecular descriptors, it was discussed whether one may find a way to share information on physico-chemical properties such as $\log P$ or solubility, or even on biological activities or toxicities, without disclosing chemical structures. From the almost 30 million existing substances in the Chemical Abstracts Service database, only few databases with such properties or biological activities are openly available for academic research. Although the “organic chemical universe (under constraints defining chemical stability and synthetic feasibility)” contains about 10^{63} substances, most drug molecules have molecular weights (MW) < 360 Da, and can be obtained from smaller lead compounds with $MW \approx 150$ Da by adding substituents. The huge number mentioned above was found assuming that there are about 30 C, N, O, or S atoms linked in a molecule with up to 4

rings and 10 branch points. Reymond and coworkers [175] generated a database of all the 13.9 million virtually possible such organic compounds with MW < 160 Da. Stereoisomers account for 75% of the structures, and heterocycles predominate [175].

On the other hand, the large pharmaceutical companies have proprietary databases including hundreds of thousands of existing compounds. The problem is compounded by the fact that several thousands of large-volume chemical are manufactured around the world and reach the environment, yet the toxicity (including teratogenic, endocrine disruption, fertility and other biological effects) is known only for a small part of these chemicals. Large amounts of plasticizers in polymers, pesticides or fertilizers are probably involved in alarming signals about decreasing spermatogenesis and these signals lead to the widespread negative image of *chemistry* and *synthetic*, as opposed to *organic* or *natural* (as if everything material around us was not chemical). There is therefore an urgent need for computing (at least approximately) the toxicities of these structures by QSTR methods.

Among the views provided at this Symposium, some of them expressed the hope that perhaps chemical identity might be kept apart from physico-chemical or biological properties [176,177,178,179,180], but other authors believed that so far there is no guarantee that one could not retrieve the structure from a collection of properties [181]. This is an open and challenging question.

Future Directions

It was seen in this monograph that a continuous effort existed during the last decades for devising new molecular descriptors and computational techniques for 2D-QSAR to precede the more elaborate 3D-QSAR. The success rate so far is satisfactory, but the field progresses so rapidly that based on data about protein conformations and receptors gathered from X-ray diffraction or from NMR in solution, there is high hope that many more virtual drug-like ligands (enzyme inhibitors or activators) could be designed by computer-assisted methods.

What would greatly facilitate the area of CADD using TIs would be the development of simple descriptors (supplementing the information on covalent bonding) for charges, hydrogen-bond donors and acceptors. A few starting efforts have been published but till now no real success in this area was demonstrated.

The challenge described in the preceding paragraph is only one of the possible applications of better future molecular descriptors.

Bibliography

Primary Literature

1. Devillers J, Balaban AT (eds) (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam
2. Templeton NS, Lasic DD (eds) (2000) Gene therapy. Therapeutic mechanisms and strategies. Marcel Dekker, New York
3. Ilies MA, Seitz WA, Hohnson BH, Ezell EL, Miller AL, Thompson EB, Balaban AT (2006) J Med Chem 49:3872
4. Kubinyi H (2002) Drug design course (free download from his home page www.kubinyi.de). See also "Why drugs fail" Herman Skolnik Award Lecture, ACS Meeting, San Francisco, September 2006 on that web page
5. Maggiora GM (2006) J Chem Inf Model 46:1535
6. Balaban AT, Pompe M (2007) J Phys Chem 111:2448; erratum, ibid. 4871
7. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley, New York
8. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim
9. Balaban AT (2000) Quantitative structure-activity relationships and computational methods in drug discovery. In: Meyers RA (ed) Encyclopedia of analytical chemistry. Wiley, Chichester, p 7288
10. Randić M (1998) Topological indices. In: Schleyer PVR et al (eds) Encyclopedia of computational chemistry. Wiley, Chichester, p 3018
11. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth
12. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
13. Kier LB, Hall LH (1999) Molecular structure description: the electrotopological state. Academic Press, San Diego
14. Bonchev D (1983) Information theoretic indices for characterization of chemical structure. Wiley, Chichester
15. Hammett LP (1940) Physical organic chemistry, 2nd edn, 1970. McGrawHill, New York
16. Hammett LP (1935) Chem Rev 17:125
17. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Nature 194:258
18. Hansch C, Fujita T (1964) J Am Chem Soc 86:1616
19. Fujita T, Iwasa J, Hansch C (1964) J Am Chem Soc 86:5175
20. Hansch C (1969) Acc Chem Res 2:232
21. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR, hydrophobic, electronic and steric constants. Am. Chem. Soc., Washington
22. Hansch C, Hoekman D, Gao H (1996) Chem Rev 96:1045
23. Hansch C, Gao H, Hoekman D (1998) In: Devillers J (ed) Comparative QSAR. Taylor and Francis, Washington, p 285
24. Hansch C, Leo A (1979) Substituent constants for correlation analysis in chemistry and biology. Wiley/Interscience, New York
25. Hansch C, Leo A (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. Am. Chem. Soc., Washington
26. Kubinyi H (1993) QSAR: Hansch analysis and related approaches. VCH Publishers, Weinheim
27. Verloop A (1976) In: Ariens EJ (ed) Drug design, vol 3. Academic, New York, p 133

28. Verloop A, Tipker J (1977) In: Buisman JAK (ed) Biological activity and chemical structure. Elsevier, Amsterdam, p 63
29. Verloop A, Tipker J (1977) In: Jerman-Blazić D (ed) QSAR in drug design and toxicology, Hadzi D21–23. Elsevier, Amsterdam, p 97
30. Kubinyi H (1998) Quantitative structure-activity relationships in drug design. In: Schleyer PVR et al (eds) Encyclopedia of computational chemistry. Wiley, Chichester, p 2309
31. Kubinyi H (2006) In: Sener EA, Yalcin I (eds) QSAR and molecular modeling in rational design of bioactive molecules. CADD Society, Ankara, p 30
32. Balaban AT (1976) (ed) Chemical applications of graph theory. Academic Press, London
33. Trinajstić N (1992) Chemical graph theory, 2nd edn. CRC Press, Boca Raton
34. Ivanciuc O, Balaban AT (1998) Graph theory in chemistry. In: Schleyer PVR et al (eds) Encyclopedia of computational chemistry. Wiley, Chichester, p 1169
35. Rouvray DH, Balaban AT (1979) In: Wilson RJ, Beineke LW (eds) Applications of graph theory. Academic Press, London, p 177
36. Balaban AT (1995) J Chem Inf Comput Sci 35:339
37. Ivanciuc O, Ivanciuc T, Diudea MV (1997) SAR QSAR Environ Res 7:63
38. Ivanciuc O, Ivanciuc T (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam
39. Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) In: Gutman I (ed) Graph theoretical matrices in chemistry. University of Kragujevac (Serbia)
40. Wiener H (1947) J Am Chem Soc 69(17):2636
41. Hosoya H (1971) Bull Chem Soc Jpn 44:2332
42. Platt JR (1947) J Chem Phys 15:419
43. Platt JR (1952) J Phys Chem 56:151
44. Gordon M, Scantlebury GR (1967) J Chem Soc B:1
45. Gutman I, Ruscic B, Trinajstić N, Wilcox CF (1975) J Chem Phys 62:3399
46. Schultz HP, Schultz TP (1998) J Chem Inf Comput Sci 38:853
47. Balaban AT (1979) Theor Chim Acta 5:239
48. Bonchev D, Balaban AT, Mekenyan O (1980) J Chem Inf Comput Sci 20:106
49. Bonchev D, Mekenyan O, Balaban AT (1989) J Chem Inf Comput Sci 29:91
50. Bonchev D (1989) Theochem 185:155
51. Balaban AT, Bertelsen S, Basak SC (1994) MATCH Commun Math Comput Chem 30:55
52. Bonchev D (2001) J Mol Graphics Model 20:65
53. Randić M (1991) J Math Chem 7:155
54. Randić M (1975) J Am Chem Soc 97:6609
55. Kier LB, Hall LH (1976) J Pharm Sci 65:1806
56. Kier LB (1989) Quant Struct-Act Relat 8:221
57. Kier LH, Hall LB (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam, pp 307, 455, 491
58. Bonchev D, Trinajstić N (1977) J Chem Phys 67:4517
59. Basak SC (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam, p 563
60. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) J Comput Chem 5:581
61. Balaban AT, Balaban TS (1991) J Math Chem 8:383
62. Balaban AT (1983) Pure Appl Chem 55:199
63. Balaban AT (1982) Chem Phys Lett 89:399
64. Balaban AT, Ionescu-Pallas N, Balaban TS (1985) MATCH Commun Math Comput Chem 17:121
65. Balaban AT (1986) MATCH Commun Math Comput Chem 21:115
66. Ivanciuc O, Ivanciuc T, Balaban AT (1998) J Chem Inf Comput Sci 38:395
67. Balaban AT, Quintas LV (1983) MATCH Commun Math Comput Chem 14:213
68. Lovasz L, Pelikan J (1973) Periodica Math Hung 3:175
69. Medeleanu M, Balaban AT (1998) J Chem Inf Comput Sci 38:1038
70. Randić M, Vračko M, Novič M (2001) In: Diudea MV (ed) QSPR/QSAR studies by molecular descriptors. Nova Sci. Publ., Huntington, p 147
71. Plavšić D, Nikolić S, Trinajstić N, Mihalić Z (1993) J Math Chem 12:235
72. Ivanciuc O, Balaban TS, Balaban AT (1993) J Math Chem 12:309
73. Khadikar PV, Deshpande NV, Kale PP, Dobrynin AA, Gutman I, Dömötör G (1995) J Chem Inf Comput Sci 35:547
74. Diudea MV, Minailiuc O, Balaban AT (1991) J Comput Chem 12:527
75. Balaban AT, Diudea MV (1993) J Chem Inf Comput Sci 33:421
76. Diudea MV (1997) J Chem Inf Comput Sci 37:292–300
77. Diudea MV (1997) MATCH Commun Math Comput Chem 35:163
78. Diudea MV, Vizitiu AE, Janežič D (2007) J Chem Inf Model 47:864
79. Rücker G, Rücker C (1993) J Chem Inf Comput Sci 33:683
80. Rücker G, Rücker C (1999) J Chem Inf Comput Sci 39:788
81. Rücker G, Rücker C (1994) J Chem Inf Comput Sci 34:534
82. Balaban AT, Beteringhe A, Constantinescu T, Filip PA, Ivanciuc O (2007) J Chem Inf Model 47:716; Vukičević D, Beteringhe A, Constantinescu T, Pompe M, Balaban AT (2008) Chem Phys Lett 464:155
83. Filip PA, Balaban TS, Balaban AT (1987) J Math Chem 1:61
84. Basak SC, Gute BD (2001) J Mol Graphics Model 20:95
85. Estrada E (2001) Chem Phys Lett 336:248
86. Estrada E (2003) J Phys Chem A 107:7482
87. Estrada E (2004) J Phys Chem A 108:5468
88. Matamala AR, Estrada E (2005) J Phys Chem A 109:9890
89. Matamala AR, Estrada E (2005) Chem Phys Lett 410:343
90. Estrada E, Matamala AR (2007) J Chem Inf Comput Sci 47:794
91. Estrada E, Gutierrez Y (2001) MATCH Commun Math Comput Chem 44:155
92. Esterada E (2000) SAR QSAR Environ Des 11:55; Estrada E, Gutierrez Y, Gonzales H (2000) J Chem Inf Comput Sci 40:1386
93. Estrada E, Pena A, Garcia-Domenech R (1998) J Comput-Aided Mol Des 12:583
94. Estrada E, Quincoces JA, Patlewicz G (2004) Mol Divers 8:21
95. Randić M (1991) J Chem Inf Comput Sci 31:311
96. Randić M (1991) New J Chem 15:517
97. Randić M (2001) J Mol Graphics Model 20:19
98. Randić M, Pompe M (2001) J Chem Inf Comput Sci 41:573
99. Pompe M, Randić M (2006) J Chem Inf Comput Sci 46:2
100. Randić M, Pompe M (2007) Acta Chim Slov 54:605
101. Toropov AA, Toropova AP, Mukhamedzhnova DV, Gutman I (2005) Indian J Chem 44A:1545
102. Toropov AA, Benfenati E (2007) Comput Biol Chem 31:57
103. Toropov AA, Benfenati E (2007) Eur J Med Chem 42:606

104. Roy K, Toropov AA, Raska I Jr (2007) *QSAR & Comp Sci* 26:460
105. Gálvez J, García R, Salabert MT, Soler R (1994) *J Chem Inf Comput Sci* 34:520
106. Gálvez J, de Julian-Ortiz JV, García-Domenech R (2001) *J Mol Graph Model* 20:84
107. Grassy G, Calas B, Yasri A, Lahana R, Woo J, Iyer S, Kaczorek M, Floc'h R, Buelow R (1998) *Nat Biotechnol* 16:748
108. Gorse D, Lahana R (2000) *Curr Opin Chem Biol* 4:287; Gorse A-D (2006) *Curr Top Med Chem* 6:3; Iyer S, Lahana R, Buelow R (2002) *Curr Pharm Des* 8:2217; Gorse D, Rees A, Kaczorek M, Lahana R (1999) *Drug Disc Today* 4:257
109. Grassy G, Kaczorek M, Lahana R, Yasri A (2006) *US Patent* 7,024,311
110. Kubinyi H, Folkers G, Martin YC (eds) (1998) *3D QSAR in drug design: Ligand-protein interactions and molecular similarity*, vols 9–11. Kluwer, Dordrecht
111. Balaban AT (ed) (1997) *From chemical topology to three-dimensional geometry*. Plenum, New York
112. Balaban AT (1997) *J Chem Inf Comput Sci* 37:645
113. Kubinyi H, Folkers G, Martin YC (eds) (1998) *3D QSAR in drug design: Recent advances*, vols. 12–14. Kluwer, Dordrecht
114. Cramer RD III, Patterson DE, Bunce JD (1988) *J Am Chem Soc* 110:5959
115. Cramer RD III, Patterson DE, Bunce JD (1989) In: Fauchère JL (ed) *Quantitative structure-activity relationships in drug design*. Alan R Liss, New York, p 161
116. Cramer RD III, DePriest SA, Patterson DE, Hecht P (1993) In: Kubinyi H (ed) *3D QSAR in Drug Design: Theory, methods and applications*. ESCOM, Leiden, p 583
117. Böhm M, Klebe G (2002) *J Med Chem* 45:1585
118. Martin YC, Kim KH, Lin CT (1996) In: Charton M (ed) *Advances in quantitative structure-activity relationships*. JAI Press, Greenwich
119. Todeschini R, Lasagni M, Marengo E (1994) *J Chemom* 8:263
120. Todeschini R, Gramatica P, Marengo E, Provenzano R (1995) *Chemom Intell Lab Syst* 27:221
121. Todeschini R, Vighi M, Provenzano R, Finizio A, Gramatica P (1996) *Chemosphere* 32:1527
122. Todeschini R, Gramatica P (1997) *Quant Struct-Act Relat* 16:113, 120
123. Todeschini R, Gramatica P (1997) *SAR QSAR Environ Res* 7:89
124. Gramatica P, Consonni V, Todeschini R (1999) *Chemosphere* 38:1371
125. Gramatica P, Corradi M, Consonni V (2000) *Chemosphere* 41:763
126. Ferguson AM, Heritage T, Jonathon P, Pack SE, Phillips L, Rogan J, Snaith PJ (1997) *J Comp-Aided Mol Des* 11:143
127. Fontaine F, Pastor M, Sanz F (2004) *J Med Chem* 47:2805
128. Johnson M, Maggiora GM (eds) (1990) *Concepts and applications of molecular similarity*. Wiley, New York; Maggiora GM, Shanmugasundaram V (2004) *Methods Mol Biol* 275:1
129. Horvath D, Mao B (2003) *QSAR Comb Sci* 22:498
130. Mezey PG (1993) *Shape in chemistry. An introduction to molecular shape and topology*. Wiley, New York
131. Carbo-Dorca R, Mezey PG (eds) (1998) *Advances in molecular similarity*, vol 2. JAI Press, Stamford
132. Karelson M, Lobanov VS, Katritzky AR (1996) *Chem Rev* 96:1027
133. Zyrianov Y (2005) *J Chem Inf Model* 45:657
134. Burden FR (1989) *J Chem Inf Comput Sci* 29:225
135. Basak SC, Balaban AT, Grunwald G, Gute BD (2000) *J Chem Inf Comput Sci* 40:891; Basak SC, Gute BD, Balaban AT (2004) *Croat Chem Acta* 77:331
136. Basak SC, Gute BD, Mills D (2006) *Arkivoc* ix:157
137. Bertz SH (1988) *Discret Appl Math* 19:65
138. Balaban AT (2002) In: Rouvray DH, King RB (eds) *Topology in chemistry: Discrete mathematics of molecules*. Horwood Publishing Ltd., Chichester, p 89
139. Ivanciuc O, Ivanciuc T, Cabrol-Bass D, Balaban AT (2000) *MATCH Commun Math Comput Chem* 42:155
140. Balaban AT (2002) *MATCH Commun Math Comput Chem* 45:5
141. Basak SC, Mills D, Mumtaz MM, Balasubramanian K (2003) *Indian J Chem* 42A:1385; Basak SC, Mills D, Balaban AT, Gute BD (2001) *J Chem Inf Comput Sci* 41:671
142. Katritzky AR, Lobanov VS, Karelson M (1995) *Chem Soc Rev* 24:279
143. Perun TJ, Propst CL (1989) *Computer-aided drug design. Methods and applications*. Marcel Dekker, New York
144. Karelson M, Maran U, Wang Y, Katritzky AR (2000) *Coll Czech Chem Commun* 64:1551
145. Jurs PC (1996) *Computer software applications in chemistry*, 2nd edn. Wiley, New York
146. Zupan J, Gasteiger J (1999) *Neural networks in chemistry and drug design*. Wiley, Weinheim
147. Devillers J (1996) *Neural networks in QSAR and drug design*. Academic Press, London
148. Jorissen RN, Gilson MK (2005) *J Chem Inf Model* 45:549
149. Devillers J (1996) *Genetic algorithms in molecular modeling*. Academic Press, London
150. Goldberg DE (1989) *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, Reading
151. Liu R, So SS (2001) *J Chem Inf Comput Sci* 41:1633
152. Gola J, Obrezanova O, Champness E, Segall M (2006) *QSAR Comb Sci* 25:1172
153. Katritzky AR, Kuanar M, Slavov S, Dobchev DA, Fara DC, Karelson M, Acree WE, Solov'ev VP, Varnek A (2006) *Bioorg Med Chem* 14:4888
154. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) *J Chem Inf Model* 45:1376
155. Balaban AT, Catana C (1994) *SAR QSAR Environm Res* 2:1
156. Mekenyan O, Bonchev D, Balaban AT (1988) *J Math Chem* 2:347
157. Charton M (2003) *J Comput-Aided Mol Des* 17:197; Charton M, Charton B (2003) *J Comput-Aided Mol Des* 17:211
158. Mekenyan O, Pavlov T, Grancharov V, Todorov M, Schmieder P, Veith G (2005) *J Chem Inf Model* 45:283
159. Schultz HP, Schultz EB, Schultz TP (1995) *J Chem Inf Comput Sci* 35:864
160. Pyka A (1993) *J Planar Chromatog Mod TLC* 6:282
161. Pyka A (1997) *J Serb Chem Soc* 62:251; Gutman I, Pyka A, *ibid.* 261
162. Pyka A (1999) *J Liq Chromatog Relat Technol* 22:41
163. Randić M, Zupan J (2001) *J Chem Inf Comput Sci* 41:550
164. Randić M, Balaban AT, Basak SC (2001) *J Chem Inf Comput Sci* 41:593
165. Labanowski JK, Motoc I, Dammkoehler RA (1991) *Comput Chem* 15:47
166. Bögel H, Dettman J, Randić M (1997) *Croat Chem Acta* 70:827
167. Stankevich IV, Skvortsova MI, Zefirov NS (1995) *Theochem* 342:173
168. Hosoya H, Gotoh M, Murajami M, Ikeda S (1999) *J Chem Inf Comput Sci* 39:192

169. Kier LB, Hall LB (1977) *Eur J Med Chem Chim Ther* 12:307
170. Estrada E (1999) *Chem Phys Lett* 312:556
171. Skvortsova MI, Fedyaev KS, Palyulin VA, Zefirov NS (2003) *Internet Electron J Mol Des* 2:70
172. Skvortseva MI, Baskin II, Slovokhovotova OL, Palyulin VA, Zefirov NS (1993) *J Chem Inf Comput Sci* 33:630
173. Gordeeva EV, Molchanova MS, Zefirov NS (1990) *Tetrahedron Comput Meth* 3:389
174. García GC, Luque-Ruiz I, Gómez MA, Doncel AC, Plaza AG (2005) *J Chem Inf Model* 45:231
175. Fink T, Bruggesser, Raymond J-L (2005) *Angew Chem Int Ed Engl* 44:1504
176. Bologa C, Allu TK, Olah M, Kappler MA, Oprea TI (2005) *J Comput Aided Mol Des* 19:625
177. Balaban AT (2005) *J Comput Aided Mol Des* 19:651
178. Clement OO, Güner OF (2005) *J Comput Aided Mol Des* 19:731
179. Faulon J-L, Brown WM, Martin S (2005) *J Comput Aided Mol Des* 19:637
180. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) *J Comput Aided Mol Des* 19:693
181. Filimonov D, Poroikov V (2005) *J Comput Aided Mol Des* 19:705

Books and Reviews

- Bohacek RS, McMartin C, Guida WC (1996) *Med Res Revs* 16:3
- Bonchev D, Rouvray DH (eds) (2005) *Complexity in chemistry, biology, and ecology*. Springer, New York
- Bonchev D, Buck GA (2007) *J Chem Inf Model* 47:909
- Estrada E, Uriarte E (2001) *Curr Med Chem* 8:1573
- Diudea MV (ed) (2000) *QSPR/QSAR studies by molecular descriptors*. Nova Science Press, New York
- Diudea MV, Florescu MS, Khadikar PV (2006) *Molecular topology and its applications*. Eficon, Bucharest
- Holtje HD, Sippl W (eds) (2001) *Rational approach to drug design*. Prous Sci, Barcelona
- Mannhold R, Kubinyi H, Timmermann H (eds) (1997) *Molecular modeling. Methods and principles in medicinal chemistry*, vol 5. VCH, Weinheim
- Ooms F (2000) *Curr Med Chem* 7:141
- Pogliani L (2000) *Chem Rev* 100:3827
- Randić M (1998) In: Schleyer PvR et al (eds) *Encyclopedia of computational chemistry*, vol 5. Wiley, Chichester, p 3018
- Richon AB, Young SS: An introduction of QSAR methodology. <http://www.netsci.org/Science/Compchem/feature19.html>
- Snyder JP, Snyder FD (2000) In: Gundertofte K, Jørgensen FS (eds) *Molecular modeling: prediction of bioactivity*. Kluwer, New York
- van de Waterbeemd H (1995) *Chemometric methods in drug design*. VCH, Weinheim

Article Outline

Glossary
 Definition of the Subject
 Introduction
 The Dynamic – or Stochastic – Game Model
 The Dynamic – or Stochastic – Game: Results
 Global Climate Change – Issues, Models
 Global Climate Change – Results
 Future Directions
 Bibliography

Glossary

Players The agents who take actions. These actions can be – depending on application – the choice of capital stock, greenhouse emissions, level of savings, level of Research & Development expenditures, price level, quality and quantity of effort, etc.

Strategies Full contingent plans for the actions that players take. Each strategy incorporates a choice of action not just once but rather a choice of action for every possible decision node for the player concerned.

Payoffs The utility or returns to a player from playing a game. These payoffs typically depend on the strategies chosen – and the consequent actions taken – by the player herself as well as those chosen by the other players in the game.

Game horizon The length of time over which the game is played, i. e., over which the players take actions. The horizon may be finite – if there are only a finite number of opportunities for decision-making – or infinite – when there are an infinite number of decision-making opportunities.

Equilibrium A vector of strategies, one for each player in the game, such that no player can unilaterally improve her payoffs by altering her strategy, if the others' strategies are kept fixed.

Climate change The consequence to the earth's atmosphere of economic activities such as the production and consumption of energy that result in a build-up of greenhouse gases such as carbon dioxide.

Dynamic Games with an Application to Climate Change Models

PRAJIT K. DUTTA

Department of Economics, Columbia University,
 New York, USA

Definition of the Subject

The study of dynamic games is an important topic within game theory. Dynamic games involve the study of problems that are a) inherently dynamic in nature (even without a game-theoretic angle) and b) are naturally studied from a strategic perspective. Towards that end the struc-

ture generalizes dynamic programming – which is the most popular model within which inherently dynamic but non-strategic problems are studied. It also generalizes the model of repeated games within which strategic interaction is often studied but which structure cannot handle dynamic problems. A large number of economic problems fit these two requirements.

In this paper we examine the dynamic game model. The structure is discussed in detail as well as its principal results. Then the paper introduces a leading important application, the economics of climate change. It is shown that the problem is best studied as a dynamic commons game. Some recent models and associated results are then discussed.

We begin the analysis with a recall of the familiar model of repeated games (whose main results have been presented elsewhere in this volume). That is followed by the generalization of that framework to the model of dynamic – also known as stochastic or Markovian – games. These games may be thought of as “repeated games with a state variable”. The presence of a state variable allows the analysis of situations where there is a fundamental dynamic intrinsic to the problem, a situation – or “state” – that changes over time often on account of the players’ past actions. (In contrast to repeated games where an identical stage game is played period after period.) Such a state variable maybe capital stock, level of technology, national or individual wealth or even environmental variables such as the size of natural resources or the stock of greenhouse gases. To provide a concrete illustration of the dynamic game concepts and results, this paper will provide a fairly detailed overview of ongoing research by a number of authors on the very current and important topic of the economics of global climate change.

Section “[The Dynamic – or Stochastic – Game Model](#)” recalls the repeated games structure, introduces the subject of dynamic games and presents the dynamic games model. Section “[The Dynamic – or Stochastic – Game: Results](#)” presents – mostly with proofs – the main results from the theory of dynamic games. Section “[Global Climate Change – Issues, Models](#)” then introduces the problem of climate change, argues why the dynamic game framework is appropriate for studying the problem and presents a family of models that have been recently studied by Dutta and Radner – and in a variant by Dockner, Long and Sorger. Finally, Sect. “[Global Climate Change – Results](#)” presents the main results of these analyzes of the climate change problem. Future directions for research are discussed in Sect. “[Future Directions](#)” while references are collected in Sect. “[Bibliography](#)”.

Introduction

In this paper we examine the dynamic game model. The structure is discussed in detail as well as its principal results. Then the paper introduces a leading important application, the economics of climate change. It is shown that the problem is best studied as a dynamic commons game. Some recent models and associated results are then discussed.

The Dynamic – or Stochastic – Game Model

The most familiar model of dynamic interaction is the Repeated Game model (described elsewhere in this volume). In that set-up players interact every period for many periods – finite or infinite in number. At each period they play exactly the same game, i. e., they pick from exactly the same set of actions and the payoff consequence of any given action vector is identical. Put differently, it is as if there is an intrinsic static set-up, a “state” of the system that never changes. The only thing that changes over time is (potentially) every player’s response to that fixed state, i. e., players (can) treat the game dynamically if they so wish but there is no inherent non-strategic reason to do so. An impartial “referee” choosing on behalf of the players to achieve some optimization aim indeed would pick the same action every period.

Things change in a set-up where the state can change over time. That is the structure to which we now turn. This set-up was introduced by Shapley [26] under the name of Stochastic Games. It has since also been called Markovian Games – on account of the Markovian structure of the intrinsic problem – or Dynamic Games. We will refer to the set-up (for the most part) as Dynamic Games.

Set-Up

There are I players, and time is discrete. Each period the players interact by picking an action. Their action interaction take place at a given state which state changes as a consequence of the action interaction. There is a payoff that each player receives in each period based on the action vector that was picked and the state.

The basic variables are:

Definition

- t Time period $(0, 1, 2, \dots, T)$.
- i Players $(1, \dots, I)$.
- $s(t)$ State at the beginning of period t , $s(t) \in S$.
- $a_i(t)$ Action taken by player i in period, $a_i(t) \in A_i$.
- $a(t)$ $(a_1(t), a_2(t), \dots, a_I(t))$ vector of actions taken in period t .
- $\pi_i(t)$ $\pi_i(s(t), a(t))$ payoff of player i in period t .

- $q(t)$ $q(s(t+1) | s(t), a(t))$ conditional distribution of state at the beginning of period $t+1$.
- δ The discount factor, $\delta \in [0, 1)$.

The state variable affects play in two ways as stated above. In any given period, the payoff to a player depends not only on the actions that she and other players take but it also depends on the state in that period. Furthermore, the state casts a shadow on future payoffs in that it evolves in a Markovian fashion with the state in the next period being determined – possibly stochastically – by the state in the current period and the action vector played currently.

The initial value of the state, $s(0)$, is exogenous. So is the discount factor δ and the game horizon, T . Note that the horizon can be finite or infinite. All the rest of the variables are endogenous, with each player controlling its own endogenous variable, the actions. Needless to add, both state as well as action variables can be multi-dimensional and when we turn to the climate change application it will be seen to be multi-dimensional in natural ways.

Example 1 S infinite – The state space can be countably or uncountably infinite. It will be seen that the infinite case, especially the uncountably infinite one, has embedded within it a number of technical complications and – partly as a consequence – much less is known about this case.

S finite – In this case, imagine that we have a repeated game like situation except that there are a finite number of stage games any one of which gets played at a time.

Example 2 When the number of players is one, i. e., $I = 1$, then we have a dynamic programming problem. When the number of states is one, i. e., $\#(S) = 1$, then we have a repeated game problem. (Alternatively, repeated games constitute the special case where the conditional distribution brings a state s always back to itself, regardless of action.) Hence these two very familiar models are embedded within the framework of dynamic games.

Histories and Strategies

Preliminaries – A *history* at time t , $h(t)$, is a list of prior states and action vectors up to time t (but not including $a(t)$)

$$h(t) = s(0), a(0), s(1), a(1), \dots, s(t).$$

Let the set of histories be denoted $H(t)$. A *strategy* for player i at time t , $\sigma_i(t)$, is a complete conditional plan that specifies a choice of action for every history. The choice may be probabilistic, i. e., may be an element of $P(A_i)$, the set of distributions over A_i . So a strategy at time t is

$$\sigma_i(t): H(t) \longrightarrow P(A_i).$$

A strategy for the entire game for player i , σ_i , is a list of strategies, one for every period: $\sigma_i = \sigma_i(0), \sigma_i(1), \dots, \sigma_i(t), \dots$. Let $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_I)$ denote a vector of strategies, one for each player.

A particular example of a strategy for player i is a pure strategy σ_i where $\sigma_i(t)$ is a deterministic choice (from A_i). This choice may, of course, be conditional on history, i. e., may be a map from $H(t)$ to A_i . Another example of a strategy for player i is one where the player's choice $\sigma_i(t)$ may be probabilistic but the conditioning variables are not the entire history but rather only the current state. In other words such a strategy is described by a map from S to $P(A_i)$ – and is called a Markovian strategy. Additionally, when the map is independent of time, the strategy is called a stationary Markovian strategy, i. e., a stationary Markovian strategy for player i is described by a mapping $f_i: S \longrightarrow P(A_i)$.

Example 3 Consider, for starters, a pure strategy vector σ , i. e., a pure strategy choice for every i . Suppose further that q the conditional distribution on states is also deterministic. In that case, there is, in a natural way, a unique history that is generated by σ :

$$h(t; \sigma) = s(0), a(0; \sigma), s(1; \sigma), a(1; \sigma), \dots, s(t; \sigma)$$

where $a(\tau; \sigma) = \sigma(\tau; h(\tau; \sigma))$ and $s(\tau+1; \sigma) = q(s(\tau+1) | s(\tau; \sigma), a(\tau; \sigma))$. This unique history associated with the strategy vector σ is also called the *outcome path* for that strategy. To every such outcome path there is an associated lifetime payoff

$$R_i(\sigma) = \sum_{t=0}^T \delta^t \pi_i(s(t; \sigma), a(t; \sigma)). \quad (1)$$

If σ is a mixed strategy, or if the conditional distribution q , is not deterministic, then there will be a joint distribution on the set of histories $H(t)$ generated by the strategy vector σ and the conditional distribution q in the obvious way. Moreover, there will be a marginal distribution on the state and action in period t , and under that marginal, an expected payoff $\pi_i(s(t; \sigma), a(t; \sigma))$. Thereafter lifetime payoffs can be written exactly as in Eq. 1.

Consider the game that remains after every history $h(t)$. This remainder is called a *subgame*. The restriction of the strategy vector σ to the subgame that starts after history $h(t)$, is denoted $\sigma | h(t)$.

Equilibrium

A strategy vector σ^* is said to be a Nash Equilibrium (or NE) of the game if

$$R_i(\sigma^*) \geq R_i(\sigma_i, \sigma_{-i}^*), \quad \text{for all } i, \sigma_i. \quad (2)$$

A strategy vector σ^* is said to be a Subgame Perfect (Nash) Equilibrium of the game – referred to in short as SPE – if not only is Eq. 2 true for σ^* but it is true for every restriction of the strategy vector σ^* to every subgame $h(t)$, i. e., is true for $\sigma^* \mid h(t)$ as well. In other words, σ^* is a SPE if

$$R_i(\sigma^* \mid h(t)) \geq R_i(\sigma_i, \sigma_{-i}^* \mid h(t)), \quad \text{for all } i, \sigma_i, h(t). \quad (3)$$

As is well-known, not all NE satisfy the further requirement of being a SPE. This is because a NE only considers the outcome path associated with that strategy vector σ^* – or, when the outcome path is probabilistic, only considers those outcome paths that have a positive probability of occurrence. That follows from the inequality Eq. 2. However, that does not preclude the possibility that players may have no incentive to follow through with σ^* if some zero probability history associated with that strategy is reached. (Such a history may be reached either by accident or because of deviation/experimentation by some player.) In turn that may have material relevance because how players behave when such a history is reached will have significance for whether or not a player wishes to deviate against σ^* . Eq. 3 ensures that – even after a deviation – σ^* will get played and that deviations are unprofitable.

Recall the definition of a stationary Markovian strategy (SMS) above. Associated with that class of strategies is the following definition of equilibrium. A stationary Markov strategy vector f^* is a Markov Perfect Equilibrium (MPE) if

$$R_i(f^*) \geq R_i(f_i, f_{-i}^*), \quad \text{for all } i, f_i.$$

Hence, a MPE restricts attention to SMS both on and off the outcome path. Furthermore, it only considers – implicitly – histories that have a positive probability of occurrence under f^* . Neither “restriction” is a restriction when T is infinite because when all other players play a SMS player i has a stationary dynamic programming problem to solve in finding his most profitable strategy and – as is well-known – he loses no payoff possibilities in restricting himself to SMS as well. And that best strategy is a best strategy on histories that have zero probabilities of occurrence as well as histories that have a positive probability of occurrence. In particular therefore, when T is infinite, a MPE is also a SPE.

The Dynamic – or Stochastic – Game: Results

The main questions that we will now turn to are:

1. Is there always a SPE in a dynamic – or stochastic – game?

2. Is there a characterization for the set of SPE akin to the Bellman optimality equation of dynamic programming? If yes, what properties can be deduced of the SPE payoff set?
3. Is there a Folk Theorem for dynamic games – akin to that in Repeated Games?
4. What are the properties of SPE outcome paths?

The answers to questions 1–3 are very complete for finite dynamic games, i. e., games where the state space S is finite. The answer is also complete for questions 1 and 2 when S is countably infinite but when the state space is uncountably infinite, the question is substantively technically difficult and there is reason to believe that there may not always be a SPE. The finite game arguments for question 3 is conceptually applicable when S is (countably or uncountably) infinite provided some technical difficulties can be overcome. That and extending the first two answers to uncountably infinite S remain open questions at this point. Not a lot is known about Question 4.

Existence

The first result is due to Parthasarathy [23] and applies to the infinite horizon model, i. e., where $T = \infty$. When T is finite, the result and arguments can be modified in a straightforward way as will be indicated in the remarks following the proof.

Theorem 1 *Suppose that S is countable, A_i are finite sets, and the payoff functions π_i are bounded. Suppose furthermore that T is infinite. Then there is a MPE (and hence a SPE).*

Proof The proof will be presented by way of a fixed point argument. The domain for the fixed point will be the set of stationary Markovian strategies:

$$M_i = \{f_i: S \rightarrow P(A_i), \quad \text{s.t. for all } s, \sum_{a_i} f_i(a_i; s) = 1, f_i(a_i; s) \geq 0\}.$$

□

Properties of M_i : In the pointwise convergence topology, M_i is compact. That this is so follows from a standard diagonalization argument by way of which a subsequence can be constructed from any sequence of SMS f_i^n such that the subsequence, call it $f_i^{n'}$ has the property that $f_i^{n'}(s)$ converges to some $f_i^0(s)$ for every s . Clearly, $f_i^0 \in M_i$. The diagonalization argument requires S to be countable and A_i to be finite.

M_i is also clearly convex since its elements are probability distributions on A_i at every state.

The mapping for which we shall seek a fixed point is the best response mapping:

$$B_i(f) = \{g_i \in M_i : R_i(g_i, f_{-i}) \geq R_i(f_i, f_{-i}), \text{ for all } f_i.\}$$

Since the best response problem for player i is a stationary dynamic programming problem, it follows that there is an associated value function for the problem, say v_i , such that it solves the optimality equation of dynamic programming

$$v_i(s) = \max_{\lambda_i} \left\{ \pi_i(s, \lambda_i, f_{-i}(s)) + \delta \sum_{s'} v_i(s') q(s' | s, \lambda_i, f_{-i}(s)) \right\} \quad (4)$$

where

$$\begin{aligned} \pi_i(s, \lambda_i, f_{-i}(s)) \\ = \sum_{a_{-i}} \left[\sum_{a_i} \pi_i(s, a_i, a_{-i}) \lambda_i(a_i) \right] f_i(a_{-i}, s) \end{aligned} \quad (5)$$

where $\lambda_i(a_i)$ is the probability of player i picking action a_i whilst $f_i(a_{-i}, s)$ is the product probability of players other than i picking the action vector a_{-i} . Similarly,

$$\begin{aligned} q(s' | s, \lambda_i, f_{-i}(s)) \\ = \sum_{a_{-i}} \left[\sum_{a_i} q(s' | s, a_i, a_{-i}) \lambda_i(a_i) \right] f_i(a_{-i}, s). \end{aligned} \quad (6)$$

Additionally, it follows that the best best response, i. e., g_i , solves the optimality equation, i. e.,

$$v_i(s) = \pi_i(s, g_i, f_{-i}(s)) + \delta \sum_{s'} v_i(s') q(s' | s, g_i, f_{-i}(s)) \quad (7)$$

where $\pi_i(s, g_i, f_{-i}(s))$ and $q(s' | s, g_i, f_{-i}(s))$ have the same interpretations as given by Eqs. 5 and 6.

Properties of B_i : B_i is a correspondence that is convex-valued and upper hemi-continuous. That B_i is convex-valued follows from the fact that we are operating in the set of mixed strategies, that every convex combination of mixed strategies is itself a mixed strategy and that every convex combination of best responses is also a best response.

To show that B_i is upper hemi-continuous, consider a sequence of other players' strategies f_{-i}^n , an associated best response sequence of player i , g_i^n with value function sequence v_i^n . Note that each of these best responses g_i^n satisfies the Eqs. 6 and 7 (for the value function v_i^n). By diagonalization there exist subsequences and subsequential

pointwise convergent limits: $f_{-i}^n \rightarrow f_{-i}^0$, $g_i^n \rightarrow g_i^0$, and $v_i^n \rightarrow v_i^0$. It suffices to show that

$$v_i^0(s) = \max_{\lambda_i} \left\{ \pi_i(s, \lambda_i, f_{-i}^0(s)) + \delta \sum_{s'} v_i^0(s') q(s' | s, \lambda_i, f_{-i}^0(s)) \right\} \quad (8)$$

and

$$v_i^0(s) = \pi_i(s, g_i^0, f_{-i}^0(s)) + \delta \sum_{s'} v_i^0(s') q(s' | s, g_i^0, f_{-i}^0(s)). \quad (9)$$

Equation. 9 will be proved by using the analog of Eq. 7, i. e.,

$$v_i^n(s) = \pi_i(s, g_i^n, f_{-i}^n(s)) + \delta \sum_{s'} v_i^n(s') q(s' | s, g_i^n, f_{-i}^n(s)). \quad (10)$$

Clearly the left-hand side of Eq. 10 converges to the left-hand side of Eq. 9. Lets check the right-hand side of each equation. Evidently

$$\begin{aligned} \sum_{a_{-i}} \left[\sum_{a_i} \pi_i(s, a_i, a_{-i}) g_i^n(a_i) \right] f_i^n(a_{-i}; s) \\ \rightarrow \sum_{a_{-i}} \left[\sum_{a_i} \pi_i(s, a_i, a_{-i}) g_i^0(a_i) \right] f_i^0(a_{-i}; s) \end{aligned}$$

since each component of the sum converges and we have a finite sum. Finally,

$$\begin{aligned} & \left| \sum_{s'} v_i^n(s') q(s' | s, g_i^n, f_{-i}^n(s)) \right. \\ & \quad \left. - \sum_{s'} v_i^0(s') q(s' | s, g_i^0, f_{-i}^0(s)) \right| \\ & \leq \left| \sum_{s'} [v_i^n(s') - v_i^0(s')] q(s' | s, g_i^n, f_{-i}^n(s)) \right| \quad (11) \\ & \quad + \left| \sum_{s'} v_i^0(s') [q(s' | s, g_i^n, f_{-i}^n(s)) \right. \\ & \quad \left. - q(s' | s, g_i^0, f_{-i}^0(s))] \right|. \end{aligned}$$

The first term in the right-hand side of the inequality above goes to zero by the dominated convergence theo-

rem. The second term can be re-written as

$$\sum_{s'} v_i^0(s') q(s' | s, a_i, a_{-i}) \cdot \sum_{a_{-i}} \sum_{a_i} [g_i^n(a_i) f_i^n(a_{-i}; s) - g_i^0(a_i) f_i^0(a_{-i}; s)]$$

and goes to zero because each of the finite number of terms in the summation over action probabilities goes to zero. Hence the RHS of Eq. 10 converges to the RHS of Eq. 9 and the proposition is proved.

Remark 1 Note that the finiteness of A_i is crucial. Else, the very last argument would not go through, i. e., knowing that $[g_i^n(a_i) f_i^n(a_{-i}; s) - g_i^0(a_i) f_i^0(a_{-i}; s)] \rightarrow 0$ for every action vector \mathbf{a} would not guarantee that the sum would converge to zero as well.

Remark 2 If the horizon were finite one could use the same argument to prove that there exists a Markovian strategy equilibrium, though not a stationary Markovian equilibrium. That proof would combine the arguments above with backward induction. In other words, one would first use the arguments above to show that there is an equilibrium at every state in the last period T . Then the value function so generated, v_i^T , would be used to show that there is an equilibrium in period $T - 1$ using the methods above thereby generating the relevant value function for the last two periods, v_i^{T-1} . And so on.

The natural question to ask at this point is whether the restriction of countable finiteness of S can be dropped (and – eventually – the finiteness restriction on A_i). The answer, unfortunately, is not easily. The problems are twofold:

1. *Sequential Compactness of the Domain Problem* – If S is uncountably infinite, then it is difficult to find a domain M_i that is sequentially compact. In particular, diagonalization arguments do not work to extract candidate strategy and value function limits.
2. *Integration to the Limit Problem* – Note as the other players change their strategies, f_{-i}^n , continuation payoffs to player i change in two ways. They change first because the value function v_i^n changes, i. e., $v_i^n \neq v_i^m$ if $n \neq m$. Second, the expected continuation value changes because the measure over which the value function is being integrated, $q(s' | s, \lambda_i, f_{-i}^n(s))$, itself changes, i. e., $q(s' | s, \lambda_i, f_{-i}^n(s)) \neq q(s' | s, \lambda_i, f_{-i}^m(s))$. This is the well-known – and difficult – integration to the limit problem: simply knowing that v_i^n “converges” to v_i^0 in some sense – such as pointwise – and knowing that the integrating measure q^n “converges” to q^0

in some sense – such as in the weak topology – does not, in general, imply that

$$\int v_i^n dq^n \rightarrow \int v_i^0 dq^0. \quad (12)$$

(Of course in the previous sentence q^n is a more compact stand-in for $q(s' | s, \lambda_i, f_{-i}^n(s))$ and q^0 for $q(s' | s, \lambda_i, f_{-i}^0(s))$.) There are a limited number of cases where Eq. 12 is known to be true. These results typically require q^n to converge to q^0 in some strong sense. In the dynamic game context what this means is that very strong convergence restrictions need to be placed on the transition probability q . This is the underlying logic behind results reported in [7,20,22,24].

Such strong convergence properties are typically not satisfied when q is deterministic – which case comprises the bulk of the applications of the theory. Indeed simply imposing continuity when q is deterministic appears not to be enough to generate an existence result. Harris et al. [16] and Dutta and Sundaram [14] contain results that show that there may not be a SPE in finite horizon dynamic games when the transition function q is continuous. Whether other often used properties of q and π_i – such as concavity and monotonicity – can be used to rescue the issue remains an open question.

Characterization

The Bellman optimality equation has become a workhorse for dynamic programming analysis. It is used to derive properties of the value function and the optimal strategies. Moreover it provides an attractive and conceptually simple way to view a multiple horizon problem as a series of one-stage programming problems by exploiting the recursive structure of the optimization set-up. A natural question to ask, since dynamic games are really multi-player versions of dynamic programming, is whether there is an analog of the Bellman equation for these games. Abreu, Pearce and Stachetti – APS (1988), in an important and influential paper, showed that this is indeed the case for repeated games. They defined an operator, hereafter the APS operator, whose largest fixed point is the set of SPE payoffs in a repeated game and whose every fixed point is a subset of the set of SPE payoffs. (Thereby providing a necessary and sufficient condition for SPE equilibrium payoffs in much the same way that the unique fixed point of the Bellman operator constitutes the value function for a dynamic programming problem.) As with the Bellman equation, the key idea is to reduce the multiple horizon problem to a (seemingly) static problem.

In going from repeated to dynamic games there are some technical issues that arise. We turn now to that analysis pointing out along the way where the technical pitfalls are. Again we start with the infinite horizon model, i.e., where $T = \infty$. When T is finite, the result and arguments can be modified in a straightforward way as will be indicated in a remark following the proof.

But first, some definitions. Suppose for now that S is countable.

APS Operator – Consider a compact-valued correspondence, W defined on domain S which takes values that are subsets of \mathbb{R}^I . Define the APS operator on W , call it LW , as follows:

$$LW(s) = \left\{ \begin{array}{l} v \in \mathbb{R}^I: \exists \hat{f} \in P(A) \text{ and} \\ w: S \times A \times S \rightarrow W, \\ \text{uniformly bounded, s.t. } v_i = \pi_i(s, \hat{f}) \\ + \delta \sum_{s'} w_i(s, \hat{f}, s') q(s' | s, \hat{f}) \\ \geq \pi_i(s, a_i, \hat{f}_{-i}) + \delta \sum_{s'} w_i(s, a_i, \hat{f}_{-i}, s') \\ q(s' | s, a_i, \hat{f}_{-i}), \text{ for all } a_i, i \end{array} \right\} \quad (13)$$

where, as before,

$$\pi_i(s, \hat{f}) = \sum_{a-i} \left[\sum_{a_i} \pi_i(s, a_i, a_{-i}) \hat{f}_i(a_i) \right] \hat{f}_{-i}(a_{-i}; s) \quad (14)$$

and

$$q(s' | s, \hat{f}) = \sum_{a-i} \left[\sum_{a_i} q(s' | s, a_i, a_{-i}) \hat{f}_i(a_i) \right] \hat{f}_{-i}(a_{-i}; s) \quad (15)$$

Finally, let V^* denote the SPE payoffs correspondence, i.e., $V^*(s)$ is the set of SPE payoffs starting from initial state s . Note that the correspondence is non-empty by virtue of Theorem 1.

Theorem 2 Suppose that S is countable, A_i are finite sets, and the payoff functions π_i are bounded. Suppose furthermore that T is infinite. Then a) V^* is a fixed point of the APS operator, i.e., $LV^* = V^*$. Furthermore, b) consider any other fixed point, i.e., a correspondence \tilde{V} such that $L\tilde{V} = \tilde{V}$. Then it must be the case that $\tilde{V} \subset V^*$. Finally, c) there is an algorithm that generates the SPE correspondence, V^* .

Proof of a: Suppose that $v^* \in V^*(s)$, i.e., is a SPE payoff starting from s . Then, by definition, there is a first-period play, f^* and a continuation strategy after every one-period

history $h(1)$, $\sigma^*(h(1))$, such that $\{f^*, \sigma^*(h(1))\}$ is a SPE. By definition, then, the payoffs associated with each history-dependent strategy, $\sigma^*(h(1))$, call them $w_i(s, f^*, s')$, satisfy Eq. 13. (Note that π_i bounded implies that the lifetime payoffs $w_i(s, f^*, s')$ are uniformly bounded. In other words, $v^* \in LV^*(s)$. On the other hand, suppose that $v^* \in LV^*(s)$. Then, by definition, there is a first-period play, f^* , and SPE payoffs, $w(s, f^*, s')$, that together satisfy Eq. 13. Let the SPE strategy associated with $w(s, f^*, s')$ be $\sigma^*(h(1))$. It is not difficult to see that the concatenated strategy – $f^*, \sigma^*(h(1))$ – forms a SPE. Since the associated lifetime payoff is v^* , it follows that $v^* \in V^*(s)$. \square

Proof of b: Suppose that $L\tilde{V} = \tilde{V}$. Let $v \in L\tilde{V}$. Then, from Eq. 13, there is \tilde{f} , and SPE payoffs, $w(s, \tilde{f}, s')$ that satisfy the equation. In turn, since $w(s, \tilde{f}, s') \in \tilde{V}(s') = L\tilde{V}(s')$ there is an associated $\tilde{f}(s')$ and $w(s', \tilde{f}, s'')$ for which Eq. 13 holds. By repeated application of this idea, we can create a sequence of strategies for periods $t = 0, 1, 2, \dots$ – $\tilde{f}, \tilde{f}(s'), \tilde{f}(s, s'), \dots$ such that at each period Eq. 13 holds. Call the strategy so formed, ϕ . This strategy can then not be improved upon by a single-period deviation. A standard argument shows that if a strategy cannot be profitably deviated against in one period then it cannot be profitably deviated against even by deviations in multiple periods. (This idea of “unimprovability” is already present in dynamic programming. Within the context of repeated games, it was articulated by Abreu [1].) \square

Proof of c: Note two properties of the APS operator:

Lemma 1 LW is a compact-valued correspondence (whenever W is compact-valued).

Proof Consider Eq. 13. Suppose that $v^n \in LW(s)$ for all n , with associated \hat{f}^n and w^n . By diagonalization, there exists a subsequence s.t. $v^n \rightarrow v^0$, $\hat{f}^n \rightarrow \hat{f}^0$ and $w^n \rightarrow w^0$. This argument uses the countability of S and the finiteness of A_i . From Eq. 14 evidently $\pi_i(s, \hat{f}^n) \rightarrow \pi_i(s, \hat{f}^0)$ and similarly from Eq. 15 $\sum_{s'} w_i(s, \hat{f}^n, s') q(s' | s, \hat{f}^n)$ goes to $\sum_{s'} w_i(s, \hat{f}^0, s') q(s' | s, \hat{f}^0)$. Hence the inequality in Eq. 13 is preserved and $v^0 \in LW(s)$. \square

It is not difficult to see that – on account of the boundedness of π_i – if W has a uniformly bounded selection, then so does LW . Note that the operator is also monotone in the set-inclusion sense, i.e., if $W'(s) \subset W(s)$ for all s then $LW' \subset LW$.

The APS algorithm finds the set of SPE payoffs by starting from a particular starting point, an initial set $W^0(s)$ that is taken to be the set of all feasible payoffs from initial state s . (And hence the correspondence W^0 is so

defined for every initial state.) Then define, $W^1 = LW^0$. More generally, $W^{n+1} = LW^n$, $n \geq 0$. It follows that $W^1 \subset W^0$. This is because W^1 requires a payoff that is not only feasible but additionally satisfies the incentive inequality of Eq. 13 as well. From the monotone inclusion property above it then follows that, more generally, $W^{n+1} \subset W^n$, $n \geq 0$. Furthermore, $W^n(s)$ is a non-empty, compact set for all n (and s). Hence, $W^\infty(s) = \bigcap_n W^n(s) = \lim_{n \rightarrow \infty} W^n(s)$ is non-empty and compact.

Let us now show that W^∞ is a fixed point of the APS operator, i. e., that $LW^\infty = W^\infty$.

Lemma 2 $LW^\infty = W^\infty$, or, equivalently,
 $L(\lim_{n \rightarrow \infty} W^n) = \lim_{n \rightarrow \infty} LW^n$.

Proof Clearly, by monotonicity, $L(\lim_{n \rightarrow \infty} W^n) \subset \lim_{n \rightarrow \infty} LW^n$. So consider a $v \in LW^n(s)$, for all n . By Eq. 13 there is at each n an associated first-period play f^n and a continuation payoff $w^n(s, f^n, s')$ such that the inequality is satisfied and

$$v_i = \pi_i(s, f^n) + \delta \sum_{s'} w_i^n(s, f^n, s') q(s' | s, f^n).$$

By the diagonalization argument, and using the countability of S , we can extract a (subsequential) limit $f^\infty = \lim_{n \rightarrow \infty} f^n$ and $w^\infty = \lim_{n \rightarrow \infty} w^n$. Clearly, $w^\infty \in W^\infty$. Since equalities and inequalities are maintained in the limit, equally clearly

$$v_i = \pi_i(s, f^\infty) + \delta \sum_{s'} w_i^\infty(s, f^\infty, s') q(s' | s, f^\infty)$$

and

$$v_i \geq \pi_i(s, a_i, f_{-i}^\infty) + \delta \sum_{s'} w_i^\infty(s, a_i, f_{-i}^\infty, s') q(s' | s, a_i, f_{-i}^\infty), \text{ for all } a_i, i$$

thereby proving that $v \in L(\lim_{n \rightarrow \infty} W^n)(s)$. The lemma is proved. \square

Since the set of SPE payoffs, $V^*(s)$, is a subset of $W^0(s)$ – and $LV^*(s) = V^*(s)$ – it further follows $V^*(s) \subset W^\infty(s)$, for all s . From the previous lemma, and part b), it follows that $V^*(s) \supset W^\infty(s)$, for all s . Hence, $V^* = W^\infty$. Theorem 2 is proved. \square

A few remarks are in order.

Remark 1 If the game horizon T is finite, there is an immediate modification of the above arguments. In the algorithm above, take W^0 to be the set of SPE payoffs in the one-period game (with payoffs π_i for player i). Use the

APS operator thereafter to define $W^{n+1} = LW^n$, $n \geq 0$. It is not too difficult to show that W^n is the set of SPE payoffs for a game that lasts $n + 1$ periods (or has n remaining periods after the first one).

Remark 2 Of course an immediate corollary of the above theorem is that the set of SPE payoffs $V^*(s)$ is a compact set for every initial state s . Indeed one can go further and show that V^* is in fact an upper hemi-continuous correspondence. The arguments are very similar to those used above – plus the Maximum Theorem.

Remark 3 Another way to think of Theorem 2 is that it is also an existence theorem. Under the conditions outlined in the result, the SPE equilibrium set has been shown to be non-empty. Of course this is not a generalization of Theorem 1 since Theorem 2 does not assert the existence of a MPE.

Remark 4 When the state space A_i is infinite or the state space S is uncountably infinite we run into technical difficulties. The complications arise from not being able to take limits. Also, as in the discussion of the Integration to the Limit problem, integrals can fail to be continuous thereby rendering void some of the arguments used above.

Folk Theorem

The folk theorem for Repeated Games – Fudenberg and Maskin [15] following up on earlier contributions – is very well-known and the most cited result of that theory. It proves that the necessary conditions for a payoff to be a SPE payoff – feasibility and individual rationality – are also (almost) sufficient provided the discount factor δ is close enough to 1. This is the result that has become the defining result of Repeated Games. For supporters, the result and its logic of proof are a compelling demonstration of the power of reciprocity, the power of long-term relationships in fostering cooperation through the lurking power of “punishments” when cooperation breaks down. It is considered equally important and significant that such long-term relationships and behaviors are sustained through implicit promises and threats which therefore do not violate any legal prohibitions against explicit contracts that specify such behavior. For detractors, the “anything goes” implication of the Folk Theorem is a clear sign of its weakness – or the weakness of the SPE concept – in that it robs the theory of all predictive content. Moreover there is a criticism, not entirely correct, that the strategies required to sustain certain behaviors are so complex that no player in a “real-world” setting could be expected to implement them.

Be that as it may, the Folk Theorem question in the context of Dynamic Games then is: is it the case that feasibility and individual rationality are also (almost) enough to guarantee that a payoff is a SPE payoff at high enough δ ? Two sets of obstacles arise in settling this question. Both emanate from the same source, the fact that the state does not remain fixed in the play of the games, as it does in the case of Repeated Games. First, one has to think long and hard as to how one should define individual rationality. Relatedly, how does one track feasibility? In both cases, the problem is that what payoff is feasible and individually rational depends on the state and hence changes after every history $h(t)$. Moreover, it also changes with the discount factor δ . The second set of problems stems from the fact that a deviation play can unalterably change the future in a dynamic game – unlike a repeated game where the basic game environment is identical every period. Consequently one cannot immediately invoke the logic of repeated game folk theorems which basically work because any deviation has only short-term consequences while the punishment of the deviation is long-term. (And so if players are patient they will not deviate.)

Despite all this, there are some positive results that are around. Of these, the most comprehensive is one due to Dutta [9]. To set the stage for that result, we need a few crucial preliminary results. For this sub-section we will assume that S is finite – in addition to A_i .

Feasible Payoffs Role of Markovian Strategies – Let $F(s, \delta)$ denote the set of “average” feasible payoffs from initial state s and for discount factor δ . By that I mean

$$F(s, \delta) = \{v \in \mathbb{R}^I : \exists \text{ strategy } \sigma \\ \text{s.t. } v = (1 - \delta) \sum_{t=0}^T \delta^t \pi_i(s(t; \sigma), a(t; \sigma))\}.$$

Let $\Phi(s, \delta)$ denote the set of “average” feasible payoffs from initial state s and for discount factor δ that are generated by pure stationary Markovian strategies – PSMS. Recall that a SMS is given by a map f_i from S to the probability distributions over A_i , so that at state $s(t)$ player i chooses the mixed strategy $f_i(s(t))$. A pure SMS is one where the map f_i is from S to A_i . In other words,

$$\Phi(s, \delta) = \{v \in \mathbb{R}^I : \exists \text{ PSMS } f \\ \text{s.t. } v = (1 - \delta) \sum_{t=0}^T \delta^t \pi_i(s(t; f), a(t; f))\}.$$

Lemma 3 Any feasible payoff in a dynamic game can be generated by averaging over payoffs to stationary Markov strategies, i. e., $F(s, \delta) = \text{co}\Phi(s, \delta)$, for all (s, δ) .

Proof Note that $F(s, \delta) = \text{co}[\text{extreme points } F(s, \delta)]$. In turn, all extreme points of $F(s, \delta)$ are generated by an optimization problem of the form: $\max_{\sigma} \sum_{i=1}^I \alpha_i v_i(s, \delta)$. That optimization problem is a dynamic programming problem. Standard results in dynamic programming show that the optimum is achieved by some stationary Markovian strategy. \square

Let $F(s)$ denote the set of feasible payoffs under the long-run average criterion. The next result will show that this is the set to which discounted average payoffs converge:

Lemma 4 $F(s, \delta) \rightarrow F(s)$, as $\delta \rightarrow 1$, for all s .

Proof Follows from the fact that a) $F(s) = \text{co}\Phi(s)$ where $\Phi(s)$ is the set of feasible long-run average payoffs generated by stationary Markovian strategies, and b) $\Phi(s, \delta) \rightarrow \Phi(s)$. Part b) exploits the finiteness of S (and A_i). \square

The lemmas above simplify the answer to the question: What is a feasible payoff in a dynamic game? Note that they also afford a dimensional reduction in the complexity and number of strategies that one needs to keep track of to answer the question. Whilst there are an uncountably infinite number of strategies – even with finite S and A_i – including the many that condition on histories in arbitrarily complex ways – the lemmas establish that all we need to track are the finite number of PSMS. Furthermore, whilst payoffs do depend on δ , if the discount factor is high enough then the set of feasible payoffs is well-approximated by the set of feasible long-run average payoffs to PSMS.

One further preliminary step is required however. This has to do with the fact that while $v \in F(s, \delta)$ can be exactly reproduced by a period 0 average over PSMS payoffs, after that period continuation payoffs to the various component strategies may generate payoffs that could be arbitrarily distant from v . This, in turn, can be problematic since one would need to check for deviations at every one of these (very different) payoffs. The next lemma addresses this problem by showing that there is an averaging over the component PSMS that is ongoing, i. e., happens periodically and not just at period 0, but which, consequently, generates payoffs that after all histories stays arbitrarily close to v .

For any two PSMS f^1 and f^2 denote a *time-cycle strategy* as follows: for T^1 periods play proceeds along f^1 , then it moves for T^2 periods to f^2 . After the elapse of the $T^1 + T^2$ periods play comes back to f^1 for T^1 periods and f^2 for T^2 periods. And so on. Define $\lambda^1 = T^1/(T^1 + T^2)$. In the obvious way, denote a general time-cycle strategy to be one that cycles over any finite number of PSMS f^k

where the proportion of time spent at strategy f^k is λ^k and allows the lengths of time to depend on the initial state at the beginning of the cycle.

Lemma 5 *Pick any $v \in \cap_s F(s)$. Then for all $\varepsilon > 0$ there is a time cycle strategy such that its long-run average payoff is within ε of v after all histories.*

Proof Suppose that $v = \sum_k \lambda^k(s) v^k(s)$ where $v^k(s)$ is the long-run average payoff to the k th PSMS when the initial state is s . Ensure that T^k is chosen such that a) the average payoff over those periods under that PSMS – $1/T^k \sum_{t=0}^{T^k-1} \pi_i(s(t; f^k), a(t; f^k))$ – is within ε of $v^k(s)$ for all s . And b) that $T^k(s)/\sum_l T^l(s)$ is arbitrarily close to $\lambda^k(s)$ for all s . \square

Since $\Phi(s, \delta) \rightarrow \Phi(s)$ it further follows that the above result also holds under discounting:

Lemma 6 *Pick any $v \in \cap_s F(s)$. Then for all $\varepsilon > 0$ there is a time cycle strategy and a discount cut-off $\delta(\varepsilon) < 1$ such that the discounted average payoffs to that strategy are within ε of v for all $\delta > \delta(\varepsilon)$ and after all histories.*

Proof Follows from the fact that $(1-\delta)/(1-\delta^T) \sum_{t=0}^{T-1} \delta^t \pi_i(s(t; f^k), a(t; f^k))$ goes to $\frac{1}{T^k} \sum_{t=0}^{T^k-1} \pi_i(s(t; f^k), a(t; f^k))$ as $\delta \rightarrow 1$. \square

Individually Rational Payoffs Recall that a min-max payoff is a payoff level that a player can guarantee by playing a best response. In a Repeated Game that is defined at the level of the component stage game. Since there is no analog of that in a dynamic game, the min-max needs to be defined over the entire game – and hence is sensitive to initial state and discount factor:

$$m_i(s, \delta) = \min_{\sigma_{-i}} \max_{\sigma_i} R_i(\sigma \mid s, \delta).$$

Evidently, given (s, δ) , in a SPE it cannot be that player i gets a payoff $v_i(s, \delta)$ that is less than $m_i(s, \delta)$. Indeed that inequality must hold at all states for a strategy to be a SPE, i.e., for all $s(t)$ it must be the case that $v_i(s(t), \delta) \geq m_i(s(t), \delta)$. But, whilst necessary, even that might not be a sufficient condition for the strategy to be a SPE. The reason is that if player i can deviate and take the game to, say, s' at $t+1$, rather than $s(t+1)$, he would do so if $v_i(s(t), \delta) < m_i(s', \delta)$ since continuation payoffs from s' have to be at least as large as the latter level and this deviation would be worth essentially that continuation when δ is close to 1. So sufficiency will require a condition such as $v_i(s(t), \delta) > \max_s m_i(s, \delta)$ for all $s(t)$. Call such a strategy *dynamically Individually Rational*. From the previous lemmas, and the fact that $m_i(s, \delta) \rightarrow m_i(s)$, as $\delta \rightarrow 1$, where

$m_i(s)$ is the long-run average min-max level for player i the following result is obvious. The min-max limiting result is due to Mertens and Neyman [21].

Lemma 7 *Pick any $v \in \cap_s F(s)$ such that $v_i > \max_s m_i(s)$ for all s . Then there is a time-cycle strategy which is dynamically Individually Rational for high δ .*

We are now ready to state and prove the main result:

Theorem 3 (Folk Theorem) *Suppose that S and A_i are finite sets. Suppose furthermore that T is infinite and that $\cap_s F(s)$ has dimension I (where I is the number of players). Pick any $v \in \cap_s F(s)$ such that $v_i > \max_s m_i(s)$ for all s . Then, for all $\varepsilon > 0$, there is a discount cut-off $\delta(\varepsilon) < 1$ and a time-cycle strategy that for $\delta > \delta(\varepsilon)$ is a SPE with payoffs that are within ε of v .*

Proof Without loss of generality, let us set $\max_s m_i(s) = 0$ for all i . From the fact that $\cap_s F(s)$ has dimension I it follows that we can find I payoff vectors in that set – v^i , $i = 1, \dots, I$ – such that for all i a) $v^i \gg 0$, b) $v_j^i > v_j^i$, $j \neq i$, and c) $v_i > v_i^i$. That we can find these vectors such that b) is satisfied follows from the dimensionality of the set. That we can additionally get the vectors to satisfy a) and c) follows from the fact that it is a convex set and hence an appropriate “averaging” with a vector such as v achieves a) while an “averaging” with i ’s worst payoff achieves c). \square

Now consider the following strategy: Norm – Start with a time-cycle strategy that generates payoffs after all histories that are within ε of v . Choose a high enough δ as required. Continue with that strategy if there are no deviations against it. Punishment – If there is, say if player i deviates, then min-max i for T periods and thereafter proceed to the time-cycle strategy that yields payoffs within ε of v^i after all histories. Re-start the punishment whenever there is a deviation.

Choose T in such a fashion that the payoff to the min-max period plus v_j^i is strictly less than v_i . That ensures there is no incentive to deviate against the norm provided the punishment is carried out. That there is incentive for players $j \neq i$ to punish player i follows from the fact that $v_j^i > v_j^j$ the former payoff being what they get from punishing and the latter from not punishing i . That there is incentive for player i not to deviate against his own punishment follows from the fact that re-starting the punishment only lowers his payoffs. The theorem is proved.

A few remarks are in order.

Remark 1 If the game horizon T is finite, there is likely a Folk Theorem along the lines of the result proved for

Repeated Games by Benoit and Krishna [4]. To the best of my knowledge it remains, however, an open question.

Remark 2 When the state space A_i is infinite or the state space S is uncountably infinite we again run into technical difficulties. There is an analog to Lemmas 3 and 4 in this instance and under appropriate richer assumptions the results can be generalized – see Dutta (1993). Lemmas 5–7 and the Folk Theorem itself does use the finiteness of S to apply uniform bounds to various approximations and those become problematical when the state space is infinite. It is our belief that nevertheless the Folk Theorem can be proved in this setting. It remains, however, to be done.

Dynamics

Recall that the fourth question is: what can be said about the dynamics of SPE outcome paths? The analogy that might be made is to the various convergence theorems – sometimes also called “turnpike theorems” – that are known to be true in single-player dynamic programming models. Now even within those models – as has become clear from the literature of the past twenty years in chaos and cycles theory for example – it is not always the case that there are regularities exhibited by the optimal solutions. Matters are worse in dynamic games.

Even within some special models where the single-player optima are well-behaved, the SPE of the corresponding dynamic game need not be. A classic instance is the neo-classical aggregative growth model. In that model, results going back fifty years show that the optimal solutions converge monotonically to a steady-state, the so-called “golden rule”. (For references, see Majumdar, Mitra and Nishimura (2000).) However, examples can be constructed – and may be found in Dutta and Sundaram (1996) and Dockner, Long and Sorger (1998) – where there are SPE in these models that can have arbitrarily complex state dynamics which for some range of discount factor values descend into chaos. And that may happen with Stationary Markov Perfect Equilibrium. (It would be less of a stretch to believe that SPE in general can have complex dynamics. The Folk Theorem already suggests that it might be so.)

There are, however, many questions that remain including the breadth of SPE that have regular dynamics. One may care less for complex dynamic SPE if it can be shown that the “good ones” have regular dynamics. What also remains to be explored is whether adding some noise in the transition equation can remove most complex dynamics SPE.

Global Climate Change – Issues, Models

Issues

The dramatic rise of the world’s population in the last three centuries, coupled with an even more dramatic acceleration of economic development in many parts of the world, has led to a transformation of the natural environment by humans that is unprecedented in scale. In particular, on account of the greenhouse effect, *global warming* has emerged as a central problem, unrivaled in its potential for harm to life as we know it on planet Earth. Seemingly the consequences are everywhere: melting and break-up of the world’s ice-belts whether it be in the Arctic or the Antarctic; heat-waves that set all-time temperature highs whether it be in Western Europe or sub-Saharan Africa; storms increased in frequency and ferocity whether it be Hurricane Katrina or typhoons in Japan or flooding in Mumbai. In addition to Al Gore’s eminently readable book, “An Inconvenient Truth”, two authoritative recent treatments are the Stern Review on the Economics of Climate Change, October, 2006 and the IPCC Synthesis Report, November, 2007. Here are three – additional – facts drawn from the IPCC Report:

1. Eleven of the last twelve years (1995–2006) have been amongst the twelve warmest years in the instrumental record of global surface temperatures (since 1850).
2. If we go on with “Business as Usual”, by 2100 global sea levels will probably have risen by 9 to 88 cm and average temperatures by between 1.5 and 5.5°C. Various factors contribute to global warming, but the major one is an increase in greenhouse gases (GHGs) – primarily, carbon dioxide – so called because they are transparent to incoming shortwave solar radiation but trap outgoing longwave infrared radiation. Increased carbon emissions due to the burning of fossil fuel is commonly cited as the principal immediate cause of global warming. A third relevant fact is:
3. Before the Industrial Revolution, atmospheric CO₂ concentrations were about 270–280 parts per million (ppm). They now stand at almost 380 ppm, and have been rising at about 1.5 ppm annually.

The IPCC Synthesis (2007) says “Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising global average sea level.” (IPCC Synthesis Report [17]).

It is clear that addressing the global warming problem will require the coordinated efforts of the world’s nations. In the absence of an international government, that coordination will have to be achieved by way of an inter-

national environmental treaty. For a treaty to be implemented, it will have to align the incentives of the signatories by way of rewards for cutting greenhouse emissions and punishments for not doing so. For an adequate analysis of this problem one needs a dynamic and fully strategic approach. A natural methodology for this then is the theory of Subgame Perfect (Nash) equilibria of dynamic games – which we have discussed at some length in the preceding sections.

Although there is considerable uncertainty about the exact costs of global warming, the two principal sources will be a rise in the sea-level and climate changes. The former may wash away low-lying coastal areas such as Bangladesh and the Netherlands. Climate changes are more difficult to predict; tropical countries will become more arid and less productive agriculturally; there will be an increased likelihood of hurricanes, fires and forest loss; and there will be the unpredictable consequences of damage to the natural habitat of many living organisms. On the other hand, emission abatement imposes its own costs. Higher emissions are typically associated with greater GDP and consumer amenities (via increased energy usage). Reducing emissions will require many or all of the following costly activities: cutbacks in energy production, switches to alternative modes of production, investment in more energy-efficient equipment, investment in R&D to generate alternative sources of energy, etc.

The principal features of the global warming problem are:

- *The Global Common* – although the sources of carbon buildup are localized, it is the total stock of GHGs in the global environment that will determine the amount of warming.
- *Near-irreversibility* – since the stock of greenhouse gases depletes slowly, the effect of current emissions can be felt into the distant future.
- *Asymmetry* – some regions will suffer more than others.
- *Nonlinearity* – the costs can be very nonlinear; a rise in one degree may have little effect but a rise in several degrees may be catastrophic.
- *Strategic Setting* – Although the players (countries) are relatively numerous, there are some very large players, and blocks of like-minded countries, like the US, Western Europe, China, and Japan. That warrants a strategic analysis.

The theoretical framework that accommodates all of these features is an *asymmetric dynamic commons* model with the global stock of greenhouse gases as the (common) state variable. The next sub-section will discuss a few models which have most of the above characteristics.

Models

Before presenting specific models, let us briefly relate the climate change problem to the general dynamic game model that we have seen so far, and provide a historical outline of its study. GHGs form – as we saw above – a global common. The study of global commons is embedded in dynamic commons game (DCG). In such a game the state space S is a single-dimensional variable with a “commons” structure meaning that each player is able to change the (common) state. In particular, the transition function is of the form

$$s(t+1) = q \left(s(t) - \sum_{i=1}^I a_i(t) \right).$$

The first analysis of a DCG may be found in [18]. That paper considered the particular functional form in which $q(s(t) - \sum_{i=1}^I a_i(t)) = [s(t) - \sum_{i=1}^I a_i(t)]^\alpha$ for a fixed fraction α . (And, additionally, Levhari and Mirman assumed the payoffs π_i to be logarithmic.) Consequently, the paper was able to derive in closed form a (linear) MPE and was able to analyze its characteristics.

Subsequently several authors – Sundaram [31], Sobel [27], Benhabib and Radner [3], Rustichini [25], Dutta and Sundaram (1992, [14]), Sorger [28] – studied this model in great generality, without making the specific functional form assumption of Levhari and Mirman, and established several interesting qualitative properties relating to existence of equilibria, welfare consequences and dynamic paths.

More recently in a series of papers by Dutta and Radner on the one hand and Dockner and his co-authors on the other, the DCG model has been directly applied to environmental problems including the problem of global warming. We shall describe the Dutta and Radner work in detail and also discuss some of the Dockner, Long and Sorger research. In particular, the transition equation is identical in the two models (and described below). What is different is the payoff functions.

We turn now to a simplified climate change model to illustrate the basic strategic ideas. The model is drawn from Dutta and Radner [12]. In the basic model there is no population growth and no possibility of changing the emissions producing technologies in each country. (Population growth is studied in Dutta and Radner [11] while certain kinds of technological changes are allowed in Dutta and Radner [10]. These models will be discussed later.) However, the countries may differ in their “sizes”, their emissions technologies, and their preferences.

There are I countries. The emission of (a scalar index of) greenhouse gases during period t by country i is de-

noted by $a_i(t)$. [Time is discrete, with $t = 0, 1, 2, \dots$, ad inf.] Let $A(t)$ denote the global (total) emission during period t ;

$$A(t) = \sum_{i=1}^I a_i(t). \quad (16)$$

The total (global) stock of greenhouse gases (GHGs) at the beginning of period t is denoted by $g(t)$. (Note, for mnemonic purposes we are denoting the state variable – the amount of “gas” – g .) The law of motion – or transition function q in the notation above – is

$$g(t+1) = A(t) + \sigma g(t), \quad (17)$$

where σ is a given parameter ($0 < \sigma < 1$). We may interpret $(1 - \sigma)$ as the fraction of the beginning-of-period stock of GHG that is dissipated from the atmosphere during the period. The “surviving” stock, $\sigma g(t)$, is augmented by the quantity of global emissions, $A(t)$, during the same period.

Suppose that the payoff of country i in period t is

$$\pi_i(t) = h_i[a_i(t)] - c_i g(t). \quad (18)$$

The function h_i represents, for example, what country i 's gross national product would be at different levels of its own emissions, holding the global level of GHG constant. This function reflects the costs and benefits of producing and using energy as well as the costs and benefits of other activities that have an impact on the emissions of GHGs, e. g., the extent of forestation. It therefore seems natural to assume that h_i is a strictly concave C^2 function that reaches a maximum and then decreases thereafter.

The parameter $c_i > 0$ represents the marginal cost to the country of increasing the global stock of GHG. Of course, it is not the stock of GHG itself that is costly, but the associated climatic conditions. As discussed below, in a more general model, the cost would be nonlinear.

Histories, strategies – Markovian strategies – and outcomes are defined in exactly the same way as in the general theory above – and will, hence, not be repeated. Thus associated with each strategy vector σ is a total discounted payoff for each player

$$v_i(\sigma, g_0) \equiv \sum_{t=0}^{\infty} \delta^t \pi_i(t; \sigma, g_0).$$

Similarly, SPE and MPE can be defined in exactly the same way as in the general theory.

The linearity of the model is undoubtedly restrictive in several ways. It implies that the model is unable to ana-

lyze catastrophes or certain kinds of feedback effects running back from climate change to economic costs. It has, however, two advantages: first, its conclusions are simple, can be derived in closed-form and can be numerically calibrated; hence may have a chance of informing policy-makers. Second, there is little consensus on what is the correct form of non-linearity in costs. Partly the problem stems from the fact that some costs are not going to be felt for another fifty to hundred years and forecasting the nature of costs on that horizon length is at best a hazardous exercise. Hence, instead of postulating one of many possible non-linear cost functions, all of which may turn out to be incorrect for the long-run, one can opt instead to work with a cost function which may be thought of as a linear approximation to any number of actual non-linear specifications.

Dockner, Long and Sorger (1998) impose linearity in the emissions payoff function h (whereas in Dutta and Radner it is assumed to be strictly concave) while their cost to g is strictly convex (as opposed to the above specification in which it is linear). The consequent differences in results we will discuss later.

Global Climate Change – Results

In this section we present two sets of results from the Dutta and Radner [12] paper. The first set of results characterize two benchmarks – the global Pareto optima, and a simple MPE, called “Business As Usual” and compares them. The second set of results then characterizes the entire SPE correspondence and – relatedly – the best and worst equilibria. Readers are referred to that paper for further results from this model and for a numerical calibration of the model. Furthermore, for the results that are presented, the proofs are merely sketched.

Global Pareto Optima

Let $x = (x_i)$ be a vector of positive numbers, one for each country. A *Global Pareto Optimum* (GPO) corresponding to x is a profile of strategies that maximizes the weighted sum of country payoffs,

$$v = \sum_i x_i v_i, \quad (19)$$

which we shall call *global welfare*. Without loss of generality, we may take the weights, x_i , to sum to I .

Theorem 4 *Let $\hat{V}(g)$ be the maximum attainable global welfare starting with an initial GHG stock equal to g . That*

function is linear in g :

$$\begin{aligned}\hat{V}(g) &= \hat{u} - wg, \\ w &= \frac{1}{1 - \delta\sigma} \sum_i x_i c_i, \\ \hat{u} &= \frac{\sum_i x_i h_i(\hat{a}_i) - \delta w \hat{A}}{1 - \delta}.\end{aligned}\quad (20)$$

The optimal strategy is to pick a constant action – emission – every period and after all histories, \hat{a}_i where its level is determined by

$$x_i h'_i(\hat{a}_i) = \delta w. \quad (21)$$

Proof We shall show by dynamic programming arguments that the Pareto-optimal value function is of the form $\hat{V} = \sum_{i=1}^I x_i [\hat{u}_i - w_i g]$. We need to be able to find the constants \hat{u}_i to satisfy:

$$\begin{aligned}\sum_{i=1}^I x_i [\hat{u}_i - w_i g] &= \max_{a_1, \dots, a_I} \sum_{i=1}^I x_i \\ &\cdot \left[h_i(a_i) - c_i g + \delta (\hat{u}_i - w_i (\sigma g + \sum_{j=1}^I a_j)) \right].\end{aligned}\quad (22)$$

Collecting terms that need maximization we can reduce the equation above to

$$\begin{aligned}\sum_{i=1}^I x_i \hat{u}_i \\ = \max_{a_1, \dots, a_I} \sum_{i=1}^I x_i \left[h_i(a_i) - \delta w_i \sum_{j=1}^I a_j \right] + \delta \sum_{i=1}^I x_i \hat{u}_i.\end{aligned}\quad (23)$$

It is clear that the solution to this system is the same for all g ; call this (first-best) solution \hat{a}_i . Elementary algebra reveals that

$$\hat{u}_i = \frac{h_i(\hat{a}_i) - \delta w_i \sum_{j=1}^I \hat{a}_j}{1 - \delta} \quad \text{and} \quad w_i = \frac{c_i}{1 - \delta\sigma}.$$

It is also obvious that $x_i h'_i(\hat{a}_i) = \delta w$, where $w = \sum_{i=1}^I x_i w_i$. \square

Theorem 4 states that, independently of the level of GHG, g , each country should emit an amount \hat{a}_i . The fact that the optimal emission is constant follows from the linearity of the model in g . Notice that on account of the linearity in the gas buildup equation – Eq. 17 – a unit of emission in period t can be analyzed in isolation as a surviving unit of size σ in period $t + 1$, σ^2 in period $t + 2$, σ^3

in period $t + 2$, and so on. On account of the linearity in cost, these surviving units add $(\sum_i x_i c_i) \times \delta\sigma$ in period $t + 1$, $(\sum_i x_i c_i) \times (\delta\sigma)^2$ in period $t + 2$, and so on, i. e., the marginal lifetime cost is

$$\frac{1}{1 - \delta\sigma} \sum_i x_i c_i,$$

or w , and that marginal cost is independent of g .

A Markov-Perfect Equilibrium: “Business as Usual”

This MPE shares the feature that the equilibrium emission rate of each country is constant in time, and it is the unique MPE with this property. We shall call it the “Business-as-Usual” equilibrium. Note that in this equilibrium each country takes account of the incremental damage to itself caused by an incremental increase in its emission rate, but does not take account of the damage caused to other countries.

Theorem 5 (Business-as-Usual Equilibrium) Let g be the initial stock of GHG. For each country i , let a_i^* be determined by

$$\begin{aligned}h'_i(a_i^*) &= \delta w_i, \\ w_i &= \frac{c_i}{1 - \delta\sigma},\end{aligned}\quad (24)$$

and let its strategy be to use a constant emission equal to a_i^* in each period; then this strategy profile is a MPE, and country i 's corresponding payoff is

$$\begin{aligned}V_i^*(g) &= u_i^* - w_i g, \\ u_i^* &= \frac{h_i(a_i^*) - \delta w_i A^*}{1 - \delta}.\end{aligned}\quad (25)$$

The intuition for the existence of an MPE with constant emissions is similar to the analogous result for the GPO solution. (And indeed for that reason the proof will be omitted.) As long as other countries do not make their emissions contingent on the level of GHGs, country i has a constant marginal lifetime cost to emissions. And that marginal cost is independent of g .

Comparison of the GPO and Business as Usual

The preceding results enable us to compare the emissions in the GPO with those in the Business-as-Usual MPE:

$$\begin{aligned}\text{GPO: } h'_i(\hat{a}_i) &= \frac{\delta \sum_j x_j c_j}{x_i (1 - \delta\sigma)}, \\ \text{BAU: } h'_i(a_i^*) &= \frac{\delta c_i}{1 - \delta\sigma}.\end{aligned}\quad (26)$$

Since

$$x_i c_i < \sum_j x_j c_j ,$$

it follows that

$$\frac{\delta c_i}{1 - \delta \sigma} < \frac{\delta \sum_j x_j c_j}{x_i (1 - \delta \sigma)} .$$

Since h_i is concave, it follows that

$$a_i^* > \hat{a}_i . \quad (27)$$

Note that this inequality holds except in the trivial case in which all welfare weights are zero (except one). This result is known as the tragedy of the commons – whenever there is some externality to emissions, countries tend to over-emit in equilibrium. In turn, all this follows from the fact that in the BAU equilibrium each country only considers its own marginal cost and ignores the cost imposed on other countries on account of its emissions; in the GPO solution that additional cost is, of course, accounted for. It follows that the GPO is strictly Pareto superior to the MPE for an open set of welfare weights x_i (and leads to a strictly lower steady-state GHG level for all welfare weights).

One can contrast these results with those in Dockner, Long and Sorger (1998) that studies a model in which the benefits are linear in emission – i. e., h_i is linear – but convex in costs $c_i(\cdot)$. The consequence of linearity in the benefit function h is that the GPO and BAU solutions have a “most rapid approach” (MRAP) property – if $(1 - \sigma)g$, the depreciated stock in the next period, is less than a most preferred g^* , it is optimal to jump the system to g^* . Else it is optimal to wait for depreciation to bring the stock down to g^* . In other words, linearity in benefits implies a “one-shot” move to a desired level of gas g^* , which is thereafter maintained, while linearity in cost (as in the Dutta and Radner model) implies a constant emission rate. What is unclear in the Dockner, Long and Sorger model is why the multiple players would have the same target steady-state g^* . It would appear natural that, with asymmetric payoffs, each player would have a different steady-state. The existence of a MRAP equilibrium would appear problematical consequently. The authors impose a condition that implies that there is not too much asymmetry.

All SPE

We now turn to the second set of results – a full characterization of SPE in Dutta and Radner [12]. We will show that the SPE payoff correspondence has a surprising simplicity; the set of equilibrium payoffs at a level g is a simple

linear translate of the set of equilibrium payoffs from some benchmark level, say, $g = 0$. Consequently, it will be seen that the set of emission levels that can arise in equilibrium from level g is identical to those that can arise from equilibrium play at a GHG level of 0. Note that the fact that the set of equilibrium possibilities is invariant to the level of g is perfectly consistent with the possibility that, in a particular equilibrium, emission levels vary with g . However, the invariance property will make for a particularly simple characterization of the best and worst equilibria.

Let $\Xi(g)$ denote the set of equilibrium payoff vectors with initial state g , i. e., each element of $\Xi(g)$ is the payoff to some SPE starting from g .

Theorem 6 *The equilibrium payoff correspondence Ξ is linear; there is a compact set $U \subset \mathbb{R}^I$ such that for every initial state g*

$$\Xi(g) = U - \{w_1 g, w_2 g, \dots, w_I g\}$$

where $w_i = c_i/(1 - \sigma\delta)$, $i = 1, \dots, I$. In particular, consider any SPE, any period t and any history of play up until t . Then the payoff vector for the continuation strategies must necessarily be of the form

$$v - (w_1 g_t, w_2 g_t, \dots, w_I g_t) .$$

The theorem is proved by way of a bootstrap argument. We presume that a (candidate) payoff set has this invariance and show that the linear structure of the model confirms the conjecture. Consequently, we generate another candidate payoff set – which is also state-invariant. Then we look for a fixed point of that operator. In other words, we employ the APS operator to generate the SPE correspondence. Since that has already been discussed in the previous section, it is skipped here.

We will now use the above result to characterize the best – and the worst – equilibria in the global climate change game. Consider the *second-best problem* (from initial state g and for a given vector of welfare weights $x = (x_i; i = 1, \dots, I)$), i. e., the problem of maximizing a weighted sum of *equilibrium payoffs*:

$$\max \sum_{i=1}^I x_i V_i(g) , \quad V(g) \in \Xi(g) .$$

Note that we consider all possible equilibria, i. e., we consider equilibria that choose to condition on current and past GHG levels as well as equilibria that do not. The result states that the best equilibrium *need not* condition on GHG levels:

Theorem 7 *There exists a constant emission level $\bar{a} \equiv \bar{a}_1, \bar{a}_2, \dots, \bar{a}_I$ – such that no matter what the initial level of GHG, the second-best policy is to emit at the constant rate \bar{a} . In the event of a deviation from this constant emissions policy by country i , play proceeds to i 's worst equilibrium. Furthermore, the second-best emission rate is always strictly lower than the BAU rate, i. e., $\bar{a} < a^*$. Above a critical discount factor (less than 1), the second-best rate coincides with the GPO emission rate \hat{a} .*

The theorem is attractive for three reasons: first, it says that the best possible equilibrium behavior is no more complicated than BAU behavior; so there is no argument for delaying a treaty (to cut emissions) merely because the status quo is simple. Second, the cut required to implement the second-best policy is an across the board cut – independently of anything else, country i should cut its emissions by the amount $a_i^* - \bar{a}_i$. Third, the second-best is exactly realized at high discount factors, rather than asymptotically approached as the discount factor tends to 1.

Sanctions will be required if countries break with the second-best policy and without loss of generality we can restrict attention to the worst such sanction. We turn now to a characterization of this worst equilibrium (for, say, country i). One definition will be useful for this purpose:

Definition 1 An i -less second-best equilibrium is the solution to a second-best problem in which the welfare weight of i is set equal to zero, i. e., $x_i = 0$.

By the previous theorem, every such problem has a solution in which on the equilibrium path, emissions are a constant. Denote that emission level $a(x_{-i})$:

Theorem 8 *There exists a “high” emission level $\bar{a}(i)$ (with $\sum_{j \neq i} \bar{a}_j(i) > \sum_{j \neq i} a_j^*$) and an i -less second-best equilibrium $a(x_{-i})$ such that country i 's worst equilibrium is:*

1. Each country emits at rate $\bar{a}_j(i)$ for one period (no matter what g is), $j = 1, \dots, I$.
2. From the second period onwards, each country emits at the constant rate $a_j(x_{-i})$, $j = 1, \dots, I$.

And if any country k deviates at either stages 1 or 2, play switches to k 's worst equilibrium from the very next period after the deviation.

Put another way, for every country i , a sanction is made up of two emission rates, $\bar{a}(i)$ and $a(x_{-i})$. The former imposes immediate costs on country i . The way it does so is by increasing the emission levels of countries $j \neq i$. The effect of this is a temporary increase in incremental GHG but due to the irreversibility of gas accumulation, a permanent increase in country i 's costs, enough of an increase to wipe

out any immediate gains that the country might have obtained from the deviation. Of course this additional emission also increases country j 's costs. For the punishing countries, however, this increase is offset by the subsequent permanent change, the switch to the emission vector $a(x_{-i})$, which permanently increases their quota at the expense of country i 's.

Generalizations

The models discussed thus far are base-line models and do not deal with two important issues relating to climate change – technological change and capital accumulation. Technological change is important because that opens access to technologies that do not currently exist, technologies that may have considerably lower “emissions to energy” ratios, i. e., cleaner technologies. Capital accumulation is important because an important question is whether or not curbing GHGs is inimical to growth. The position articulated by both developing countries like Indian and China as well as by developed economies like the United States is that it is: placing curbs on emissions would restrict economic activities and hence restrain the competitiveness of the economy.

In Dutta and Radner [10,13] the following modification was made to the model studied in the previous section. It was presumed that the actual emission level associated with energy usage e_i is $f_i e_i$ where f_i is an index of (un)cleanliness – or *emission factor* – higher values implying larger emissions for the same level of energy usage. It was presumed that the emission factor could be changed at cost but driven no lower than some minimum μ_i . In other words,

$$0 \leq e_i(t), \quad (28)$$

$$\mu_i \leq f_i(t+1) \leq f_i(t). \quad (29)$$

Capital accumulation and population growth is also allowed in the model but taken to be exogenous. The dynamics of those two variables are governed by:

$$g(t) = \sigma g(t-1) + \sum_{i=1}^I f_i(t) e_i(t), \quad (30)$$

$$K_i(t+1) = H[K_i(t)], \quad K_i(t) \nearrow \text{ and unbounded in } t, \quad (31)$$

$$P_i(t+1) = \psi_i P_i(t) + (1 - \psi_i) \Psi, \quad P_i(t) \leq \Psi. \quad (32)$$

The output (gross-domestic product) of country i in period t is

$$h_i[K_i(t), P_i(t), e_i(t)],$$

where the function h_i has all of the standard properties mentioned above. The damage due to the stock of GHG, $g(t)$, is assumed to be (in units of GDP):

$$c_i P_i(t) g(t).$$

The cost of reducing the emission factor from $f_i(t)$ to $f_i(t+1)$ is assumed to be:

$$\varphi_i[f_i(t) - f_i(t+1)].$$

Immediately it is clear that the state variable now encompasses not just the common stock g but, additionally, the emission factor profile as well as the sizes of population and capital stock. In other words, $s = (g, f, K, P)$. Whilst this significant increase in dimensionality might suggest that it would be difficult to obtain clean characterizations, the papers show that there is some separability. The MPE “Business as Usual” has a separable structure – energy usage $e_i(t)$ and emission factor choice $f_i(t+1)$ – depend solely on country i ’s capital stock and population alone. It varies by period – unlike in the base-line model discussed above – as the exogenous variables vary. Furthermore, the emission factor $f_i(t+1)$ stays unchanged till the population and capital stock cross a threshold level beyond which the cleanest technology μ_i gets picked. (This bang-bang character follows from the linearity of the model.)

The Global Pareto Optimal solution has similar features – the energy usage in country i is directly driven by the capital stock and population of that country. Furthermore the emission factor choice follows the same bang-bang character as for the MPE. However, there is a tragedy of the common in that in the MPE (versus the Pareto optimum) the energy usage is higher – at every state – and the switch to the cleanest technology happens later.

Future Directions

Within the general theory of dynamic games there are several open questions and possible directions for future research to take. On the existence question, there needs to be a better resolution of the case where the state space S is uncountably infinite. This is not just a technical curiosity. In applications, typically, in order to apply calculus techniques, we take the state variable to be a subset of some real space. The problem is difficult but one hopes that ancillary assumptions – such as concavity and monotonicity –

will be helpful. These assumptions come “cheaply” because they are routinely invoked in economic applications.

The characterization result via APS techniques has a similar technical difficulty blocking its path, as the existence question. The folk theorem needs to be generalized as well to the S infinite case. Here it is our belief though that the difficulty is not conceptual but rather one where the appropriate result needs to be systematically worked out. As indicated above, the study of the dynamics of SPE paths is in its infancy and much remains to be done here.

Turning to the global climate change application, this is clearly a question of utmost social importance. The subject here is very much in the public consciousness yet academic study especially within economics is only a few years old. Many questions remain: generalizing the models to account for technological change and endogenous capital accumulation, examination of a carbon tax, of cap and trade systems for emission permits, of an international bank that can selectively foster technological change, ... There are – as should be immediately clear – enough interesting important questions to exhaust many dissertations and research projects!

Bibliography

1. Abreu D (1988) On the theory of infinitely repeated games with discounting. *Econometrica* 56:383–396
2. Abreu D, Pearce D, Stachetti E (1990) Towards a general theory of discounted repeated games with discounting. *Econometrica* 58:1041–1065
3. Benhabib J, Radner R (1992) The joint exploitation of a productive asset: A game-theoretic approach. *Econ Theory* 2:155–190
4. Benoit J-P, Krishna V (1987) Finitely repeated games. *Econometrica* 53:905–922
5. Dockner E, Long N, Sorger G (1996) Analysis of Nash equilibria in a class of capital accumulation games. *J Econ Dyn Control* 20:1209–1235
6. Dockner E, Nishimura K (1999) Transboundary boundary problems in a dynamic game model. *Jpn Econ Rev* 50:443–456
7. Duffie D, Geanakoplos J, Mas-Colell A, McLennan A (1994) Stationary Markov equilibria. *Econometrica* 62-4:745–781
8. Dutta P (1991) What do discounted optima converge to? A theory of discount rate asymptotics in economic models. *J Econ Theory* 55:64–94
9. Dutta P (1995) A folk theorem for stochastic games. *JET* 66:1–32
10. Dutta P, Radner R (2004) Self-enforcing climate change treaties. *Proc Nat Acad Sci USA* 101-14:5174–5179
11. Dutta P, Radner R (2006) Population growth and technological change in a global warming model. *Econ Theory* 29:251–270
12. Dutta P, Radner R (2008) A strategic model of global warming model: Theory and some numbers. *J Econ Behav Organ* (forthcoming)
13. Dutta P, Radner R (2008) Choosing cleaner technologies: Global warming and technological change, (in preparation)

14. Dutta P, Sundaram R (1993) How different can strategic models be? *J Econ Theory* 60:42–61
15. Fudenberg D, Maskin E (1986) The Folk theorem in repeated games with discounting or incomplete information. *Econometrica* 54:533–554
16. Harris C, Reny P, Robson A (1995) The existence of subgame perfect equilibrium in continuous games with almost perfect information: A case for extensive-form correlation. *Econometrica* 63:507–544
17. Inter-Governmental Panel on Climate Change (2007) *Climate Change, the Synthesis Report*. IPCC, Geneva
18. Levhari D, Mirman L (1980) The great fish war: An example using a dynamic cournot-Nash solution. *Bell J Econ* 11:322–334
19. Long N, Sorger G (2006) Insecure property rights and growth: The role of appropriation costs, wealth effects and heterogeneity. *Econ Theory* 28:513–529
20. Mertens J-F, Parthasarathy T (1987) *Equilibria for Discounted Stochastic Games*. Research Paper 8750, CORE. University Catholique de Louvain
21. Mertens, Neyman (1983)
22. Nowak A (1985) Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. *J Optim Theory Appl* 45:591–603
23. Parthasarathy T (1973) Discounted, positive and non-cooperative stochastic games. *Int J Game Theory* 2–1:
24. Rieder U (1979) Equilibrium plans for non-zero sum Markov games. In: Moeschlin O, Pallasche D (ed) *Game theory and related topics*. North-Holland, Amsterdam
25. Rustichini A (1992) Second-best equilibria for games of joint exploitation of a productive asset. *Econ Theory* 2:191–196
26. Shapley L (1953) *Stochastic Games*. In: *Proceedings of National Academy of Sciences*, Jan 1953
27. Sobel M (1990) Myopic solutions of affine dynamic models. *Oper Res* 38:847–53
28. Sorger G (1998) Markov-perfect Nash equilibria in a class of resource games. *Econ Theory* 11:79–100
29. Stern N (2006) *Review on the economics of climate change*. HM Treasury, London. www.sternreview.org.uk
30. Stern Review on the Economics of Climate Change, October, (2006)
31. Sundaram R (1989) Perfect equilibrium in a class of symmetric dynamic games. *J Econ Theory* 47:153–177

Dynamics of Cellular Automata in Non-compact Spaces

ENRICO FORMENTI¹, PETR KŮRKA²

¹ Laboratoire I3S – UNSA/CNRS UMR 6070, Université de Nice Sophia Antipolis, Sophia Antipolis, France

² Center for Theoretical Study, Academy of Sciences and Charles University, Prague, Czech Republic

Article Outline

Glossary

Definition of the Subject

Introduction

Dynamical Systems
Cellular Automata
Submeasures
The Cantor Space
The Periodic Space
The Toeplitz Space
The Besicovitch Space
The Generic Space
The Space of Measures
The Weyl Space
Examples
Future Directions
Acknowledgments
Bibliography

Glossary

Almost equicontinuous CA A CA which has at least one equicontinuous configuration.

Attraction basin The set of configurations whose orbit is eventually attracted by an attractor.

Attractor A closed invariant set which attracts all orbits in some of its neighborhood.

Besicovitch pseudometrics A pseudometric that quantifies the upper-density of differences.

Blocking word A word that interrupts the information flow. A configuration containing an infinite number of blocking words both to the right and to the left is equicontinuous in the Cantor topology.

Equicontinuous CA A CA in which all configurations are equicontinuous.

Equicontinuous configuration A configuration for which nearby configurations remain close.

Expansive CA Two distinct configurations, no matter how close, eventually separate during the evolution.

Generic space The space of configurations for which upper-density and lower-density coincide.

Sensitive CA In any neighborhood of any configuration there exists a configuration such that the orbits of the two configurations eventually separate.

Spreading set A clopen invariant set propagating both to the left and to the right.

Toeplitz space The space of regular quasi-periodic configurations.

Weyl pseudometrics A pseudometric that quantifies the upper density of differences with respect to all possible cell indices.

Definition of the Subject

In topological dynamics, the assumption of compactness is usually adopted as it has far reaching consequences.

Each compact dynamical system has an almost periodic point, contains a minimal subsystem, and each trajectory has a limit point. Nevertheless, there are important examples of non-compact dynamical systems like linear systems on \mathbb{R}^n and the theory should cover these examples as well. The study of dynamics of cellular automata (CA) in the compact Cantor space of symbolic sequences starts with Hedlund [6] and is by now a firmly established discipline (see e. g., ► [Topological Dynamics of Cellular Automata](#)). The study of dynamics of CA in non-compact spaces like Besicovitch or Weyl spaces is more recent and provides an interesting alternative perspective.

The study of dynamics of cellular automata in non-compact spaces has at least two distinct origins. The first concerns the study of dynamical properties on peculiar countable dense sub-spaces of the Cantor space (the space of finite configuration or the space of spatially periodic configurations, for instance). The idea is that on those spaces, some properties are easier to prove than on the full Cantor space. Once a property is proved on such a sub-space, one can try to lift it to the original Cantor space by using denseness. Another advantage is that the configurations on these spaces are easily representable on computers. Indeed, computer simulations and practical applications of CA usually take place in these subspaces.

The second origin is connected to the question of suitability of the classical Cantor topology for the study of chaotic behavior of CA and of symbolic systems in general. We briefly recall the motivations. Consider sensitivity to initial conditions for a CA in the Cantor topology. The shift map σ , which is a very simple CA, is sensitive to initial conditions since small perturbations far from the central region are eventually brought to the central part. However, from an algorithmic point of view, the shift map is very simple. We are inclined to regard a system as chaotic if its behavior cannot easily be reconstructed. This is not the case of the shift map whose chaoticity is more an artifact of the Cantor metric, rather than an intrinsic property of the system. Therefore, one may want to define another metric in which sensitive CA not only transport information (like the shift map) but also build/destroy new information at each time step.

This basic requirement stimulated the quest for alternative topologies to the classical Cantor space. This led first to the Besicovitch topology and then to the Weyl topology in Cattaneo et al. [4] used to investigate almost periodic real functions (see [1,8]). Both these pseudometrics can be defined starting from suitable semi-measures on the set \mathbb{Z} of integers. This way of construction had a *Pandora effect* opening the way to many new interesting topological spaces. Some of them are reported in this pa-

per; others can be found in Cervelle and Formenti (► [Algorithmic Complexity and Cellular Automata](#)).

Each topology focuses on some peculiar aspects of the dynamics under study but all of them have a common denominator, namely non-compactness.

Introduction

A given CA over an alphabet A can be regarded as a dynamical system in several topological spaces: Cantor configuration space C_A , the space \mathcal{M}_A of shift-invariant Borel probability measures on $A^{\mathbb{Z}}$, the Weyl space \mathcal{W}_A , the Besicovitch space \mathcal{B}_A , the generic space \mathcal{G}_A , the Toeplitz space \mathcal{T}_A and the periodic space \mathcal{P}_A . We refer to various topological properties of these systems by prefixing the name of the space in question. Basic results correlate various dynamical properties of CA in these spaces.

The Cantor topology corresponds to the point of view of an observer who can distinguish only a finite central part of a configuration and sites outside this central part of the configuration are not taken into account. The Besicovitch and Weyl topologies, on the other hand, correspond to a god-like position of someone who sees whole configurations and can distinguish the frequency of differences. In the Besicovitch topology, the centers of configurations still play a distinguished role, as the frequencies of differences are computed from the center. In the Weyl topology, on the other hand, no site has a privileged position. Both Besicovitch and Weyl topologies are defined by pseudometrics. Different configurations can have zero distance and the topological space consists of equivalence classes of configurations which have zero distance.

The generic space \mathcal{G}_A is a subspace of the Besicovitch space of those configurations, in which each finite word has a well defined frequency. These frequencies define a Borel probability measure on the Cantor space of configurations, so we have a projection from the generic space \mathcal{G}_A to the space \mathcal{M}_A of Borel probability measures equipped with the weak* topology. This is a natural space for investigating the dynamics of CA on random configurations.

The Toeplitz space \mathcal{T}_A consists of regular quasi-periodic configurations. This means that each pattern repeats periodically but different patterns have different periods. The Besicovitch and Weyl pseudometrics are actually metrics on the Toeplitz space and moreover they coincide on \mathcal{T}_A .

Dynamical Systems

A **dynamical system** is a continuous map $F: X \rightarrow X$ of a nonempty metric space X to itself. The n th it-

eration $F^n: X \rightarrow X$ of F is defined by $F^0(x) = x$, $F^{n+1}(x) = F(F^n(x))$. A point $x \in X$ is **fixed**, if $F(x) = x$. It is **periodic**, if $F^n(x) = x$ for some $n > 0$. The least positive n with this property is called the **period** of x . The **orbit** of x is the set $\mathcal{O}(x) = \{F^n(x): n > 0\}$. A set $Y \subseteq X$ is **positively invariant**, if $F(Y) \subseteq Y$ and **strongly invariant** if $F(Y) = Y$. A point $x \in X$ is **equicontinuous** ($x \in \mathcal{E}_F$) if the family of maps F^n is equicontinuous at x , i.e. $x \in \mathcal{E}_F$ iff

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall y \in B_\delta(x))(\forall n > 0) \\ (d(F^n(y), F^n(x)) < \varepsilon).$$

The system (X, F) is **almost equicontinuous** if $\mathcal{E}_F \neq \emptyset$ and **equicontinuous**, if

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in X)(\forall y \in B_\delta(x)) \\ (\forall n > 0)(d(F^n(y), F^n(x)) < \varepsilon).$$

For an equicontinuous system $\mathcal{E}_F = X$. Conversely, if $\mathcal{E}_F = X$ and if X is compact, then F is equicontinuous; this needs not be true in the non-compact case. A system (X, F) is **sensitive** (to initial conditions), if

$$(\exists \varepsilon > 0)(\forall x \in X)(\forall \delta > 0)(\exists y \in B_\delta(x)) \\ (\exists n > 0)(d(f^n(y), f^n(x)) \geq \varepsilon).$$

A sensitive system has no equicontinuous point. However, there exist systems with no equicontinuity points which are not sensitive. A system (X, F) is **positively expansive**, if

$$(\exists \varepsilon > 0)(\forall x \neq y \in X)(\exists n \geq 0)(d(f^n(x), f^n(y)) \geq \varepsilon).$$

A positively expansive system on a perfect space is sensitive. A system (X, F) is (topologically) **transitive**, if for any nonempty open sets $U, V \subseteq X$ there exists $n \geq 0$ such that $F^{-n}(U) \cap V \neq \emptyset$. If X is perfect and if the system has a dense orbit, then it is transitive. Conversely, if (X, F) is topologically transitive and if X is compact, then (X, F) has a dense orbit. A system (X, F) is **mixing**, if for any nonempty open sets $U, V \subseteq X$ there exists $k > 0$ such that for every $n \geq k$ we have $F^{-n}(U) \cap V \neq \emptyset$. An ε -chain (from x_0 to x_n) is a sequence of points $x_0, \dots, x_n \in X$ such that $d(F(x_i), x_{i+1}) < \varepsilon$ for $0 \leq i < n$. A system (X, F) is **chain-transitive**, if for any $\varepsilon > 0$ and any $x, y \in X$ there exists an ε -chain from x to y .

A strongly invariant closed set $Y \subseteq X$ is **stable**, if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x \in X, (d(x, Y) < \delta) \\ \implies \forall n > 0, d(F^n(x), Y) < \varepsilon).$$

A strongly invariant closed stable set $Y \subseteq X$ is an **attractor**, if

$$\exists \delta > 0, \forall x \in X, (d(x, Y) < \delta \implies \lim_{n \rightarrow \infty} d(F^n(x), Y) = 0).$$

A set $W \subseteq X$ is **inward**, if $F(\overline{W}) \subseteq W^\circ$. In compact spaces, attractors are exactly Ω -limits $\Omega_F(W) = \bigcap_{n>0} F^n(W)$ of inward sets.

Theorem 1 (Knudsen [10]) *Let (X, F) be a dynamical system and $Y \subseteq X$ a dense, F -invariant subset.*

- (1) (X, F) is sensitive iff (Y, F) is sensitive.
- (2) (X, F) is transitive iff (Y, F) is transitive.

Recall that a space X is **separable**, if it has a countable dense set.

Theorem 2 (Blanchard, Formenti, and Kůrka [2]) *Let (X, F) be a dynamical system on a non-separable space. If (X, F) is transitive, then it is sensitive.*

Cellular Automata

For a finite alphabet A , denote by $|A|$ the number of its elements, by $A^* := \bigcup_{n \geq 0} A^n$ the set of words over A , and by $A^+ := \bigcup_{n > 0} A^n = A^* \setminus \{\lambda\}$ the set of nonempty words. The length of a word $u \in A^n$ is denoted by $|u| := n$. We say that $u \in A^*$ is a subword of $v \in A^*$ ($u \sqsubseteq v$) if there exists k such that $v_{k+i} = u_i$ for all $i < |u|$. We denote by $u_{[i,j]} = u_i \dots u_{j-1}$ and $u_{[i,j]} = u_i \dots u_j$ subwords of u associated to intervals. We denote by $A^{\mathbb{Z}}$ the set of A -**configurations**, or doubly-infinite sequences of letters of A . For any $u \in A^+$ we have a periodic configuration $u^\infty \in A^{\mathbb{Z}}$ defined by $(u^\infty)_{k|u|+i} = u_i$ for $k \in \mathbb{Z}$ and $0 \leq i < |u|$. The **cylinder** of a word $u \in A$ located at $l \in \mathbb{Z}$ is the set $[u]_l = \{x \in A^{\mathbb{Z}}: x_{[l, l+|u|)} = u\}$. The cylinder set of a set of words $U \subseteq A^+$ located at $l \in \mathbb{Z}$ is the set $[U]_l = \bigcup_{u \in U} [u]_l$.

A **subshift** is a nonempty subset $\Sigma \subseteq A^{\mathbb{Z}}$ such that there exists a set $D \subseteq A^+$ of **forbidden words** and $\Sigma = \Sigma_D := \{x \in A^{\mathbb{Z}}: \forall u \sqsubseteq x, u \notin D\}$. A subshift Σ_D is of **finite type** (SFT), if D is finite. A subshift is uniquely determined by its **language**

$$\mathcal{L}(\Sigma) := \bigcup_{n \geq 0} \mathcal{L}^n(\Sigma),$$

$$\text{where } \mathcal{L}^n(\Sigma) := \{u \in A^n: \exists x \in \Sigma, u \sqsubseteq x\}.$$

A **cellular automaton** is a map $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ defined by $F(x)_i = f(x_{[i-r, i+r]})$, where $r \geq 0$ is a radius and $f: A^{2r+1} \rightarrow A$ is a local rule. In particular the **shift map**

$\sigma: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ is defined by $\sigma(x)_i := x_{i+1}$. A local rule extends to the map $f: A^* \rightarrow A^*$ by $f(u)_i = f(u_{[i, i+2r]})$ so that $|f(u)| = \max\{|u| - 2r, 0\}$.

Definition 3 Let $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ be a CA.

- (1) A word $u \in A$ is m -blocking, if $|u| \geq m$ and there exists offset $d \leq |u| - m$ such that $\forall x, y \in [u]_0, \forall n > 0, F^n(x)_{[d, d+m]} = F^n(y)_{[d, d+m]}$.
- (2) A set $U \subseteq A^+$ is spreading, if $[U]$ is F -invariant and there exists $n > 0$ such that $F^n([U]) \subseteq \sigma^{-1}([U]) \cap \sigma([U])$.

The following results will be useful in the sequel.

Proposition 4 (Formenti, Kůrka [5]) Let $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ be a CA and let $U \subseteq A^+$ be an invariant set. Then $\Omega_F([U])$ is a subshift iff U is spreading.

Theorem 5 (Hedlund [6]) Let $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ be a CA with local rule $f: A^{2r+1} \rightarrow A$. Then F is surjective iff $f: A^* \rightarrow A^*$ is surjective iff $|f^{-1}(u)| = |A|^{2r}$ for each $u \in A^+$.

Submeasures

A pseudometric on a set X is a map $d: X \times X \rightarrow [0, \infty)$ which satisfies the following conditions:

1. $d(x, y) = d(y, x)$,
2. $d(x, z) \leq d(x, y) + d(y, z)$.

If moreover $d(x, y) > 0$ for $x \neq y$, then we say that d is a metric. There is a standard method to create pseudometrics from submeasures. A bounded submeasure (with bound $M \in \mathbb{R}^+$) is a map $\varphi: \mathcal{P}(\mathbb{Z}) \rightarrow [0, M]$ which satisfies the following conditions:

1. $\varphi(\emptyset) = 0$,
2. $\varphi(U) \leq \varphi(U \cup V) \leq \varphi(U) + \varphi(V)$ for $U, V \subseteq \mathbb{Z}$.

A bounded submeasure φ on \mathbb{Z} defines a pseudometric $d_\varphi: A^{\mathbb{Z}} \times A^{\mathbb{Z}} \rightarrow [0, \infty)$ by $d_\varphi(x, y) := \varphi(\{i \in \mathbb{Z}: x_i \neq y_i\})$. The Cantor, Besicovich and Weyl pseudometrics on $A^{\mathbb{Z}}$ are defined by the following submeasures:

$$\begin{aligned} \varphi_C(U) &:= 2^{-\min\{|i|: i \in U\}}, \\ \varphi_B(U) &:= \limsup_{l \rightarrow \infty} \frac{|U \cap [-l, l]|}{2l}, \\ \varphi_W(U) &:= \limsup_{l \rightarrow \infty} \sup_{k \in \mathbb{Z}} \frac{|U \cap [k, k+l]|}{l}. \end{aligned}$$

The Cantor Space

The Cantor metric on $A^{\mathbb{Z}}$ is defined by

$$d_C(x, y) = 2^{-k} \quad \text{where} \quad k = \min\{|i|: x_i \neq y_i\},$$

so $d_C(x, y) < 2^{-k}$ iff $x_{[-k, k]} = y_{[-k, k]}$. We denote by $C_A = (A^{\mathbb{Z}}, d_C)$ the metric space of two-sided configurations with metric d_C . The cylinders are clopen sets in C_A . All Cantor spaces (with different alphabets) are homeomorphic. The Cantor space is compact, totally disconnected and perfect, and conversely, every space with these properties is homeomorphic to a Cantor space. Literature about CA dynamics in Cantor spaces is really huge. In this section, we just recall some results and definitions which will be used later.

Theorem 6 (Kůrka [11]) Let (C_A, F) be a CA with radius r .

- (1) (C_A, F) is almost equicontinuous iff there exists a r -blocking word for F
- (2) (C_A, F) is equicontinuous iff all sufficiently long words are r -blocking.

Denote by \mathcal{E}_F the set of equicontinuous points of F . The sets of **equicontinuous directions** and **almost equicontinuous directions** of a CA (C_A, F) (see Sablik [15]) are defined by

$$\begin{aligned} \mathcal{E}(F) &= \left\{ \frac{p}{q}: p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} = A^{\mathbb{Z}} \right\}, \\ \mathfrak{A}(F) &= \left\{ \frac{p}{q}: p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} \neq \emptyset \right\}. \end{aligned}$$

The Periodic Space

Definition 7 The **periodic space** $\mathcal{P}_A = \{x \in A^{\mathbb{Z}}: \exists n > 0, \sigma^n(x) = x\}$ over an alphabet A consists of shift periodic configurations with Cantor metric d_C .

All periodic spaces (with different alphabets) are homeomorphic. The periodic space is not compact, but it is totally disconnected and perfect. It is dense in C_A . If (C_A, F) is a CA, Then $F(\mathcal{P}_A) \subseteq \mathcal{P}_A$. We denote by $F_P: \mathcal{P}_A \rightarrow \mathcal{P}_A$ the restriction of F to \mathcal{P}_A , so (\mathcal{P}_A, F_P) is a (non-compact) dynamical system. Every F_P -orbit is finite, so every point $x \in \mathcal{P}_A$ is F_P -eventually periodic.

Theorem 8 Let F be a CA over alphabet A .

- (1) (C_A, F) is surjective iff (\mathcal{P}_A, F_P) is surjective.
- (2) (C_A, F) is equicontinuous iff (\mathcal{P}_A, F_P) is equicontinuous.

- (3) (C_A, F) is almost equicontinuous iff (P_A, F_P) is almost equicontinuous.
 (4) (C_A, F) is sensitive iff (P_A, F_P) is sensitive.
 (5) (C_A, F) is transitive iff (P_A, F_P) is transitive.

Proof (1a) Let F be surjective, let $y \in P_A$ and $\sigma^n(y) = y$. There exists $z \in F^{-1}(y)$ and integers $i < j$ such that $z_{[inr, inr+r)} = z_{[jnr, jnr+r)}$. Then $x = (z_{[inr, jnr)})^\infty \in P_A$ and $F_P(x) = y$, so F_P is surjective.

(1b) Let F_P be surjective, and $u \in A^+$. Then u^∞ has F_P -preimage and therefore u has preimage under the local rule. By Hedlund Theorem, (C_A, F) is surjective.

(2a) Since $P_A \subset C_A$, the equicontinuity of F implies trivially the equicontinuity of F_P .

(2b) Let F_P be equicontinuous. There exist $m > r$ such that if $x, y \in P_A$ and $x_{[-m, m]} = y_{[-m, m]}$, then $F^n(x)_{[-r, r]} = F^n(y)_{[-r, r]}$ for all $n \geq 0$. We claim that all words of length $2m+1$ are $(2r+1)$ -blocking with offset $m-r$. If not, then for some $x, y \in A^\mathbb{Z}$ with $x_{[-m, m]} = y_{[-m, m]}$, there exists $n > 0$ such that $F^n(x)_{[-r, r]} \neq F^n(y)_{[-r, r]}$. For periodic configurations $x' = (x_{[-m-nr, m+nr]})^\infty, y' = (y_{[-m-nr, m+nr]})^\infty$ we get $F^n(x')_{[-r, r]} \neq F^n(y')_{[-r, r]}$ contradicting the assumption. By Theorem 6, F is C -equicontinuous.

(3a) If (C_A, F) is almost equicontinuous, then there exists a r -blocking word u and $u^\infty \in P_A$ is an equicontinuous configuration for (P_A, F_P) .

(3b) The proof is analogous as (2b).

(4) and (5) follow from the Theorem 1 of Knudsen. \square

The Toeplitz Space

Definition 9 Let A be an alphabet

- (1) The **Besicovitch pseudometric** on $A^\mathbb{Z}$ is defined by

$$d_B(x, y) = \limsup_{l \rightarrow \infty} \frac{|\{j \in [-l, l]: x_j \neq y_j\}|}{2l}.$$

- (2) The **Weyl pseudometric** on $A^\mathbb{Z}$ is defined by

$$d_W(x, y) = \limsup_{l \rightarrow \infty} \max_{k \in \mathbb{Z}} \frac{|\{j \in [k, k+l): x_j \neq y_j\}|}{l}.$$

Clearly $d_B(x, y) \leq d_W(x, y)$ and

$$\begin{aligned} d_B(x, y) < \varepsilon &\iff \exists l_0 \in \mathbb{N}, \forall l \geq l_0, \\ &\quad |\{j \in [-l, l]: x_j \neq y_j\}| < (2l+1)\varepsilon, \\ d_W(x, y) < \varepsilon &\iff \exists l_0 \in \mathbb{N}, \forall l \geq l_0, \forall k \in \mathbb{Z}, \\ &\quad |\{j \in [k, k+l): x_j \neq y_j\}| < l\varepsilon. \end{aligned}$$

Both d_B and d_W are symmetric and satisfy the triangle inequality, but they are not metrics. Distinct configurations $x, y \in A^\mathbb{Z}$ can have zero distance. We construct a set of **regular quasi-periodic** configurations, on which d_B and d_W coincide and are metrics.

Definition 10

- (1) The **period** of $k \in \mathbb{Z}$ in $x \in A^\mathbb{Z}$ is $r_k(x) := \inf\{p > 0: \forall n \in \mathbb{Z}, x_{k+np} = x_k\}$. We set $r_k(x) = \infty$ if the defining set is empty.
 (2) $x \in A^\mathbb{Z}$ is **quasi-periodic**, if $r_k(x) < \infty$ for all $k \in \mathbb{Z}$.
 (3) A **periodic structure** for a quasi-periodic configuration x is a sequence of positive integers $\mathfrak{p} = (p_i)_{i \in \mathbb{Z}}$, such that $p_i | p_{i+1}$ (p_i divides p_{i+1}), and for every $k \in \mathbb{Z}$, $r_k(x) | p_i$ for some i .
 (4) A quasi-periodic configuration $x \in A^\mathbb{Z}$ is **regular**, if for some periodic structure \mathfrak{p} of x we have $\lim_{i \rightarrow \infty} q_i(x)/p_i = 0$, where $q_i(x) := |\{k \in [0, p_i): r_k(x) \nmid p_i\}|$ ($r_k(x)$ does not divide p_i).

Clearly every σ -periodic configuration is quasi-periodic and has a finite periodic structure.

Proposition 11

- (1) If x, y are regular quasi-periodic configurations, then $d_W(x, y) = d_B(x, y)$.
 (2) If $x \neq y$ are quasi-periodic configurations, then $d_W(x, y) \geq d_B(x, y) > 0$.

Proof (1) We must show $d_W(x, y) \leq d_B(x, y)$. Let $\mathfrak{p}^x, \mathfrak{p}^y$ be the periodic structures for x and y , and let $p_i = k_i^x p_i^x = k_i^y p_i^y$ be the lowest common multiple of p_i^x and p_i^y . Then $\mathfrak{p} = (p_i)_i$ is a periodic structure for both x and y . For each $i > 0$ and for each $k \in \mathbb{Z}$ we have

$$\begin{aligned} &|\{j \in [k - p_i, k + p_i): x_j \neq y_j\}| \\ &\leq 2k_i^x q_i^x + 2k_i^y q_i^y + |\{j \in [-p_i, p_i): x_j \neq y_j\}|, \end{aligned}$$

$$\begin{aligned} d_W(x, y) &\leq \lim_{i \rightarrow \infty} \max_{k \in \mathbb{Z}} |\{j \in [k - p_i, k + p_i): x_j \neq y_j\}| \\ &\leq \lim_{i \rightarrow \infty} \left(\frac{2k_i^x q_i^x}{2k_i^x p_i^x} + \frac{2k_i^y q_i^y}{2k_i^y p_i^y} \right. \\ &\quad \left. + \frac{|\{j \in [-p_i, p_i): x_j \neq y_j\}|}{2p_i} \right) \\ &= d_B(x, y). \end{aligned}$$

- (2) Since $x \neq y$, there exists i such that for some $k \in [0, p_i)$ and for all $n \in \mathbb{Z}$ we have $x_{k+np_i} = x_k \neq y_k = y_{k+np_i}$. It follows $d_B(x, y) \geq 1/p_i$. \square

Definition 12 The **Toeplitz space** \mathcal{T}_A over A consists of all regular quasi-periodic configurations with metric $d_B = d_W$.

Toeplitz sequences are constructed by filling in periodic parts successively. For an alphabet A put $\tilde{A} = A \cup \{*\}$.

Definition 13

- (1) The **p-skeleton** $S_p(x) \in \tilde{A}^{\mathbb{Z}}$ of $x \in A^{\mathbb{Z}}$ is defined by

$$S_p(x)_k = \begin{cases} x_k & \text{if } \forall n \in \mathbb{Z}, x_{k+np} = x_k \\ * & \text{otherwise.} \end{cases}$$

- (2) The **sequence of gaps** of $x \in \tilde{A}^{\mathbb{Z}}$ is the unique increasing integer sequence $(t_i)_{a < i < b}$ such that $x_{t_i} = *, x_k \neq *$ for $t_i < k < t_{i+1}$ and $t_{-1} < 0 \leq t_0$.
- (3) Let $x, y \in \tilde{A}^{\mathbb{Z}}$ and let (t_i) be the sequence of gaps of x . The **amalgamation** $T(x, y) \in \tilde{A}^{\mathbb{Z}}$ of x, y is

$$T(x, y)_i = \begin{cases} x_i & \text{if } x_i \neq * \\ y_j & \text{if } x_i = * \text{ \& } i = t_j. \end{cases}$$

The p -skeleton is p -periodic. If p is its smallest period, we say that p is an **essential** period of x . The sequence of gaps may be two-way infinite (then $a = -\infty, b = \infty$), one-way infinite ($a = -\infty, b < \infty$ or $-\infty < a, b < \infty$), finite ($-\infty < a < b < \infty$) or even empty when $x \in A^{\mathbb{Z}}$. If it is nonempty then it must be defined at least on -1 or 0 .

Proposition 14 Let $2 := \{0, 1\}$ be the binary alphabet and $[0, 1]$ the real unit interval (with standard metric). There exists an isometric embedding $f: [0, 1] \rightarrow \mathcal{T}_2$ such that $f(0) = 0^\infty$ and $f(1) = 1^\infty$.

Proof Consider a map $h: 2^* \rightarrow \tilde{2}^{\mathbb{Z}}$ defined by $h(\lambda) = *^\infty, h(0) = (0*)^\infty, h(1) = (*1)^\infty, h(x_0 \dots x_{n-1} x_n) = T(h(x_0 \dots x_{n-1}), h(x_n))$. Thus

$$\begin{aligned} h(0) &= 0*0*0*0*0*0*0*0*0*0*0 \dots \\ h(1) &= *1*1*1*1*1*1*1*1*1*1*1 \dots \\ h(01) &= 0*010*010*010*010*010 \dots \\ h(10) &= 01*101*101*101*101*101 \dots \\ h(011) &= 0*0101010*0101010 \dots \\ h(100) &= 010101*1010101 \dots \\ h(0111) &= 0*010101010101010 \dots \\ h(1000) &= 01010101010101 \dots \end{aligned}$$

For $x \in 2^{\mathbb{N}}$, the limit (in the Cantor topology) $h(x) = \lim_{n \rightarrow \infty} h(x_0 \dots x_n)$ exists. If no suffix of x is 1^∞ , then $h(x) \in 2^{\mathbb{Z}}$, otherwise $h(x)$ contains exactly one star and we replace it by 1 . Thus for each $x \in 2^{\mathbb{N}}$, $h(x) \in 2^{\mathbb{Z}}$ is a regular quasi-periodic sequence. It can be verified directly that $h(0111 \dots) = (01)^\infty = h(1000 \dots)$. Using an

easily verifiable formula $T(x, T(y, z)) = T(T(x, y), z)$, we get

$$\begin{aligned} h(x_0 \dots x_n 01^\infty) &= T(h(x_0 \dots x_n), h(01^\infty)) \\ &= T(h(x_0 \dots x_n), h(10^\infty)) \\ &= h(x_0 \dots x_n 10^\infty). \end{aligned}$$

For a real number $x \in [0, 1]$ with binary expansion $x = \sum_{i=0}^{\infty} x_i 2^{-i-1}$ put $f(x) = h(x_0 x_1 x_2 x_3 \dots)$. If $2^n x$ is an integer for some n , then x has two binary expansions, and $f(x)$ is the same for both expansions. If $|x - y| \leq 2^{-m}$, then $x_{[0, m)} = y_{[0, m)}$; therefore $d_W(f(x), f(y)) \leq 2^{-m}$ and $f: [0, 1] \rightarrow \mathcal{T}_A$ is continuous. For dyadic numbers x, y of the form $k/2^m$ we verify $d_B(f(x), f(y)) = |x - y|$, so f is an isometry. \square

Proposition 15 The Toeplitz space \mathcal{T}_A of regular quasi-periodic sequences is pathwise connected and infinite-dimensional.

Proof Assume that the alphabet A contains letters $0, 1$ and consider the map $f: [0, 1] \rightarrow A^{\mathbb{Z}}$ from Proposition 14. For $a \in A$ set $a \cdot 0 = 0, a \cdot 1 = a$. Given $u \in 2^{\mathbb{Z}}$ the map $g_u: [0, 1] \rightarrow 2^{\mathbb{Z}}$ defined by $g_u(x)_i = u_i f(x)_i$ is continuous, $g_u(0) = 0$ and $g_u(1) = u$. Thus \mathcal{T}_A is pathwise connected. To show that \mathcal{T}_A is infinite-dimensional, construct for any n an embedding $f_n: [0, 1]^n \rightarrow X_W$ of an n -dimensional cube by

$$f_n(x_1, \dots, x_n) = f(x_1)_0 \dots f(x_n)_0 f(x_1)_1 \dots f(x_n)_1 \dots$$

Thus \mathcal{T}_A is at least n -dimensional and therefore infinite-dimensional. \square

Proposition 16 Let $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ be a CA with radius r .

- (1) If $x \in A^{\mathbb{Z}}$ is a quasi-periodic with periodic structure p , then $F(x)$ is quasi-periodic with periodic structure p .
- (2) If x is regular quasi-periodic, then $F(x)$ is regular.

Proof (1) For $k \in \mathbb{Z}$ denote by $m := \min\{i: \forall j \in [k - r, k + r], r_j(x) | p_i\}$. Then p_m is a period of $F(x)_k$.

(2) We have

$$\begin{aligned} q_i(F(x)) &:= |\{k < p_i: r_k(F(x)) \not\mid p_i\}| \\ &\leq (2r + 1) \cdot |\{k < p_i: r_k(x) \not\mid p_i\}| \\ &= (2r + 1) \cdot q_i(x), \end{aligned}$$

so

$$\lim_{i \rightarrow \infty} \frac{q_i(F(x))}{p_i} \leq (2r + 1) \cdot \lim_{i \rightarrow \infty} \frac{q_i(x)}{p_i} = 0.$$

\square

For a CA F we denote by $F_{\mathcal{T}}$ the restriction of F to \mathcal{T}_A .

Theorem 17 Let F be a CA.

- (1) (C_A, F) is surjective iff $(\mathcal{T}_A, F_{\mathcal{T}})$ is surjective.
- (2) If $\mathfrak{A}(F) \neq \emptyset$, then $(\mathcal{T}_A, F_{\mathcal{T}})$ is almost equicontinuous.
- (3) if $\mathfrak{C}(F) \neq \emptyset$, then $(\mathcal{T}_A, F_{\mathcal{T}})$ is equicontinuous.
- (4) If (C_A, F) is chain-transitive, then $(\mathcal{T}_A, F_{\mathcal{T}})$ is chain-transitive.
- (5) $(\mathcal{T}_A, F_{\mathcal{T}})$ is injective iff it is surjective.

Proof (1) The proof is the same as in Theorem 8(1).

(2) Assume first that F is almost equicontinuous, so there exists $m > r$ and $u \in A^{2m+1}$ such that for any $x, y \in [u]_{-m}$, $F^n(x)_{[-r, r]} = F^n(y)_{[-r, r]}$ for all $n > 0$. We show that u^∞ is \mathcal{T} -equicontinuous. For a given $\varepsilon > 0$ set $\delta = \varepsilon/(4m - 2r + 1)$. If $d_{\mathcal{T}}(y, x) < \delta$, then there exists l_0 such that for all $l \geq l_0$, $|\{i \in [-l, l] : x_i \neq y_i\}| < (2l + 1)\delta$. For $k(2m + 1) \leq j < (k + 1)(2m + 1)$, $F^n(y)_j$ can differ from $F^n(x)_j$ only if y differs from x in some $i \in [k(2m + 1) - (m - r), (k + 1)m + (m - r))$. Thus a change $x_i \neq y_i$ can cause at most $2m + 1 + 2(m - r) = 4m - 2r + 1$ changes $F^n(y)_j \neq F^n(x)_j$. We get

$$|\{i \in [-l, l] : F^n(x)_i \neq F^n(y)_i\}| \leq 2l\delta(4m - 2r + 1) \leq 2l\varepsilon.$$

This shows that $F_{\mathcal{T}}$ is almost equicontinuous. In the general case that $\mathfrak{A}(F) \neq \emptyset$, we get that $F_{\mathcal{T}}^q \sigma^p$ is almost equicontinuous for some $p \in \mathbb{Z}$, $q \in \mathbb{N}^+$. Since σ is \mathcal{T} -equicontinuous, $F_{\mathcal{T}}^q$ is almost equicontinuous and therefore $(\mathcal{T}_A, F_{\mathcal{T}})$ is almost equicontinuous.

(3) The proof is the same as in (2) with the only modification that all $u \in A^m$ are $(2r + 1)$ -blocking.

(4) The proof of Proposition 8 from [2] works in this case too.

(5) The proof of Proposition 12 of [3] works in this case also. \square

The Besicovitch Space

On $A^{\mathbb{Z}}$ we have an equivalence $x \approx_B y$ iff $d_B(x, y) = 0$. Denote by \mathcal{B}_A the set of equivalence classes of \approx_B and by $\pi_B : A^{\mathbb{Z}} \rightarrow \mathcal{B}_A$ the projection. The factor of d_B is a metric on \mathcal{B}_A . This is the **Besicovitch space** on alphabet A . Using prefix codes, it can be shown that every two Besicovitch spaces (with different alphabets) are homeomorphic. By Proposition 11 each equivalence class contains at most one quasi-periodic sequence.

Proposition 18 \mathcal{T}_A is dense in \mathcal{B}_A .

The proof of Proposition 9 of [3] works also for regular quasi-periodic sequences.

Theorem 19 (Blanchard, Formenti and Kůrka [2]) *The Besicovitch space is pathwise connected, infinite-dimen-*

sional, homogenous and complete. It is neither separable nor locally compact.

The properties of path-connectedness and infinite dimensionality is proved analogously as in Proposition 15. To prove that \mathcal{B}_A is neither separable nor locally compact, Sturmian configurations have been used in [2]. The completeness of \mathcal{B}_A has been proved by Marcinkiewicz [14]. Every cellular automaton $F : A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ is uniformly continuous with respect to d_B , so it preserves the equivalence \approx_B . If $d_B(x, y) = 0$, then $d_B(F(x), F(y)) = 0$. Thus a cellular automaton F defines a uniformly continuous map $F_B : \mathcal{B}_A \rightarrow \mathcal{B}_A$.

Theorem 20 (Blanchard, Formenti and Kůrka [2]) *Let F be a CA on A .*

- (1) (C_A, F) is surjective iff (\mathcal{B}_A, F_B) is surjective.
- (2) If $\mathfrak{A}(F) \neq \emptyset$ then (\mathcal{B}_A, F_B) is almost equicontinuous.
- (3) if $\mathfrak{C}(F) \neq \emptyset$, then (\mathcal{B}_A, F_B) is equicontinuous.
- (4) If (\mathcal{B}_A, F_B) is sensitive, then (C_A, F) is sensitive.
- (5) No cellular automaton (\mathcal{B}_A, F_B) is positively expansive.
- (6) If (C_A, F) is chain-transitive, then (\mathcal{B}_A, F_B) is chain-transitive.

Theorem 21 (Blanchard, Cervelle and Formenti [3])

- (1) No CA (\mathcal{B}_A, F_B) is transitive.
- (2) A CA (\mathcal{B}_A, F_B) has either a unique fixed point and no other periodic point, or it has uncountably many periodic points.
- (3) If a surjective CA has a blocking word, then the set of its F_B -periodic points is dense in \mathcal{B}_A .

The Generic Space

For a configuration $x \in A^{\mathbb{Z}}$ and word $v \in A^+$ set

$$\underline{\Phi}_v(x) = \liminf_{n \rightarrow \infty} |\{i \in [-n, n) : x_{[i, i+|v|)} = v\}|/2n,$$

$$\overline{\Phi}_v(x) = \limsup_{n \rightarrow \infty} |\{i \in [-n, n) : x_{[i, i+|v|)} = v\}|/2n.$$

For every $v \in A^+$, $\underline{\Phi}_v, \overline{\Phi}_v : A^{\mathbb{Z}} \rightarrow [0, 1]$ are continuous in the Besicovitch topology. In fact we have

$$|\overline{\Phi}_v(x) - \overline{\Phi}_v(y)| \leq d_B(x, y) \cdot |v|,$$

$$|\underline{\Phi}_v(x) - \underline{\Phi}_v(y)| \leq d_B(x, y) \cdot |v|.$$

Define the generic space (over the alphabet A) as

$$\mathcal{G}_A = \{x \in A^{\mathbb{Z}} : \forall v \in A^+, \underline{\Phi}_v(x) = \overline{\Phi}_v(x)\}.$$

It is a closed subspace of \mathcal{B}_A . For $v \in A^+$ denote by $\Phi_v : \mathcal{G}_A \rightarrow [0, 1]$ the common value of $\underline{\Phi}$ and $\overline{\Phi}$.

Using prefix codes, one can show that all generic spaces (with different alphabets) are homeomorphic. The generic space contains all uniquely ergodic subshifts, in particular all Sturmian sequences and all regular Toeplitz sequences. Thus the proofs in Blanchard Formenti and K urka [2] can be applied to the generic space too. In particular the generic space is homogenous. If we regard the alphabet $A = \{0, \dots, m-1\}$ as the group $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$, then for every $x \in \mathcal{G}_A$ there is an isometry $H_x: \mathcal{G}_A \rightarrow \mathcal{G}_A$ defined by $H_x(y) = x + y$. Moreover, \mathcal{G}_A is pathwise connected, infinite-dimensional and complete (as a closed subspace the full Besicovitch space). It is neither separable nor locally compact. If $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ is a cellular automaton, then $F(\mathcal{G}_A) \subseteq \mathcal{G}_A$. Thus, the restriction of F_B to \mathcal{G}_A defines a dynamical system (\mathcal{G}_A, F_G) . See also Pivato for a similar approach.

Theorem 22 Let $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ be a CA.

- (1) (C_A, F) is surjective iff (\mathcal{G}_A, F_G) is surjective.
- (2) If $\mathfrak{A}(F) \neq \emptyset$, then (\mathcal{G}_A, F_G) is almost equicontinuous.
- (3) if $\mathfrak{E}(F) \neq \emptyset$, then (\mathcal{G}_A, F_G) is equicontinuous.
- (4) If (\mathcal{G}_A, F_G) is sensitive, then (C_A, F) is sensitive.
- (5) If F is C -chain transitive, then F is G -chain transitive.

The proofs are the same as the proofs of corresponding properties in [2].

The Space of Measures

By a **measure** we mean a **Borel shift-invariant probability measure** on the Cantor space $A^{\mathbb{Z}}$ (see ► [Ergodic Theory of Cellular Automata](#)). This is a countably additive function μ on the Borel sets of $A^{\mathbb{Z}}$ which assigns 1 to the full space and satisfies $\mu(U) = \mu(\sigma^{-1}(U))$. A measure on $A^{\mathbb{Z}}$ is determined by its values on cylinders $\mu(u) := \mu([u]_n)$ which does not depend on $n \in \mathbb{Z}$. Thus a measure can be identified with a map $\mu: A^* \rightarrow [0, 1]$ subject to **bilateral Kolmogorov compatibility conditions**

$$\sum_{a \in A} \mu(ua) = \sum_{a \in A} \mu(au) = \mu(u), \quad \mu(\lambda) = 1.$$

Define the distance of two measures

$$d_{\mathcal{M}}(\mu, \nu) := \sum_{u \in A^+} |\mu(u) - \nu(u)| \cdot |A|^{-2|u|}.$$

This is a metric which yields the topology of weak* convergence on the compact space $\mathcal{M}_A := \mathcal{M}_{\sigma}(A^{\mathbb{Z}})$ of shift-

invariant Borel probability measures. A CA $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ with local rule f determines a continuous and affine map $F_{\mathcal{M}}: \mathcal{M}_A \rightarrow \mathcal{M}_A$ by

$$(F_{\mathcal{M}}(\mu))(u) = \sum_{v \in f^{-1}(u)} \mu(v).$$

Moreover F and $F\sigma$ determine the same dynamical system on \mathcal{M}_A : $F_{\mathcal{M}} = (F\sigma)_{\mathcal{M}}$.

For $x \in \mathcal{G}_A$ denote by $\Phi^x: A^* \rightarrow [0, 1]$ the function $\Phi^x(v) = \Phi_v(x)$. For every $x \in \mathcal{G}_A$, Φ^x is a shift-invariant Borel probability measure. The map $\Phi: \mathcal{G}_A \rightarrow \mathcal{M}_A$ is continuous with respect to the Besicovich and weak* topologies. In fact we have

$$\begin{aligned} d_{\mathcal{M}}(\Phi^x, \Phi^y) &\leq d_B(x, y) \sum_{u \in A^+} |u| \cdot |A|^{-2|u|} \\ &= d_B(x, y) \sum_{n>0} n \cdot |A|^{-n} \\ &= d_B(x, y) \cdot |A|/(|A| - 1)^2. \end{aligned}$$

By a theorem of Kamae [9], Φ is surjective. Every shift-invariant Borel probability measure has a generic point. It follows from the ergodic theorem that if μ is a σ -invariant measure, then $\mu(\mathcal{G}_A) = 1$ and for every $v \in A^*$, the measure of v is the integral of its density Φ_v ,

$$\mu(v) = \int \Phi_v(x) d\mu.$$

If F is a CA, we have a commutative diagram $\Phi F_G = F_{\mathcal{M}}\Phi$.

$$\begin{array}{ccc} \mathcal{G}_A & \xrightarrow{F_G} & \mathcal{G}_A \\ \Phi \downarrow & & \downarrow \Phi \\ \mathcal{M}_A & \xrightarrow{F_{\mathcal{M}}} & \mathcal{M}_A \end{array}$$

Theorem 23 Let F be a CA over A .

- (1) (C_A, F) is surjective iff $(\mathcal{M}_A, F_{\mathcal{M}})$ is surjective.
- (2) If (\mathcal{G}_A, F_G) has dense set of periodic points, then $(\mathcal{M}_A, F_{\mathcal{M}})$ has dense set of periodic points.
- (3) If $\mathfrak{A}(F) \neq \emptyset$, then $(\mathcal{M}_A, F_{\mathcal{M}})$ is almost equicontinuous.
- (4) If $\mathfrak{E}(F) \neq \emptyset$, then $(\mathcal{M}_A, F_{\mathcal{M}})$ is equicontinuous.

Proof (1) See K urka [13] for a proof.

(2) This holds since $(\mathcal{M}_A, F_{\mathcal{M}})$ is a factor of (\mathcal{G}_A, F_G) .

(3) It suffices to prove the claim for the case that F is almost equicontinuous. In this case there exists a blocking word

$u \in A^+$ and the Dirac measure δ_u defined by

$$\delta_u(v) = \begin{cases} 1/|u| & \text{if } v \sqsubseteq u \\ 0 & \text{if } v \not\sqsubseteq u \end{cases}$$

is equicontinuous for $(\mathcal{M}_A, F_{\mathcal{M}})$.

(4) If (C_A, F) is equicontinuous, then all sufficiently long words are blocking and there exists $d > 0$ such that for all $n > 0$, and for all $x, y \in A^{\mathbb{Z}}$ such that $x_{[-n-d, n+d]} = y_{[-n-d, n+d]}$ we have $F^k(x)_{[-n, n]} = F^k(y)_{[-n, n]}$ for all $k > 0$. Thus there are maps $g_k: A^* \rightarrow A^*$ such that $|g_k(u)| = \max\{|u| - 2d, 0\}$ and for every $x \in A^{\mathbb{Z}}$ we have $F^k(x)_{[-n, n]} = F^k(x_{[-n-kd, n+kd]}) = g_k(x_{[-n-d, n+d]})$, where f is the local rule for F . We get

$$\begin{aligned} d_{\mathcal{M}}(F_{\mathcal{M}}^k(\mu), F_{\mathcal{M}}^k(\nu)) &= \sum_{n=1}^{\infty} \sum_{u \in A^n} \left| \sum_{v \in f^{-k}(u)} (\mu(v) - \nu(v)) \right| \cdot |A|^{-2n} \\ &= \sum_{n=1}^{\infty} \sum_{u \in A^n} \left| \sum_{v \in g_k^{-1}(u)} (\mu(v) - \nu(v)) \right| \cdot |A|^{-2n} \\ &\leq \sum_{n=1}^{\infty} \sum_{v \in A^{n+2d}} |\mu(v) - \nu(v)| \cdot |A|^{-2n} \\ &\leq |A|^{4d} \cdot d_{\mathcal{M}}(\mu, \nu). \end{aligned} \quad \square$$

The Weyl Space

Define the following equivalence relation on $A^{\mathbb{Z}}$: $x \approx_W y$ iff $d_W(x, y) = 0$. Denote by \mathcal{W}_A the set of equivalence classes of \approx_W and by $\pi_W: A^{\mathbb{Z}} \rightarrow \mathcal{W}_A$ the projection. The factor of d_W is a metric on \mathcal{W}_A . This is the Weyl space on alphabet A . Using prefix codes, it can be shown that every two Weyl spaces (with different alphabets) are homeomorphic. The Toeplitz space is not dense in the Weyl space (see Blanchard, Cervelle and Formenti [3]).

Theorem 24 (Blanchard, Formenti and Kůrka [2]) *The Weyl space is pathwise connected, infinite-dimensional and homogenous. It is neither separable nor locally compact. It is not complete.*

Every cellular automaton $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ is continuous with respect to d_W , so it preserves the equivalence \approx_W . If $d_W(x, y) = 0$, then $d_W(F(x), F(y)) = 0$. Thus a cellular automaton F defines a continuous map $F_W: \mathcal{W}_A \rightarrow \mathcal{W}_A$. The shift map $\sigma: \mathcal{W}_A \rightarrow \mathcal{W}_A$ is again an isometry, so in \mathcal{W}_A many topological properties are preserved if F is composed with a power of the shift. This is true for example for equicontinuity, almost continuity and sensitivity. If $\pi: \mathcal{W}_A \rightarrow \mathcal{B}_A$ is the (continuous) projection and F a CA,

then the following diagram commutes.

$$\begin{array}{ccc} \mathcal{W}_A & \xrightarrow{F_W} & \mathcal{W}_A \\ \pi \downarrow & & \downarrow \pi \\ \mathcal{B}_A & \xrightarrow{F_B} & \mathcal{B}_A \end{array}$$

Theorem 25 (Blanchard, Formenti and Kůrka [2]) *Let F be a CA on A .*

- (1) (C_A, F) is surjective iff (\mathcal{W}_A, F_W) is surjective.
- (2) If $\mathfrak{A}(F) \neq \emptyset$, then (\mathcal{W}_A, F_W) is almost equicontinuous.
- (3) If $\mathfrak{E}(F) \neq \emptyset$, then (\mathcal{W}_A, F_W) is equicontinuous.
- (4) If (C_A, F) is chain-transitive, then (\mathcal{W}_A, F_W) is chain-transitive.

Theorem 26 (Blanchard, Cervelle and Formenti [3]) *No CA is (\mathcal{W}_A, F_W) is transitive.*

Theorem 27 *Let Σ be a subshift attractor of finite type for F (in the Cantor space). Then there exists $\delta > 0$ such that for every $x \in \mathcal{W}_A$ satisfying $d_W(x, \Sigma) < \delta$, $F^n(x) \in \Sigma$ for some $n > 0$.*

Thus a subshift attractor of finite type is a \mathcal{W} -attractor. Example 2 shows that it need not be \mathcal{B} -attractor. Example 3 shows that the assertion need not hold if Σ is not of finite type.

Proof Let $U \subseteq A^{\mathbb{Z}}$ be a C -clopen set such that $\Sigma = \Omega_F(U)$. Let U be a union of cylinders of words of length q . Set $\widetilde{\Omega}_{\sigma}(U) = \bigcap_{n \in \mathbb{Z}} \sigma^n(U)$. By a generalization of a theorem of Hurd [7] (see ▶ [Topological Dynamics of Cellular Automata](#)), there exists $m > 0$ such that $\Sigma = F^m(\widetilde{\Omega}_{\sigma})$. If $d_W(x, \Sigma) < 1/q$ then there exists $l > 0$ such that for every $k \in \mathbb{Z}$ there exists a nonnegative $j < l$ such that $\sigma^{k+j}(x) \in U$. It follows that there exists $n > 0$ such that $F^n(x) \in \widetilde{\Omega}_{\sigma}(U)$ and therefore $F^{n+m}(x) \in \Sigma$. \square

Examples

Example 1 The identity rule $\text{Id}(x) = x$.

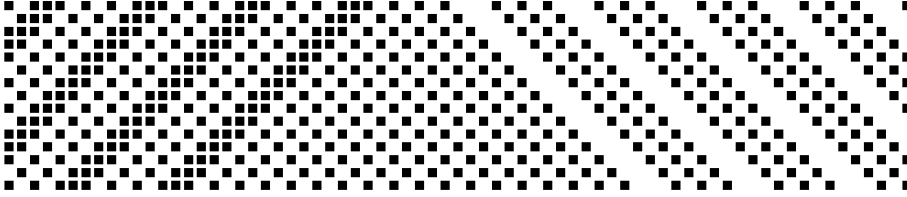
$(\mathcal{B}_A, \text{Id}_{\mathcal{B}})$ and $(\mathcal{W}_A, \text{Id}_{\mathcal{W}})$ are chain-transitive (since both \mathcal{B}_A and \mathcal{W}_A are connected). However, (C_A, Id) is not chain-transitive. Thus the converse of Theorem 20(6) and of Theorem 25(4) does not hold.

Example 2 The product rule ECA128 $F(x)_i = x_{i-1} \cdot x_i \cdot x_{i+1}$.

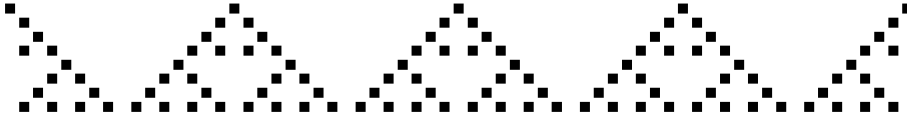
(C_A, F) , $(\mathcal{B}_A, F_{\mathcal{B}})$ and (\mathcal{W}_A, F_W) are almost equicontinuous and the configuration 0^{∞} is equicontinuous in all these versions. By Theorem 27, $\{0^{\infty}\}$ is a \mathcal{W} -attractor. However, contrary to a mistaken Proposition 9 in [2], $\{0^{\infty}\}$ is



Dynamics of Cellular Automata in Non-compact Spaces, Figure 1
The product ECA184



Dynamics of Cellular Automata in Non-compact Spaces, Figure 2
The traffic ECA184



Dynamics of Cellular Automata in Non-compact Spaces, Figure 3
The sum ECA90

not \mathcal{B} -attractor. For a given $0 < \varepsilon < 1$ define $x \in A^{\mathbb{Z}}$ by $x_i = 1$ iff $3^n(1 - \varepsilon) < |i| \leq 3^n$ for some $n \geq 0$. Then $d_{\mathcal{B}}(x, 0^\infty) = \varepsilon$ but x is a fixed point, since $d_{\mathcal{B}}(F(x), x) = \lim_{n \rightarrow \infty} 2n/3^n = 0$ (see Fig. 1).

Example 3 The traffic ECA184 $F(x)_i = 1$ iff $x_{[i-1, i]} = 10$ or $x_{[i, i+1]} = 11$.

No $F^q \sigma^p$ is C -almost equicontinuous, so $\mathfrak{A}(F) = \emptyset$. However, if $d_{\mathcal{W}}(x, 0^\infty) < \delta$, then $d_{\mathcal{W}}(F^n(x), 0^\infty) < \delta$ for every $n > 0$, since F conserves the number of letters 1 in a configuration. Thus 0^∞ is a point of equicontinuity in (\mathcal{T}_A, F_T) , (\mathcal{B}_A, F_B) , and $(\mathcal{W}_A, F_{\mathcal{W}})$. This shows that item (2) of Theorems 17, 20 and 25 cannot be converted. The maximal C -attractor $\Omega_F = \{x \in A^{\mathbb{Z}} : \forall n > 0, 1(10)^n 0 \not\sqsubseteq x\}$ is not SFT. We show that it does not \mathcal{W} -attracts points from any of its neighborhood. For a given even integer $q > 2$ define $x \in A^{\mathbb{Z}}$ by

$$x_i = \begin{cases} 0 & \text{if } \exists n \geq 0, i = qn + 1 \\ 1 & \text{if } \exists n < 0, i = qn \\ ((01)^\infty)_i & \text{otherwise.} \end{cases}$$

Then $d_{\mathcal{W}}(F^k(x), \Omega_F) = 1/q$ for all $k > 0$ (see Fig. 2, where $q = 8$).

Example 4 The sum ECA90 $F(x)_i = (x_{i-1} + x_{i+1}) \bmod 2$.

Both (\mathcal{B}_A, F_B) and $(\mathcal{W}_A, F_{\mathcal{W}})$ are sensitive (Cattaneo et al. [4]). For a given $n > 0$ define a configuration z by

$z_i = 1$ iff $i = k2^n$ for some $k \in \mathbb{Z}$. Then $F^{2^{n-1}}(z) = (01)^\infty$. For any $x \in A^{\mathbb{Z}}$, we have $d_{\mathcal{W}}(x, x + z) = 2^{-n}$ but $d_{\mathcal{W}}(F^{2^{n-1}}(x), F^{2^{n-1}}(x + z)) = 1/2$. The same argument works for (\mathcal{B}_A, F_B) .

Example 5 The shift ECA170 $F(x)_i = x_{i+1}$.

Since the system has fixed points 0^∞ and 1^∞ , it has uncountable number of periodic points. However, the periodic points are not dense in \mathcal{B}_A ([3]).

Future Directions

One of the promising research directions is the connection between the generic space and the space of Borel probability measures which is based on the factor map Φ . In particular Lyapunov functions based on particle weight functions (see K urka [12]) work both for the measure space \mathcal{M}_A and the generic space \mathcal{G}_A . The potential of Lyapunov functions for the classification of attractors has not yet been fully explored. This holds also for the connections between attractors in different topologies. While the theory of attractors is well established in compact spaces, in noncompact spaces there are several possible approaches. Finally, the comparison of entropy properties of CA in different topologies may be revealing for classification of CA.

There is even a more general approach to different topologies for CA based on the concept of submeasure on \mathbb{Z} . Since each submeasure defines a pseudometric, it would

be interesting to know, whether CA are continuous with respect to any of these pseudometrics, and whether some dynamical properties of CA can be derived from the properties of defining submeasures.

Acknowledgments

We thank Marcus Pivato and Francois Blanchard for careful reading of the paper and many valuable suggestions. The research was partially supported by the Research Program Project “Sycomore” (ANR-05-BLAN-0374).

Bibliography

Primary Literature

1. Besicovitch AS (1954) Almost periodic functions. Dover, New York
2. Blanchard F, Formenti E, Kůrka P (1999) Cellular automata in the Cantor, Besicovitch and Weyl spaces. *Complex Syst* 11(2):107–123
3. Blanchard F, Cerveille J, Formenti E (2005) Some results about the chaotic behaviour of cellular automata. *Theor Comput Sci* 349(3):318–336
4. Cattaneo G, Formenti E, Margara L, Mazoyer J (1997) A shift-invariant metric on $S^{\mathbb{Z}}$ inducing a nontrivial topology. *Lecture Notes in Computer Science*, vol 1295. Springer, Berlin
5. Formenti E, Kůrka P (2007) Subshift attractors of cellular automata. *Nonlinearity* 20:105–117
6. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. *Math Syst Theory* 3:320–375
7. Hurd LP (1990) Recursive cellular automata invariant sets. *Complex Syst* 4:119–129
8. Iwanik A (1988) Weyl almost periodic points in topological dynamics. *Colloquium Mathematicum* 56:107–119
9. Kamae J (1973) Subsequences of normal sequences. *Isr J Math* 16(2):121–149
10. Knudsen C (1994) Chaos without nonperiodicity. *Am Math Mon* 101:563–565
11. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. *Ergod Theory Dyn Syst* 17:417–433
12. Kůrka P (2003) Cellular automata with vanishing particles. *Fundamenta Informaticae* 58:1–19
13. Kůrka P (2005) On the measure attractor of a cellular automaton. *Discret Continuous Dyn Syst* 2005(suppl):524–535
14. Marcinkiewicz J (1939) Une remarque sur les espaces de a.s. Besicovitch. *C R Acad Sc Paris* 208:157–159
15. Sablik M (2006) étude de l'action conjointe d'un automate cellulaire et du décalage: une approche topologique et ergodique. Ph D thesis, Université de la Méditerranée

Books and Reviews

- Besicovitch AS (1954) Almost periodic functions. Dover, New York
 Kitchens BP (1998) Symbolic dynamics. Springer, Berlin
 Kůrka P (2003) Topological and symbolic dynamics. *Cours spécialisés*, vol 11. Société Mathématique de France, Paris
 Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge

Dynamics and Evaluation: The Warm Glow of Processing Fluency

PIOTR WINKIELMAN, DAVID E. HUBER
 Department of Psychology,
 University of California, San Diego, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Psychological Evidence
 for the Role of Fluency in Evaluation
 Computational Mechanisms
 Modeling Fluency-Affect Interactions:
 The Influence of Specific Variables
 Neural Basis of Fluency-Affect Connection
 Future Directions
 Bibliography

Glossary

Evaluative responses Reactions reflecting an implicit or explicit assessment of stimulus' goodness or badness. Such reactions can be captured with judgments (liking, preference, choice), behavior (approach, avoidance, desire to continue or terminate), and physiology (peripheral and central). Strong evaluative reactions can develop into full-blown affective states, such as moods and emotion.

Fluency A general term used to describe efficiency of processing on perceptual and conceptual levels. Fluent processing is fast, error-free, and easy – a quality reflected in several dynamical properties (see next). The level of processing fluency can be monitored and influence evaluative as well as cognitive processes.

Processing dynamics Processing dynamic refers to content non-specific parameters characterizing a system's behavior at the level of individual units (e.g., neurons) or networks. Those dynamical parameters include (i) coherence of the signals within the system – the extent to which the signals arriving at a given unit from other units consistently dictate the same state, (ii) settling time – amount of time for the system to achieve a steady state, (iii) volatility – the number of units changing state, (iv) signal-to-noise ratio – overall strength of signals in the network, and (v) or differentiation – ratio of strongly activated to weakly activated units.

Mere-exposure effect Empirical observation that simple repetition (mere-exposure) of an initially neutral stimulus enhances people's liking for it. Because a typical result of simple repetition is enhancement of processing fluency (more efficient re-processing), this classic empirical phenomenon inspired research into evaluative consequences of changes in processing dynamics, and led to investigation of other variables that change processing dynamics, such as priming, duration, clarity, contrast, prototypicality or symmetry.

Definition of the Subject

A major goal of contemporary psychology, cognitive science and neuroscience is to understand the interactions of cognition and emotion – the mutual influences between thinking and feeling [11,41,101]. Sometimes such interactions are quite dramatic – as when emotions rob an individual of sound judgment (e.g., crimes of passions) or when emotions inspire an individual to transgress self-interest (e.g., acts of compassion). But most cognition-emotion interactions are more subtle. For example, seeing a well-balanced piece of art or a symmetric design may invoke a sense of aesthetic pleasure. Simply recognizing a familiar acquaintance on a street may evoke a sense of warm glow. A difficult to hear cell phone conversation may evoke a subtle sense of annoyance. And, in the first few days of being in a foreign country, people's faces just look "weird". This contribution deals with the nature of cognition-emotion interactions in these more subtle *everyday evaluative responses* – basic liking/disliking reactions that can be captured by people's preference judgments, behaviors, and physiological measures. Our special focus is on the role of non-specific, dynamical aspect of processing in the formation of evaluative responses. We show how the mechanisms underlying such responses can be understood using a combination of psychological experimentation and mathematical modeling grounded in physical models.

Introduction

Overview

The structure of our contribution is roughly as follows. First, we distinguish various sources of evaluative responses – non-specific processing dynamics and specific feature-based information. Next, we describe empirical work suggesting that evaluative reactions to processing dynamics can explain several common preference phenomena. Then, we describe some computational models of mechanisms underlying dynamics-affect connection. Finally, we discuss some neural underpinnings of the dy-

namical-affect connection and close with suggestions for future work.

Dynamical and Featural Information

You walk down a busy street and scan the passing faces. Some you like, some you do not. Why? Psychologists explore this question by focusing on the "how" and "what" of processing. The "how" refers to non-specific dynamical information about the quality of processing, and the "what" refers to specific feature-based information. Let us distinguish these two sources and briefly characterize their relationship.

During any kind of processing, sometimes even before any specific features are extracted from the stimulus, the mental system has access to a nonspecific source of information – the dynamics accompanying the processing of the stimulus. Historically, interest in the non-specific dynamics of processing originated in the field of metacognition [37,47,50]. This work highlighted that people monitor not only the content ("what") of mental representations, but also the "how" of processing, including such non-specific parameters as processing speed, ease, representation strength, volatility, and the degree of match between the incoming information and stored representations. Although there are substantial differences between these various parameters, it is common to refer to these non-specific aspects of processing with the general term of "*fluency*" (for reviews see [34,77]). As we describe shortly, the central idea guiding this review is that high fluency is typically associated with positive evaluations.

Evaluation is also obviously influenced by "what" is extracted – the stimulus' specific *features*. Thus, the overall positivity of a response to a face of a passing stranger will also depend on detection of such features (e.g., a smile or a symmetrical appearance), and on the perceivers idiosyncratic appraisals of these features (e.g., interest in meeting a stranger). There are many available reviews of the experimental and modeling work on featural processing in evaluation and emotion, so we will not repeat it here [1,3,76]. But we want to highlight a couple of things about the relation between dynamical and featural aspects of processing. First, both sources of information are available simultaneously, with each contributing to the net evaluative reaction. For example, positivity from detecting a smile can combine with positivity from high fluency of recognition. Second, the dynamical and featural sources can play off each other. For example, the same feature, such as symmetry, might create a positive reaction because of its implications (e.g., good health), but also make the face easier to recognize. In other words, a feature might not only cre-

ate an evaluative reaction directly, but also indirectly, via its influence on the processing dynamics.

How is it Going? Linking Dynamics and Affect

The idea that dynamical aspects of information processing have affective implications has been discussed in several domains of research. The major proposals focus on the role of dynamics as cue to the quality of the internal state of the system, or as a cue to the quality of an external stimulus.

Feedback About Quality of Internal Processing

At least since Simon [81], psychologists assume that one function of affect to provide information about the internal state of the system. Thus, unless there is an obvious external cause for feeling good or bad, the presence of negative affect conveys that something is internally “wrong”, whereas positive affect conveys that things are going “right” (e.g., [8]). More specifically, affect can provide information about the current state of cognitive operations. Thus, high fluency of a perceptual or a conceptual process indicates progress toward, for example, successful recognition of the stimulus or a successful solution of a task. Besides informing the organism that processing is going well, positive affect triggered by high fluency may play a motivational function and reinforce the successful strategy [62,91]. On the other hand, low fluency can be a signal of cognitive error or incompatibility, and play a motivational role in revision of a processing strategy [12,18]. These ideas converge with classic observations that mental states characterized by low coherence, such as cognitive dissonance, are unpleasant, as indicated by self-report as well as physiological measures of affect [24].

Feedback About Quality of External Stimuli

Processing dynamics can also have affective consequences because it informs (probabilistically) whether an external stimulus is good or bad. For example, it's known, at least since Titchener [90], that familiar stimuli elicit a “warm glow.” Conversely, illusions of familiarity (oldness) can be produced through unobtrusive inductions of positive affect [20,60]. One reason for this warmth-familiarity link could be biological predispositions for caution in encounters with novel, and thus potentially harmful, stimuli Zajonc [101]. Other accounts suggest that familiarity is just a learned, “fast and frugal” heuristic for easily identifying choices that are in truth objectively better [21]. Similarly, as we discuss next, dynamics could offer a probabilistic cue regarding other valued properties of external stimuli, such as symmetry, prototypicality, etc.

Psychological Evidence for the Role of Fluency in Evaluation

So far, we have focused on theoretical reasons for the dynamics-affect connection. The specific empirical research on the role of dynamical information in affect has centered around five related variables: (i) repetition/mere exposure, (ii) priming, (iii) contrast, clarity, duration, (iv) symmetry and (v) prototypicality. As we show, all these preference phenomena are consistent with the notion that high processing fluency enhances evaluations (for more comprehensive review see [66,99]).

Mere-Exposure/Repetition

The “mere exposure effect” (MEE) is the observation that simple repetition enhances liking for an initially neutral stimulus [101]. Interestingly, all that is required for the MEE is that the stimulus is “merely” shown, however briefly or incidentally, to the individual – no reinforcement is required and the presentation can be even subliminal (for reviews see [5]). The reader has probably experienced this phenomenon many times. Thus, most melodies and paintings “grow on you” with repeated exposure, faces that are simply familiar tend to generate a “warm glow,” and advertisers try to increase sales by simply repeating product's name or image. Anecdotes aside, empirical evidence for the mere exposure effects is quite robust. For example in a study by Monahan, Murphy, and Zajonc [51], participants were subliminally exposed to 25 pictures of novel ideographs, and were later asked to report their tonic mood. For some participants, each of the 25 ideographs was different, while for other participants, 5 different ideographs were repeated 5 times each. The results showed that participants who were subliminally exposed to repeated ideographs reported being in a better mood than participants exposed to 25 different ideographs. Additional evidence for the positivity of reactions from the mere exposure effect comes from studies that used facial electromyography (EMG). This technique relies on the observation that positive affective responses manifest themselves in incipient smiles, as reflected by higher activity over the cheek region – zygomaticus major – whereas negative affective responses manifest themselves in incipient frowns, as reflected by higher activity over the brow region – corrugator supercilii [7]. Harmon-Jones and Allen [25] observed that repeatedly presented stimuli elicited stronger EMG activity over the “smiling” region of the participants' face (cheek), indicative of positive affect, without changing the activity over the “frowning” region (brow).

There is now good evidence that the mere-exposure effect reflects changes in processing fluency – the ease of recognition (e.g., [6,34,36,45,79]). Stimulus repetition speeds up stimulus recognition and enhances judgments of stimulus clarity and presentation duration, which are indicative of processing facilitation (e.g., [22,33]).

Priming

Based on the just mentioned research, we may expect that any variable that facilitates processing should result in increased liking, even under conditions of a single exposure. Several studies confirmed this possibility. In one of these studies (see Study 1 in [65]) participants were exposed to pictures of everyday objects (e.g., a desk, bird, or plane). The processing fluency of these target pictures was facilitated or inhibited by subliminal presentation of visual contours (e.g., [2]). Some target pictures were preceded by matched contours (e.g., contour of a desk followed by a picture of the desk), whereas others were preceded by mismatched contours (e.g., contour of a desk followed by a picture of a bird). Some participants were asked to indicate how much they liked the target pictures; other participants were asked to press a button as soon as they could recognize the object in the picture, thus providing an independent measure of processing ease. The data showed that pictures preceded by matched contours were recognized faster, indicating higher fluency, and were liked more than pictures preceded by mismatched contours.

Importantly, Winkielman and Cacioppo [95] provided evidence for the positivity of reactions caused by priming using the earlier-mentioned technique of facial electromyography (fEMG). High fluency was associated with stronger activity over the zygomaticus region (indicative of positive affect), but was not associated with stronger activity of the corrugator region (indicative of negative affect). This effect occurred in the first 3 seconds after the presentation of the stimulus, which was several seconds before participants made their overt judgments. This suggests a quick link between high fluency and positive affect.

Contrast, Clarity, and Duration

High contrast and clarity have repeatedly been identified as characteristics of aesthetically appealing objects (e.g., [86]). According to our proposal, these properties trigger liking because they facilitate processing. In one study (see Study 2 in [65]) we manipulated fluency through different degrees of figure-ground contrast, taking advantage of the observation that high contrast decreases identification speed [9]. Participants liked the same

stimulus more when it was presented with higher contrast, and hence could be processed more fluently. In another study (see Study 3 in [65]) we manipulated fluency through subtle increases in presentation duration, taking advantage of the observation that longer presentation durations facilitate the extraction of information [44]. As expected, participants evaluated the same stimulus more positively when it was presented for a longer duration, even if they were unaware that duration was manipulated. Winkielman and Cacioppo [95] replicated these results and also found corresponding changes in EMG activity, which suggests that high fluency elicits positive affect on the physiological level.

Symmetry

Humans and non-human animals show a widespread preference for symmetry [67]. This is often attributed to the biological value of symmetry as a signal of mate quality (e.g., [89]). However, we propose that symmetry is appealing at least partly because it facilitates information processing. After all, symmetrical stimuli are structurally simpler, and thus more fluent, than non-symmetrical stimuli. Support for this comes from studies on preference and fluency of abstract shapes [64]. These researchers asked participants to make preference judgments and also same-different judgments for symmetrical and asymmetrical shapes. The results showed that symmetrical shapes are not only more appealing, but also easier to identify than comparable asymmetrical shapes. This finding is compatible with earlier studies by Palmer and his colleagues showing that symmetry is preferred, as long as it facilitates information processing. Specifically, Palmer [58] presented the same symmetrical dot patterns (such that overall amount of information was held constant) in one of three orientations – vertically, diagonally, or horizontally – and asked participants to rate the figural goodness of each of the patterns. He found that dot patterns presented in the vertically symmetrical orientation received the highest figural goodness ratings, followed by those presented in the horizontally symmetrical orientation, with those presented in the diagonally symmetrical orientation receiving the lowest figural goodness ratings. Importantly, the figural goodness ratings paralleled earlier work by Palmer and Hemenway [59] on ease of symmetry detection: symmetry in the dot patterns presented in vertically symmetrical orientations was detected the fastest, followed by the symmetry in the horizontally symmetrical orientations, with the symmetry of the dot patterns presented in diagonally symmetrical orientations being the most difficult to detect. Since each of the patterns in the three orientations con-

tained the same amount of information, this result suggests that symmetry makes any given stimulus more appealing because it facilitates the ability of the perceiver to detect redundant information and, as such, to more easily identify the stimulus.

Prototypicality

Another robust source of preference is prototypicality or “averageness” – in the sense of a stimulus providing the “best representation” of the category, or fitting its central tendency [67]. People show prototypicality preference for living objects, such as faces, fish, dogs and birds, and also for nonliving objects, such as color patches, furniture, wristwatches and automobiles [23,40]. This effect, known since Galton [19], has also been explained as reflecting evolved predisposition to interpret prototypicality as a cue to mate quality [88]. However, there is a more straightforward dynamical explanation. Given that prototypes are the most representative members of their categories, they are also fluent, as reflected in accuracy and speed of classification [66]. This raises the possibility that prototypes are liked *because* they are fluent. Winkelman, Halberstadt, Fazendeiro, and Catty [96] examined this idea in a series of three experiments. Participants first learned a category of random dot patterns (Experiment 1) or of common geometric patterns (Experiment 2) and then were presented with novel patterns varied across different levels of prototypicality. Participants classified these patterns into their respective categories as quickly as possible (measure of fluency), and also rated the attractiveness of each. A close relationship between fluency, attractiveness, and the level of prototypicality was observed. Both fluency and attractiveness increased with prototypicality. Importantly, when fluency was statistically controlled, the relation between prototypicality and attractiveness dropped by half (though it remained significant). This suggests that processing facilitation is important to, but not the sole cause of the “beauty-in-averageness” effect. Finally, Experiment 3 showed that viewing prototypical, rather than non-prototypical patterns elicited significantly greater EMG activity, suggesting that viewing prototypes involves genuine affective reactions.

In combination, the above studies, based on manipulations of repetition, figure-ground contrast, presentation duration, symmetry, and prototypicality consistently show that high perceptual fluency leads to more positive evaluations of the perceived stimuli. However, verbal descriptions of fluency are often vague and so fluency is often difficult to quantify – what exactly does it mean that one stimulus is more fluent than another? The answer to this

question has been provided by computational models inspired by physical phenomena.

Computational Mechanisms

There is surprisingly little research on the role of dynamical parameters in cognition and emotion [55,61]. One notable exception is the Neural network approach, or connectionism, in which cognition is viewed in terms of the passage of activation among simple, neuron-like units organized in large, densely interconnected networks [73]. The individual units function as simple processors that can influence each other through connections, which vary in strength and sign (facilitatory or inhibitory). This massively interconnected and parallel architecture gives the neural network approach a certain neurophysiological realism and makes it suitable for a wide variety of applications. For more biological applications, one can conceptualize the network units as actual neurons, whereas for more psychological applications, one can treat the units as blocks of neurons or functional sub-systems [57]. Many different neural network architectures have been proposed that utilize dynamical parameters. Below we primarily focus on a proposal by Lewenstein and Nowak [42], which illustrates the role of dynamical parameters in learning and recognition using a simple attractor neural network [26]. Although this is in some regards an overly simplified model, the conceptual framework of the attractor network has been successfully expanded to more complicated applications, such as the plasticity-stability dilemma [52], and more realistic biological assumptions [54,84]. We address some of these more complex models later.

Fluency in a Hopfield Network

In a typical Hopfield network, representations are encoded as attractors of the network, i. e. states into which the network dynamics converge. The processing of information with the network can be seen as a gradual, evolving process, during which each neuron adjusts to the signal coming from other neurons. Because neurons are reciprocally connected, and because there are a large number of paths connecting one neuron to another, activation can reverberate dynamically through the network over simulated time steps until the network settles on the identified representation. For example, when presented with a to-be-recognized pattern, the network goes through a series of adjustments and after some time approaches a stable state, an attractor, corresponding to the “recognition” of a particular pattern.

Lewenstein and Nowak [42] proposed that a typical Hopfield model can be extended with a simple control

mechanism, which allows the network to monitor the dynamics of its own processing. Such a control mechanism can measure a variety of dynamical parameters, such as settling time, volatility, signal strength, coherence, and so on. These formally related properties can then be used by the network to roughly monitor the quality of its own processing (e.g., is it going well?) as well as estimate the characteristics of the stimuli being processed (e.g., is it familiar).

Studies with this model focused on how monitoring the dynamical properties of cognition can allow the network to estimate proximity to its closest attractor during the recognition process. This, in turn, allows the network to estimate the likelihood that the presented pattern is “known”, without requiring full specification for the manner in which the attractor is known. Specifically, two key dynamical properties were identified. The first property is the network’s “volatility”, or the proportion of neurons changing their state at a given point. When the incoming, “to-be-recognized” pattern matches or closely approximates a known pattern, corresponding to one of the attractors (memories), the network is characterized by a relatively small proportion of neurons changing their state. When the incoming pattern is novel, and thus does not approximate one of the attractors, the network is characterized by a large number of neurons changing their state. The second key dynamical property is the coherence of the signals received by the neurons. In the vicinity of an attractor (old pattern), the signals arriving from other neurons at a given neuron are consistent in that they belong to the same pattern. However, when the network is far from an attractor (new pattern), the signals arriving from other neurons at a given neuron dictate conflicting states and may provide partial matches to a variety of other patterns. A closely related criterion is the signal-to-noise ratio. In the vicinity of the attractor (old pattern), signals from other neurons typically add up, resulting in a relatively large summary signal dictating the state of a given neuron. However, far from an attractor (new pattern), signals from other neurons cancel each other, resulting in a relatively weak summary signal dictating the state of a given neuron. As a consequence, the processing of “old” patterns is characterized by a higher signal-to-noise ratio than the processing of “new” patterns.

Extension to Graded Representations and Incremental Change

Traditional Hopfield networks use simulated neurons that are either “on” or “off”, with no graded signal between these states. More realistic simulated neurons use a con-

tinuous range of intermediary values, allowing a graded measure for the magnitude and speed of settling into attractor states (e.g., [57]). However, because many applications are focused on learning and representational change, large simulated time steps are used and settling occurs in less than 10 time steps, which makes it difficult to measure relatively subtle differences in settling time. For such applications, fluency is measured rather indirectly as differentiation – the magnitude of the most active units [42]. Providing a more direct measure of fluency based on speed of processing, we have used neural simulations with millisecond time steps, which allows measurement of the time needed to achieve peak activation [28,30]. Not only does this provide a measure of choice preference, but it can be used to indicate reaction times [29]. In these real-time simulations, habituation dynamics are implemented such that activation achieves a peak value, but then falls to a lower value with continued processing. Because well learned representations include stronger connections, and because activation is the driving force behind habituation, familiar representations reach a peak value more quickly as habituation occurs more quickly [31].

Fast Fluency

The modeling work on fluency also shed light on the puzzling phenomenon when the system responds affectively to a pattern before it is fully recognized (“preference without inference”, Zajonc [101]). Processing speed, volatility, differentiation, and the onset of habituation are all measurements of fluency that allow the network to estimate whether a pattern is “new” or “old” (i.e., proximity to its closest attractor) prior to explicit identification of the pattern. For instance, it is possible to determine the familiarity of incoming stimuli by monitoring how frequently a mere 10% of the neurons change their state during the very first time step [42]. Similarly, a fast familiarity signal can be based on the early differentiation [53]. It is also worth noting that checking the coherence of incoming signals makes it possible to estimate not only the global novelty of the whole pattern, but also the novelty of fragments in the perceived pattern, such as elements of an object or objects in a scene [103]. As discussed earlier, because familiarity is affectively positive, all these mechanisms explain how one can “like” something before even knowing what it is.

Fluency and Self-Regulation

In addition to quick feedback about the valence of the incoming stimulus, the early pre-recognition of familiarity may be used to control the recognition process, so that known stimuli are processed differently than new ones.

This may be achieved by linking the outcome of pre-recognition based on monitoring the system dynamics to a control parameter (e. g., network's overall noise level) that influences the later stages of the recognition process. A number of specific models that involve a feedback loop between pre-recognition and the noise level have been proposed. For example, in the original model by Lewenstein and Nowak [42], unknown patterns raised the noise level, preventing false "recognition" of unfamiliar patterns – a common problem for neural networks. In another example, by monitoring its own early dynamics a network can switch between recognizing known patterns and learning novel patterns [104]. Yet another implementation of this control mechanism allows a network to recognize the emotional quality of the stimulus in the pre-recognition process and use this emotional pre-recognition to facilitate the recognition of stimuli that are relevant to this emotion [102]. For an extensive model of how such loops are used in self-regulation, see Nowak and Vallacher [55] and also Vallacher and Nowak [91].

Modeling Fluency-Affect Interactions: The Influence of Specific Variables

So far, we have discussed computational models of fluency in terms of more general principles. In this section, we show that such models can be used to precisely specify the processing dynamics that underlie affective responses in several concrete empirical phenomena discussed earlier. To recall, experimental psychological research found that positive affect can be enhanced by repetition, priming, figure-ground contrast, presentation duration, symmetry, and prototypicality. How does this work computationally?

Repetition

Drogosz and Nowak [14] used a dynamic attractor neural network to simulate the effect of repetition on liking and explicit recognition. Specifically, they modeled the results of a study by Seamon, Marsh, and Brody [78] who exposed participants to 50 repetitions of polygons, presented at very brief exposure times ranging from 2 to 48 milliseconds. As in other mere exposure experiments, participants showed an increased preference for repeated polygons, even those presented at 2 and 8 milliseconds. Moreover, their preference increased with increasing exposure times, but reached asymptote at 24 milliseconds. In contrast, explicit recognition was at chance at low durations (2 and 8 milliseconds), and then gradually increased up to 90% recognition at 48 milliseconds. The model by Drogosz and Nowak [14] showed that the relationship between preference and recognition as a function of exposure time can be

simulated by assuming that the affective response represents a non-specific signal about the early dynamics of the network, as indexed by the estimated proportion of change in the first time step, whereas the recognition response represents a stabilization of the network on a specific pattern, which takes approximately 6 time steps. A psychological interpretation that can be attached to these simulation data is that at very short presentation durations, the participants only have access to the non-specific fluency signal, which elicits positive affect and influences their preference judgments. With progressively longer presentation duration, the fluency signal (affective response) increases only marginally, whereas the recognition response continues to grow until it reaches nearly perfect performance. The above simulations show that many prior exposures to a pattern establish a relatively strong memory for this pattern, whereas few prior exposures establish a relatively weak memory for the pattern. Test patterns with relatively stronger memories (i. e., stronger attractors) are processed with higher processing fluency (less volatility, more coherent signals) than test patterns with weaker or no memories. These differential fluency signals are picked up early on, as indicated by the simulation, and precede the extraction of stimulus information. Because the fluency signal is hedonically marked, it allows for evaluative responses prior to stimulus recognition, as initially reported by Kunst-Wilson and Zajonc [38].

Computational models of this type can also help us conceptualize the results of studies that used all novel patterns and manipulated the fluency of processing through procedures like figure-ground contrast, presentation duration, symmetry, prototypicality, and priming. To account for these effects, the model requires only minimal modifications. Specifically, the above simulations were carried out in attractor networks composed of neurons with binary states, where a state of the neuron corresponds either to the presence or the absence of the feature preferred by that neuron [26]. However, the same "fluency" criteria (volatility, coherence, differentiation) apply to networks with continuous neurons, where the state of a neuron encodes the degree to which a feature is present or activated [27,57].

Duration, Clarity and Contrast

The influence of these liking-enhancing variables can be conceptualized as reflecting a process in which patterns presented for longer time, greater clarity, and higher contrast are represented by more extreme values of activation. All this leads to stronger signals in the network, more differentiated states of the neurons, and faster settling.

Symmetry

This highly valued feature is easily incorporated because the representation of symmetrical patterns is stronger. This is due to simplicity and redundancy (e. g., in faces the left side of the symmetrical faces is identical to the right) and position-independence in recognition (e. g., symmetrical face looks the same from different angles). In contrast, the representation of asymmetrical features is weaker due to complexity and position-dependence [15,35].

Prototypicality

The effects of prototypicality (responsible for the ‘beauty-in-averages’ effect) can result from converging exemplars creating a strong attractor for a prototype. As a result, the recognition of a prototype pattern typically involves faster settling time, and less volatility [97]. Recent computational models of fluency using a support vector machine (a nonlinear classifier) have also shown that prototypical faces are located further from a face/non-face classification boundary, which allows for more efficient categorization [72].

Priming

In neural networks, priming corresponds either to the pre-activation of neurons that encode the pattern (activation-based priming) or to temporary changes in weights between the neurons (weight-based priming). The effects of the prime and the actual target sum up in determining the state of neurons. This results in more extreme values of activation (i. e., better differentiation) of the neurons for primed versus non-primed patterns.

As mentioned previously, fluency might be better captured by models that simulate the real-time, millisecond by millisecond processing of information. With such a model, we have explained a variety of empirical priming phenomena by including habituation dynamics (decrease in responding as a result of repetition or strong activation). In this model the presentation of a minimal prime (i. e., a brief duration or a single stimulus) immediately prior to a target induces a positive priming effect through pre-activation that boosts fluency. This is similar to the above mentioned pre-activation model. However, our model also predicted that presentation of an excessive prime (i. e., a long duration or repeated stimulus) immediately prior to a target would eliminate the positive priming effect, or perhaps even induce a negative priming effect. This occurs because habituation to the prime produces a disfluency in processing the target (i. e., the response to the target occurs slowly). This transition from positive to neg-

ative priming as a function of prime duration explained a variety of priming phenomena in the domain of word identification [28] and recognition memory [31].

We recently demonstrated in several experiments that this fluency-disfluency dynamic also applies to the domain of evaluation, more specifically, the appearance and disappearance of evaluative priming effects [32]. These experiments explored predictions for situations that should produce or eliminate priming as a function of prime duration, prime-target similarity and target salience. More specifically, when the prime-target similarity is low, such as with extremely valenced prime words and ideograph targets that have no meaning, habituation to the prime does not produce habituation to the target, and so empirically there is no correction effect even with long duration primes. Furthermore, when the target is itself minimal (e. g., a subliminally presented target word), then there is only an assimilative pre-activation effect because, again, habituation to the prime does not cause a change in target processing. In short, the fluency-based priming models can explain not only when the evaluative priming occurs, but also when it disappears.

In sum, the just discussed computational models show that manipulations such as repetition, priming, presentation duration, figure-ground contrast, clarity, prime-target similarity, and prototypicality change fluency in the network dynamics. These changes in fluency can trigger an affective response via the monitoring mechanisms discussed earlier.

Neural Basis of Fluency–Affect Connection

The assumptions guiding the just discussed psychological and computational models are consistent with neuroscience evidence. So far, this work has focused on low-level perceptual effects of stimulus repetition, response of higher-level value-coding system to previously exposed stimuli, and effects of processing coherence and conflict.

Perceptual Response

There is much evidence that novel stimuli elicit a non-specific, undifferentiated activity, which gradually decreases with repetition [82,85]. More specifically, single cell recording and neuroimaging studies suggest that stimulus repetition tends to decrease non-specific activation and leads to more selective firing [13,71]. One interpretation of these data is that stimulus familiarization leads to a gradual differentiation of the neurons that represent the incoming stimulus from neurons that do not represent the stimulus [48,54].

Response of Value-Coding Regions

A specific example of neuroscience research that examined connections between novelty and evaluation comes from Elliot and colleagues. Elliot, Dolan, and Frith [17] reported an fMRI study on the neural substrates of delayed matching to sample, as compared with delayed non-matching to sample. In such a task, participants are initially shown an item and then are subsequently shown a pair of items whereby they have to identify either the item they saw or the one that they did not see. Significantly more activity occurred in the ventromedial OFC in the matching condition (while reprocessing an old item) as compared to the non-matching condition (while processing a novel item). This conclusion is consistent with results of an earlier PET study of the subliminal mere exposure effect by Elliot and Dolan [16]. In their study, when participants made preference judgments, repeated stimuli activated the medial PFC, an area closely connected to the medial OFC and that is also known for its role in reward processing. Notably, these neuroimaging studies showing activation of neural circuits involved in reward to repeated stimuli fit nicely with the earlier reviewed Harmon-Jones and Allen [25] EMG study showing greater zygomaticus activation to merely exposed items. Taken together, they highlight the multilevel hedonic consequences of mere exposure and, more generally, high processing fluency.

Processing Coherence and Conflict

There is also work on the neural basis of mechanisms involved in successful and unsuccessful integration of different cognitive representations [10]. Neuroimaging evidence highlights a particular role of the anterior cingulate cortex (ACC) [18,39]. Though originally thought of as primarily a “cognitive” structure, more recent studies suggest that enhanced ACC due to cognitive conflict is accompanied by negative affect and enhanced arousal [10]. If so, the ACC could provide a neural substrate by which processing coherence on the level of multiple representations translates into negative affect.

Future Directions

Several issues remain essential for further studies. First, we primarily focused on the role of perceptual sources of dynamical information. However, dynamical information is available across the entire processing spectrum, from simple pattern matching to semantic coherence of high-order conceptual content. Though there is some psychological work available on this issue, there is very little computational and neural work that deals with specific mechanisms [97].

Second, emerging evidence in psychology suggests that the impact of fluency is moderated by processing expectations – how much speed, effort, coherence is expected given the stimulus [93]. For example, the simulations by Drogosz and Nowak [14] discussed earlier were conducted using very similar patterns, as is typical in the mere-exposure studies. Accordingly, the absolute processing fluency of a given pattern was a reliable indicator of its “oldness.” However, for the fluency signal to be informative in a more realistic situation, in which stimuli differ widely in overall signal strength, the network needs to scale the absolute value of the fluency signal for the particular pattern against the expected value [93]. A comparison between an observed value and an expected value can be derived with a computational likelihood ratio model, which is a class of model that has proven remarkably successful in explaining recognition memory based on familiarity [48,80]. Developing a similar Bayesian approach to fluency would likewise provide a precise account for the role of expectations in the study of affect. Finally, this contribution has emphasized the role of non-specific dynamical information in evaluation and has been silent on the role of specific stimulus features. However, we know very little about the interaction the proposed fluency based “how” and the content based “what” of processing, and exploring this interaction may prove a useful direction for future research.

Bibliography

Primary Literature

1. Anderson NH (1981) Foundations of information integration theory. Academic Press, New York
2. Bar M, Biederman I (1998) Subliminal visual priming. *Psychol Sci* 9:464–469
3. Beeman M, Ortony A, Monti LA (1995) Emotion-cognition interactions. In: Arbib MA (ed) The handbook of brain theory and neural networks. MIT Press, Cambridge, pp 360–363
4. Berlyne DE (1974) Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Hemisphere, Washington
5. Bornstein RF (1989) Exposure and affect: Overview and meta-analysis of research, 1968-1987. *Psychol Bull* 106:265–289
6. Bornstein RF, D’Agostino PR (1994) The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Soc Cogn* 12:103–128
7. Cacioppo JT, Bush LK, Tassinary LG (1992) Microexpressive facial actions as a function of affective stimuli: Replication and extension. *Personality Soc Psychol Bull* 18:515–526
8. Carver CS, Scheier MF (1990) Origins and functions of positive and negative affect: A control-process view. *Psychol Rev* 97:19–35

9. Checkosky SF, Whitlock D (1973) The effects of pattern goodness on recognition time in a memory search task. *J Exp Psychol* 100:341–348
10. Critchley HD (2005) Neural mechanisms of autonomic, affective, and cognitive integration. *J Comp Neurol* 493:154–166
11. Damasio AR (1994) *Descartes' error: Emotion, reason and the human brain*. Grosset/Putnam, New York
12. Derryberry D, Tucker DM (1994) Motivating the focus of attention. In: Niedenthal PM, Kitayama S (eds) *The heart's eye*. Academic Press, San Diego, pp 167–196
13. Desimone R, Miller EK, Chelazzi L, Lueschow A (1995) Multiple memory systems in the visual cortex. In: Gazzaniga MS (ed) *The cognitive neurosciences*. MIT Press, Cambridge, pp 475–490
14. Drogosz M, Nowak A (2006) A neural model of mere exposure: The EXAC mechanism. *Pol Psychol Bull* 37:7–15
15. Enquist M, Arak A (1994) Symmetry, beauty and evolution. *Nature* 372:169–172
16. Elliott R, Dolan R (1998) Neural response during preference and memory judgments for subliminally presented stimuli: A functional neuroimaging study. *J Neurosci* 18:4697–4704
17. Elliot R, Dolan RJ, Frith CD (2000) Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cereb Cortex* 10:308–317
18. Fernandez-Duque D, Baird JA, Posner MI (2000) Executive attention and metacognitive regulation. *Conscious Cogn* 9:288–307
19. Galton F (1878) Composite portraits. *J Anthropol Inst G B Irel* 8:132–144
20. Garcia-Marques T, Mackie DM (2000) The positive feeling of familiarity: Mood as an information processing regulation mechanism. In: Bless H, Forgas J (eds) *The message within: The role of subjective experience in social cognition and behavior*. Psychology Press, Philadelphia, pp 240–261
21. Gigerenzer G (2007) *Gut feelings: The intelligence of the unconscious*. Viking Press, New York
22. Haber RN, Hershenson M (1965) The effects of repeated brief exposures on growth of a percept. *J Exp Psychol* 69:40–46
23. Halberstadt J, Rhodes G (2000) The attractiveness of nonface averages: Implications for an evolutionary explanation of the attractiveness of average faces. *Psychol Sci* 4:285–289
24. Harmon-Jones E (2000) A cognitive dissonance theory perspective on the role of emotion in the maintenance and change of beliefs and attitudes. In: Frijda NH, Manstead ARS, Bem S (eds) *Emotion and Beliefs*. Cambridge University Press, Cambridge, pp 185–211
25. Harmon-Jones E, Allen JB (2001) The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality Soc Psychol Bull* 27:889–898
26. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 79:2554–2558
27. Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Natl Acad Sci* 81:3088–3092
28. Huber DE (2008) Immediate priming and cognitive aftereffects. *J Exp Psychol Gen* 137:324–347
29. Huber DE, Cousineau D (2004) A race model of perceptual forced choice reaction time. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Erlbaum Associates, Hillsdale, pp 687–692
30. Huber DE, O'Reilly RC (2003) Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cogn Sci A Multidiscip J* 27:403–430
31. Huber DE, Clark TF, Curran T, Winkelman P. Effects of repetition priming on recognition memory: Testing a perceptual fluency-disfluency model. *J Exp Psychol Learn Mem Cogn* (in press)
32. Huber DE, Winkelman P, Parsa A, Chun WY. Too much of a good thing: Testing a Bayesian model of evaluative priming. Submitted
33. Jacoby LL (1983) Perceptual enhancement: Persistent effects of an experience. *J Exp Psychol Learn Mem Cogn* 9:21–38
34. Jacoby LL, Kelley CM, Dywan J (1989) Memory attributions. In: Roediger HL, FIM Craik (eds) *Varieties of memory and consciousness: Essays in honour of Endel Tulving*. Erlbaum, Hillsdale, pp 391–422
35. Johnstone RA (1994) Female preference for symmetrical males as a by-product of selection for mate recognition. *Nature* 372:172–175
36. Klinger MR, Greenwald AG (1994) Preferences need no inferences?: The cognitive basis of unconscious mere exposure effects. In: Niedenthal PM, Kitayama S (eds) *The heart's eye*. Academic Press, San Diego, pp 67–85
37. Koriat A (2000) The feeling of knowing: Some metatheoretical implications for consciousness and control. *Conscious Cogn* 9:149–171
38. Kunst-Wilson WR, Zajonc RB (1980) Affective discrimination of stimuli that cannot be recognized. *Science* 207:557–558
39. Lane RD, Reiman EM, Axelrod B, Yun L, Holmes A, Schwartz GE (1998) Neural correlates of levels of emotional awareness: Evidence of an interaction between emotion and attention in the anterior cingulate cortex. *J Cogn Neurosci* 10:525–535
40. Langlois JH, Roggman LA (1990) Attractive faces are only average. *Psychol Sci* 1:115–121
41. LeDoux JE (1996) *The Emotional Brain*. Touchstone, New York
42. Lewenstein M, Nowak A (1989) Recognition with self-control in neural networks. *Phys Rev* 40:4652–4664
43. Losch ME, Cacioppo JT (1990) Cognitive dissonance may enhance sympathetic tonus, but attitudes are changed to reduce negative affect rather than arousal. *J Exp Soc Psychol* 26: 289–304
44. Mackworth JF (1963) The duration of the visual image. *Canadian J Psychol* 17:62–81
45. Mandler G, Nakamura Y, Van Zandt BJ (1987) Nonspecific effects of exposure on stimuli that cannot be recognized. *J Exp Psychol Learn Mem Cogn* 13:646–648
46. Martindale C, Moore K (1988) Priming, prototypicality, and preference. *J Exp Psychol Hum Percept Perform* 14:661–670
47. Mazzoni G, Nelson TO (1998) Metacognition and cognitive neuropsychology: Monitoring and control processes. Lawrence Erlbaum, Mahwah
48. McClelland JL, Chappell M (1998) Familiarity breeds differentiation: A Bayesian approach to the effects of experience in recognition memory. *Psychol Rev* 105:724–760
49. Metcalfe J (1993) Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff Amnesia. *Psychol Rev* 100:3–22

50. Metcalfe J, Shimamura AP (1994) *Metacognition: Knowing about knowing*. MIT Press, Cambridge
51. Monahan JL, Murphy ST, Zajonc RB (2000) Subliminal mere exposure: Specific, general, and diffuse effects. *Psychol Sci* 6:462–466
52. Murre JMJ, Phaf RH, Wolters G (1992) CALM: Categorizing and learning module. *Neural Netw* 5:55–82
53. Norman KA, O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol Rev* 110:611–646
54. Norman KA, O'Reilly RC, Huber DE (2000) Modeling hippocampal and neocortical contributions to recognition memory. Poster presented at the Cognitive Neuroscience Society Meeting, San Francisco
55. Nowak A, Vallacher RR (1998) *Dynamical social psychology*. Guilford Press, New York
56. Oatley K, Johnson-Laird P (1987) Towards a cognitive theory of emotions. *Cogn Emot* 1:29–50
57. O'Reilly RC, Munakata Y (2000) *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT Press, Cambridge
58. Palmer SE (1991) Goodness, gestalt, groups, and Garner: Local symmetry subgroups as a theory of figural goodness. In: Pomerantz JR, Lockhead GR (eds) *Perception of Structure*. APA, Washington
59. Palmer SE, Hemenway K (1978) Orientation and symmetry: Effects of multiple, near, and rotational symmetries. *J Exp Psychol Hum Percept Perform* 4:691–702
60. Phaf RH, Roteveel M (2005) Affective modulation of recognition bias. *Emotion* 5(3):309–318
61. Port RT, van Gelder T (1995) *Mind as motion: Exploration in the dynamics of cognition*. MIT Press, Cambridge
62. Posner MI, Keele SW (1968) On the genesis of abstract ideas. *J Exp Psychol* 77:353–363
63. Ramachandran VS, Hirstein W (1999) The science of art: A neurological theory of aesthetic experience. *J Conscious Stud* 6:15–51
64. Reber R, Schwarz N (2006) Perceptual fluency, preference, and evolution. *Pol Psychol Bull* 37:16–22
65. Reber R, Winkielman P, Schwarz N (1998) Effects of perceptual fluency on affective judgments. *Psychol Sci* 9:45–48
66. Reber R, Schwarz N, Winkielman P (2004) Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality Soc Psychol Rev* 8:364–382
67. Rhodes G (2006) The evolution of facial attractiveness. *Annu Rev Psychol* 57:199–226
68. Rhodes G, Tremewan T (1996) Averageness, exaggeration, and facial attractiveness. *Psychol Sci* 7:105–110
69. Rhodes G, Proffitt F, Grady JM, Sumich A (1998) Facial symmetry and the perception of beauty. *Psychon Bull Rev* 5:659–669
70. Roediger HL (1990) Implicit memory: Retention without remembering. *American Psychol* 45:1043–1056
71. Rolls ET, Baylis GC, Hasselmo ME, Nalwa V (1989) The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Exp Brain Res* 76:153–164
72. Rosen LH, Bronstad PM, Griffin AM, Hoss RA, Langlois JH. Children, adults, and a computational model identify attractive faces more fluently: Evidence in support of averageness theory of facial attractiveness (under review)
73. Rumelhart DE, McClelland JL (1986) *Parallel Distributed Processes: Exploration in Microstructure of Cognition*. MIT Press, Cambridge
74. Schachter SE, Singer J (1962) Cognitive, social and physiological determinants of emotional state. *Psychol Rev* 69:379–399
75. Schacter DL (1992) Understanding implicit memory: A cognitive neuroscience approach. *American Psychol* 47:559–569
76. Schwarz N (1998) Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality Soc Psychol Rev* 2:87–99
77. Schwarz N, Clore GL (1996) Feelings and phenomenal experiences. In: Higgins ET, Kruglanski AW (eds) *Social Psychology: Handbook of Basic Principles*. The Guilford Press, New York
78. Seamon JG, Marsh RL, Brody N (1984) Critical importance of exposure duration for affective discrimination of stimuli that are not recognized. *J Exp Psychol Learn Mem Cogn* 10:465–469
79. Seamon JG, McKenna PA, Binder N (1998) The mere exposure effect is differentially sensitive to different judgment tasks. *Conscious Cogn* 7:85–102
80. Shiffrin RM, Steyvers M (1997) A model for recognition memory: REM: Retrieving effectively from memory. *Psychon Bull Rev* 4(2):145–166
81. Simon HA (1967) Motivational and emotional controls of cognition. *Psychol Rev* 74:29–39
82. Skarda CA, Freeman WJ (1987) How brains make chaos in order to make sense of the world. *Behav Brain Sci* 10:161–195
83. Smith ER (1998) Mental representation and memory. In: Gilbert DT, Fiske ST, Lindzey G (eds) *The Handbook of Social Psychology* pp 269–322; The McGraw-Hill Companies, Boston
84. Smith ER (2000) Subjective experience of familiarity: Functional basis in connectionist memory. In: Bless H, Forgas JP (eds) *The message within: The role of subjective experience in social cognition and behavior*. Psychology Press, Philadelphia, pp 109–124
85. Sokolov EN (1963) *Perception and the orienting reflex*. MacMillan, NY
86. Solso RL (1997) *Cognition and the visual arts*. MIT Press, Cambridge
87. Squire LR (1992) *Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans*. *Psychol Rev* 99:195–231
88. Symons D (1979) *Evolution of human sexuality*. Oxford University Press, New York
89. Thornhill R, Gangestad SW (1993) Human facial beauty: Averageness, symmetry, and parasite resistance. *Hum Nat* 4:237–269
90. Titchener EB (1910) *A textbook of psychology*. Macmillan, New York
91. Vallacher RR, Nowak A (1999) The dynamics of self-regulation. In: Jr. Wyer RS (ed) *Perspectives on behavioral self-regulation*. Lawrence Erlbaum Associates, Mahwah, pp 241–259
92. Whittlesea BWA (1993) Illusions of familiarity. *J Exp Psychol Learn Mem Cogn* 19:1235–1253
93. Whittlesea BWA, Williams LD (2001) The Discrepancy-Attribution Hypothesis: I. The Heuristic Basis of Feelings of Familiarity. *J Exp Psychol Learn Mem Cogn* 27:3–13
94. Winkielman P, Berntson GG, Cacioppo JT (2001) The psychophysiological perspective on the social mind. In: Tesser A,

- Schwarz N (eds) Blackwell Handbook of Social Psychology: Intra-individual Processes. Blackwell, Oxford, pp 89–108
95. Winkielman P, Cacioppo JT (2001) Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation leads to positive affect. *J Personality Soc Psychol* 81:989–1000
 96. Winkielman P, Halberstadt J, Fazendeiro T, Catty S (2006) Prototypes are attractive because they are easy on the mind. *Psychol Sci* 17:799–806
 97. Winkielman P, Hooda P, Munakata Y (2004) Neural network model of fluency for average patterns. Unpublished manuscript. University of Denver
 98. Winkielman P, Schwarz N (2001) How pleasant was your childhood? Beliefs about memory shape inferences from experienced difficulty of recall. *Psychol Sci* 2:176–179
 99. Winkielman P, Schwarz N, Fazendeiro T, Reber R (2003) The hedonic marking of processing fluency: Implications for evaluative judgment. In: Musch J, Klauer KC (eds) *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*. Lawrence Erlbaum, Mahwah, pp 189–217
 100. Zajonc RB (1968) Attitudinal effects of mere exposure. *J Personality Soc Psychol Monogr Suppl* 9:1–27
 101. Zajonc RB (1998) Emotions. In: Gilbert DT, Fiske ST, Lindzey G (eds) *The Handbook of Social Psychology*. McGraw-Hill, Boston, pp 591–632
 102. Zochowski M, Lewenstein M, Nowak A (1993) A memory which tentatively forgets. *J Phys A* 26:2099–2112
 103. Zochowski M, Lewenstein M, Nowak A (1994) Local noise in neural networks with self-control. *International J Neural Syst* 5:287–298
 104. Zochowski M, Lewenstein M, Nowak A (1995) SMARTNET – A neural net with self-controlled learning. *Network* 6:93

Books and Reviews

- Reber R, Schwarz N, Winkielman P (2004) Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality Soc Psychol Rev* 8:364–382
- Winkielman P, Schwarz N, Fazendeiro T, Reber R (2003) The hedonic marking of processing fluency: Implications for evaluative judgment. In: Musch J, Klauer KC (eds) *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*. Lawrence Erlbaum, Mahwah, pp 189–217

Dynamics on Fractals

RAYMOND L. ORBACH

Department of Physics, University of California,
Riverside, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Fractal and Spectral Dimensions

Nature of Dynamics on Fractals – Localization

Mapping of Physical Systems onto Fractal Structures

Relaxation Dynamics on Fractal Structures

Transport on Fractal Structures

Future Directions

Bibliography

Glossary

Fractal Fractal structures are of two types: deterministic fractals and random fractals. The former involves repeated applications of replacing a given structural element by the structure itself. The process proceeds indefinitely, leading to *dilation symmetry*: if we magnify part of the structure, the enlarged portion looks the same as the original. Examples are the Mandelbrot–Given fractal and the Sierpinski gasket. A random fractal obeys the same properties (e.g. dilation symmetry), but only in terms of an ensemble average. The stereotypical example is the percolating network. Fractals can be constructed in any dimension, d , with for example, $d = 6$ being the mean-field dimension for percolating networks.

Fractal dimension The fractal dimension represents the “mass” dependence upon length scale (measuring length). It is symbolized by D_f , with the number of sites on a fractal as a function of the measurement length L being proportional to L^{D_f} , in analogy to a homogeneous structure embedded in a dimension d having mass proportional to the volume spanned by a length L proportional to L^d .

Spectral (or fracton) dimension The spectral (or fracton) dimension refers to the dynamical properties of fractal networks. It is symbolized by \tilde{d}_s and can be most easily thought of in terms of the density of states of a dynamical fractal structure (e.g. vibrations of a fractal network). Thus, if the excitation spectrum is measured as a function of energy ω , the density of states for excitations of a fractal network would be proportional to $\omega^{(\tilde{d}_s-1)}$, in analogy to a homogeneous structure embedded in a dimension d having a density of states proportional to $\omega^{(d-1)}$.

Localization exponent Excitations on a fractal network are in general *strongly localized* in the sense that wave functions fall off more rapidly than a simple exponential, $\Psi(r) \sim \exp[-\{r/\Lambda(\omega)\}^{d_\phi}]$ where $\Lambda(\omega)$ is an energy dependent localization length, and the exponent d_ϕ is in general greater than unity.

Definition of the Subject

The dynamical properties of fractal networks are very different from homogeneous structures, dependent upon

strongly localized excitations. The thermal properties of fractals depend upon a “spectral dimension” \tilde{d}_s less than the “Euclidean” or embedding dimension d . Fractal randomness introduces statistical properties into relaxation dynamics. Transport takes place via hopping processes, reminiscent of Mott’s variable range-rate hopping transport for electronic impurity states in semi-conductors. Fractal dynamics serve as a guide for the behavior of random structures where the short length scale excitations are localized.

Introduction

The study of the dynamics on and of fractal networks [29, 30, 31] is not an arcane investigation, with little application to real physical systems. Rather, there are many examples of real materials that exhibit fractal dynamics. Examples are the vibrational properties of silica aerogels [21, 45], and magnetic excitations in diluted antiferromagnets [18, 19, 43]. Beyond these explicit physical realizations, one can learn from the very nature of the excitations on fractal structures how localized excitations in homogeneous materials behave. That is, the dynamics of fractal networks serve as a calculable model for transport of localized excitations in random structures that are certainly not mass fractals. Examples are thermal transport in glasses above the so-called “plateau” temperature [2, 20, 26], and the lifetime of high-energy lattice vibrations in a-Si [32, 33, 36, 37]. As we shall show in subsequent sections, the former is an example of vibrational hopping-type transport (similar in nature to Mott’s variable-range hopping [22, 23] for localized electronic states), while the latter exhibits counter-intuitive energy dependencies for vibrational lifetimes.

The following Sect. “[Fractal and Spectral Dimensions](#)” describes in brief terms the concept of the fractal (or mass) dimension, D_f [7, 8, 9, 10, 15, 40, 41, 46], and the spectral (or fracton) dimension \tilde{d}_s [5, 14, 17, 34, 35]. The latter is related to the anomalous diffusion characteristics of fractals (the so-called “ant in a labyrinth” introduced by de Gennes [8]), and is at the heart of the thermal properties of fractal structures. A conjecture introduced by Alexander and Orbach [5] suggested that the mean-field value (exact at $d = 6$) for \tilde{d}_s for percolating networks, $\tilde{d}_s = 4/3$, might be universal, independent of dimension. It is now generally regarded that this conjecture is only approximate, though a very good approximation for percolating networks in $2 = d < 6$. Were it to be exact, it would mean that the dynamical properties of percolating networks could be expressed in terms of their geometrical properties.

Section “[Nature of Dynamics on Fractals–Localization](#)” introduces the nature of the wave function for an excitation embedded in a fractal structure. In general, random self-similar structures generate “super-localized” wave functions, falling off faster than exponential. We will characterize the envelope of the wave function by the functional form [1]

$$\Psi(r) \sim \exp[-\{r/\Lambda(\omega)\}^{d_\phi}],$$

where $\Lambda(\omega)$ is an energy-dependent localization length, and the exponent d_ϕ is in general greater than unity. For excitations on fractal structures, the localization length $\Lambda(\omega)$ can be calculated analytically, allowing for explicit computation of physical phenomena. This allows analytic expressions for scattering lifetimes and excitation transport.

Mapping of the diffusion problem onto the secular equation for scalar lattice vibrations allows the extraction of the density of vibrational states [5]. It is given by the simple relationship $D(\omega) \sim \omega^{(\tilde{d}_s-1)}$ displaying the utility of the spectral dimension \tilde{d}_s . A similar though slightly more complicated mapping can be applied to diluted Heisenberg antiferromagnets. The silica aerogels [13, 16, 42, 44, 45], and $\text{Mn}_x\text{Zn}_{1-x}\text{F}_2$ [43] and $\text{RbMn}_x\text{Mg}_{1-x}\text{F}_3$ [18, 19] are realizations of such systems, respectively, and have been extensively investigated experimentally.

The random nature of fractal structures introduces statistical complexity into relaxation processes. The direct (single vibration) [38] and Raman (scattering of two vibrations) [24, 39] spin-lattice relaxation processes are discussed for random fractal networks in Sect. “[Relaxation Dynamics on Fractal Structures](#)”. The theoretical results are complex, related to the “Devil’s staircase” for Raman relaxation processes. Experiments probing these rather bizarre behaviors would be welcome.

Given the “super-localization” of excitations on fractal networks [1, 42], transport properties need to be expressed according to the same concepts as introduced by Mott [22, 23] for the insulating phase of doped semi-conductors—the so-called “variable range-rate hopping”. This is developed in Sect. “[Transport on Fractal Structures](#)”, and applied to the thermal transport of fractal structures [2, 20, 26]. The “lessons learned” from these applications suggest an interpretation of the thermal transport of glassy materials above the so-called “plateau” temperature range. This instance of taking the insights from fractal structures, and applying it to materials that are certainly not fractal yet possess localized vibrational states, is an example of the utility of the analysis of fractal dynamics.

Finally, some suggestions, hinted at above, for future experimental investigations are briefly discussed in Sect. “Future Directions”.

Fractal and Spectral Dimensions

Fractals can be defined by the nature of the symmetry they exhibit: self-similar geometry or equivalently dilation symmetry [7,8,9,10,15,40,41,46]. This symmetry is most assuredly not translational, leading to a very different set of behaviors for physical realizations that shall be explored in subsequent Sections. Stated most simply, self-similar structures “look the same” on any length scale. Conversely, one cannot extract a length scale from observation of a self-similar structure. There are clearly limits to self-similarity for physical realizations. For example, at the atomic level, dilation symmetry must come to an end. And for finite structures, there will be a largest length scale. There are other length scales that bracket the self-similar regime. Thus, for percolating networks, an example of a random fractal [38,39], the percolation correlation length, $\xi(p)$, where p is the concentration of occupied sites or bonds, defines an upper limit to the regime of fractal geometry. In general, for percolating networks, the range of length scales ℓ over which the structure is fractal is sandwiched between the atomic length a , and the percolation correlation length $\xi(p)$: $a < \ell < \xi(p)$.

Fractal structures can be deterministic or random. The former is a result of applying a structural rule indefinitely, replacing an element of a design by the design itself. Examples are the Mandelbrot–Given fractal and the Sierpinski gasket. The latter arises from a set of probabilistic rules. A simple example is the percolating network where a site or bond (e.g. in $d = 2$, a square grid) is occupied randomly with probability p . The resulting clusters contain a random number of sites or bonds, where the distribution function for the number of finite clusters of size s is denoted by $n_s(p)$. There exists a critical value $p = p_c$ where a connected cluster extends over all space (the “infinite” cluster). For $p < p_c$ only finite clusters exist, with the largest cluster spanned by the length scale $\xi(p)$. $\xi(p)$ diverges at $p = p_c$ as $|p - p_c|^{-\nu}$. For $p > p_c$ finite clusters and the infinite cluster coexist, the number of finite clusters vanishing at p approaches unity. The probability that a site belongs to the infinite cluster varies with occupation probability p as $(p - p_c)^\beta$.

The characteristic length $\xi(p)$ for $p > p_c$ is a measure of the largest finite cluster size. Hence, for length scales $\ell > \xi(p)$ the systems looks homogeneous. Thus, the fractal regime is sandwiched between the lattice constant a and the percolation correlation length $\xi(p)$. The remarkable

utility of the percolating network becomes clear from these properties: there is a crossover length scale between fractal and homogeneous structural behavior that can be chosen at will by choosing a value for the site or bond occupation probability $p > p_c$. Excitations of structures that map onto percolating networks can change their nature from fractal to continuous as a function of their length scale ℓ , the transition occurring when $\ell_c \sim \xi(p)$. This property makes the percolating network the “fruit fly” of random structures. The flexibility of adjusting the crossover length ℓ_c enables a fit between those properties calculated for a percolating network with the properties of those physical systems that map onto percolating networks.

The fractal dimension, D_f , is defined from the dependence of the “mass” on length scale. For fractals, the number of points or bonds on the infinite cluster within a length scale ℓ (hence, the “mass”), for $\ell \ll \xi(p)$, depends upon ℓ as $M(\ell) \sim \ell^{D_f}$. From above it is straight forward to show $D_f = d - (\beta/\nu)$. For a percolation network [29] in $d = 2$, $D_f = 91/48$. For $d = 3$, $D_f = 2.48 \pm 0.09$. In mean field, $d = 6$, $D_f = 4$.

The spectral dimension, \tilde{d}_s , follows from the analysis of diffusion on a fractal network, as first postulated by de Gennes [14]: “An ant parachutes down onto an occupied site of the infinite cluster of a percolating network. At every time unit, the ant makes one attempt to jump to one of its adjacent sites. If that site is occupied, it moves there. If it is empty, the ant stays at its original site. What is the ensemble-averaged square distance that the ant travels in time t ?” Gefen et al. [17] found that $\langle r^2(t) \rangle \sim t^{2+\theta}$ where θ depends upon the scaling exponent μ (conductivity exponent), the probability that a site belongs to the infinite cluster scaling exponent β , and the scaling exponent ν for $\xi(p)$: $\theta = (\mu - \beta)/\nu$. Alexander and Orbach [5] defined the spectral (or fracton) dimension as $\tilde{d}_s = 2D_f/(2 + \theta)$. The importance of \tilde{d}_s can be seen from the calculation of the probability of finding a diffusing particle at the starting point at time t , $P_0(t)$, which for compact diffusion is the inverse of the number of visited sites in time t , $V(t) \sim t^{(\tilde{d}_s/2)}$. For scalar elasticity, the vibrational density of states

$$D(\omega) = -(2\omega/\pi) \text{Im} \tilde{P}_0(-\omega^2 + i0^+) \sim \omega^{(\tilde{d}_s-1)},$$

where $\tilde{P}_0(\omega)$ is the Laplace transform of $P_0(t)$. Noting that for homogeneous systems, $D(\omega) \sim \omega^{(d-1)}$, the name spectral dimension for \tilde{d}_s becomes evident. Alexander and Orbach [5] named the vibrational excitations of fractal structures, fractons, in analogy with the vibrational excitations of homogeneous structures, phonons. Hence, they termed \tilde{d}_s the fracton dimension.

Alexander and Orbach noted that for percolation structures in the mean-field limit, $d = 6$, $\tilde{d}_s = 4/3$ precisely. At that time (1982), values for \tilde{d}_s for $d < 6$ appeared close to that value, and they conjectured that $\tilde{d}_s = 4/3$ for $2 \leq d \leq 6$. We now know that this is only approximate (but remarkably close [24]): $\tilde{d}_s = 1.325 \pm 0.002$ in $d = 2$, and 1.317 ± 0.003 in $d = 3$. Were the conjecture exact, a numerical relationship between structure and transport would be exact [$\theta = (3D_f/2) - 2$], dictated solely by fractal geometry.

For percolating networks, the structure appears fractal for length scales $\ell \leq \xi(p)$ and homogeneous for $\ell \geq \xi(p)$. The vibrational excitations in the fractal regime are termed fractons, and in the homogeneous regime phonons. We define the crossover frequency ω_c as that for which the excitation length scale equals $\xi(p)$. Then $\omega_c \sim (p - p_c)^{(vD_f/\tilde{d}_s)}$. Continuity leads to a phonon velocity $v(p) \sim (p - p_c)^v [(D_f/\tilde{d}_s) - 1]$. This then leads to the dispersion relations:

$$[\ell \gg \xi(p), \omega \ll \omega_c, \text{ phonon regime}] \omega \sim v(p)k,$$

and

$$[\ell \ll \xi(p), \omega \gg \omega_c, \text{ fracton regime}] \omega \sim k^{(D_f/\tilde{d}_s)}.$$

In the latter case, as we shall show in the next Section, vibrational excitations in the fractal regime are localized, so that k should not be thought of as a wave vector but rather the inverse of the localization length $\Lambda(\omega)$.

Nature of Dynamics on Fractals – Localization

The conductance of a d dimensional percolating network of size L , $G(L)$, can be shown to be proportional to L^β where the exponent $\beta = (D_f/\tilde{d}_s)(\tilde{d}_s - 2)$. In the Anderson sense [6], localization occurs when $\beta \leq 0$ (marginal at $\beta = 0$). The Alexander–Orbach conjecture [5], $\tilde{d}_s \approx 4/3$, leads to β well less than zero for all embedding dimensions d . One can think of this form of localization as geometrical, as opposed to the scattering localization that traditionally is associated with Anderson localization. The localization is strong, the wave function falling off faster than simple exponential.

The wave function has the form [1]: $\langle \Psi(r) \rangle \sim \exp[-\{r/\Lambda(\omega)\}^{d_\phi}]$. The exponent d_ϕ is a geometrical exponent describing the fact that an exponential decay along the fractal will be distorted when viewed in real space. In general, $1 \leq d_\phi \leq d_{\min}$, where d_{\min} is defined by $\ell \sim R^{d_{\min}}$, with ℓ being the shortest path along the network between two points separated by a Pythagorean distance R . Bunde and Roman [11] have found that for percolating networks

$d_\phi = 1$. The decay length $\Lambda(\omega)$ is the localization length from Sect. “Introduction”, $\Lambda(\omega) \sim \omega^{(-\tilde{d}_s/D_f)}$.

For a random system such as a percolating network, a given realization of $\Psi(r)$ is badly behaved, depending upon the particular choice of origin taken for the position coordinate r . The ensemble average takes the average over all realizations of $\Psi(r)$, and is denoted by $\langle \Psi(r) \rangle$. This has its dangers, as the calculation of physical properties should be performed using a specific realization for $\Psi(r)$, and then ensemble averaged. Calculation of physical properties utilizing an ensemble averaged wave function $\langle \Psi(r) \rangle$ will not yield the same result, and could be very misleading.

Specific realizations for $\Psi(r)$ for two-dimensional bond percolation vibrational networks have been calculated by Nakayama and Yakubo [25,28,47]. The fracton “core” (or largest amplitude) possesses very clear boundaries for the edges of the excitation, with an almost step-like character and a long tail. The tail extends over large distances and the amplitude oscillates in sign. This is required for orthogonality to the uniform translational mode with $\omega = 0$. Clearly these modes do not look like simple exponentials, hence the warning about the use of $\langle \Psi(r) \rangle$ as compared to $\Psi(r)$ for the calculation of matrix elements.

Mapping of Physical Systems onto Fractal Structures

The foregoing description of dynamics of fractal networks would be interesting in and of itself, but its applicability to physical systems makes it of practical importance. Further, the nature of localization, and computation of “hopping-type” transport on fractal structures, has application to materials that are not fractal. Thus, it is worthwhile to consider what systems are fractal, the nature of dynamics on such systems, and the lessons we can learn from them for structures that possess similar characteristics but are not fractal. Specifically, Mott’s [22,23] variable range-rate hopping for localized electronic impurity states has direct application to transport of vibrational energy via fracton hopping on fractal networks [2], and we shall argue to thermal transport in glassy materials above the “plateau” temperature [20,26].

The fundamental equations for fractal dynamics spring from the equation for diffusion of a random walker on a network: $dP_i/dt = \sum_{j \neq i} w_{ij}(P_j - P_i)$ where P_i is the probability that the i th site is occupied, w_{ij} is the probability per unit time for hopping from site i to site j , and $w_{ij} = w_{ji}$. Introducing new quantities $W_{ij} = w_{ij}$ for $i \neq j$, and $W_{ii} = -\sum_{j \neq i} w_{ij}$, one can rewrite the diffu-

sion equation as $dP_i/dt = \sum_j W_{ij}P_j$ with the defined relation $\sum_j W_{ij} = 0$.

For scalar elasticity, the equation of motion for atomic vibrations is $d^2u_i/dt^2 = \sum_j K_{ij}u_j$, with u_i the atomic displacement at position i , and $K_{ij}(i \neq j)$ is given by $K_{ij} = k_{ij}/m_i$ with k_{ij} the force constant connecting two sites i and j , and m_i the mass of the i th site. The diagonal element W_{ii} is obtained from the balancing of forces at the i th site: $\sum_j K_{ij} = 0$.

The only difference between these two equations is the order of the time derivatives, leading to the equivalence of the eigenvalue problems. We have used this equivalence to extract the density of vibrational states for scalar elasticity, and the definition of the spectral (or fracton) dimension in Sect. “Introduction”.

The most direct physical example of a vibrating fractal network is the silica aerogel [13,16,42,44]. Different preparation conditions can change the density of these materials to very small values, indicative of the open porous structure of these materials. Two length scales are present: the fundamental building block of the structure (the silica), and the correlation length of the gel. In between, the clusters possess a fractal structure, while at length scales beyond the correlation length the gel behaves as a homogeneous porous glass. This crossover from fractal to homogeneous is the physical example of our previous discussion for percolating networks when the length scale of excitations passes through the percolation correlation length $\xi(p)$. The thermal, transport, and scattering properties of aerogels have been well studied, and are the classic example of fractal behavior.

There are other examples of structures that map onto the eigenvalue spectrum of the diffusion equation. Ferromagnetic and antiferromagnetic materials can be diluted by non-magnetic impurities. This leads to random site dilution, mapping directly onto the site-diluted percolating network for magnetic interactions. This can be seen from the Heisenberg Hamiltonian for spin systems: $H = 1/2 \sum_{i,j} J_{ij} \mathbf{S}_i \mathbf{S}_j$, where \mathbf{S}_i is the vector spin at the i th site, and J_{ij} the isotropic exchange interaction between sites i and j . The equation of motion for the spin operator $S_i^+ = S_i^x + iS_i^y$ becomes for diluted ferromagnets $i\hbar \partial S_i^+ / \partial t = \sum_{j \neq i} J_{ij} (S_i^z S_j^+ - S_j^z S_i^+)$. For low levels of spin wave excitations, $S_i^z \approx S_j^z \approx S$, so that one obtains the linear equation of motion $i\hbar \partial S_i^+ / \partial t = \sum_{j \neq i} J_{ij} (S_j^+ - S_i^+)$. This is the same form as that for diffusion of a random walker on a network, so that, in analogy with the scalar vibrational network, the eigenvalue spectrum is the same. This mapping allows the spin wave spectrum to exhibit magnons and fracton waves obeying the same prop-

erties that phonons and fractons exhibit for scalar vibrations. For antiferromagnets, the change in sign of S_i^z for the two sublattices complicates the linearized equations of motion. They become, $i\hbar \partial S_i^+ / \partial t = \sigma_i \sum_{j \neq i} J_{ij} (S_j^+ + S_i^+)$, where $\sigma_i = -1$ for down spins and $\sigma_i = 1$ for up spins. This alters the dynamical equations for diluted antiferromagnets, and requires a separate calculation of the eigenvalue spectrum.

Neutron diffraction experiments exhibit both conventional magnon and fracton excitations. In the case of the site-diluted antiferromagnet $\text{Mn}_x\text{Zn}_{1-x}\text{F}_2$ both excitations are observed simultaneously for length scales in the vicinity of $\xi(p)$ [43].

Relaxation Dynamics on Fractal Structures

Electron spin resonance and non-radiative decay experiments can be performed in fractal materials. It is of interest to compare these dynamical properties with those found in translationally ordered structures. At first sight, for vibrational relaxation, it might be thought that the only difference would originate with the differences in the densities of vibrational states $D(\omega)$. However, the localized nature of the vibrational states (fractons) and their random positions in the fractal network, introduces properties that are specific to fractal lattices. There are two cases of interest: one-fracton relaxation, and two-fracton inelastic scattering. The former is referred to as the “direct” relaxation process; the latter the “Raman” relaxation process. In both cases, the main effect of vibrational localization is on the relaxation time profile. Different spatial sites can have very different relaxation rates because of the randomness of their distance from the appropriate fracton sites. The result is a strongly non-exponential time decay.

For one-fracton relaxation [1], a straightforward generalization of the direct relaxation process rate, $W(\omega_0, L)$, for relaxation energy ω_0 arising from interaction with a fracton of the same energy centered a distance L away, is proportional to

$$W(\omega_0, L) \sim \omega_0^{2q-1} \{ [A(\omega_0)]^{(-D_f)} \} \cdot [\coth(\beta\omega_0/2)] (1/\delta_L) \exp\{-[L/\Lambda(\omega_0)]^{d_\phi}\}.$$

Here, $\beta = 1/k_B T$ and the factor δ_L represents the energy width of the fracton state. In the homogeneous limit, where phonons represent the extended vibrational states, an energy conserving delta function would replace δ_L . There are two limits for δ_L , case (a) when the fracton relaxation rate δ caused by anharmonicity dominates; and case (b) when the electronic relaxation rate caused by the electron-fracton interaction is greater than δ . In the for-

mer case, $\delta_L = \delta$; while in the latter, $\delta_L = W(\omega_0, L)$ itself, leading to a self-consistent determination of the relaxation rate. The latter case requires explicit consideration of the L dependence of δ_L , further complicating the calculation of $W(\omega_0, L)$. The calculation of the time profile of the electronic state population and the average relaxation rate are complex, and the reader is referred to [32] for details.

For case (a), the population of the initial electronic state is found to vary as

$$(\ln t)^{[(D_f - d_\phi)/2d_\phi]} \left[t^{\{-c_1(\ln t)^{[(D_f/d_\phi)-1]}\}} \right].$$

Here, c_1 is a constant independent of time, but dependent upon ω_0 and δ . The population of the initial electronic state decays is thus found to be faster than a power law, but slower than exponential or stretched exponential.

For case (b), the population of the initial electronic state is found to vary as

$$(1/t) \left\{ (\ln t)^{[(D_f/d_\phi)-1]} \right\}.$$

The decay is slower than in case (a), and is closer to a power law.

The average relaxation rates for cases (a) and (b) have the same functional form: $\langle W \rangle \sim D(\omega_0)[(\omega_0)^{(2q-1)}] \coth(\beta\omega_0/2)$ where $q = \tilde{d}_s(d_\phi/D_f)$. The temperature dependence is the usual one found for the direct process, but the energy dependence differs, depending on the values of the dimensions and parameters for fractal networks.

Two fracton relaxation is considerably more complex [3,4]. In summary, the time profile of the initial electronic state in the long time regime begins as a stretched exponential, then crosses over into a form similar to case (a) for the one fracton decay process. In the presence of rapid electronic cross relaxation, the time profile is exponential, with a low-temperature relaxation time $\langle 1/T_1 \rangle$ proportional to $T^{\{2\tilde{d}_s[1+2(d_\phi/D_f)]-1\}}$ for Kramers transitions (between half-integer time-reversed spin states), and $T^{\{2\tilde{d}_s[1+2(d_\phi/D_f)]-3\}}$ for non-Kramers transitions (between integer time-reversed spin states).

The complexity of random systems adds an interesting consideration: the nature of the relaxation rate *at a specific site* as compared to the *average* relaxation rate calculated above. The decay profile of single-site spin-lattice relaxation is always exponential. However, the temperature dependence does not simply follow that of the average relaxation rate. Instead, it exhibits irregular statistical fluctuations reflecting the environment of the chosen site. As the temperature is raised, new relaxation channels, involving higher-frequency localized vibrations, will become activated, adding to the relaxation processes responsible

for the relaxation at low temperatures. The temperature dependence of the relaxation rate $W(T)$ at different sites will fluctuate with strong correlations in the temperature dependence between them, underlying large accumulative fluctuations.

Using a step function for the vibrational Bose functions, the relaxation rate has a T^2 temperature dependence between randomly spaced steps that occur when new relaxation channels are opened: a “devil’s staircase” structure. The full Bose function smoothes out these steps, but the qualitative features remain.

All of these features point to the complexity and richness of electronic relaxation associated with localized vibrational spectra. Their remarkable complexity is worthy of investigation.

Transport on Fractal Structures

The localization of excitations discussed in Sect. “[Fractal and Spectral Dimensions](#)” has profound influence on the transport properties of and on fractal structures. For example, in the case of the silica aerogels, the “conventional” form for glasses or amorphous materials for the thermal conductivity $\kappa(T)$ is found [12]: a rapid rise at low temperatures leveling off to a “plateau” as a function of temperature T , and then a rise in $\kappa(T)$ roughly as T^2 . At first sight, this is to be “expected” as the aerogels are random, and one could make an analogy with glassy systems. However, there is a fundamental inconsistency in such an analysis. The fracton excitations are localized, and therefore cannot contribute to transport. Thus, one would expect instead a rapid rise in $\kappa(T)$ at low temperatures where the long length scale excitations are phonon-like (and hence delocalized), flattening off to a constant value when all the extended state density of states have reached their duLong–Petit value. The specific heat for these states is then a constant, and $\kappa(T)$ would be a constant independent of temperature T .

So far, this behavior is consistent with experiment. But what happens for temperatures T greater than the plateau temperature region? How can $\kappa(T)$ increase when the only excitations are fractons and therefore localized? The solution to this conundrum lies in the anharmonic forces connecting the silica clusters in the aerogel network. The presence of the anharmonicity allows for a process of fracton hopping [2,20,26], in precisely the same fashion as in Mott’s [22,23] variable range-rate hopping for localized electronic impurity states in semiconductors. The fracton can absorb/emit a lower frequency extended state phonon, and *hop* to another fracton location, thereby transmitting excitation energy spatially. This possibility is amusing, for

it means that for fractal networks the anharmonicity *contributes* to thermal transport, whereas in translationally invariant systems the anharmonicity *inhibits* thermal transport. In fact, were there no anharmonicity in a fractal network, the thermal conductivity $\kappa(T)$ would stay at the plateau value for all T greater than the temperature T_p , the onset of the plateau, for all $T > T_p$. In this sense, anharmonicity *stands on its head*. It is essential for transport in the fracton regime, defined by $T \geq T_p \approx \hbar\omega_c/k_B$.

The localized vibrational excitation hopping contribution to the thermal conductivity is $\kappa_{\text{hop}}(T) = (k_B/2V)\Sigma_{\lambda'} R^2(\omega_{\lambda'})/\tau_{\text{sl}}(\omega_{\lambda'}, T)$. Here sl means strongly localized modes, $R(\omega_{\lambda'})$ is the hopping distance associated with the sl mode, λ' , $\tau_{\text{sl}}(\omega_{\lambda'}, T)$ the hopping lifetime of the sl mode λ' caused by anharmonic interactions at temperature T , and V the volume of the system. We have already derived the ensemble averaged wave function for fractons, $\langle\Psi(r)\rangle$, in Sect. “Fractal and Spectral Dimensions”. We adopt the sl notation [26] because the hopping transport arguments apply to strongly localized states, independent of whether or not they are fractons. The dynamical process for $\kappa_{\text{hot}}(T)$ is an sl mode at one site coupling with a low energy extended state to hop to an sl mode at another site, thereby transferring excitation energy spatially.

The evaluation of $\tau_{\text{sl}}(\omega_{\lambda'}, T)$ will depend upon the hopping distance for transport of the sl modes, $R(\omega_{\lambda'})$. The argument of Mott [22,23], derived by him for localized electronic states, can be taken over for sl vibrational modes [2]. The volume that contains at least one sl mode is given by $(4\pi/3)D(\omega_{\lambda'})\omega_{\text{sl}}[R(\omega_{\lambda'})]^3 = 1$, where $D(\omega_{\lambda'})$ is the density of sl states per unit energy. This condition assures that, for an sl mode at the origin, a second sl mode can be found within a hopping distance $R(\omega_{\lambda'})$. We find $R(\omega_{\lambda'}) \approx (\omega_{\lambda'}/\omega_{\text{sl}})\Lambda(\omega_{\lambda'})$. Thus, the most probable hopping distance is *at least* the sl localization length.

The hopping rate $1/\tau_{\text{sl}}(\omega_{\lambda'}, T)$ caused by anharmonicity, with anharmonic coupling constant C_{eff} can be found from the Golden Rule, leading to the hopping contribution to the thermal conductivity, with $(\xi_M)^3$ the volume for finding a single sl mode, $\kappa_{\text{hop}}(T) = [4^3 (C_{\text{eff}})^2 (k_B)^2 T]/[\pi^3 (v_s)^5 \rho^3 (\xi_M)^3 \langle\Lambda(\omega_{\lambda'})\rangle^2]$. Here $\langle\Lambda(\omega_{\lambda'})\rangle$ is the average sl length scale, v_s the velocity of sound, and ρ the density. There are very few undetermined parameters in this expression for $\kappa_{\text{hop}}(T)$. The anharmonic coupling constant can be found from the shift of the “boson peak” in the Raman spectra with pressure [48,49], typically a factor of ~ 25 times larger than the long length scale third-order elastic coupling constant for extended modes. Making use of experiments on amorphous a-GeS₂ one finds [26,27] $\kappa_{\text{hop}}(T) \approx 0.0065 T$ (W/mK) using the appropriate values for ρ and v_s , and $\xi_M \approx \langle\Lambda(\omega_{\lambda'})\rangle = 15 \text{ \AA}$. This contri-

bution is comparable to that observed for other network-forming glasses above the plateau value.

There is a quantum phenomenon associated with sl mode hopping. The hopping vertex is associated with a low energy extended mode coupling with an sl mode to hop to another (spatially separated) sl mode. As the temperature increases, the lifetime of the sl mode will become shorter than $1/\omega$ of the low energy extended mode. This quantum effect leads to a breakdown of the Golden Rule expression, causing a leveling off of the linear-in- T thermal conductivity above the plateau value.

The sl mode hopping contribution to $\kappa(T)$ above the plateau temperature for glasses and amorphous materials appears to provide a quantitative explanation of the observed phenomena. It is a specific example of how dynamics on a fractal network can be used to determine the properties of materials that, while certainly not fractal, have properties that map onto the behavior of fractal structures.

Future Directions

As stated in the Introduction, studying the dynamics on fractal networks is a convenient structure for analyzing the properties of localized states, with all of the frequency and temperature-dependent physical properties determined without arbitrary parameters. For example, the frequency dependence of the localization length scale, $\Lambda(\omega)$, is known precisely on fractal networks, and is used for precise calculation of the hopping transport in Sect. “Relaxation Dynamics on Fractal Structures”, *without* adjustable parameters.

If one extrapolates the properties of fractal networks to those random systems that are certainly not fractal, but which exhibit localization, the transport and relaxation properties can be understood. Examples are the thermal transport of glasses and amorphous materials above the plateau temperature, and the lifetime of high energy vibrational states in a-Si. Measurements of these properties may offer opportunities for practical devices in frequency regimes beyond extended state frequencies. For example, the lifetime of high energy vibrational states in a-Si has been shown to *increase* with *increasing vibrational energies* [36,37]. Such behavior is opposite to that expected from extended vibrational states interacting anharmonically. But it is entirely consistent if these states are localized and behave as shown for fractal networks [32,33].

The thesis presented here points to a much broader application of the consequences of fractal geometry than simply of those physical systems which exhibit such structures. The claim is made that using what has been learned from fractal dynamics can point to explication of physical

phenomena in disordered systems that are certainly not fractal, but display properties analogous to fractal structures. In that sense, the future lies in application of the ideas contained in the study of fractal dynamics to that of random systems in general. This very broad class of materials may well have properties of great practical importance, predicted from the dynamics on fractals.

Bibliography

Primary Literature

- Alexander S, Entin-Wohlman O, Orbach R (1985) Relaxation and nonradiative decay in disordered systems, I. One-fracton emission. *Phys Rev B* 32:6447–6455
- Alexander S, Entin-Wohlman O, Orbach R (1986) Phonon-Fracton anharmonic interaction: the thermal conductivity of amorphous materials. *Phys Rev B* 34:2726–2734
- Alexander S, Entin-Wohlman O, Orbach R (1986) Relaxation and non-radiative decay in disordered systems, II. Two-fracton inelastic scattering. *Phys Rev B* 33:3935–3946
- Alexander S, Entin-Wohlman O, Orbach R (1987) Relaxation and non-radiative decay in disordered systems, III. Statistical character of Raman (two-quanta) spin-lattice relaxation. *Phys Rev B* 35:1166–1173
- Alexander S, Orbach R (1982) Density of states on fractals, “fractons”. *J Phys Lett (Paris)* 43:L625–L631
- Anderson PW (1958) Absence of diffusion in certain random lattices. *Phys Rev* 109:1492–1505
- Avnir D (ed) (1989) *The Fractal Approach to Heterogeneous Chemistry*. Wiley, Chichester
- Barabási A-L, Stanley HE (1995) *Fractal Concepts in Crystal Growth*. Cambridge University Press, Cambridge
- Bunde A, Havlin S (eds) (1991) *Fractals and Disordered Systems*. Springer, Berlin
- Bunde A, Havlin S (eds) (1994) *Fractals in Science*. Springer-Verlag, Berlin
- Bunde A, Roman HE (1992) Vibrations and random walks on random fractals: anomalous behavior and multifractality. *Philos Magazine B* 65:191–211
- Calemczuk R, de Goer AM, Salce B, Maynard R, Zarembowitch A (1987) Low-temperature properties of silica aerogels. *Europhys Lett* 3:1205–1211
- Courtens E, Pelous J, Phalippou J, Vacher R, Woignier T (1987) Brillouin-scattering measurements of phonon-fracton crossover in silica aerogels. *Phys Rev Lett* 58:128–131
- de Gennes PG (1976) La percolation: un concept unificateur. *Recherche* 7:919–927
- Feder J (1988) *Fractals*. Plenum Press, New York
- Freltoft T, Kjems J, Richter D (1987) Density of states in fractal silica smoke-particle aggregates. *Phys Rev Lett* 59:1212–1215
- Gefen Y, Aharony A, Alexander S (1983) Anomalous diffusion on percolating clusters. *Phys Rev Lett* 50:77–80
- Ikeda H, Fernandez-Baca JA, Nicklow RM, Takahashi M, Iwasa I (1994) Fracton excitations in a diluted Heisenberg antiferromagnet near the percolation threshold: $\text{RbMn}_{0.39}\text{Mg}_{0.61}\text{F}_3$. *J Phys: Condens Matter* 6:10543–10549
- Ikeda H, Itoh S, Adams MA, Fernandez-Baca JA (1998) Crossover from Homogeneous to Fractal Excitations in the Near-Percolating Heisenberg Antiferromagnet $\text{RbMn}_{0.39}\text{Mg}_{0.61}\text{F}_3$. *J Phys Soc Japan* 67:3376–3379
- Jagannathan A, Orbach R, Entin-Wohlman O (1989) Thermal conductivity of amorphous materials above the plateau. *Phys Rev B* 39:13465–13477
- Kistler SS (1932) Coherent Expanded-Aerogels. *J Phys Chem* 36:52–64
- Mott NF (1967) Electrons in disordered structures. *Adv Phys* 16:49–144
- Mott NF (1969) Conduction in non-crystalline materials III. Localized states in a pseudogap and near extremities of conduction and valence bands. *Philos Mag* 19:835–852
- Nakayama T (1992) Dynamics of random fractals: large-scale simulations. *Physica A* 191:386–393
- Nakayama T (1995) Elastic vibrations of fractal networks. *Japan J Appl Phys* 34:2519–2524
- Nakayama T, Orbach R (1999) Anharmonicity and thermal transport in network glasses. *Europhys Lett* 47:468–473
- Nakayama T, Orbach RL (1999) On the increase of thermal conductivity in glasses above the plateau region. *Physica B* 263–264:261–263
- Nakayama T, Yakubo K (2001) The forced oscillator method: eigenvalue analysis and computing linear response functions. *Phys Rep* 349:239–299
- Nakayama T, Yakubo K (2003) *Fractal Concepts in Condensed Matter Physics*. Solid-State Sciences. Springer, Berlin
- Nakayama T, Yakubo K, Orbach R (1994) Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations. *Rev Mod Phys* 66:381–443
- Orbach R (1986) Dynamics of Fractal Networks. *Science* 231:814–819
- Orbach R (1996) Transport and vibrational lifetimes in amorphous structures. *Physica B* 219–220:231–234
- Orbach R, Jagannathan A (1994) High energy vibrational lifetimes in a-Si:H. *J Phys Chem* 98:7411–7413
- Rammal R (1984) Spectrum of harmonic excitations on fractals. *J Phys (Paris)* 45:191–206
- Rammal R, Toulouse G (1983) Random walks on fractal structures and percolation clusters. *J Phys Lett (Paris)* 44:L13–L22
- Scholten AJ, Dijkhuis JI (1996) Decay of high-frequency phonons in amorphous silicon. *Phys Rev B* 53:3837–3840
- Scholten AJ, Verleg PAWE, Dijkhuis JI, Akimov AV, Meltzer RS, Orbach R (1995) The lifetimes of high-frequency phonons in amorphous silicon: evidence for phonon localization. *Solid State Phenom* 44–46:289–296
- Stanley HE (1977) Cluster shapes at the percolation threshold: and effective cluster dimensionality and its connection with critical-point exponents. *J Phys A* 10:L211–L220
- Stauffer D, Aharony A (1992) *Introduction to Percolation Theory*, 2nd edn. Taylor and Francis, London/Philadelphia
- Stauffer D, Stanley HE (1995) *From Newton to Mandelbrot: A primer in Theoretical Physics*, 2nd edn. Springer, Berlin
- Takayasu H (1990) *Fractals in the Physical Sciences*. Manchester University Press, Manchester
- Tsujimi Y, Courtens E, Pelous J, Vacher R (1988) Raman-scattering measurements of acoustic superlocalization in silica aerogels. *Phys Rev Lett* 60:2757–2760
- Uemura YJ, Birgeneau RJ (1987) Magnons and fractons in the diluted antiferromagnet $\text{Mn}_x\text{Zn}_{1-x}\text{F}_2$. *Phys Rev B* 36:7024–7035
- Vacher R, Courtens E, Coddens G, Pelous J, Woignier T (1989) Neutron-spectroscopy measurement of a fracton density of states. *Phys Rev B* 39:7384–7387

45. Vacher R, Woignier T, Pelous J, Courtens E (1988) Structure and self-similarity of silica aerogels. *Phys Rev B* 37:6500–6503
46. Vicsek T (1993) *Fractal Growth Phenomena*. World Scientific, Singapore
47. Yakubo K, Nakayama T (1989) Direct observation of localized fractons excited on percolating nets. *J Phys Soc Japan* 58:1504–1507
48. Yamaguchi M, Nakayama T, Yagi T (1999) Effects of high pressure on the Bose peak in a-GeS₂ studied by light scattering. *Physica B* 263–264:258–260
49. Yamaguchi M, Yagi T (1999) Anharmonicity of low-frequency vibrations in a-GeS₂ studied by light scattering. *Europhys Lett* 47:462–467

Recommended Reading

While many of the original articles are somewhat difficult to access, the reader is nevertheless recommended to read them as there are subtle issues that tend to disappear as references are quoted and re-quoted over time. For a very helpful introduction to percolation theory, Ref. 39 is highly recommended. A thorough review of dynamics of/on fractal structures can be found in Ref. 30. Finally, a comprehensive treatment of localization and multifractals, and their relevance to dynamics of/on fractal networks, can be found in Ref. 29.

Dynamics of Hamiltonian Systems

HEINZ HANSSMANN

Mathematisch Instituut, Universiteit Utrecht,
Utrecht, The Netherlands

Article Outline

Glossary
 Definition of the Subject
 Introduction
 The Phase Space
 Systems in One Degree of Freedom
 Systems in Two Degrees of Freedom
 Symmetry Reduction
 Integrable Systems
 Perturbation Analysis
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Action angle variables In an *integrable* Hamiltonian system with n degrees of freedom the level sets of regular values of the n integrals are *Lagrangian submanifolds*. In case these level sets are compact they are (unions of) n -tori. In the neighborhood of such an invariant torus one can find a *Darboux chart* with co-ordinates (x, y)

such that the integrals (and in particular the Hamiltonian) depend only on the actions y . The values of (x, y) form a product set $\mathbb{T}^n \times \mathbb{Y}$, with $\mathbb{Y} \subseteq \mathbb{R}^n$ open. Every invariant n -torus $y = \text{const}$ in the domain of this chart is parametrized by the angles x .

Canonical 1-form The symplectic form (or canonical 2-form) is closed, satisfying $d\omega = 0$. If it is furthermore exact, i. e. of the form $\omega = d\theta$, then $\vartheta = -\theta$ is called a canonical 1-form (unique up to addition of df , $f \in C^\infty(\mathcal{P})$).

Canonical transformation An invertible transformation that preserves the Poisson bracket, turning canonical co-ordinates into canonical co-ordinates. In the symplectic case also called *symplectomorphism*.

Cantor set, Cantor dust, Cantor family, Cantor stratification Cantor dust is a separable locally compact space that is perfect, i. e. every point is in the closure of its complement, and totally disconnected. This determines Cantor dust up to homeomorphy. The term Cantor set (originally reserved for the specific form of Cantor dust obtained by repeatedly deleting “the middle third”) designates topological spaces that locally have the structure $\mathbb{R}^n \times \text{Cantor dust}$ for some $n \in \mathbb{N}$. Cantor families are parametrized by such Cantor sets. On the real line \mathbb{R} one can define Cantor dust of positive measure by excluding around each rational number p/q an interval of size $2\gamma/q^\tau$, $\gamma > 0$, $\tau > 2$. Similar *Diophantine conditions* define Cantor sets in \mathbb{R}^n . Since these Cantor sets have positive measure their (Hausdorff)-dimension is n . Where the unperturbed system is stratified according to the co-dimension of occurring (bifurcating) tori, this leads to a Cantor stratification.

Casimir function A function $f: \mathcal{P} \rightarrow \mathbb{R}$ is a Casimir element of the *Poisson algebra* $C^\infty(\mathcal{P})$ if $\{f, g\} = 0$ for all $g \in C^\infty(\mathcal{P})$. This induces a Poisson structure on the level sets $f^{-1}(a)$, $a \in \mathbb{R}$. In case every point has a small neighborhood on which the only Casimir elements are the constant functions, the Poisson structure on \mathcal{P} is non-degenerate, i. e. \mathcal{P} is a *symplectic manifold*.

Center An equilibrium of a vector field is called a center if all eigenvalues of the linearization are nonzero and on the imaginary axis.

Center-saddle bifurcation Under variation of a parameter λ a *center* and a *saddle* meet and vanish.

Chern class Let $\rho: \mathcal{R} \rightarrow \mathcal{C}$ be a torus bundle with fiber \mathbb{T}^n and denote by \mathcal{G} the locally constant sheaf of first homotopy groups $\pi_1(\mathbb{T}^n)$ of the fibers. The Chern class of the torus bundle is an element of $H^2(\mathcal{C}, \mathcal{G})$ that measures the obstruction to the existence of a global section of ρ – such a section exists if and only if

the Chern class vanishes. An example with a non-vanishing Chern class is the Hopf fibration $S^3 \rightarrow S^2$.

Conditionally periodic A motion $t \mapsto \alpha(t) \in \mathcal{P}$ is conditionally periodic if there are frequencies $\varpi_1, \dots, \varpi_k \in \mathbb{R}$ and a smooth embedding $F: \mathbb{T}^k \rightarrow \mathcal{P}$ such that $\alpha(t) = F(e^{2\pi i \varpi_1 t}, \dots, e^{2\pi i \varpi_k t})$. One can think of the motion as a superposition of the periodic motions $t \mapsto F(1, \dots, 1, e^{2\pi i \varpi_j t}, 1, \dots, 1)$. If the frequencies are rationally independent, the motion $t \mapsto \alpha(t) \in \mathcal{P}$ lies dense in $\text{im } F$ and this embedded torus is invariant. In case there are *resonances* among the frequencies the motion is restricted to a subtorus. A flow on a torus is parallel or conditionally periodic if there exist co-ordinates in which the vector field becomes constant. In the absence of resonances the flow is called *quasi-periodic*.

Conjugate co-ordinates, canonical co-ordinates Two co-ordinates Q and P of a *Darboux* chart are (canonically) conjugate if $\omega(X_Q, X_P) = \pm 1$, i.e. X_Q and X_P span for every point x in the domain U of the chart a hyperbolic plane in $T_x U$.

Darboux chart, Darboux co-ordinates, Darboux basis

In a Darboux basis $\{e_1, \dots, e_n, f_1, \dots, f_n\}$ of a symplectic vector space (V, ω) the symplectic product takes the simple form $\omega(e_i, e_j) = 0 = \omega(f_i, f_j)$ and $\omega(e_i, f_j) = \delta_{ij}$. In Darboux co-ordinates $(q_1, \dots, q_n, p_1, \dots, p_n)$ of a symplectic manifold (\mathcal{P}, ω) the symplectic form becomes $\omega = \sum dq_i \wedge dp_i$.

Degree of freedom In simple mechanical systems the *phase space* is the cotangent bundle of the configuration space and the dimension of the latter encodes “in how many directions the system can move”. For *symplectic manifolds* this notion is immediately generalized to one half of the dimension of the phase space. *Poisson spaces* are foliated by their symplectic leaves and the number of degrees of freedom is defined to be one half of the rank of the Poisson structure.

Diophantine condition, Diophantine frequency vector

A frequency vector $\varpi \in \mathbb{R}^n$ is called Diophantine if there are constants $\gamma > 0$ and $\tau > n - 1$ with

$$|\langle k | \varpi \rangle| \geq \frac{\gamma}{|k|^\tau} \quad \text{for all } k \in \mathbb{Z}^n \setminus \{0\}.$$

The Diophantine frequency vectors satisfying this condition for fixed γ and τ form a Cantor set of half lines. As the “Diophantine parameter” γ tends to zero (while τ remains fixed), these half lines extend to the origin. The complement in any compact set of frequency vectors satisfying a Diophantine condition with fixed τ has a measure of order $O(\gamma)$ as $\gamma \rightarrow 0$.

Elliptic A periodic orbit/invariant torus is elliptic if all *Floquet multipliers/exponents* are on the unit circle/imaginary axis. An elliptic equilibrium is called a *center*.

Energy shell The set of points that can be attained given a certain energy. If the phase space is a symplectic manifold, this set is given by the pre-image of that energy value under the Hamiltonian. For more general Poisson spaces this pre-image has to be intersected with a symplectic leaf.

Focus-focus type The (normal) linear behavior generated by a quartet of eigenvalues/Floquet multipliers/exponents.

Frequency-halving bifurcation In the *supercritical* case an invariant torus loses its stability, giving rise to a stable torus with one of its frequencies halved. In the *dual* or *subcritical* case stability is lost through collision with an unstable invariant torus with one of its frequencies halved. The periodic counterpart is called a period-doubling bifurcation.

Generating function A function $S = S(q, y)$ that defines a canonical transformation $(x, y) \mapsto (q, p)$ by means of $x = \partial S / (\partial y)$ and $p = \partial S / (\partial q)$. The generating function always depends on both the old and the new variables, working with dependence on e.g. x instead of y introduces a minus sign.

Group action A mapping $\Gamma: G \times \mathcal{P} \rightarrow \mathcal{P}$ is a group action of the Lie group G on the phase space \mathcal{P} if $\Gamma_e = \text{id}$ and $\Gamma_{g \cdot h} = \Gamma_g \circ \Gamma_h$ for all Lie group elements g and h . For $G = \mathbb{R}$ one can recover the generating vector field by means of $X = d/(dt) \Gamma_t$.

Hamiltonian Hopf bifurcation In a two-degree-of-freedom system a center loses its stability: the two pairs of purely imaginary eigenvalues meet in a *Krein collision*. In a three-degree-of-freedom system the same can happen to two pairs $e^{\pm \lambda}, e^{\pm \mu}$ of *Floquet multipliers* in a one-parameter family of periodic orbits – parametrized by the energy. For a quasi-periodic Hamiltonian Hopf bifurcation one needs at least four degrees of freedom.

Hamiltonian system Newton’s second law states $F = m\ddot{q}$. Suppose that F is a conservative force, with potential V . Write the equations of motion as a system of first order differential equations

$$\begin{aligned} \dot{q} &= \frac{1}{m} p \\ \dot{p} &= -\frac{\partial V}{\partial q} \end{aligned}$$

that has the total energy $H(q, p) = \langle p | p \rangle / (2m) + V(q)$ as a first integral. This can be generalized. Given

a Hamiltonian function $H(q, p)$ one has the Hamiltonian vector field

$$\dot{q} = \frac{\partial H}{\partial p}$$

$$\dot{p} = -\frac{\partial H}{\partial q}$$

with first integral H . Moreover, one can replace \mathbb{R}^{2n} by a *symplectic manifold* or by a *Poisson space*.

Hilbert basis The (smooth) invariants of an action of a compact group are all given as functions of finitely many “basic” invariants.

Hypo-elliptic An equilibrium is called hypo-elliptic if its linearization has both elliptic and hyperbolic eigenvalues, all nonzero. Similar for periodic orbits and invariant tori.

Integrable system A Hamiltonian system with n degrees of freedom is (Liouville)-integrable if it has n functionally independent commuting integrals of motion. Locally this implies the existence of a (local) torus action.

Integral (of motion) A conserved quantity.

Iso-energetic Poincaré mapping For a Hamiltonian system a *Poincaré mapping* leaves the *energy shells* invariant. Restricting to the intersection $\Sigma \cap \{H = h\}$ of the *Poincaré section* with an energy shell one obtains an iso-energetic Poincaré mapping.

Isotropic For a subspace $U < V$ of a symplectic vector space (V, ω) are equivalent: (i) U is contained in its ω -orthogonal complement, (ii) every basis of U can be extended to a *Darboux basis* of V . If U satisfies one and thus both of these conditions it is called an isotropic subspace. A submanifold $\mathcal{B} \subseteq \mathcal{P}$ of a symplectic manifold (\mathcal{P}, ω) is called isotropic if all tangent spaces $T_x \mathcal{B} < T_x \mathcal{P}$ are isotropic subspaces.

Krein collision Two pairs of purely imaginary *Floquet exponents* meet in a double pair on the imaginary axis and split off to form a complex quartet $\pm \Re \pm i \Im$. This is a consequence of a transversality condition on the linear terms at 1: -1 *resonance*; an additional non-degeneracy condition on the non-linear part ensures that a ((quasi)-periodic) *Hamiltonian Hopf bifurcation* takes place.

Lagrangian submanifold For a subspace $U < V$ of a symplectic vector space (V, ω) are equivalent: (i) U is a maximal isotropic subspace, (ii) V has a *Darboux basis* $\{e_1, \dots, e_n, f_1, \dots, f_n\}$ with $\text{span}\{e_1, \dots, e_n\} = U$. If U satisfies one and thus both of these conditions it is called a Lagrangian subspace. A submanifold $\mathcal{B} \subseteq \mathcal{P}$ of a symplectic manifold (\mathcal{P}, ω) is called Lagrangian if all tangent spaces $T_x \mathcal{B} < T_x \mathcal{P}$ are Lagrangian subspaces.

Lie–Poisson structure Let \mathfrak{g} be a Lie algebra with structure constants Γ_{ij}^k . Then $\{\mu_i, \mu_j\} = \pm \sum \Gamma_{ij}^k \mu_k$ defines two *Poisson structures* on the dual space \mathfrak{g}^* .

Local bifurcation Bifurcations of equilibria can be studied within a small neighborhood in the product of phase space and parameter space. The same is true for fixed points of a discrete dynamical system, but when suspended to a flow the corresponding bifurcating periodic orbits obtain a *semi-local* character.

Momentum mapping Let $\Gamma: G \times \mathcal{P} \rightarrow \mathcal{P}$ be a symplectic action of the Lie group G on the *symplectic manifold* \mathcal{P} . A mapping $J: \mathcal{P} \rightarrow \mathfrak{g}^*$ into the dual space of the Lie algebra of G is a momentum mapping for the action if $X_{J\xi} = \xi_{\mathcal{P}}$ for all $\xi \in \mathfrak{g}$. Here $J^\xi: \mathcal{P} \rightarrow \mathbb{R}$ is defined by $J^\xi(z) = J(z) \cdot \xi$ and $\xi_{\mathcal{P}}$ is the infinitesimal generator of the action corresponding to ξ . The momentum mapping J is called Ad^* -equivariant provided that $J \circ \Gamma_g = \text{Ad}_{g^{-1}}^* \circ J$.

Monodromy The group homomorphism $\mathcal{M}: \pi_1(C) \rightarrow SL_n(\mathbb{Z})$ that measures how much the n -torus bundle $\mathcal{R} \rightarrow C$ deviates from a *principal* torus bundle.

Non-degenerate integrable Hamiltonian system, function In *action angle variables* (x, y) the Hamiltonian H only depends on the action variables y and the equations of motion become $\dot{x} = \varpi(y)$, $\dot{y} = 0$; with frequencies $\varpi(y) = \partial H / \partial y$. The system is non-degenerate at the invariant torus $\{y = y_0\}$ if $D^2 H(y_0)$ is invertible. In this case ϖ defines near $\{y = y_0\}$ an isomorphism between the actions and the angular velocities. Other conditions ensuring that most tori have *Diophantine* frequency vectors are iso-energetic non-degeneracy or Rüssmann-like conditions on higher derivatives.

Normal frequency Given an elliptic invariant torus of a Hamiltonian system, one can define the normal linearization on the *symplectic normal bundle*. The eigenvalues of the normal linearization being $\pm i\Omega_1, \dots, \pm i\Omega_m$, one calls the Ω_j the normal frequencies. Under the exponential mapping the eigenvalues of the normal linearization of a periodic orbit are mapped to *Floquet multipliers*.

Normal linearization The linearization within the normal bundle of an invariant submanifold. In the Hamiltonian context these are often *isotropic* and one further restricts to the symplectic normal linear behavior, e. g. to identify *elliptic* and *hyperbolic* invariant tori.

Normal modes A family of periodic orbits parametrized by the energy value h , as $h \rightarrow h_0$ shrinking down to an equilibrium that has a pair of eigenvalues $\pm 2\pi i \lim_{h \rightarrow h_0} T_h^{-1}$, where T_h denotes the period(s).

Parabolic An equilibrium of a one-degree-of-freedom system is called parabolic if its linearization is nilpotent but nonzero. An invariant torus is parabolic if its symplectic *normal linearization* has a parabolic equilibrium. In particular the four *Floquet multipliers* of a parabolic periodic orbit in two degrees of freedom are all equal to 1.

Phase space By Newton's second law the equations of motion are second order differential equations. The trajectory is completely determined by the initial positions and the initial velocities, or, equivalently, the initial momenta. The phase space is the set of all possible combinations of initial positions and initial momenta.

Pinched torus The compact (un)stable manifold of a saddle in two degrees of freedom with a quartet $\pm \Re \pm i \Im$ of hyperbolic eigenvalues resembles a torus $\mathbb{T}^2 = S^1 \times S^1$ with one of the fibers $\{x\} \times S^1$ reduced to a point.

Poisson space, Poisson bracket, Poisson structure

A Poisson algebra \mathcal{A} is a real Lie algebra that is also a commutative ring with unit. These two structures are related by Leibniz' rule $\{f \cdot g, h\} = f \cdot \{g, h\} + g \cdot \{f, h\}$. A Poisson manifold \mathcal{P} has a Poisson bracket on $C^\infty(\mathcal{P})$ that makes $C^\infty(\mathcal{P})$ a Poisson algebra. If there are locally no *Casimir elements* other than constant functions this leads to a *symplectic structure* on \mathcal{P} . Poisson spaces naturally arise in *singular reduction*, this motivates to allow varieties \mathcal{P} where the Poisson bracket is defined on a suitable subalgebra \mathcal{A} of $C(\mathcal{P})$.

Given a Hamiltonian function $H \in \mathcal{A}$ one obtains for $f \in \mathcal{A}$ the equations of motion $(d/(dt))f = \{f, H\}$. For canonically *conjugate* co-ordinates (q, p) on \mathcal{P} , i. e. with $\{q_i, q_j\} = 0 = \{p_i, p_j\}$ and $\{q_i, p_j\} = \delta_{ij}$, this amounts to

$$\begin{aligned}\dot{q} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial q}.\end{aligned}$$

Poisson symmetry A symmetry that preserves the *Poisson structure*.

Principal fiber bundle A fiber bundle $F: \mathcal{P} \rightarrow C$ admitting a free Lie group action $G \times \mathcal{P} \rightarrow \mathcal{P}$ that acts transitively on the fibers $F^{-1}(c) \cong G$.

Proper degeneracy For the application of the KAM theorem to a perturbation of an *integrable system* it is necessary that the integrable system is *non-degenerate*, so that the frequencies have maximal rank as function of the actions. If there are global conditions relating the

frequencies, so that the *conditionally periodic* motion can be described by a smaller number of these, the system is properly degenerate. *Superintegrable systems* are a particular example.

Quasi-periodic A *conditionally periodic* motion that is not periodic. The closure of the trajectory is an invariant k -torus with $k \geq 2$. A parallel or conditionally periodic flow on a k -torus is called quasi-periodic if the frequencies $\varpi_1, \dots, \varpi_k$ are rationally independent.

Ramified torus bundle Let a differentiable mapping $f: \mathcal{P} \rightarrow \mathbb{R}^m$ be given. According to Sard's Lemma almost all values $a \in \mathbb{R}^m$ are regular. The connected components of the sets $f^{-1}(a)$, $a \in \text{im } f$ regular, define a foliation of an open subset \mathcal{R} of the $(n+m)$ -dimensional manifold \mathcal{P} . In the present settings the components of f are the first integrals of an *integrable* Hamiltonian system with compact level sets. Then the regular fibers are n -tori, and their union is a torus bundle. The topology of this bundle depends on the topology of the base space C , the *monodromy* and the *Chern class* of the bundle. In many examples one encounters the connected components of C are contractible, whence monodromy and Chern class are trivial. For the geometry of the bundle one also wants to know how the singular fibers are distributed: where n -tori shrink to normally elliptic $(n-1)$ -tori and where they are separated by stable and unstable manifolds of normally hyperbolic $(n-1)$ -tori. The singular fibers are not necessarily manifolds, but may be stratified into X_H -invariant strata which are possibly non-compact. One can continue and look for the singularities of these 'regular singular leaves'; the tori of dimension $\leq n-2$ and the normally *parabolic* $(n-1)$ -tori in which normally elliptic and normally hyperbolic $(n-1)$ -tori meet in a quasi-periodic *center-saddle bifurcation* or a *frequency-halving bifurcation*. The next "layer" is given by quasi-periodic *Hamiltonian Hopf bifurcations* and bifurcations of higher co-dimension.

Reduced phase space Let $\Gamma: G \times \mathcal{P} \rightarrow \mathcal{P}$ be a symplectic action of the Lie group G on the *symplectic manifold* \mathcal{P} with Ad^* -equivariant *momentum mapping* $J: \mathcal{P} \rightarrow \mathfrak{g}^*$. For a regular value $\mu \in \mathfrak{g}^*$ let the action of the isotropy group $G_\mu = \{g \in G \mid \text{Ad}_{g^{-1}}^*(\mu) = \mu\}$ on $J^{-1}(\mu)$ be free and proper. Then the quotient $J^{-1}(\mu)/G_\mu$ is again a symplectic manifold, the reduced phase space. A Γ -invariant Hamiltonian function H on \mathcal{P} leads to a reduced Hamiltonian function H_μ on the reduced phase space.

Relative equilibrium Let H be a Hamiltonian function invariant under the *symplectic action* $G \times \mathcal{P} \rightarrow \mathcal{P}$ and let $\mu \in \mathfrak{g}^*$ be a regular value of the Ad^* -

equivariant *momentum mapping* $J: \mathcal{P} \longrightarrow \mathfrak{g}^*$. Also assume that the isotropy group G_μ under the Ad^* action on \mathfrak{g}^* acts freely and properly on $J^{-1}(\mu)$. Then X_H induces a Hamiltonian flow on the *reduced phase space* $\mathcal{P}_\mu = J^{-1}(\mu)/G_\mu$. The phase curves of the given Hamiltonian system on \mathcal{P} with momentum constant $J = \mu$ that are taken by the projection $J^{-1}(\mu) \longrightarrow \mathcal{P}_\mu$ into equilibrium positions of the reduced Hamiltonian system are called *relative equilibria* or *stationary motions* (of the original system).

Remove the degeneracy A perturbation of a *superintegrable system* removes the degeneracy if it is sufficiently mild to define an intermediate system that is still integrable and sufficiently wild to make that intermediate system *non-degenerate*.

Resonance If the frequencies of an invariant torus with *conditionally periodic* flow are rationally dependent this torus divides into invariant subtori. Such resonances $\{h \mid \varpi\} = 0, h \in \mathbb{Z}^k$, define hyperplanes in ϖ -space and, by means of the frequency mapping, also in phase space. The smallest number $|h| = |h_1| + \dots + |h_k|$ is the order of the resonance. *Diophantine conditions* describe a measure-theoretically large complement of a neighborhood of the (dense!) set of all resonances.

Saddle An equilibrium of a vector field is called a saddle if the linearization has no eigenvalues on the imaginary axis. On a small neighborhood of a saddle the flow is topologically *conjugate* to its linearization.

Semi-local bifurcation Bifurcations of n -tori can be studied in a tubular neighborhood. For $n = 1$ a *Poincaré section* turns the periodic orbit into a fixed point of the *Poincaré mapping* and the bifurcation obtains a *local* character.

Simple mechanical system A quintuple $(T^*M, \langle \cdot \mid \cdot \rangle, V, T, \omega)$ consisting of the co-tangent bundle of a Riemannian manifold $(M, \langle \cdot \mid \cdot \rangle)$, a potential function $V: M \longrightarrow \mathbb{R}$, the kinetic energy $T(\alpha) = \langle \alpha \mid \alpha \rangle$ and the symplectic form $\omega = -d\vartheta$ derived from the canonical 1-form on the co-tangent bundle T^*M .

Singular reduction If $\Gamma: G \times \mathcal{P} \longrightarrow \mathcal{P}$ is a *Poisson symmetry*, then the group action on \mathcal{P} makes $\mathcal{B} = \mathcal{P}/G$ a *Poisson space* as well. Fixing the values of the resulting *Casimir functions* yields the *reduced phase space*, which turns out to have singular points where the action Γ is not free.

Solenoid Given a sequence $f_j: S^1 \longrightarrow S^1$ of coverings $f_j(\zeta) = \zeta^{\alpha_j}$ of the circle S^1 the solenoid $\Sigma_a \subseteq (S^1)^{\mathbb{N}_0}$, $a = (\alpha_j)_{j \in \mathbb{N}_0}$ consists of all $z = (\zeta_j)_{j \in \mathbb{N}_0}$ with $\zeta_j = f_j(\zeta_{j+1})$ for all $j \in \mathbb{N}_0$.

Stratification The decomposition of a topological space

into smaller pieces satisfying certain boundary conditions.

Superintegrable system A Hamiltonian system with n degrees of freedom is superintegrable if it has $n + 1$ functionally independent integrals of motion such that each of the first $n - 1$ of them commutes with all $n + 1$. Such a *properly degenerate* system admits generalized action angle co-ordinates $(x_1, \dots, x_{n-1}, y_1, \dots, y_{n-1}, q, p)$. In case the non-degeneracy condition $\det D^2H(y) \neq 0$ is satisfied almost everywhere the system is “minimally superintegrable”. In the other extreme of a “maximally superintegrable” system all motions are periodic.

Symplectic manifold A $2n$ -dimensional manifold \mathcal{P} with a non-degenerate closed two-form ω , i.e. $d\omega = 0$ and $\omega(u, v) = 0$ for all $v \in T\mathcal{P} \Rightarrow u = 0$. A diffeomorphism ψ of symplectic manifolds that respects the two-form(s) is called a *symplectomorphism*. Given a Hamiltonian function $H \in C^\infty(\mathcal{P})$ one obtains through $\omega(X_H, \cdot) = dH$ the Hamiltonian vector field X_H . For every $x \in \mathcal{P}$ there are co-ordinates (q, p) around x with $\omega = dq \wedge dp$. In these *Darboux co-ordinates* X_H reads

$$\begin{aligned}\dot{q} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial q}.\end{aligned}$$

Symplectic form A non-degenerate closed two-form.

Szyggy A constraining equation that is identically fulfilled by the elements of a *Hilbert basis*.

(Un)stable manifold In Hamiltonian systems with one degree of freedom the *stable manifold* and the unstable manifold of an equilibrium often coincide and thus consist of *homoclinic orbits*. In such a case it is called an (un)stable manifold. This carries over to the stable and the unstable manifold of a periodic orbit or an invariant torus in higher degrees of freedom if the system is *integrable*.

Definition of the Subject

Hamiltonian systems are a class of dynamical systems which can be characterized by preservation of a symplectic form. This allows to write down the equations of motion in terms of a single function, the Hamiltonian function. They were conceived in the 19th century to study physical systems varying from optics to frictionless mechanics in a unified way. This description turned out to be particularly efficient for symmetry reduction and perturbation analysis.

Introduction

The best-known Hamiltonian system is the harmonic oscillator. The second order differential equation

$$\ddot{x} + \varpi^2 x = 0$$

models the motion of a point mass attached to a massless spring (Hooke's law) and has the general solution

$$x(t) = x_0 \cos \varpi t + \frac{y_0}{\varpi} \sin \varpi t$$

with initial conditions $(x(0), \dot{x}(0)) = (x_0, y_0)$. Choosing co-ordinates

$$q = \sqrt{\varpi} x \quad \text{and} \quad p = \frac{y}{\sqrt{\varpi}} \quad (1)$$

yields the first order system

$$\frac{d}{dt} \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & \varpi \\ -\varpi & 0 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix},$$

the solutions of which move along the circles $p^2 + q^2 = \text{const}$. This last observation lies at the basis of the Hamiltonian formalism, defining

$$H(q, p) = \varpi \frac{p^2 + q^2}{2}$$

to be the Hamiltonian function of the system. The Hamiltonian vector field is then defined to be perpendicular to the gradient ∇H , ensuring that H is a conserved quantity. Passing back through (1) yields the energy

$$H(x, y) = \frac{\dot{x}^2}{2} + \varpi^2 \frac{x^2}{2}$$

of the spring system (the point mass being scaled to $m = 1$).

The conserved quantity H makes two-dimensional systems, said to have one *degree of freedom*, easy to study and many efforts are made to reduce more complex systems to this setting, often using symmetries. In such a case the system is *integrable* and can in principle be explicitly solved. Where the symmetries are only approximately preserving the system, a more geometric understanding allows to analyze how the perturbation from the integrable approximation to the original system alters the dynamics.

The Phase Space

The simplest type of example of a Hamiltonian system is that of a 1-dimensional particle with kinetic energy

$$T = \frac{1}{2} m v^2 = \frac{p^2}{2m} \quad (2)$$

and potential energy $V = V(q)$. The canonical equations derived from the Hamiltonian function $H = T + V$ are

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} = \frac{p}{m} \\ \dot{p} &= -\frac{\partial H}{\partial q} = -V'(q) \end{aligned}$$

and are equivalent with Newton's laws, where $F = -V'$ is the force field with potential V . An immediate consequence is

$$\dot{H} = \frac{\partial H}{\partial q} \dot{q} + \frac{\partial H}{\partial p} \dot{p} = 0,$$

the conservation of energy.

The same conclusion holds in \mathbb{R}^{2n} , with *canonical co-ordinates* $q_1, \dots, q_n, p_1, \dots, p_n$, where a Hamiltonian function $H: \mathbb{R}^{2n} \rightarrow \mathbb{R}$ defines the canonical equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad i = 1, \dots, n \quad (3a)$$

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad i = 1, \dots, n. \quad (3b)$$

The Hamiltonian vector field X_H defined by these equations satisfies

$$i_{X_H} \omega := \omega(X_H, _) = dH \quad (4)$$

where

$$\omega = dq_1 \wedge dp_1 + \dots + dq_n \wedge dp_n \quad (5)$$

is the canonical *symplectic form* on \mathbb{R}^{2n} .

Hamiltonian mechanics becomes conceptually easier if one abstracts from well-chosen canonical co-ordinates and considers the phase space to be a *symplectic manifold* (\mathcal{P}, ω) . A Hamiltonian function $H: \mathcal{P} \rightarrow \mathbb{R}$ then defines by means of (4) the Hamiltonian vector field X_H on \mathcal{P} in a co-ordinate free way. With Darboux's Theorem [1,2,22,26] one can always return to canonical co-ordinates, locally around any given point of \mathcal{P} .

The flow $\varphi: \mathbb{R} \times \mathcal{P} \rightarrow \mathcal{P}$ of a Hamiltonian vector field preserves the symplectic structure, i. e. $\varphi_t^* \omega = \omega$ [2,8,22,26]. An immediate consequence is Liouville's Theorem that the phase space volume $\omega^n = \omega \wedge \dots \wedge \omega$ is preserved as well. As a corollary one obtains Poincaré's Recurrence Theorem.

Theorem (Poincaré) *Let $\Omega \subseteq \mathcal{P}$ be compact and invariant under φ_t . Then every neighborhood U of every point $a \in \Omega$ has a trajectory that returns to U .*

The proof consists of “a walk in the park” – however small my shoes are, I am bound to step on my own trail if I walk forever in a park of finite size.

Hamiltonian systems have special properties that are untypical for general dissipative systems. An important aspect is that volume preservation excludes the existence of attractors. In dissipative systems the restriction of the flow to an attractor often allows to dramatically lower the dimension of the system; more generally one restricts with the same aim to the non-wandering set. In the present conservative context, if e. g. the energy level sets are compact then the non-wandering set consists of the whole phase space.

One speaks of a *simple mechanical system* [1,2,22] if the phase space $\mathcal{P} = T^*M$ is the cotangent bundle of a Riemannian manifold M , the configuration space, and the Hamiltonian $H = T + V$ is the sum of kinetic and potential energy. Furthermore, the kinetic energy

$$T(\alpha_q) = \langle \alpha_q | \alpha_q \rangle_q$$

is given by the Riemannian metric (evaluated at the base point $q \in M$ of $\alpha_q \in T_q^*M$) and the potential energy

$$V(\alpha_q) = V(q)$$

depends only on the configuration $q \in M$. The cotangent bundle has a *canonical 1-form* ϑ defined by

$$\vartheta(v_{\alpha_q}) = \alpha_q(T\pi(v_{\alpha_q})) \quad \text{for all } v_{\alpha_q} \in T_{\alpha_q}T^*M$$

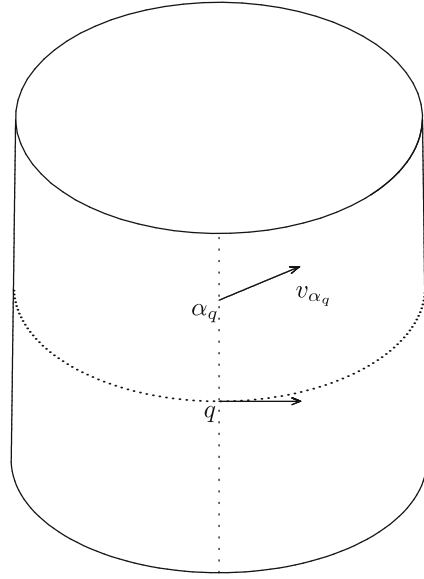
where $\pi: T^*M \rightarrow M$ is the bundle projection to the configuration space and $T\pi: TT^*M \rightarrow TM$ its derivative, see Fig. 1. From ϑ the symplectic form is obtained as the exterior derivative $\omega = -d\vartheta$.

Choosing any co-ordinates q_1, \dots, q_n on the configuration space M and completing them with the conjugate momenta p_1, \dots, p_n one automatically has a canonical co-ordinate system on the phase space $\mathcal{P} = T^*M$ in which the canonical 1-form reads

$$\vartheta = \sum_{i=1}^n p_i dq_i$$

and the symplectic form is given by (5). This freedom of choosing any co-ordinate system is lost on the whole phase space if one insists on the equations of motion to be in canonical form (3).

A significant part of the classical literature [2,12,16,31] is devoted to *generating functions*, a means to generate transformations that turn canonical co-ordinates into canonical co-ordinates, therefore called *canonical transformations*. A contemporary means to obtain canonical



Dynamics of Hamiltonian Systems, Figure 1

Projection of a vector v_{α_q} tangent to T^*M to the tangent space of M at $q = \pi(\alpha_q)$. The resulting tangent vector can be fed into $\alpha_q: T_qM \rightarrow \mathbb{R}$

transformations is to use the time-1-flow φ_1 of a well-chosen Hamiltonian function as these are better suited for implementation on a computer. Since $\varphi_1^{-1} = \varphi_{-1}$ it is as simple (or complicated) to compute the inverse of such a transformation as it is to compute the transformation itself.

In a (not necessarily canonical) co-ordinate system z_1, \dots, z_{2n} one can use the *Poisson bracket*, defined by

$$\{F, H\} := \omega(XF, XH),$$

to write down the equations of motion

$$\dot{z}_j = \{z_j, H\} = \sum_{k=1}^{2n} \frac{\partial H}{\partial z_k} \{z_j, z_k\}$$

which have the canonical form precisely when it is possible to write $(z_1, \dots, z_n) = (q_1, \dots, q_n)$ and $(z_{n+1}, \dots, z_{2n}) = (p_1, \dots, p_n)$ with Poisson bracket relations

$$\{q_i, q_j\} = 0, \quad \{p_i, p_j\} = 0, \quad \{q_i, p_j\} = \delta_{ij}.$$

While a symplectic manifold has necessarily even dimension, one can turn also manifolds of odd dimension into a Poisson space. An important example is the Poisson structure

$$\{F, H\}(z) := \langle \nabla F(z) \times \nabla H(z) | z \rangle \quad (6)$$

on \mathbb{R}^3 (with its inner product $\langle x | y \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3$) for which the equations of motion read

$$\dot{z} = \nabla H(z) \times z. \quad (7)$$

The function $R(z) = (1/2)\langle z | z \rangle$ is a *Casimir function* of the Poisson structure as it is invariant under every Hamiltonian flow on \mathbb{R}^3 since $\{R, H\} = 0$ for all (Hamiltonian) functions $H: \mathbb{R}^3 \rightarrow \mathbb{R}$. This fibrates \mathbb{R}^3 into invariant spheres $\{R = (1/2)\rho^2\}$ of radius ρ , with a singular sphere reduced to a point at the origin. The area element σ makes each sphere S_ρ^2 a symplectic manifold, and the restriction of the Poisson bracket (6) to S_ρ^2 satisfies

$$\{F, H\} = \frac{1}{\rho^2} \sigma(XF, XH).$$

A similar (though in general slightly more complicated) decomposition into symplectic leaves exists for every Poisson manifold [2,22,27].

A first measure for the complexity of a Hamiltonian system is the number of degrees of freedom. For a simple mechanical system this is the dimension of the configuration space, and accordingly one defines this number as $(1/2)\dim \mathcal{P}$ for a symplectic manifold \mathcal{P} . On a Poisson space this number is related to the rank of the Poisson bracket, given by $\text{rank}(\{z_j, z_k\})_{j,k=1,\dots,m}$ in local coordinates z_1, \dots, z_m . This even number is upper semi-continuous and coincides at each point with the dimension of the symplectic leaf passing through that point. Hence, the number of degrees of freedom is defined as one half of the maximal rank of the Poisson structure.

Systems in One Degree of Freedom

For the simple mechanical systems consisting of a 1-dimensional particle with kinetic energy (2) moving in a potential $V = V(q)$ the trajectories coincide with the level curves $\{H = h\}$ of $H = T + V$. Finding the time parametrizations of the trajectories is thereby reduced to the quadrature

$$\int \frac{dq}{\sqrt{2m(h - V(q))}},$$

and correspondingly one speaks of an integrable system. Where this time parametrization is not important one can always multiply the Hamiltonian by a strictly positive function, in the present one-degree-of-freedom situation one has then the extra choice [6,13] of performing any co-ordinate transformation and still writing the equations in canonical form (with respect to the transformed Hamiltonian function).

The phase portraits in a Poisson space \mathcal{P} with one degree of freedom can be obtained by intersecting the level sets of the energy with the symplectic leaves. In particular, a point where the rank of the Poisson structure drops from 2 to 0 is an equilibrium for every Hamiltonian system on \mathcal{P} . Equilibria on regular symplectic leaves are called regular equilibria.

There are four types of linearizations of regular equilibria in one degree of freedom. In canonical co-ordinates these are given by quadratic Hamiltonian functions H_2 .

Elliptic $H_2(q, p) = (\varpi/2)(p^2 + q^2)$, the nearby motion is periodic with frequency close to ϖ . This is the harmonic oscillator.

Hyperbolic $H_2(q, p) = (a/2)(p^2 - q^2)$, the equilibrium is a *saddle*.

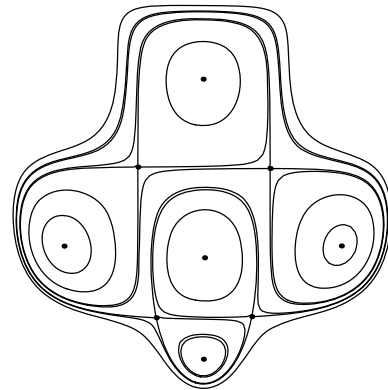
Parabolic $H_2(q, p) = (a/2)p^2$, the higher order terms of the Hamiltonian around the equilibrium determine the character of the flow.

Vanishing $H_2(q, p) \equiv 0$, the linearization contains no information and the flow is given by the higher order terms.

Generic Hamiltonian systems on a symplectic surface have only elliptic and hyperbolic equilibria, see Fig. 2. Where the system depends on external parameters or is defined on a family of symplectic leaves one may also encounter parabolic equilibria. The phenomenon of vanishing linearization is of co-dimension three.

As an example consider the phase space \mathbb{R}^3 with Poisson structure (6) and Hamiltonian energy function

$$H(z) = \sum_{i=1}^3 \frac{a_i}{2} z_i^2 + b_i z_i$$



Dynamics of Hamiltonian Systems, Figure 2
Typical recurrent flow in one degree of freedom

depending on the external parameters $(a, b) \in \mathbb{R}^3 \times \mathbb{R}^3$. On each sphere S_ρ^2 the points with minimal and maximal energy are elliptic equilibria. All equilibria occur where S_ρ^2 touches the quadric $\{H = h\}$ of constant energy. Where this happens with coinciding curvature along a line the equilibrium is parabolic. For $a, b \in \mathbb{R}^3$ in general position a *center-saddle bifurcation* occurs as ρ passes through such a value. If e. g. the three conditions $a_1 = a_2, b_1 = b_2 = 0$ hold, then the curvatures at

$$z = \begin{pmatrix} 0 \\ 0 \\ -\frac{b_i}{a_i} \end{pmatrix}$$

coincide along all lines and the linearization at this equilibrium vanishes. The origin $z = 0$ is for all values of $a, b \in \mathbb{R}^3$ a (singular) equilibrium.

Systems in Two Degrees of Freedom

While all recurrent motion is periodic in one degree of freedom, the flow can have a stochastic (or chaotic) character already in two degrees of freedom. The level sets $\{H = h\}$ of the Hamiltonian are now 3-dimensional invariant hypersurfaces, and complicated dynamics is known to be possible from dimension three on. Still, being Hamiltonian imposes certain restrictions.

Leaving aside equilibria with vanishing eigenvalues, there are the following types of linearizations of regular equilibria in two degrees of freedom, with quadratic Hamiltonian H_2 in canonical co-ordinates.

Elliptic The quadratic part

$$H_2(q, p) = \alpha \frac{p_1^2 + q_1^2}{2} + \varpi \frac{p_2^2 + q_2^2}{2}, \text{ with } |\alpha| \leq |\varpi|, \quad (8)$$

is the superposition of two harmonic oscillators. For $\alpha/\varpi \in \mathbb{Q}$ the motion is periodic, but for irrational frequency ratio the trajectories spin densely around invariant tori. In the case $\varpi = -\alpha$ of 1 : -1 resonance one can add a nilpotent part

$$\frac{p_1^2 + q_1^2 + p_2^2 + q_2^2}{4} - \frac{p_1 q_2 + p_2 q_1}{2}$$

whence the linear flow becomes unbounded.

Hypo-elliptic The linear vector field with quadratic Hamiltonian

$$H_2(q, p) = a \frac{p_1^2 - q_1^2}{2} + \varpi \frac{p_2^2 + q_2^2}{2}$$

has one pair of eigenvalues on the real axis and one pair of eigenvalues on the imaginary axis. One also speaks of a saddle-center equilibrium.

Hyperbolic The linearization has no eigenvalues on the imaginary axis. In case the spectrum consists of two pairs of real eigenvalues the standard form of the Hamiltonian is

$$H_2(q, p) = a \frac{p_1^2 - q_1^2}{2} + b \frac{p_2^2 - q_2^2}{2}$$

and one speaks of a saddle-saddle equilibrium (or real saddle). In the alternative case of a complex quartet $\pm\alpha \pm i\varpi$ one has

$$H_2(q, p) = \alpha(p_1 q_1 + p_2 q_2) + \varpi(p_1 q_2 - p_2 q_1)$$

and speaks of a *focus-focus* equilibrium (or complex saddle), since the flow on the stable manifold spirals into the equilibrium and the flow on the unstable manifold spirals away from the equilibrium.

The 1 : -1 resonance is special in that it typically triggers off a *Hamiltonian Hopf bifurcation* during which an elliptic equilibrium becomes hyperbolic, of focus-focus type. Other bifurcations occur where eigenvalues vanish under variation of an external parameter (possibly parametrizing 4-dimensional symplectic leaves in a higher-dimensional Poisson space).

In generic Hamiltonian systems all equilibria are isolated and the periodic orbits form 1-parameter families. In one degree of freedom these families extend between (elliptic) equilibria and/or the (un)stable manifolds of (hyperbolic) equilibria, see Fig. 2. In two degrees of freedom the 1-parameter families of periodic orbits are special solutions and one may hope that they similarly organize the dynamics. The normal linear behavior of a periodic orbit with period τ is determined by its Floquet multipliers, the eigenvalues of the linearization $D\varphi_\tau$ of the flow.

The simplest way to find periodic orbits in a Hamiltonian system, e. g. as a starting point for continuation, is to look for the *normal modes* of an equilibrium.

Theorem (Liapunov) Let $\lambda_\pm = \pm i\alpha$ be a purely imaginary pair of eigenvalues of the linearization $A = DXH(z)$ at an equilibrium $z \in \mathcal{P}$ of a Hamiltonian system XH on a symplectic manifold \mathcal{P} for which no integer multiple $k\lambda_\pm$, $k \in \mathbb{N}$, is an (other) eigenvalue of A . Then XH admits a 1-parameter family $(\gamma_\varepsilon)_{0 < \varepsilon \leq \varepsilon_0}$ of periodic orbits that approach z as $\varepsilon \rightarrow 0$ with periods $T_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} 2\pi/\alpha$. The union

$$\{z\} \cup \bigcup_{0 < \varepsilon \leq \varepsilon_0} \gamma_\varepsilon$$

forms a smooth 2-dimensional submanifold of \mathcal{P} with boundary γ_{e_0} that is diffeomorphic to the closed disk in \mathbb{R}^2 .

For a proof see [1,23,26]; this result immediately generalizes to n degrees of freedom.

If the Hessian $D^2H(z)$ is positive (or negative) definite, the non-resonance condition $\lambda \neq k\lambda_{\pm}$ for the remaining eigenvalues of A can be dropped, but the resulting families of periodic orbits may no longer form manifolds through z and only form cones with z as vertex instead. In two degrees of freedom the $1 : -1$ and $1 : -2$ resonances provide examples where the normal mode is lacking, but for $1 : -k$ resonant equilibria with $k \geq 3$ it turns out to be generic for the normal modes to exist.

The normal mode of a hypo-elliptic equilibrium in two degrees of freedom is a hyperbolic periodic orbit with a real pair $a, (1/a) \neq \pm 1$ of Floquet multipliers. The periodic orbits born at an elliptic equilibrium also inherit their normal behavior from the “second” pair of eigenvalues and have a pair $e^{\pm i\varpi} \neq \pm 1$ of Floquet multipliers on the unit circle. Coupling two generic systems with one degree of freedom with a sufficiently weak interaction yields periodic orbits that are globally organized in this fashion, each periodic orbit being the superposition of a center or a saddle in one subsystem and a periodic orbit in the other subsystem.

The periodic orbits that are neither elliptic nor hyperbolic are isolated within the 1-parameter families of periodic orbits. The arcs in between consist either completely of elliptic or completely of hyperbolic periodic orbits and may be parametrized by the values of the energy. The following types of bifurcations are triggered off by parabolic periodic orbits, for which the *normal linear* behavior is governed by a Floquet matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$.

The periodic center-saddle bifurcation Under variation of the energy an elliptic and a hyperbolic periodic orbit meet at a parabolic periodic orbit with all Floquet multipliers equal to 1. No periodic orbit remains when further in(or de)creasing the energy. In a suitable projection

$$(H, I): \mathcal{P} \longrightarrow \mathbb{R}^2 \xrightarrow{H} \mathbb{R}$$

the family of periodic orbits forms a fold. See also Fig. 4 below.

The Hamiltonian period-doubling bifurcation Under variation of the energy an elliptic periodic orbit turns hyperbolic (or *vice versa*) when passing through a parabolic periodic orbit with Floquet multipliers -1 . Furthermore a family of periodic orbits with twice the period emerges from the parabolic periodic orbit,

inheriting the normal linear behavior from the initial periodic orbit. See also Fig. 5 below.

For proofs of this and the following see [20,23].

In generic systems with two degrees of freedom these are the only occurring bifurcations of periodic orbits. In three and more degrees of freedom there is “more space” in the normal direction and also periodic Hamiltonian Hopf bifurcations may occur. A prominent example is the gyroscopic stabilization of the rotating Lagrange top “standing up”, cf. [2,8,22].

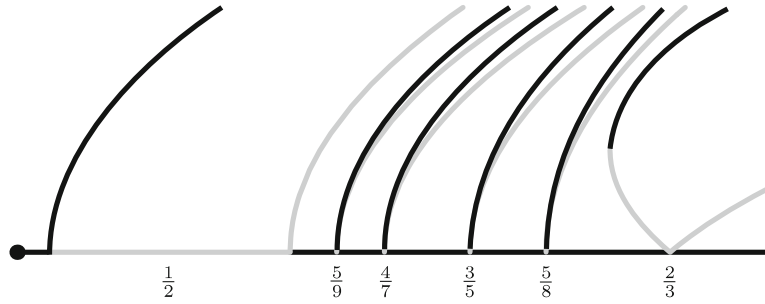
Along arcs of elliptic periodic orbits the pair $e^{\pm i\varpi}$ of Floquet multipliers passes regularly through roots of unity. Generically this happens on a dense set of parametrizing energy values, but for fixed denominator ℓ in $e^{\pm i\varpi} = e^{\pm 2\pi i k/\ell}$ the corresponding energy values are again isolated. The cases $\ell = 1$ and $\ell = 2$ correspond to the above bifurcations so the “first” case is $\ell = 3$.

Two arcs of hyperbolic periodic orbits emerge at elliptic orbits with Floquet multipliers $e^{\pm 2\pi i k/3}$, both with three times the period. One extends for lower and the other for higher energy values. The periodic orbit with Floquet multipliers $e^{\pm 2\pi i k/3}$ momentarily loses its (normal) stability due to these approaching unstable orbits.

In the case of Floquet multipliers $e^{\pm 2\pi i k/\ell}$ with $\ell \geq 5$ again two arcs of periodic orbits with ℓ times the period emerge, but now one is elliptic and the other hyperbolic, and they both extend to the same side of energy values (either lower or higher). Furthermore there is no momentary loss of (normal) stability as these families detach. For $\ell = 4$ both the $\ell = 3$ and the $\ell \geq 5$ scenarios can occur.

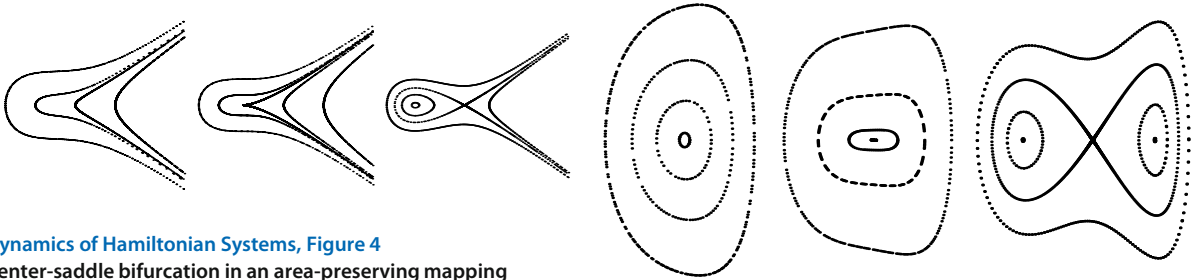
It is at isolated values of the energy that the two arcs of ($\ell \geq 5$)-times periodic orbits (or ($\ell = 4$)-times where appropriate) change from extending along lower energy values to extending along higher energy values. For definiteness let us consider a sub-arc between such values where all these exist for higher energy values. Where this sub-arc contains periodic orbits with Floquet multipliers $e^{\pm 2\pi i k/3}$ (or $e^{\pm 2\pi i k/4}$ where appropriate) the family of 3-times periodic orbits that extends for lower energy values typically vanishes in a nearby periodic center-saddle bifurcation. The family of elliptic periodic orbits born in this bifurcation then extends along higher energy values as well. A possible arc of elliptic periodic orbits is sketched in Fig. 3. For more details on periodic solutions of Hamiltonian systems see the entry ► [Periodic Orbits of Hamiltonian Systems](#) in this encyclopedia.

To visualize the flow on the 3-dimensional *energy shells* $\{H = h\}$ one uses iso-energetic Poincaré-sections, i. e. surfaces $\Sigma_h \subset \{H = h\}$ that are everywhere transverse to the vector field X_H . For recurrent points $z \in \Sigma_h$



Dynamics of Hamiltonian Systems, Figure 3

Subharmonic branching along a normal mode of an elliptic equilibrium in a response diagram with axes along normal frequency and amplitude. Where normal and internal frequency of the periodic orbit emanating from the equilibrium have a ratio $\varpi = 2\pi(k/\ell)$ with integers $k, \ell \in \mathbb{Z}$ two periodic orbits with ℓ times the period branch off, one elliptic and one hyperbolic (shown in gray). For $\ell = 3$ the distance between these widens as the two periodic orbits come into existence in a 3-periodic center-saddle bifurcation “before” the resonance after which the hyperbolic 3-periodic orbit passes through the normal mode at the $(2/3)$ -resonance (shown reflected as the vertical depicts the amplitude). For $\ell = 2$ the resonance leads to a gap within the elliptic normal mode, filled by hyperbolic periodic orbits and with boundaries marked by two Hamiltonian period-doubling bifurcations



Dynamics of Hamiltonian Systems, Figure 4

Center-saddle bifurcation in an area-preserving mapping

the (first) return time

$$\tau(z) := \min \{ T > 0 \mid \varphi_T(z) \in \Sigma_h \}$$

allows to define the *iso-energetic Poincaré-mapping*

$$P_h: \Sigma_h \longrightarrow \Sigma_h \\ z \mapsto \varphi_{\tau(z)}(z)$$

The phase space volume $\omega^2 = \omega \wedge \omega$ coming from the symplectic structure induces an area element σ on Σ_h that is preserved by P_h . On the other hand, every area-preserving mapping can be realized as an iso-energetic Poincaré-mapping of a Hamiltonian system.

A periodic orbit γ of XH with energy h corresponds to a fixed point $x \in \Sigma_h$ (or to a periodic point of period $k \in \mathbb{N}$ if γ has $k - 1$ intersections with Σ_h before returning to x). Two of the Floquet multipliers of γ are equal to 1, reflecting that periodic orbits form 1-parameter families in Hamiltonian systems and that moving the initial condition within γ yields that same periodic orbit with a translated time parametrization. The remaining two Floquet multipliers are the eigenvalues of the linearization $DP_h(z)$ of the iso-energetic Poincaré-mapping at a fixed point z .

Dynamics of Hamiltonian Systems, Figure 5

Period-doubling bifurcation in an area-preserving mapping

Because of area-preservation $\det DP_h(z) = 1$, so one eigenvalue of $DP_h(z)$ has to be the inverse of the other eigenvalue. For an elliptic periodic orbit both eigenvalues lie on the unit circle where the inverse equals the complex conjugate. For hyperbolic γ both eigenvalues are real and one sometimes makes the distinction between the direct hyperbolic case of positive eigenvalues and the inverse hyperbolic case of negative eigenvalues. See Fig. 4 for the center-saddle bifurcation triggered off by the double eigenvalue $+1$ and Fig. 5 for the period-doubling bifurcation triggered off by the double eigenvalue -1 .

Symmetry Reduction

When a dynamical system admits a symmetry group it is possible to simplify the dynamics. This reduction process is especially rewarding in Hamiltonian systems, where Noether's theorem yields for every continuous symmetry a conserved quantity. One even has the choice between first fixing the values of the conserved quantities and then

reducing what is left of the symmetry, or first reducing the symmetry and then fixing the remaining conserved quantities.

Let $H \in C^\infty(\mathcal{P})$ be a Hamiltonian function on the symplectic manifold (\mathcal{P}, ω) and G a compact Lie group (i. e. the group operation is smooth) with Lie algebra \mathfrak{g} (the tangent space $T_e G$ at the neutral element $e \in G$, provided with the Lie bracket that measures the non-commutativity of the group operation). Results for more general groups do exist, but some kind of compactness, e. g. that the *group action* be proper, is always needed. Assume that the group action

$$\begin{aligned} G \times \mathcal{P} &\longrightarrow \mathcal{P} \\ (g, z) &\mapsto gz \end{aligned} \quad (9)$$

preserves both the Hamiltonian function H and the symplectic form ω . Then G also preserves the Hamiltonian vector field XH and the resulting flow commutes with the group action (9), i. e. $\varphi_t \circ g = g \circ \varphi_t$ for all $(t, g) \in \mathbb{R} \times G$. In fact, this allows to combine the flow $\varphi: \mathbb{R} \times \mathcal{P} \longrightarrow \mathcal{P}$ of XH and the action (9) to the action

$$\begin{aligned} (\mathbb{R} \times G) \times \mathcal{P} &\longrightarrow \mathcal{P} \\ ((t, g), z) &\mapsto \varphi_t(gz) \end{aligned} \quad (10)$$

of the Lie group $\mathbb{R} \times G$ on \mathcal{P} .

Reduction aims to find a phase space of smaller dimension on which the dynamics can be studied. Identifying points $z \sim gz$ that are transformed into each other by group elements leads to the quotient space \mathcal{P}/G with canonical projection

$$\mathcal{P} \longrightarrow \mathcal{P}/G.$$

As $H(z) = H(gz)$ for all $g \in G$ the Hamiltonian induces a function on \mathcal{P}/G , again denoted by H . However, the symplectic structure on \mathcal{P} does not induce a symplectic structure on the quotient space. What can be transferred from \mathcal{P} to \mathcal{P}/G is the Poisson structure.

The symplectic form is preserved by G , whence the mappings $z \mapsto gz$ are canonical transformations, satisfying

$$\{F \circ g, H \circ g\} = \{F, H\} \circ g \quad \text{for all } F, H \in C^\infty(\mathcal{P}).$$

Thus, the Poisson bracket of G -invariant functions is again G -invariant, defining a Poisson bracket on \mathcal{P}/G . The Casimir functions on this Poisson space correspond to those G -invariant functions on \mathcal{P} that are conserved quantities for every G -invariant Hamiltonian function.

As an example consider the free rigid body with a fixed point, subject only to its own inertia. The configuration

space is $SO(3)$, all (rigid) rotations about the fixed point, with a Riemannian metric provided by the mass distribution. On the phase space $T^*SO(3)$ the Hamiltonian $H = T$ is given by the resulting kinetic energy.

The mass distribution and hence the Hamiltonian are invariant under rotations, making $SO(3)$ a symmetry group. Using the left trivialization

$$\begin{aligned} \lambda: T^*SO(3) &\longrightarrow SO(3) \times T_e^*SO(3) \\ \alpha_g &\mapsto (g, T_e^*L_g(\alpha_g)) \end{aligned}$$

defined by means of the left translation

$$\begin{aligned} L_g: SO(3) &\longrightarrow SO(3) \\ h &\mapsto gh \end{aligned}$$

the group action reads

$$\begin{aligned} SO(3) \times (SO(3) \times \mathbb{R}^3) &\longrightarrow SO(3) \times \mathbb{R}^3 \\ (g, (h, \ell)) &\mapsto (gh, \ell) \end{aligned}$$

and reveals $T^*SO(3)/SO(3) \cong \mathbb{R}^3$. Here $\ell \in \mathbb{R}^3 \cong \mathfrak{so}(3)^* = T_e^*SO(3)$ consists of the three components of the angular momentum with respect to a set of axes fixed in the body, whence one calls $\lambda(\alpha_g) = (g, \ell)$ body coordinates. Choosing the body set of axes along the principal axes of the rigid body the Hamiltonian takes the form

$$H(g, \ell) = H(\ell) = \frac{\ell_1^2}{2I_1} + \frac{\ell_2^2}{2I_2} + \frac{\ell_3^2}{2I_3}$$

where I_1, I_2, I_3 are the principal moments of inertia. The Poisson bracket relations inherited from $T^*SO(3)$ are

$$\{\ell_1, \ell_2\} = -\ell_3, \quad \{\ell_2, \ell_3\} = -\ell_1, \quad \{\ell_3, \ell_1\} = -\ell_2 \quad (11)$$

whence the Poisson structure on \mathbb{R}^3 is almost (6) considered in Sect. “The Phase Space”, differing only by a minus sign. The Casimir function $R(\ell) = (1/2)(\ell_1^2 + \ell_2^2 + \ell_3^2)$ measures (half of the square of) the length of the angular momentum, fixing this conserved quantity yields the symplectic leaves of the quotient space \mathbb{R}^3 .

The alternative approach to symmetric Hamiltonian systems first fixes the conserved quantities. For $\xi \in \mathfrak{g}$ the 1-parameter subgroup $\{\exp(s\xi) \mid s \in \mathbb{R}\}$ of G yields a conserved quantity

$$J^\xi: \mathcal{P} \longrightarrow \mathbb{R} \quad (12)$$

by Noether’s theorem, cf. [2,6,23]. The Hamiltonian vector field X_{J^ξ} has the flow $(s, z) \mapsto \exp(s\xi) \cdot z$ provided by the 1-parameter subgroup.

Here and from now on the assumption is made that the (mild) conditions for Noether's theorem are fulfilled, making (9) a Hamiltonian group action. What has to be avoided is that the flow provided by the 1-parameter subgroup is only locally Hamiltonian, see [1,27,28] for more details.

In the example of the free rigid body a conserved quantity (12) is the component of the angular momentum along an axis fixed in space. Collecting these conserved quantities by means of

$$\begin{aligned} J: \mathcal{P} &\longrightarrow \mathfrak{g}^* \\ z &\mapsto J(z): \mathfrak{g} \longrightarrow \mathbb{R} \\ \xi &\mapsto J^\xi(z) \end{aligned} \quad (13)$$

yields the *momentum mapping*. For the rigid body this amounts to fixing a set of axes in space and assigning to $\alpha_g \in T_g^*SO(3)$ the three components $\mu \in \mathbb{R}^3 \cong \mathfrak{so}(3)^*$ of the angular momentum with respect to these axes. Correspondingly, the right trivialization reads

$$\begin{aligned} \varrho: T^*SO(3) &\longrightarrow SO(3) \times \mathbb{R}^3 \\ \alpha_g &\mapsto (g, \mu) \end{aligned}$$

where $\mu = T_e^*R_g(\alpha_g)$ is obtained differentiating the right translation $R_g(h) = hg$. In space co-ordinates $\varrho(\alpha_g) = (g, \mu)$ the group action takes the form

$$\begin{aligned} SO(3) \times (SO(3) \times \mathbb{R}^3) &\longrightarrow SO(3) \times \mathbb{R}^3 \\ (g, (h, \mu)) &\mapsto (hg, g(\mu)) \end{aligned}$$

and one sees that the momentum mapping

$$SO(3) \times \mathbb{R}^3 \longrightarrow \mathbb{R}^3 \quad (14)$$

intertwines between the $SO(3)$ -actions on the phase space and on \mathbb{R}^3 .

To formulate the corresponding equivariance property of the momentum mapping in the general case the co-adjoint action

$$\begin{aligned} G \times \mathfrak{g}^* &\longrightarrow \mathfrak{g}^* \\ (g, \mu) &\mapsto \text{Ad}_{g^{-1}}^*(\mu) = \mu \circ \text{Ad}_g^{-1} \end{aligned} \quad (15)$$

is needed, here $\text{Ad}_g: \mathfrak{g} \longrightarrow \mathfrak{g}$ is the derivative $T_e(L_g R_{g^{-1}})$ of the inner automorphism $h \mapsto ghg^{-1} = L_g R_{g^{-1}} h$ at the neutral element. The momentum mapping J of the Hamiltonian group action (9) is now called equivariant if it intertwines between (9) and (15), i. e.

$$J(gz) = \text{Ad}_{g^{-1}}^* J(z) \quad \text{for all } (g, z) \in G \times \mathcal{P}.$$

In this case the assignment $\xi \mapsto J^\xi \in C^\infty(\mathcal{P})$ of conserved quantities to Lie algebra elements $\xi \in \mathfrak{g}$ turns out to be

a Lie algebra homomorphism, cf. [1,22,28]. Providing \mathfrak{g}^* with the *Lie-Poisson bracket* (defined in terms of the commutator on \mathfrak{g}) makes (13) a canonical mapping, see [27,28] for more details. The identification $\mathfrak{so}(3)^* \cong \mathbb{R}^3$ turns the Lie-Poisson bracket into (6) and indeed $\{\mu_1, \mu_2\} = \mu_3$ etc. for the spatial representation of the angular momentum. Working with right group actions (instead of acting from the left) would yield a minus sign here and a plus sign in (11).

Let $\mu \in \mathfrak{g}^*$ be a regular value of the equivariant momentum mapping (13). Then $J^{-1}(\mu)$ is a submanifold of \mathcal{P} and the (compact) isotropy subgroup

$$G_\mu = \left\{ g \in G \mid \text{Ad}_{g^{-1}}^*(\mu) = \mu \right\}$$

of the co-adjoint action leaves this manifold invariant as

$$J(gz) = \text{Ad}_{g^{-1}}^* J(z) = \mu \quad \text{for all } (g, z) \in G_\mu \times J^{-1}(\mu)$$

whence the restriction

$$\begin{aligned} G_\mu \times J^{-1}(\mu) &\longrightarrow J^{-1}(\mu) \\ (g, z) &\mapsto gz \end{aligned} \quad (16)$$

defines a Lie group action on the manifold $J^{-1}(\mu)$. Passing to the quotient $J^{-1}(\mu)/G_\mu$ is called *symplectic point reduction*.

Theorem (Marsden and Weinstein) *Let μ be a regular value of J and assume that the action (16) is free. Then the reduced phase space $\mathcal{P}_\mu = J^{-1}(\mu)/G_\mu$ is a symplectic manifold, with ω_μ uniquely determined by $\omega|_{J^{-1}(\mu)} = \omega_\mu \circ \pi_\mu$ where π_μ is the quotient projection. A G -invariant Hamiltonian function $H \in C^\infty(\mathcal{P})$ induces $H_\mu \in C^\infty(\mathcal{P}_\mu)$ and π_μ intertwines between the flows of X_H and X_{H_μ} .*

For a proof see [1,2,8,28].

In the example of the free rigid body the momentum mapping $SO(3) \times \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ is surjective and all values are regular, with inverse images $SO(3) \times \{\mu\}$. For $\mu \neq 0$ the isotropy subgroup $G_\mu \cong S^1$ consists of all rotations in \mathbb{R}^3 about the axis along μ and the reduced phase space

$$SO(3) \times \{\mu\}/G_\mu \cong SO(3)/S^1 \cong S^2$$

can be identified with the sphere of radius $|\mu|$ as

$$\begin{aligned} SO(3) \times \{\mu\} &\longrightarrow S_{|\mu|}^2 \\ (g, \mu) &\mapsto g(\mu) \end{aligned}$$

performs the reduction. The symplectic form on $S_{|\mu|}^2$ is a multiple

$$\frac{-1}{|\mu|^2} \cdot \sigma$$

of the area element σ (the minus sign is the "same" as in (11)). The isotropy subgroup of $\mu = 0$ is the whole

$SO(3)$ and the quotient space consists of the origin $\ell = g(0)$ in \mathbb{R}^3 .

As illustrated in this example, symplectic point reduction provides the symplectic leaves of the Poisson space \mathcal{P}/G obtained by directly reducing the action (9) on \mathcal{P} . A link between these two procedures of symmetry reduction is symplectic orbit reduction. By equivariance of (13) the inverse image $J^{-1}(G(\mu))$ of the co-adjoint orbit

$$G(\mu) = \left\{ \text{Ad}_{g^{-1}}^* \mu \mid g \in G \right\} \quad (17)$$

is invariant under the whole group G as

$$J(gz) = \text{Ad}_{g^{-1}}^* J(z) \in G(\mu) \text{ for all } (g, z) \in G \times J^{-1}(G(\mu)).$$

Hence, the restriction of (9) to $J^{-1}(G(\mu))$ defines a group action and the quotient spaces $J^{-1}(G(\mu))/G$ form a partition of \mathcal{P}/G (since

$$G = \bigcup_{\mu \in \mathfrak{g}^*} G(\mu)$$

and two orbits are either disjoint or equal). On the other hand, the two quotients $J^{-1}(G(\mu))/G$ and $J^{-1}(\mu)/G_\mu$ are symplectomorphic, see [27,28] for more details. In the example of the free rigid body, the inverse image of the co-adjoint orbit $G(\mu)$ is the sphere bundle in $T^*SO(3)$ of radius $|\mu|$ which is given by $SO(3) \times S_{|\mu|}^2$ both in body and space co-ordinates.

The group action (9) is free if every $z \in \mathcal{P}$ is moved away from z by every group element, i.e. all isotropy subgroups

$$G_z = \{ g \in G \mid gz = z \}$$

of (9) (not of the co-adjoint action) are trivial. For free actions the Poisson space \mathcal{P}/G is a Poisson manifold. Non-trivial isotropy groups typically lead to singularities. This can already be seen in the example of S^1 acting on $\mathbb{R}^4 = T^*\mathbb{R}^2 = \mathbb{R}^2 \times \mathbb{R}^2$ by simultaneous rotation in both planes, for the general theory see [8,27] and references therein.

The group $S^1 = SO(2)$ of planar rotations is commutative, whence the isotropy subgroups of the co-adjoint action coincide with the whole group and the generator $l = q_1 p_2 - q_2 p_1$ of the action

$$\begin{pmatrix} q \\ p \end{pmatrix} \mapsto \begin{pmatrix} \cos \rho & -\sin \rho & 0 & 0 \\ \sin \rho & \cos \rho & 0 & 0 \\ 0 & 0 & \cos \rho & -\sin \rho \\ 0 & 0 & \sin \rho & \cos \rho \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} \quad (18)$$

may be fixed both before and after passing to the quotient. The ring of S^1 -invariant functions is generated by the polynomials

$$x = \frac{q_1^2 + q_2^2}{2}, \quad y = \frac{p_1^2 + p_2^2}{2}, \quad z = p_1 q_1 + p_2 q_2 \quad \text{and } l$$

which are restricted by the relations

$$x \geq 0, \quad y \geq 0, \quad \text{and} \quad R_l(x, y, z) = 0$$

where

$$R_l(x, y, z) = \frac{1}{2} z^2 - 2xy + \frac{1}{2} l^2$$

defines the syzygy between these invariants. In particular, the S^1 -invariant Hamiltonian function H may be written as a function $H = H_l(x, y, z)$.

This allows to refrain from local co-ordinates and use (x, y, z) as (global) variables on the reduced phase spaces

$$\mathcal{P}_l = \{(x, y, z) \in \mathbb{R}^3 \mid x \geq 0, y \geq 0, R_l(x, y, z) = 0\} \quad (19)$$

which foliate the Poisson space

$$\mathbb{R}^4/S^1 = \bigcup_{l \in \mathbb{R}} \mathcal{P}_l \times \{l\} \subseteq \mathbb{R}^4$$

with Poisson brackets given in Table 1. Fixing the Casimir function l , the resulting Poisson structure on \mathbb{R}^3 can also be written as

$$\{F, H\} = \langle \nabla F \times \nabla H \mid \nabla R_l \rangle \quad (20)$$

(generalizing (6)), revealing R_l to be the second Casimir function.

The fixed point $(q, p) = 0$ of (18) has the whole group S^1 as isotropy group, all other isotropy groups of (18) are trivial. Thus, for $l \neq 0$ the positive sheets \mathcal{P}_l of the two-sheeted hyperboloids $R_l^{-1}(0)$ yield symplectic leaves of the Poisson space \mathbb{R}^4/S^1 . For $l = 0$ the reduced phase space \mathcal{P}_0 is stratified into two symplectic strata of dimensions 2 and 0, the positive part $x + y > 0$ of the double cone $R_0^{-1}(0)$ and the vertex of this cone, respectively.

Dynamics of Hamiltonian Systems, Table 1
Poisson structure on \mathbb{R}^4

$\{\downarrow, \rightarrow\}$	x	y	z	l
x	0	z	$2x$	0
y	$-z$	0	$-2y$	0
z	$-2x$	$2y$	0	0
l	0	0	0	0

Hence, the vertex is automatically an equilibrium of the reduced Hamiltonian system.

Where an action (9) of a discrete group G preserves Hamiltonian H and symplectic form ω there is no resulting conserved quantity, but one still can pass to the quotient \mathcal{P}/G . For instance, on \mathbb{R}^2 the Hamiltonians

$$H(q, p) = \frac{p^2}{2} \pm \frac{q^4}{24} + \frac{\lambda q^2}{2}$$

admit the symmetry group $G = \{\pm \text{id}\}$ that also preserves the symplectic form $dq \wedge dp$. The reduced Poisson space is the cone (19) with $l = 0$.

A symmetry may preserve only the equations of motion, but neither the Hamiltonian function nor the symplectic form. An example on \mathbb{R}^2 is given by

$$H(q, p) = \frac{p^3}{6} + \frac{pq^3}{6} + \lambda p + \mu pq$$

satisfying $H(q, -p) = -H(q, p)$ and $dq \wedge d(-p) = -dq \wedge dp$. The quotient may be realized as

$$\mathbb{R}^2 / \{p \mapsto -p\} = \{ (q, p) \in \mathbb{R}^2 \mid p \geq 0 \}$$

and inherits outside the boundary the symplectic structure $dq \wedge dp$. The q -axis is invariant under the flow, with equation of motion $\dot{q} = \partial H / \partial p$.

Integrable Systems

Let (\mathcal{P}, ω) be a symplectic manifold of dimension $\dim \mathcal{P} = 2n$ and $H \in C^\infty(\mathcal{P})$ a Hamiltonian function. The Hamiltonian vector field XH is completely (or Liouville) integrable if there are n functions $F_1, \dots, F_n \in C^\infty(\mathcal{P})$ with $\{F_i, H\} = 0$, $i = 1, \dots, n$ and $\{F_i, F_j\} = 0$, $i, j = 1, \dots, n$ (the F_i are *integrals* in involution) that are functionally independent outside a set of measure zero [1] or on an open and dense set [8], i. e. the closed set

$$\{z \in \mathcal{P} \mid \det(X_{F_1}(z), \dots, X_{F_n}(z)) = 0\} \quad (21)$$

is small, e. g. of non-zero co-dimension. Examples abound, among them are all linear systems, all systems with one degree of freedom and uncoupled superpositions of completely integrable systems. Since one may choose $F_1 := H$ a system in two degrees of freedom is integrable as soon as it has a conserved quantity F_2 . Where there is an S^1 -symmetry, as in the (planar) two body problem and the geodesic flow on a surface of revolution, the momentum mapping $J := F_2$ provides this conserved quantity.

The simplifying assumptions usually made when modeling e. g. a mechanical system often introduce extra symmetries. Consequently, some of the problems from classical mechanics, like the Lagrange top, turned out to be integrable. The continuous efforts of the 19th century lead

to more integrable systems, like the geodesic flow on a triaxial ellipsoid and the Kovalevskaya top. Eventually it became clear that integrable systems are the exception and non-integrable systems are the rule, with as most prominent example the three body problem. However, the discovery of the Toda lattice renewed interest and the list of known integrable systems is still growing.

Theorem (Liouville) *Let the components of $F: \mathcal{P} \rightarrow \mathbb{R}^n$ be n integrals in involution of the Hamiltonian $H = F_1$. For a regular value $c \in \text{im } F \subseteq \mathbb{R}^n$ a compact connected component \mathcal{R}_c of $F^{-1}(c)$ is an XH -invariant manifold that is diffeomorphic to the n -torus $\mathbb{T}^n = \mathbb{R}^n / \mathbb{Z}^n$. The subset $\mathcal{R}_c \subseteq \mathcal{P}$ has an open neighborhood \mathcal{U} on which XH admits action angle variables $(x, y) \in \mathbb{T}^n \times \mathbb{Y}$, $\mathbb{Y} \subseteq \mathbb{R}^n$ open, i. e. the diffeomorphism*

$$\begin{aligned} \mathcal{U} &\longrightarrow \mathbb{T}^n \times \mathbb{Y} \\ z &\mapsto (x(z), y(z)) \end{aligned}$$

turns the symplectic structure ω into $\sum dx_i \wedge dy_i$ and $y: \mathcal{U} \rightarrow \mathbb{Y}$ factors through $F: \mathcal{U} \rightarrow \mathbb{R}^n$.

For a proof see [1,2,8,26].

The last statement makes F independent of the angle variable x . In particular, $H = H(y)$ for the Hamiltonian function (though not $H = H(y_1)$ as in Hamilton–Jacobi theory) and the equations of motion read

$$\begin{aligned} \dot{x} &= \varpi(y) := DH(y) \\ \dot{y} &= 0 \end{aligned}$$

whence the flow $\varphi_t(x, y) = (x + t \cdot \varpi(y), y)$ is easily computed. Thus, constructing action angle variables of an integrable system is equivalent to explicitly solving the equations of motion. The term “completely integrable” indicates that this can be achieved by solving algebraic equations and indefinite integrals.

The invariant torus \mathcal{R}_c is Lagrangean, i. e. an *isotropic* submanifold (the symplectic structure vanishes on \mathcal{R}_c) and maximal with that property. The dimension of *Lagrangean submanifolds* is always equal to the number of degrees of freedom. The flow on \mathcal{R}_c is *conditionally periodic*, in the angle variable x it is parallel with frequency vector $\varpi(0)$, assuming that $y = 0 \in \mathbb{Y}$ is the action value of \mathcal{R}_c .

Denote by $C \subseteq \text{im } F \subseteq \mathbb{R}^n$ the regular values of F , the complement of the image of (21) under F in $\text{im } F$. Assume from now on that all level sets $F^{-1}(c)$, $c \in C$ are compact, e. g. because $F: \mathcal{P} \rightarrow \mathbb{R}^n$ is proper or all energy level sets are compact. Putting $\mathcal{R} := F^{-1}(C)$ yields an n -torus bundle with fibers \mathcal{R}_c , and assuming that all

$F^{-1}(c)$ are connected, the restriction

$$F: \mathcal{R} \longrightarrow C \quad (22)$$

can be used as projection mapping of this bundle. In case $\mathcal{P} = T^*M$ is the cotangent bundle of a configuration space M the \mathbb{T}^n -bundle is trivial if and only if there are global action angle variables. The first obstruction for a torus bundle to be trivial is *monodromy*. For instance, if the inverse image $S = \mathcal{P}^{\setminus \mathcal{R}}$ of the singular values of F contains an $(n-2)$ -parameter family of invariant $(n-2)$ -tori with normal linear behavior of focus-focus type, then the bundle (22) has non-trivial monodromy.

Associated to (22) is the homology bundle $H_1(\mathcal{R}/C, \mathbb{Z})$ of \mathcal{R} over C with fiber $H_1(\mathcal{R}_c, \mathbb{Z}) \cong \mathbb{Z}^n$, supplying the lattice that has to be divided out of $H_1(\mathcal{R}_c, \mathbb{R}) \cong \mathbb{R}^n$ to obtain the particular torus \mathcal{R}_c at $c \in C$. For each path $\gamma: [0, 1] \longrightarrow C$ the lift to $H_1(\mathcal{R}, \mathbb{Z})$ yields a bijective orientation-preserving \mathbb{Z} -linear mapping between the lattices at $\gamma(0)$ and $\gamma(1)$. For a closed loop the two lattices coincide and the mapping is represented by a matrix $\mathcal{M}(\gamma) \in SL_n(\mathbb{Z})$ with integer coefficients and determinant 1. This discrete object remains invariant under homotopies, and the resulting homomorphism

$$\mathcal{M}: \pi_1(C) \longrightarrow SL_n(\mathbb{Z})$$

from the first homotopy group of C into $SL_n(\mathbb{Z})$ is the monodromy of the n -torus bundle \mathcal{R} . In case $\mathcal{M} \equiv \text{id}$ one can uniquely move a chosen basis $Y_1(c), \dots, Y_n(c)$ of $H_1(\mathcal{R}_c, \mathbb{Z})$ at some chosen point $c \in C$ to all other fibers of the bundle $H_1(\mathcal{R}, \mathbb{Z})$, using paths with $\gamma(0) = c$. This yields Hamiltonian vector fields

$$X_{y_1} = Y_1 \circ F, \dots, X_{y_n} = Y_n \circ F$$

on \mathcal{R} for which the Hamiltonian functions $y_1, \dots, y_n \in C^\infty(\mathcal{R})$ are global action variables. The remaining task is to also find global angle variables.

The global actions y_i define Hamiltonian vector fields with periodic flows. This yields a free Hamiltonian group action

$$\mathbb{T}^n \times \mathcal{R} \longrightarrow \mathcal{R}$$

and makes (22) a *principal* torus bundle.

Now principal torus bundles are classified by their *Chern class* in $H^2(C, \mathbb{Z}^n)$, a discrete invariant measuring the obstruction to the existence of a global section

$$\sigma: C \longrightarrow \mathcal{R}$$

(i. e. a mapping satisfying $F \circ \sigma = \text{id}$). Such a global section yields in every fiber \mathcal{R}_c the desired “origin” $x = 0$ of

the angle variables and then one can let the group \mathbb{T}^n act. For the resulting globally defined x, y to be action angle variables the equation

$$dx \wedge dy = \sum_{i=1}^n dx_i \wedge dy_i = \omega$$

has to be fulfilled, if necessary adapting the section σ accordingly. The cohomology class $[\sigma^*\omega] \in H^2(C, \mathbb{R})$ is a continuous invariant that vanishes if and only if this can be achieved. In the particular case that ω is exact, being defined by means of a canonical 1-form, one has $[\sigma^*\omega] = \sigma^*[\omega] = 0$. See [10] for more details.

A special but important situation occurs if the globally defined Hamiltonian vector fields

$$X_{F_2}, \dots, X_{F_n} \quad (23)$$

already have periodic flows themselves (and only $F_1 = H$ has to be replaced by a local action y_1 in the construction of action angle variables). This defines an action

$$\mathbb{T}^{n-1} \times \mathcal{P} \longrightarrow \mathcal{P} \quad (24)$$

globally on the phase space (the restriction of which to the regular part \mathcal{R} is free, giving \mathcal{R} the structure of a principal $(n-1)$ -torus bundle). Reducing this symmetry yields a one-degree-of-freedom problem on (the symplectic leaves of) the base space in

$$\sigma: \mathcal{P} \longrightarrow \mathcal{P}/\mathbb{T}^{n-1}$$

with set Σ of singular values. Constructing action angle variables amounts to finding the time parametrizations of the (relative) periodic trajectories together with the areas encircled by these. Let $\mathcal{E} \subseteq \mathcal{P}/\mathbb{T}^{n-1}$ denote the set of (relative) regular equilibria.

The singular part $S = \mathcal{P}^{\setminus \mathcal{R}}$ is the union of the energy level sets containing points of $\sigma^{-1}(\Sigma)$ – here the $n-1$ vector fields (23) are linearly dependent – and those containing points of $\sigma^{-1}(\mathcal{E})$ – where $XH(z)$ is a linear combination of the linear independent vector fields (23). This makes \mathcal{P} a *ramified torus bundle*, with regular fibers in \mathcal{R} forming n -parameter families of Lagrangean n -tori, the distribution of which is determined by the collection S of singular fibers.

In case the action (24) is free, the set Σ is empty and σ makes the whole phase space \mathcal{P} a principal \mathbb{T}^{n-1} -bundle. The isotropic $(n-1)$ -tori reconstructed from \mathcal{E} behave similar to the periodic orbits in a two-degrees-of-freedom system described in Sect. “[Systems in Two Degrees of Freedom](#)”. Thus, the families of Lagrangean tori shrink down

to elliptic $(n-1)$ -tori and are separated by the (un)stable manifolds of $((n-1)$ -parameter families of) hyperbolic $(n-1)$ -tori, and the $(n-1)$ -tori may undergo bifurcations. However, these bifurcations are more involved than those of periodic orbits for three reasons.

- Normal-internal resonances $\langle k | \varpi \rangle = \ell \alpha$ between the internal frequencies $\varpi_1, \dots, \varpi_{n-1}$ and the *normal frequency* α of elliptic $(n-1)$ -tori with $k \in \mathbb{Z}^{n-1}$ and $\ell = 1, 2$ are dense, triggering off quasi-periodic center-saddle bifurcations and *frequency-halving bifurcations*.
- The occurring bifurcations may be degenerate and typically have co-dimensions up to $n-1$. This is a genericity condition on H , within the “universe” of integrable Hamiltonian systems on \mathcal{P} .
- It is also generic for heteroclinic bifurcations re-connecting the (un)stable manifolds of $(n-1)$ -tori to involve parabolic $(n-1)$ -tori.

In case the action (24) has non-trivial isotropy groups, the invariant tori reconstructed from Σ of dimensions $n-2, n-3, \dots, 2, 1, 0$ (the latter two being periodic orbits and equilibria) and their (un)stable manifolds form $\sigma^{-1}(\Sigma)$.

The description of \mathcal{P} as a ramified n -torus bundle still applies when some (or all) of the vector fields (23) do not have periodic flows and some (or all) of the action variables y_2, \dots, y_n are only locally defined. The Lagrangean tori in \mathcal{R} form n -parameter families and the singular fibers in S determine how these families fit together. At the $(n-1)$ -parameter families of elliptic $(n-1)$ -tori the Lagrangean tori shrink down in the same way as periodic orbits shrink down to centers in one degree of freedom. Different families of Lagrangean tori are separated by $(n-1)$ -parameter families of hyperbolic $(n-1)$ -tori and their (un)stable manifolds.

This picture is repeated in how the $(n-1)$ -tori shrink down to $(n-2)$ -parameter families of (partially) elliptic $(n-2)$ -tori and are separated by $(n-2)$ -parameter families of (partially) hyperbolic $(n-2)$ -tori and (part of) their (un)stable manifolds. Furthermore there are $(n-2)$ -parameter families of hyperbolic $(n-2)$ -tori of focus-focus type, together with their (un)stable manifolds these form *pinched n -tori*. These three ways lead to invariant tori of smaller and smaller dimension until ending up with 1-parameter families of periodic orbits and isolated equilibria.

Within the family of all $(n-1)$ -tori one encounters quasi-periodic center-saddle and frequency halving bifurcations along $(n-2)$ -parameter subfamilies and more generally bifurcations of co-dimension $k \leq n-1$ along $(n-k-1)$ -parameter subfamilies. Similarly, in-

variant $(n-2)$ -tori undergoing a quasi-periodic Hamiltonian Hopf bifurcation form $(n-3)$ -parameter families and the m -parameter families of invariant m -tori have $(m-k)$ -parameter subfamilies where bifurcations of co-dimension $k \leq m$ occur. Such bifurcations are not restricted to those of *semi-local* type, but may also involve coinciding stable and unstable manifolds of different invariant tori. For instance, heteroclinic orbits between hyperbolic $(n-1)$ -tori form $(2n-2)$ -dimensional submanifolds of the phase space.

Let $F_1, \dots, F_r \in C^\infty(\mathcal{P})$ be functionally independent integrals in involution. Fixing a point $z_0 \in \mathcal{P}$, the orbit $\mathbb{T}^r(z_0)$ of the \mathbb{T}^r -action generated by

$$X_{F_1}, \dots, X_{F_r} \quad (25)$$

is an r -torus to which the vector fields (25) are tangent. Hence, the symplectic structure ω vanishes on the manifold $\mathbb{T}^r(z_0) \subseteq \mathcal{P}$ whence this torus is isotropic, implying $r \leq n$. Consequently, if a Hamiltonian system X_H on \mathcal{P} is *superintegrable*, having more functionally independent integrals of motion than degrees of freedom, these cannot be all in involution. One therefore also speaks of non-commutative integrability in this context.

The “extra” integral makes superintegrable systems even more exceptional than integrable systems, although it usually can be attributed to a non-commutative symmetry group. An example is the Euler case of a free rigid body for which the four integrals are the energy and the three components of the angular momentum mapping (14) induced by the symmetry group $SO(3)$. For the planar and spatial two body problem (the Kepler system) the symmetry groups $SO(3)$ and $SO(4)$ lead to 3 and 5 independent integrals of motion, respectively. The latter is exceptional *even within the class of superintegrable systems*; replacing the inverse square attraction by another central force (not the harmonic oscillator with its 5 independent integrals of motion due to an $SU(3)$ -symmetry) breaks the symmetry down to an $SO(3)$ with still 4 independent integrals of motions.

Theorem (Nekhoroshev, Mishchenko and Fomenko)

On the subset $\mathcal{R} \subseteq \mathcal{P}$ let $F: \mathcal{R} \rightarrow \mathbb{R}^{2n-r}$ be a submersion with compact and connected fibers (hence, a fibration). Assume that $\{F_i, F_j\} = P_{ij} \circ F$, $i, j = 1, \dots, 2n-r$ and that the matrix P with entries $P_{ij}: \mathcal{P} \rightarrow \mathbb{R}$ has rank $2(n-r)$ at all points of $F(\mathcal{P})$. Then every fiber of F is diffeomorphic to \mathbb{T}^r and the fibration F has local trivializations which are symplectic.

This formulation is taken from [11], where the geometric contents of this theorem is thoroughly described in terms

of this fibration and an associated co-fibration with fibers of dimension $2n - r$.

Thus, every fiber of F has a neighborhood \mathcal{U} with co-ordinates

$$(x, y, q, p) : \mathcal{U} \longrightarrow \mathbb{T}^r \times \mathbb{R}^r \times \mathbb{R}^{n-r} \times \mathbb{R}^{n-r}$$

such that the level sets of F coincide with the level sets of (y, q, p) and

$$\sigma|_{\mathcal{U}} = \sum_{i=1}^r dx_i \wedge dy_i + \sum_{j=1}^{n-r} dq_j \wedge dp_j.$$

These co-ordinates are Nekhoroshev's generalized action-angle variables. Where superintegrability is due to a non-commutative symmetry group G , the $2(n - r)$ parameters "live" in the co-adjoint orbits (17).

Perturbation Analysis

An important property of integrable Hamiltonian systems is their behavior under small perturbations. For a satisfactory description of e. g. mechanical systems the simplifying assumptions used to derive the model should not completely change the dynamics, some kind of "robustness" is desirable. An instance where the underlying approximation by a "simpler" system is part of the mathematical treatment is normal form theory.

Using a series expansion of the Hamiltonian function, the aim of normalization is to find co-ordinates in which the terms of the expansion look particularly simple (whence the Hamiltonian vector field takes a particularly simple form as well). This is an algorithmic procedure that inductively pushes a torus symmetry through the series. While the resulting series are typically divergent, a well-chosen truncation yields a normalized approximation for which good estimates are available. Already existing symmetries are preserved and the choices to be made when normalizing a given Hamiltonian are largely dictated ensuring that the combination of acquired and inherited symmetries render the system integrable. This does not always work, the integrability of normal forms is automatic only in two degrees of freedom, cf. [29]. For more details on normal forms in perturbation theory see the entry [► Normal Forms in Perturbation Theory](#) in this encyclopedia.

Normalization often results in a torus action (24) on the phase space \mathcal{P} , for which the components F_2, \dots, F_n of the momentum mapping

$$J : \mathcal{P} \longrightarrow \mathbb{R}^{n-1}$$

define Hamiltonian vector fields (23) with periodic flows. This allows for the detailed description given in the previous section, and the question is what remains of this description when the symmetry (24) is broken.

The most prominent part of the ramified torus bundle defined by an integrable Hamiltonian system are the families of Lagrangean tori. Persistence of invariant tori can only be expected if these have a dynamical meaning. For instance, an invariant 1-torus that consists of a union of equilibria instead of being a periodic orbit is highly unlikely to remain present in a perturbed system. Similarly, an invariant torus with conditionally periodic motion has dynamical meaning if it is the closure of a dense orbit. This excludes resonances

$$\langle k | \varpi \rangle = 0, \quad k \in \mathbb{Z}^{m \setminus \{0\}}$$

between the frequencies $\varpi_1, \dots, \varpi_n$ of the parallel flow on the invariant torus \mathbb{T} . The parallel nature of the flow implies that for a non-resonant frequency vector ϖ the time average

$$\int_{-\infty}^{+\infty} f(x(t)) dt = \int_{\mathbb{T}} f(x) dx$$

of some function $f : \mathbb{T} \longrightarrow \mathbb{R}$ along the quasi-periodic motion $x(t) = x + t\varpi$ is equal to the space average over the torus.

This space average is approximated "quickly" – taking the time average over finite intervals of time – if ϖ is *Diophantine*, satisfying the strong non-resonance condition

$$|\langle k | \varpi \rangle| \geq \frac{\gamma}{|k|^\tau} \quad \text{for all } k \in \mathbb{Z}^{m \setminus \{0\}} \quad (26)$$

with constants $\gamma > 0$ and $\tau > n - 1$. The set $\mathbb{R}_{\tau, \gamma}^n$ of (τ, γ) -Diophantine $\varpi \in \mathbb{R}^n$ has the local structure

$$\mathbb{R} \times \text{Cantor dust}; \quad (27)$$

for $\varpi \in \mathbb{R}_{\tau, \gamma}^n$ also $s\varpi \in \mathbb{R}_{\tau, \gamma}^n$ for all $s \geq 1$, and the intersection $S^{n-1} \cap \mathbb{R}_{\tau, \gamma}^n$ is perfect and totally disconnected. While (Lebesgue)-almost all frequency vectors are non-resonant, the complement of Diophantine frequency vectors is an open and dense set. Still, the relative measure of $\mathbb{R}_{\tau, \gamma}^n$ goes to 1 as $\gamma \rightarrow 0$. The celebrated KAM theorem yields persistence of Lagrangean tori with Diophantine frequency vector.

Theorem (Kolmogorov, Arnol'd and Moser) *Let $\mathbb{Y} \subseteq \mathbb{R}^n$ be an open neighborhood of the origin and consider the phase space $\mathcal{P} = \mathbb{T}^n \times \mathbb{Y}$ with symplectic structure $dx \wedge dy$. Let the Hamiltonian*

$$H_\varepsilon(x, y) = H_0(y) + \varepsilon H_1(x, y; \varepsilon)$$

be real analytic and non-degenerate, satisfying

$$\det D^2 H_0(y) \neq 0 \quad \text{for all } y \in \mathbb{Y}. \quad (28)$$

Then there exists $\varepsilon_0 > 0$ such that for all $|\varepsilon| < \varepsilon_0$ there is a canonical transformation ϕ_ε near the identity and a measure-theoretically large set $\mathbb{Y}'_\varepsilon \subseteq \mathbb{Y}$ with the property that for $p \in \mathbb{Y}'_\varepsilon$ the transformed Hamiltonian $H_\varepsilon \circ \phi_\varepsilon$ does not depend on $q \in \mathbb{T}^n$.

For a proof see [3,7,25].

The Kolmogorov non-degeneracy condition (28) expresses that the frequency mapping

$$\begin{aligned} \varpi: \mathbb{Y} &\longrightarrow \mathbb{R}^n \\ y &\mapsto DH_0(y) \end{aligned} \quad (29)$$

is locally a diffeomorphism. This allows to pull back the whole geometry defined by (26) into phase space to obtain

$$\mathbb{Y}_\varepsilon = \varpi^{-1}(\mathbb{R}_{\tau,\gamma}^n)$$

(and from this the subset $\mathbb{Y}'_\varepsilon \subseteq \mathbb{Y}_\varepsilon$ by omitting points γ -close to the boundary $\partial\mathbb{Y}$). The constant γ is chosen as a function of ε – of order $\mathcal{O}(\sqrt{\varepsilon})$ – to find an optimal balance between the (relative) measure of \mathbb{Y}'_ε and the deviation of ϕ_ε from the identity.

The subset $\mathbb{T}^n \times \mathbb{Y}'_\varepsilon \subseteq \mathcal{P}$ consists (in the transformed variables) of quasi-periodic tori since \dot{p} vanishes for $p \in \mathbb{Y}'_\varepsilon$. The theorem makes no statement for $p \in \mathbb{Y} \setminus \mathbb{Y}'_\varepsilon$. For $n \geq 3$ this leaves the possibility of Arnol'd diffusion, trajectories that venture off to distant parts of the phase space.

While the KAM theorem concerns the fate of “most” trajectories and for all times, the complementary Nekhoroshev theory concerns all trajectories and states that they stay close to the unperturbed tori for times of the order

$$\exp \left[\left(\frac{\varepsilon_0}{\varepsilon} \right)^{\frac{1}{2n}} \right].$$

Here analyticity of the Hamiltonian is a necessary ingredient, for finitely differentiable Hamiltonians one only obtains polynomial times. In the above formulation of the KAM theorem the assumption of analyticity of the Hamiltonian can be weakened without essential changes for the result, during the proof one merely has to intersperse an analytic approximation at each iteration step. The diffusion is even superexponentially slow for trajectories starting close to surviving tori, see [30] for more details on this phenomenon of exponential condensation, and see the entry ▶ [Nekhoroshev Theory](#) in this encyclopedia for more details on Nekhoroshev theory.

Where the energy level sets are transversal to the continuous direction in \mathbb{Y}'_ε one has persistence of most Lagrangean tori on each energy shell, parametrized by *Cantor dust*. The same result is obtained under the condition of iso-energetic non-degeneracy

$$\det \begin{pmatrix} D^2 H_0(y) & \nabla H_0(y) \\ DH_0(y) & 0 \end{pmatrix} \neq 0, \quad (30)$$

which is independent of Kolmogorov's condition. It is generic for an integrable system to satisfy both conditions almost everywhere. However, in applications it is a non-trivial task to actually check this and to determine the hypersurfaces in action space where these determinants vanish.

In two degrees of freedom the energy shells $\{H = h\}$ are 3-dimensional invariant manifolds for regular values h of the Hamiltonian and are separated by each Lagrangean torus. Thus, for initial conditions (q_0, p_0) in the co-ordinates provided by the KAM theorem with $p_0 \notin \mathbb{Y}'_\varepsilon$ the persistent tori parametrized by \mathbb{Y}'_ε still have dynamical consequences as the trajectory is confined between two such tori, so $|p(t) - p_0|$ admits a bound of order $\mathcal{O}(\sqrt{\varepsilon})$. As a consequence, one obtains (dynamical) stability for all elliptic equilibria in generic Hamiltonian systems with two degrees of freedom.

Indeed, if the Hessian of (8) is (positive or negative) definite, then the Hamiltonian serves as a Liapunov function. In the indefinite case one also includes third and fourth order terms in the analysis and passes to a Birkhoff normal form

$$\alpha I_1 + \varphi I_2 + \frac{\beta}{2} I_1^2 + \delta I_1 I_2 + \frac{\chi}{2} I_2^2 \quad (31)$$

with $I_1 = (1/2)(p_1^2 + q_1^2)$ and $I_2 = (1/2)(p_2^2 + q_2^2)$. This can be achieved if there are no low order resonances $\varpi = -k\alpha$, $k = 1, 2, 3$ between the two (normal) frequencies of the equilibrium. A second genericity condition on the Hamiltonian is that the linear part of (31) does not divide the quadratic part (in the I_i), ensuring that (30) holds in a whole neighborhood of the elliptic equilibrium.

The Cantor set structure defined by the Diophantine conditions (26) can be used to weaken the necessary non-degeneracy condition. Since the gaps are defined by linear inequalities, the conditions on the first derivatives of the frequency mapping (29) can be replaced by conditions on the curvature or even higher derivatives. Such Rüssmann-like conditions still guarantee that the relative measure of surviving tori tends to 1 as the perturbation strength tends to zero, but at a price. For instance, the highest derivative $L \in \mathbb{N}$ needed in

$$\left\langle \frac{\partial^{|\ell|} \omega}{\partial y} \mid |\ell| \leq L \right\rangle = \mathbb{R}^n \quad (32)$$

enters the Diophantine conditions on the frequency vector by means of the inequality $\tau > nL - 1$ on the Diophantine constant τ . For more details see the entry ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#) in this encyclopedia.

The KAM theorem is a semi-local result, valid in the neighborhood of an initial torus that admits action angle variables. A global version is obtained in [5]. The global conjugacy is glued together from convex combinations of local conjugacies using a partition of unity, the key ingredient being a unicity result on the tori obtained in the KAM theorem.

In the integrable approximation the distribution of the n -parameter families of Lagrangean tori is determined by the singular part S of the ramified torus bundle \mathcal{P} . Since S consists of families of lower dimensional tori together with their (un)stable manifolds, the persistence of (isotropic) m -tori, $m < n$, becomes important. For $m = 0$ the persistence of equilibria, together with their linear behavior (a superposition of what is possible in one and two degrees of freedom) follows from the implicit mapping theorem as it is generic for the Hamiltonian function that no equilibrium has vanishing (and neither multiple) eigenvalues. In the periodic case $m = 1$ the Diophantine condition (26) ensures $\varpi \neq 0$ and the 1-parameter families of periodic orbits persist as well, together with occurring bifurcations. In 1-parameter families all bifurcations are generically of co-dimension 1 – a genericity condition on the Hamiltonian. Bifurcations of higher co-dimension would not be expected to persist.

For hyperbolic tori the criteria remain valid almost verbatim; the key step is to pass to a center manifold. A technical difficulty is that even for analytic Hamiltonians center manifolds may only be of finite differentiability. While KAM theorems remain true in this context, the analytic context has its advantages – for instance (32) is satisfied for some $L \in \mathbb{N}$ for an analytic frequency mapping ϖ if and only if $\text{im } \varpi$ does not lie within a linear hyperplane. An alternative is therefore to prove persistence of hyperbolic tori directly, this also gives a more direct hold on their stable and unstable manifolds.

Elliptic $(n - 1)$ -tori need one extra parameter to control the normal frequency as well. Similar to the isoenergetic case one can use time re-parametrization and obtain Cantor families of persistent elliptic $(n - 1)$ -tori parametrized by Cantor dust. Where there are more than one normal frequency to control this can no longer be done in a linear way; a problem solved by Rüssmann-like conditions on the higher derivatives of the frequency vector, see [7] and references therein. In case the mapping of internal frequencies satisfies Kolmogorov’s condition, the higher order derivatives are only needed of normal fre-

quencies. Now normal frequencies α_j enter the Diophantine conditions

$$|2\pi \langle k | \varpi \rangle + \langle \ell | \alpha \rangle| \geq \frac{\gamma}{|k|^\tau} \quad (33)$$

only as combinations $\langle \ell | \alpha \rangle$ with $|\ell| \leq 2$. This allows to extend the result to finite-dimensional elliptic tori in infinitely many degrees of freedom, cf. [14,15] and the entry ► [Perturbation Theory for PDEs](#) in this encyclopedia. For hypo-elliptic tori one may deal with the hyperbolic part by means of a center manifold or use a direct approach. Such m -tori have k additional pairs of purely imaginary Floquet exponents and excitation of normal modes leads for $l = 1, \dots, k$ to $(m + l)$ -parameter families of $(n + l)$ -tori inheriting the “remaining” normal linear behavior, see [30] and references therein.

Where (lower-dimensional) m -tori undergo a semi-local bifurcation the m actions y conjugate to the toral angles x first of all have to versally unfold the bifurcation scenario. It is generic for the integrable Hamiltonian H that the m -parameter families of m -tori, $1 \leq m \leq n - 1$, do not encounter bifurcations of co-dimension higher than m , so this is possible. The curvature of the frequency mapping is then used to ensure Diophanticty of most bifurcating tori, i.e. a Rüssmann-like condition with $L = 2$ is sufficient, cf. [13]. This curvature requirement is not necessary for 2-tori; these may undergo the quasi-periodic analogues of the co-dimension one bifurcations of periodic orbits as co-dimension two bifurcations are isolated within these 2-parameter families and cannot be prevented to disappear in resonance gaps.

Small perturbations of an integrable Hamiltonian thus lead to a Cantorification of the ramified n -torus bundle as sketched in Fig. 6, the stratification of the action space into various subfamilies parametrizing the tori is replaced by a Cantor stratification. Of equal importance are those changes that make sure that the non-integrable perturbed dynamics is indeed qualitatively different from the integrable unperturbed dynamics. While the former persistence results are obtained upon genericity conditions on the unperturbed system, such changes require the perturbation to be generic.

Disintegrating Lagrangean tori lead to invariant m -tori, where $n - m$ is the number of independent resonances $\langle k | \varpi \rangle = 0$ of the (internal) frequencies. Most of these lower dimensional tori will be elliptic or hyperbolic. The new hyperbolic tori lie at the basis of a possible scenario for Arnol’d diffusion. One of the effects of a small generic perturbation is that stable and unstable manifolds of hyperbolic periodic orbits no longer coincide, but split and intersect transversely, cf. [1]. This carries over to hy-



Dynamics of Hamiltonian Systems, Figure 6

A typical decomposition of the action space of a nearly integrable system in three degrees of freedom. The 2-dimensional Cantor dust parametrizes elliptic 2-tori that vanish in quasi-periodic center-saddle bifurcations along the fold lines. Between these extend arcs of hyperbolic 2-tori, parametrized by a 2-dimensional set of the form (27). The 1-dimensional Cantor dust along the folds consists of Lebesgue density points of these 2-dimensional parameter sets in the same way that these consist of Lebesgue density points of 3-dimensional sets of the form (27) above and below the surface which parametrize Lagrangean 3-tori

hyperbolic tori. The splitting of separatrices also leads to transverse intersections of stable and unstable manifolds of neighboring hyperbolic tori in the same energy shell. These hyperbolic tori form a Cantor family, and one of the main problems is to make sure that the transition chain of hyperbolic tori and their heteroclinic connections bridges the occurring gaps, cf. [9]. The dynamics in the gaps of Cantor families of hyperbolic tori can already be studied in the perturbation near resonant singular fibers of the ramified n -torus bundle. On the center manifold these become again (resonant) regular fibers, but the full perturbed motion is superposed by the hyperbolic dynamics in the symplectic normal directions.

The nature of the gaps opened by violations of (33) in families of elliptic tori is twofold. Internal resonances $\langle k | \varpi \rangle = 0$ lead to the destruction of the torus. Boundary points of the gaps resulting from normal-internal resonances are related to quasi-periodic bifurcations. The resonance $\alpha = 2\pi \langle k | \varpi \rangle$ triggers off a (quasi-periodic) center-saddle bifurcation and resonance gaps

$$|2\pi \langle k | \varpi \rangle + 2\alpha| < \frac{\gamma}{|k|^\tau}$$

are completely filled by hyperbolic tori that terminate in frequency halving bifurcations.

The maximal tori of superintegrable systems are m -tori in $n > m$ degrees of freedom and their normal behavior vanishes, both linearly and non-linearly. The strat-

egy when studying a perturbation of such a *properly degenerate* system is to find an intermediate system that is also integrable, but non-degenerately so. The perturbation H_1 of a superintegrable Hamiltonian H_0 removes the degeneracy if the perturbed Hamiltonian $H_\varepsilon = H_0 + \varepsilon H_1$ can be written in the form

$$H = H_0 + \varepsilon \tilde{H}_1 + \varepsilon^2 H_2$$

where $H_0 + \varepsilon \tilde{H}_1$ is a non-degenerate integrable Hamiltonian. Since \tilde{H}_1 is defined in terms of H_1 (e. g. as its average along the unperturbed flow defined by H_0) every genericity condition on the intermediate system puts genericity conditions on the perturbation H_1 . This first step of a normal form procedure lies also at the basis of Nekhoroshev theory for superintegrable systems, see [11] for more details.

Future Directions

Many Hamiltonian systems modeling real phenomena have symmetries, and the conditions of the regular reduction theorem of Marsden and Weinstein are often not fulfilled. The regularity assumptions could be successfully removed, cf. [8,24,27], and progress is still made in weakening the compactness conditions. Since the flow $\varphi: \mathbb{R} \times \mathcal{P} \rightarrow \mathcal{P}$ itself is also a group action some condition has to exclude too general situations. It is for this reason that the symmetry (9) is studied in Sect. “Symmetry Reduction” and not the larger symmetry (10).

In an integrable system, action angle variables define a \mathbb{T}^n -action in the neighborhood of a given Lagrangean torus. Globally one only has the \mathbb{R}^n -action defined by the commuting flows of the integrals F_1, \dots, F_n . The flow of $X_H = X_{F_1}$ is the actual object of study, and in general the flows of the vector fields (23) may be as complicated.

On the topological level, the \mathbb{T}^n -bundle \mathcal{R} has a monodromy mapping \mathcal{M} and the result of [10] explained in Sect. “Integrable Systems” allows one to uniquely characterize \mathcal{R} by its Chern class only if \mathcal{M} is trivial, such bundles are isomorphic if and only if their Chern classes coincide. An extension of this characterization to the case of non-trivial monodromy has been derived in [18,19]. The topological bundle \mathcal{R} is uniquely characterized by its affine structure and Chern class, and the symplectic bundle \mathcal{R} by its affine structure and Lagrange class.

The global version of the KAM theorem provides a Cantorification of the global bundle \mathcal{R} . This makes also non-local properties subject to the perturbation analysis, for instance showing a discrete invariant like monodromy to persist. Globally on the phase space, perturbation not only of the Hamiltonian H but also of the symplectic struc-

ture ω becomes a well-defined problem. For instance, if $\omega = -d\vartheta$ is exact one may add a small non-exact closed 2-form $\varepsilon\sigma$ whence $\omega + \varepsilon\sigma$ is a non-exact symplectic form on the phase space.

A Lagrangean torus with $n - 1$ independent resonances consists of periodic orbits. When the torus breaks up under the perturbation, only finitely many of these are expected to survive. At the same time the trivial normal behavior of these periodic orbits changes, resulting in hyperbolic and elliptic periodic orbits. The latter can serve as starting points for the construction of *solenoids*, cf. [4,21]. This construction should carry over to elliptic tori, where the “encircling” tori emerge from normal-internal resonances and might also result in solenoids that are limits of tori with varying dimension.

The results in [13] address persistence of Diophantine tori involved in a bifurcation and the corresponding gaps trigger off new phenomena, cf. [17]. Internally resonant tori involved in a quasi-periodic bifurcation may result in large dynamical instabilities, especially where multiple parabolic resonances are encountered. The effect is further amplified for tangent (or flat) parabolic resonances, which fail to satisfy the iso-energetic non-degeneracy condition.

The high co-dimensions of bifurcations that may be encountered within families of isotropic tori makes it necessary to study Hamiltonian bifurcations that have been left aside since they generically do not occur for periodic orbits. The coupling of the three types of co-dimension one bifurcations is unavoidable where the resonance gaps defined by (33) with $\ell \neq 0$ intersect. Already for an equilibrium with linearization

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

the nearby dynamics is extremely complicated, with all possible resonances of equilibria in two degrees of freedom occurring in a versal unfolding.

A perturbed superintegrable system can lead to the combination of two bifurcations in both the fast and the slow dynamics. With two different time scales e. g. the dynamics triggered off by two simultaneous violations of (33) appears to be of $(1 + 1)$ -degree-of-freedom rather than having truly 2 degrees of freedom. This might render this problem more accessible.

For more details on Hamiltonian perturbation theory see the entry [► Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#) in this encyclopedia and references therein. An important subject are variational methods, which can be used to obtain periodic solutions (but also

more complicated dynamics); see the entry [► Periodic Orbits of Hamiltonian Systems](#) on this subject and also the entry [► \$n\$ -Body Problem and Choreographies](#) for a striking application to celestial mechanics.

Acknowledgments

The author thanks Henk Broer, Marius Crainic, Hans Duistermaat, Luuk Hoevenaars, Hessel Posthuma and Ferdinand Verhulst for fruitful discussions and helpful remarks.

Bibliography

Primary Literature

1. Abraham R, Marsden JE (1978) Foundations of Mechanics, 2nd edn. Benjamin
2. Arnol'd VI (1989) Mathematical Methods of Classical Mechanics, 2nd edn. Springer
3. Arnol'd VI, Avez A (1968) Ergodic problems of classical mechanics. Benjamin
4. Birkhoff BD (1935) Nouvelles recherches sur les systèmes dynamiques. Mem Pont Acad Sci Novi Lyncae 3(1):85–216
5. Broer HW, Cushman R, Fassò F, Takens F (2007) Geometry of KAM tori for nearly integrable Hamiltonian systems. Ergod Theory Dyn Syst 27(3):725–741
6. Broer HW, Dumortier F, van Strien SJ, Takens F (1991) Structures in dynamics. Finite-dimensional deterministic studies. Stud Math Phys 2
7. Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-Periodic Motions in Families of Dynamical Systems: Order amidst Chaos. Lect Notes Math, vol 1645. Springer, Berlin
8. Cushman RH, Bates LM (1997) Global Aspects of Classical Integrable Systems. Birkhäuser
9. Delshams A, de la Llave R, Martínez-Seara T (2006) A Geometric Mechanism for Diffusion in Hamiltonian Systems. Overcoming the Large Gap Problem: Heuristics and Rigorous Verification on a Model. Mem AMS 179(844):1–141
10. Duistermaat JJ (1980) On global action-angle coordinates. Comm Pure Appl Math 33:687–706
11. Fassò F (2005) Superintegrable Hamiltonian systems: Geometry and Perturbation. In: Gaeta G (ed) Symmetry and Perturbation Theory, Cala Gonone 2004. Acta Appl Math 87:93–121
12. Goldstein H (1980) Classical Mechanics. Addison-Wesley
13. Hanßmann H (2007) Local and Semi-Local Bifurcations in Hamiltonian Dynamical Systems – Results and Examples. Lect Notes Math, vol 1893. Springer, Berlin
14. Kappeler T, Pöschel J (2003) KdV & KAM. Erg Math Grenzgeb 3(45)
15. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. Lect Notes Math, vol 1556. Springer, Berlin
16. Landau LD, Lifshitz EM (1960) Mechanics. In: Course of Theoretical Physics 1. Pergamon
17. Litvak-Hinzenon A, Rom-Kedar V (2002) Parabolic resonances in 3 degree of freedom near-integrable Hamiltonian systems. Phys D 164:213–250
18. Lukina O (to appear) Torus bundles in integrable Hamiltonian systems. Ph D thesis. Rijksuniversiteit Groningen, Groningen

19. Lukina O, Takens F, Broer HW (2008) Global properties of integrable Hamiltonian systems. Reprint, Rijksuniversiteit Groningen, Groningen
20. MacKay RS (1993) Renormalisation in Area-preserving Maps. World Scientific
21. Markus L, Meyer KR (1980) Periodic orbits and solenoids in generic Hamiltonian dynamical systems. *Am J Math* 102(1):25–92
22. Marsden JE, Rañiu TS (1994) Introduction to Mechanics and Symmetry. Springer
23. Meyer KR, Hall GR (1992) Introduction to Hamiltonian Dynamical Systems and the *N*-Body Problem. *Appl Math Sci* 90
24. Montaldi J, Rañiu TS (2005) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ Press
25. Moser J (1973) Stable and Random Motions in Dynamical Systems: With Special Emphasis on Celestial Mechanics. Princeton Univ Press
26. Moser J, Zehnder EJ (2005) Notes on Dynamical Systems. Courant Lect Notes Math 12, AMS
27. Ortega J-P, Rañiu TS (2004) Momentum Maps and Hamiltonian Reduction. *Prog Math* 222
28. Rañiu TS, Tudoran R, Sbano L, Sousa Dias E, Terra G (2005) A Crash Course in Geometric Mechanics. In: Montaldi J, Rañiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ Press, pp 23–156
29. Sanders JA, Verhulst F, Murdock J (2007) Averaging Methods in Nonlinear Dynamical Systems, 2nd edn. Springer
30. Sevryuk MB (2003) The classical KAM theory at the dawn of the twenty-first century. *Mosc Math J* 3(3):1113–1144
31. Siegel CL, Moser JK (1971) Lectures on Celestial Mechanics. Springer
- (eds) Handbook of Dynamical Systems, vol 3. North-Holland (to appear)
- Buono P-L, Laurent-Polz F, Montaldi J (2005) Symmetric Hamiltonian Bifurcations. In: Montaldi J, Rañiu TS (eds) Geometric mechanics and symmetry. The Peyresq Lecture. Cambridge University Press, Cambridge
- Cannas da Silva A, Weinstein A (1999) Geometric models for non-commutative algebras. *Berkeley Math Lect Notes* 10
- Ciocci MC, Litvak-Hinenzon A, Broer HW (2005) Survey on dissipative KAM theory including quasi-periodic bifurcation theory. In: Montaldi J, Rañiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 303–355
- Cushman R (1992) A Survey of Normalization Techniques Applied to Perturbed Keplerian Systems. *Dyn Rep new ser* 1:54–112
- Dubrovin BA, Krichever IM, Novikov SP (2001) Integrable systems I. In: Novikov SP (ed) Dynamical systems IV, Symplectic geometry and its applications. Springer, pp 177–332
- Duistermaat JJ (1984) Bifurcations of periodic solutions near equilibrium points of Hamiltonian systems. In: Salvadori L (ed) Bifurcation Theory and Applications Montecatini 1983. *Lect Notes Math* 1057:57–105
- Efstathiou K (2005) Metamorphoses of Hamiltonian systems with symmetries. *Lect Notes Math* 1864
- Efstathiou K, Sadovskii D (2005) No Polar Coordinates. In: Montaldi J, Rañiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 211–301
- Eliasson LH, Kuksin SB, Marmi S, Yoccoz J-C (2002) Marmi S, Yoccoz J-C (eds) Dynamical Systems and Small Divisors, Cetraro 1998. *Lect Notes Math* 1784
- Gallavotti G (1983) The elements of mechanics. Springer
- Gallavotti G (1994) Twistless KAM tori, quasi flat homoclinic intersections, and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems. A review. *Rev Math Phys* 6(3):343–411
- Guillemin V, Sternberg S (1984) Symplectic techniques in physics. Cambridge Univ. Press
- Guillemin V, Sternberg S (1990) Variations on a theme by Kepler. *Am Math Soc Coll Pub* 42
- Haller G (1999) Chaos near Resonance. *Appl Math Sci* 138
- Hitchin NJ, Segal GB, Ward RS (1999) Integrable systems. Twistors, loop groups, and Riemann surfaces, Oxford 1997. Clarendon/Oxford Univ. Press
- Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser
- Holm DD (2005) The Euler-Poincaré variational framework for modeling fluid dynamics. In: Montaldi J, Rañiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 157–209
- Karasev MV, Maslov VP (1993) Nonlinear Poisson Brackets: Geometry and Quantization. *Transl Math Monogr* 119
- Kuksin SB (2000) Analysis of Hamiltonian PDEs. *OLS Math Appl* 19
- Libermann P, Marle C-M (1987) Symplectic Geometry and Analytical Mechanics. Reidel
- Markus L, Meyer KR (1974) Generic Hamiltonian dynamical systems are neither integrable nor ergodic. *Mem Am Math Soc* 144:1–52
- Marsden JE (1992) Lectures on Mechanics. Cambridge Univ. Press
- Marsden JE, Misiołek G, Ortega JP, Perlmutter M, Rañiu TS (2007) Hamiltonian reduction by stages. *Lect Notes Math* 1913

Books and Reviews

Arnol'd VI, Givental' AB (2001) Symplectic geometry. In: Novikov SP (ed) Dynamical systems IV, Symplectic geometry and its applications. Springer, pp 1–138

Arnol'd VI, Kozlov VV, Neishtadt AI (1988) Mathematical Aspects of Classical and Celestial Mechanics. In: Arnol'd VI (ed) Dynamical Systems III. Springer

Arnol'd VI, Novikov SP (1994) Dynamical systems VII, Integrable Systems. Springer

Bartsch T, Szulkin A (2005) Hamiltonian systems: periodic and homoclinic solutions by variational methods. In: Cañada A, Drábek P, Fonda A (eds) Handbook of differential equations: ordinary differential equations, vol 2. Elsevier, pp 77–146

Bibikov YuN (1979) Local theory of nonlinear analytic ordinary differential equations. *Lect Notes Math* 702

Birkhoff BD (1966) Dynamical Systems. *Am Math Soc* (1927); reprint 1966

Bolsinov AV, Fomenko AT (2004) Integrable Hamiltonian Systems – Geometry, Topology, Classification. Chapman & Hall/CRC

Bridges TJ, Furter JE (1993) Singularity Theory and Equivariant Symplectic Maps. *Lect Notes Math* 1558

Broer HW, Hasselblatt B, Takens F (2007) Handbook of Dynamical Systems, vol 3. North-Holland (to appear)

Broer HW, Hoveijn I, Lunter G, Vegter G (2003) Bifurcations in Hamiltonian systems: Computing Singularities by Gröbner Bases. *Lect Notes Math* 1806

Broer HW, Sevryuk MB (2007) KAM Theory: quasi-periodicity in dynamical systems. In: Broer HW, Hasselblatt B, Takens F

- Marsden J, Montgomery R, Rañiu TS (1990) Reduction, symmetry, and phases in mechanics. *Mem Am Math Soc* 88(436):1–110
- Mawhin J, Willem M (1989) *Critical Point Theory and Hamiltonian Systems*, 2nd edn. Appl Math Sci 74
- McDuff D, Salamon D (1995) *Introduction to symplectic topology*. Clarendon/Oxford Univ. Press
- van der Meer JC (1985) The Hamiltonian Hopf bifurcation. *Lect Notes Math* 1160
- Meyer KR (1999) Periodic solutions of the N -body problem. *Lect Notes Math* 1719
- Mielke A (1991) *Hamiltonian and Lagrangian Flows on Center Manifolds – with Applications to Elliptic Variational Problems*. *Lect Notes Math* 1489
- Novikov SP (2001) *Dynamical systems IV, Symplectic geometry and its applications*. Springer
- Olshanetsky MA, Perelomov AM, Reyman AG, Semenov-Tian-Shansky MA (1994) Integrable systems II. In: Arnol'd VI, Novikov SP (eds) *Dynamical systems VII, Integrable Systems*. Springer, pp 87–259
- Oxtoby JC (1971) Measure and category. A survey of the analogies between topological and measure spaces. *Grad Texts Math* 2
- Poincaré H (1957) *Les Méthodes Nouvelles de la Mécanique Céleste*. Gauthier–Villars (1892,1893,1899); reprint (1957) Dover
- Rabinowitz PH (2002) Variational methods for Hamiltonian systems. In: Hasselblatt B, Katok A (eds) *Handbook of dynamical systems*, vol 1A. North-Holland, pp 1091–1127
- Rink B, Tuwankotta T (2005) Stability in Hamiltonian Systems. In: Montaldi J, Rañiu TS (eds) *Geometric Mechanics and Symmetry: the Peyresq Lectures*. Cambridge Univ. Press, pp 1–22
- Souriau J-M (1970) *Structure des systèmes dynamiques*. Dunod. English edition: Souriau JM (1997) *Structure of dynamical systems, a symplectic view of physics*. *Prog Math* 149
- Thirring W (1978) *A course in mathematical physics I. Classical dynamical systems*. Springer
- Trofimov VV, Fomenko AT (1994) Geometric and Algebraic Mechanisms of the Integrability of Hamiltonian Systems on Homogeneous Spaces and Lie Algebras. In: Integrable systems II. In: Arnol'd VI, Novikov SP (eds) *Dynamical systems VII, Integrable Systems*. Springer, pp 261–333

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation

KHALID SAEED

Worcester Polytechnic Institute, Worcester, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Alternative Economic Models and Their Limitations](#)

[A System Dynamics Model of Resource Allocation, Production and Entitlements](#)

[Replicating Income Distribution Patterns Implicit in Models of Alternative Economic Systems](#)

[Feedback Loops Underlying Wage and Income Patterns](#)

[Possibilities for Poverty Alleviation](#)

[Conclusion](#)

[Future Directions](#)

[Appendix](#)

[Bibliography](#)

Glossary

Absentee owners Parties not present on land and capital resources owned by them.

Artisan owners Parties using own labor together with land and capital resources owned by them to produce goods and services.

Behavioral relations Causal factors influencing a decision.

Capital intensive A process or industry that requires large sums of financial resources to produce a particular good.

Capital Machinery equipment, cash and material inputs employed for the production of goods and services.

Capitalist sector A subeconomy in which all resources are privately owned and their allocation to production and renting activities is exclusively carried out through a price system.

Capitalist system An economic system in which all resources are in theory privately owned and their allocation to production and renting activities is exclusively carried out through a price system.

Commercial Pertaining to buying and selling with intent to make profit.

Controlling feedback A circular information path that counters change.

Corporate pertaining to a profit maximizing firm.

Economic dualism Side-by-side existence of multiple subeconomies.

Economic sector A collection of production units with common characteristics.

Entrepreneurship Ability to take risk to start a new business.

Feedback loops Circular information paths created when decisions change information that affects future decisions.

Financial market A mechanism that allows people to easily buy and sell commodities, financial instruments and other fungible items of value at low transaction costs and at prices that reflect efficient markets.

Household income Income accrued to a household from wages, profits and rents received by all its members.

Institutionalist economic models Models attributing performance of economies to institutional relationships and advocating selective government intervention to change the behavior that creates dysfunctions.

Iron law of wages David Ricardo's most well-known argument about wages "naturally" tending towards a minimum level corresponding to the subsistence needs of the workers.

Keynesian A belief that the total spending in the economy is influenced by a host of economic decisions – both public and private.

Labor intensive A process or industry with significant labor costs.

Labor productivity Output per worker or worker-hour.

Labor Economically active persons in an economy.

Marginal factor cost The incremental costs incurred by employing one additional unit of input.

Marginal revenue product The additional income generated by using one more unit of input.

Market economy An economy which relies primarily on interactions between buyers and sellers to allocate resources.

Marxist economic theory A theory highlighting exploitive mechanisms in an economic system and advocating central governance.

Marxist system A centrally run economic system emphasizing in theory the Marxist axiom "from each according to ability to each according to need".

Model An abstract representation of relationships in a real system.

Neoclassical economic theory A theory highlighting constructive market forces in an economic system and advocating consumer sovereignty and a price system as invisible sources of governance.

Non-linear A system whose behavior can't be expressed as a sum of the behaviors of its parts.

Opportunity cost Real value of resources used in the most desirable alternative, or the amount of one commodity foregone when more of another is consumed.

Ordinary differential equation A relation that contains functions of only one independent variable, and one or more of its derivatives with respect to that variable.

Output elasticity Change in output caused by addition of one unit of a production factor.

Perfect market A hypothetical economic system that has a large number of buyers and sellers – all price takers trading a homogeneous product – with complete information on the prices being asked and offered in

other parts of the market; and with perfect freedom of entry to and exit from the market.

Political economy Interaction of political and economic institutions, and the political environment.

Production factor A resource input such as land, labor, or capital contributing to production of output.

Productivity The amount of output created (in terms of goods produced or services rendered) per unit input used.

Purchasing power parity The value of a fixed basket of goods and services based on the ratio of a country's price levels relative to a country of reference.

Revisionist economic models Models recognizing both constructive and exploitive forces and advocating government intervention against exploitation.

Sector A collection of production units with common characteristics.

Self-employment Work for a self-owned production unit without a defined wage.

System dynamics A methodology for studying and managing complex feedback systems, such as one finds in business and other social systems.

Subeconomy A collection of production units and households with common characteristics.

Theories of value How people positively and negatively value things and concepts, the reasons they use in making their evaluations, and the scope of applications of legitimate evaluations across the social world.

Unearned income Income received as rents.

Wage employment Work for a defined wage.

Definition of the Subject

Poverty is perhaps the most widely written about subject in economic development, although there is little agreement over its causes and how to alleviate it. The undisputed facts about poverty are that it is pervasive, growing and that the gap between the rich and the poor is widening.

It is widely believed that the governments – irrespective of their ideological inclinations – have the responsibility to intervene to help the poor. Poverty alleviation is also the key mandate of International Bank for Reconstruction and Development (World Bank) and the many civil society organizations. World Bank places poverty line at purchasing power parity of \$1 per day, which has improved a bit in terms of percentage below over the past three decades, except in Africa, but remains large in terms of head count. This threshold is however unrealistic since it targets absolutely basket cases. A poverty line at purchasing power parity of \$3 per day, which is close to average purchasing power per capita in the poor countries

shows that both poverty head count and gap between rich and poor have been expanding across board. World Bank web site at <http://iresearch.worldbank.org/PovcalNet/jsp/index.jsp> allows making such computations for selected countries, regions, years and poverty lines.

Neoclassical economic theory does not explicitly address the process of income distribution among households, although it often views income distribution as shares of profits and wages. In most economic surveys and censuses, however, income distribution is invariably measured in terms of shares of various percentages of the households. The fact that more than 80% of the income is claimed by fewer than 20% of the households who also own most of the capital resources in almost all countries of the world, the theory and the measurement have some common ground. Neoclassical theory has, however, shed little light on the process of concentration of wealth and how can this dysfunction be alleviated.

System dynamics, although rarely used for the design of public policy for addressing poverty, allows us to construct and experiment with models of social systems to understand their internal trends and test policy combinations for changing them. In this paper I have used system dynamics modeling to understand the process of concentration of wealth and re-evaluate the on-going poverty alleviation effort.

The model, which subsumes resource allocation, production and entitlements, explains the many manifestations of income distribution in a market economy. It generates multiple patterns of asset ownership, wage and employment assumed in neo-classical, Marxist and revisionist perspectives on economic growth while it allows ownership to change through the normal course of buying and selling transactions based on rational thought, information-bound criteria. Privately owned resources can be employed through hiring wage-labor, rented out or used for self-employment. In addition to the labor market conditions, the wage rate depends also on the opportunity cost of accepting wage employment as workers may be either self-employed or wage-employed. Since this opportunity cost varies with the capital resources owned by the workers, which may support self-employment, the wage rate is strongly affected by the distribution of ownership. Thus, ownership can become concentrated or widely distributed depending on legal and social norms governing transactions in the economy, which the model replicates. Extended experimentation with this model serves as a basis to identify critical policy instruments that make best use of the system potential for resource constrained growth and poverty alleviation through widening participation in the market and improving income distribution.

Introduction

The opening up of the major centrally planned economies of the world has brought to the fore problems concerning the psychological deprivation, inefficiencies of resource allocation and production, and the lack of dynamism experienced in the working of central planning in a socialist system. The accompanying enthusiasm for free market in a capitalist system has, however, hidden many of the dysfunctional aspects of this alternative. It should be recognized that both systems emerged from time-specific and geography-specific empirical evidence. Since their underlying models treat as given specific economic patterns, the institutional roles and the legal norms associated with each system have inherent weaknesses, which create dysfunctions when implemented in different environmental contexts [43,64]. Thus, neither model may furnish an adequate basis for the design of policies for sustainable economic development and poverty alleviation. A search is, therefore, necessary for an organizational framework that might explain the internal trends inherent in each model as special modes of a complex system subsuming the variety of behavioral patterns recognized by specific models before an effective policy for change can be conceived [52].

Using as an experimental apparatus a formal model of the decision structure affecting wage determination, saving and investment behavior, and the disbursement of income, presented earlier in [53], this paper seeks to identify the fundamental economic relations for creating a dynamic and sustainable market system that may also increase opportunities for the poor, whose market entry is often limited by their financial ability and social position [58], to participate in the economy and be entitled to the value it creates. System dynamics provides the technical framework to integrate the various behavioral relations in the system [13,63].

Notwithstanding the many objections to the abstract models of orthodox economics, which are difficult to identify in the real world [28,46], the model of this paper draws on neo-classical economics to construct a basic structure for growth and market clearing. This structure is, however, progressively modified by relaxing its simplifying assumptions about aggregation of sub-economies, wage determination, ownership, income disbursement, saving and investment behavior, financial markets, and technological differentiation between sub-economies to realize the many growth and income distribution patterns addressed in a variety of economic growth models.

The modified model I finally create represents a real world imperfect market in which expectations formed under bounded rational conditions govern the decisions of

the economic actors [59], as recognized in the pioneering works of Kaldor (1969) [24], Kalecki (1965) [25], Wientraub (1956) [66], and Joan Robinson (1978) [45]. The model also subsumes the concept of economic dualism first recognized by Boeke (1947) [7] and developed further by Lewis (1958) [29], Sen (1966) [57], Bardhan (1973) [4] and others to represent the multiple sub-economies that co-exist especially in the developing countries. Such a model is more identifiable with respect to the real world as compared with the time and geography specific concepts propounded by the various, often controversial, theories of economic growth.

Simulation experiments with the model explore entry points into the economic system for creating an egalitarian wage and income distribution pattern through indirect policy instruments. Also explored are the functions of entrepreneurship and innovation and the mechanisms that may increase the energy of those processes toward facilitating economic growth and alleviating poverty.

The Alternative Economic Models and Their Limitations

The economic models used as the bases for designing development policies over the past several decades have ascended largely from time-specific and geography-specific experiences rather than from a careful study of the variety of behavioral patterns occurring over various time periods and across several geographic locations. Among these, the socialist and the capitalist models are most at odds. They differ in their assumptions about ownership and income distribution patterns, the basis for wage determination, the influence of technology on income growth and the functions of entrepreneurship and innovation [21,55].

The neo-classical economic theory, which is the basis for the capitalist model, is silent on the ownership of capital resources, often assuming it in default to be widely distributed [5]. Thus, the labor-wage rate may bear little relationship to the income of households, who are also recipients of profits. It is assumed that private control of productive resources is a means for market entry, which creates unlimited potential for economic growth, although private investment is not often seen to be subject to self-finance due to the assumption that financial markets are perfect. The neo-classical economic theory also postulates that short-run labor-wage rates depend on labor market conditions, while in the long run, they are determined by the marginal revenue product of labor. Neo-classical models of economic growth, however, often make the simplifying assumption that equilibrium continues to prevail in both factor and product markets over the course of

growth. Thus, only minor fluctuations may occur in wages, profits and prices in the short run, and these can be ignored.

The belief in the existence of such equilibrium is further strengthened by the Keynesian argument for the ineffectiveness of the market mechanisms due to the dependence of prices on long-term wage contracts and production plans which may not respond easily to short-run changes of the market. As a result of the belief in this theory of wage determination, technological choices that increase labor productivity are expected to have a positive effect on wage rates and household income, because they increase the marginal revenue product of labor. Entrepreneurship is viewed as important for new entry into economic activity, which is open to all, and innovation is supposed to benefit society through increased productivity. With these assumptions, the capitalist system advocates minimal government intervention in the economy. This model is widely presented in the many texts on economic development. Pioneering texts include Hirschleifer (1976) [22] and Kindelberger and Herrick (1977) [27].

Marxist economic theory, which underpins the socialist model, assumes on the other hand that ownership of capital resources is concentrated in a minority excluding the workers and that the majority of households receive no part of the profits. Thus, wage payments have a strong effect on household income. The Marxist theory views private ownership as a source of exploitation and postulates labor-wage rates determined by the consumption necessary for a worker to support production in a grossly labor surplus economy following Ricardo's iron law of wages [32,39]. The labor-wage rate is, thus, based on the real value of the commodities needed for a worker to subsist, which is more or less fixed, irrespective of the contribution of labor to the production process. In such conditions, technological choices that increase labor productivity may indeed only serve to increase the share of the surplus of product per unit of labor appropriated by the capitalists. In this model, entrepreneurship is viewed as an asocial activity and innovation seen to originate from the need to boost falling returns on capital. Attributing the development of these conditions to market failure, the socialist system assigns control of the economy to the government.

There also exist a number of revisionist models of political economy attempting to explain the nature of interdependence of the multiple sub-economies observed to co-exist in many developing countries in violation of the theoretical premises of the neo-classical model according to which all production factors must eventually move to the most efficient sector. These models of

ten attribute the development of disparities between the various sub-economies to exploitative mechanisms that tend to maintain an upper hand of the stronger influence groups. The revisionist analyses have largely led to making moral appeals for the government policy to target the poor and the disadvantaged in its development effort, which is a stated mandate of the International Bank for Reconstruction and Development (World Bank). Targeting the poor has also been advocated widely by numerous development economists over the past half century. They include such prominent economists as Myrdal (1957) [36], Lipton (1977) [30], Galbraith (1979) [15], and Sen (1999) [58].

Indeed, each economic system can often be endorsed with the help of selected historical evidence, and this has been fully exploited to fuel the traditional debate between the neo-classical and Marxist economic schools. Interesting artifacts of this debate include the normative theories of value suggested by each system to justify the various wage systems, which have little practical significance for development policy [44,62]. This is unfortunate, since contradictions of evidence should clearly indicate the existence of fundamental organizational arrangements in the economic system, which are capable of creating the multiple behavior patterns on which the various economic models are based. Once identified, such arrangements may also serve as entry points for the design of evolutionary changes in an existing pattern. To quote Professor Joan Robinson:

Each point of view bears the stamp of the period when it was conceived. Marx formed his ideas in the grim poverty of the forties. Marshal saw capitalism blossoming in peace and prosperities in the sixties. Keynes had to find an explanation for the morbid condition of 'poverty in the midst of plenty' in the period between the wars. But each has significance for other times, for in so far as each theory is valid, it throws light upon essential characteristics of the system which have always been present in it and still have to be reckoned with. [43]

Following sections of this paper experiment with a system dynamics model of an economic system, widely found in the developing countries and presented earlier in [53], to understand the variety of economic patterns experienced over time and geography under different legal and social norms. Furthermore, exploratory experimentation with this model helps to outline the basic principles of a market system that can sustain growth, create equitable distribution of benefits and facilitate innovation and pro-

ductivity improvement, all widely deemed necessary for poverty alleviation.

A System Dynamics Model of Resource Allocation, Production and Entitlements

A system dynamics model subsuming the broad decision rules that underlie resource allocation, production, and income disbursement processes of a developing country economic system was proposed in Saeed (1988) [49] and further experimented with in Saeed (1994) [53]. In this model, capital, labor, and land (which may also be assumed as a proxy for natural resources) are used as production factors. Model structure provides for the functioning of two modes of production, commercial, in which resources are employed on the basis of their profitability and which is managed by the capitalist sector of the economy; and self-employed, in which workers not employed in the commercial mode make a living. These two modes of production have been referred to variously in the literature, for example as oligopolistic and peripheral firms [16], formal and informal sectors [29], and modern and traditional subeconomies [12].

It has been assumed in the model that all workers, whether self-employed using their own or rented capital resources or employed as wage-workers by the capitalist sector, are members of a homogeneous socio-economic group with a common interest, which is to maximize consumption. This group is also the sole supplier of labor in the economy since the small number of working capitalists is ignored. On the other hand, the capitalist sector is assumed to maximize profit while it is also the sole wage-employer in the economy [2,3,57].

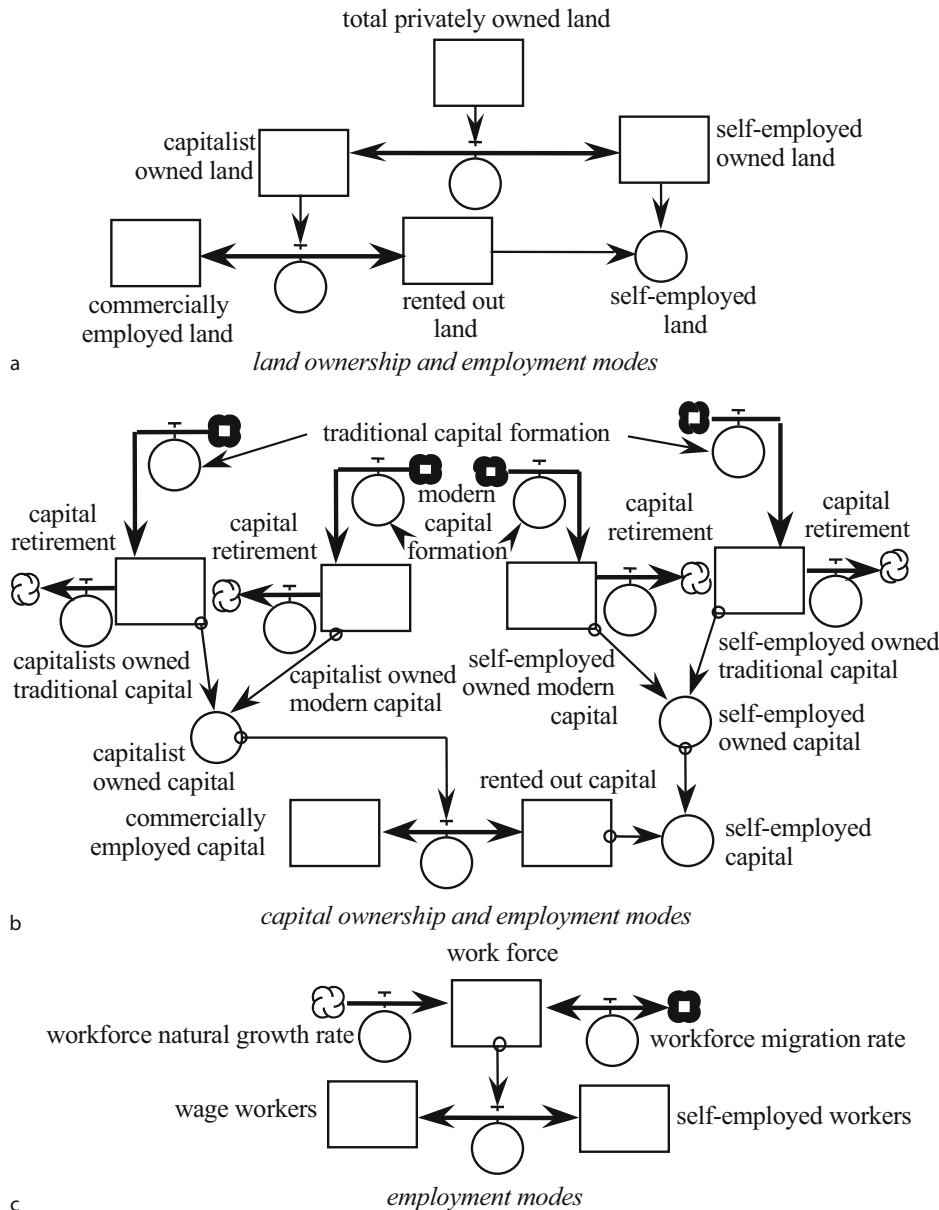
It is assumed that private ownership is protected by law, but land and capital assets can be freely bought, sold and rented by their owners. Each buying and selling transaction between the two sectors must be accompanied by a corresponding transfer of the cash value of the assets determined by the going market prices. The model also permits the appearance of technological differences between the capitalist and the self-employed sectors, when more than one technologies embodied in the type of capital used (named traditional and modern in the model) are available and the two sectors cannot employ the preferred technology with equal ease [30,41], or when the self-employed sector is burdened by excess workers not employed by the commercial sector while it lacks the financial capacity to use its preferred technology.

Figure 1 shows how workers and capital might potentially be retained and employed by the two sectors in the model. Rectangles represent stocks, valve symbols

flows and circles intermediate computations following the diagramming convention of system dynamics modeling. The size of each sector is not specified and is determined endogenously by the model, depending on assumptions about the socio-technical environment in which the system functions.

The changes in the quantities of the production factors owned or employed by each sector are governed by the

decisions of the producers and the consumers of output and by the suppliers of the production factors acting rationally according to their respective motivations within the bounds of the roles defined for them by the system [59]. The value of production is shared by households on the basis of the quantity of the production factors they contribute and the factor prices they can bargain for [10]. Income share of the workers, less any investment needed to



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 1
Potential worker and capital distribution between capitalist and self employed sectors

maintain self-employment, divided by the total workforce, determines average consumption per worker, which represents the opportunity cost of accepting wage-employment and this is the basis for negotiating a wage [57,62].

Investment and saving rates in the two sectors are decoupled through a balance of internal savings. The financial markets are segmented by sectors and the investment decisions of a sector are not independent of its liquidity position, given by the unspent balance of its savings. Thus, investment decisions depend on profitability criteria, but are constrained by the balance of internal savings of each sector [33,34]. Figure 2 shows the mechanisms of income disbursement, saving and internal finance incorporated into the model.

The saving propensity of all households is assumed not to be uniform. Since capitalist households receive incomes that are much above subsistence, their saving propensity is stable. On the other hand, the saving propensity of the worker households depends on their need to save to support investment for self-employment and on how their ab-

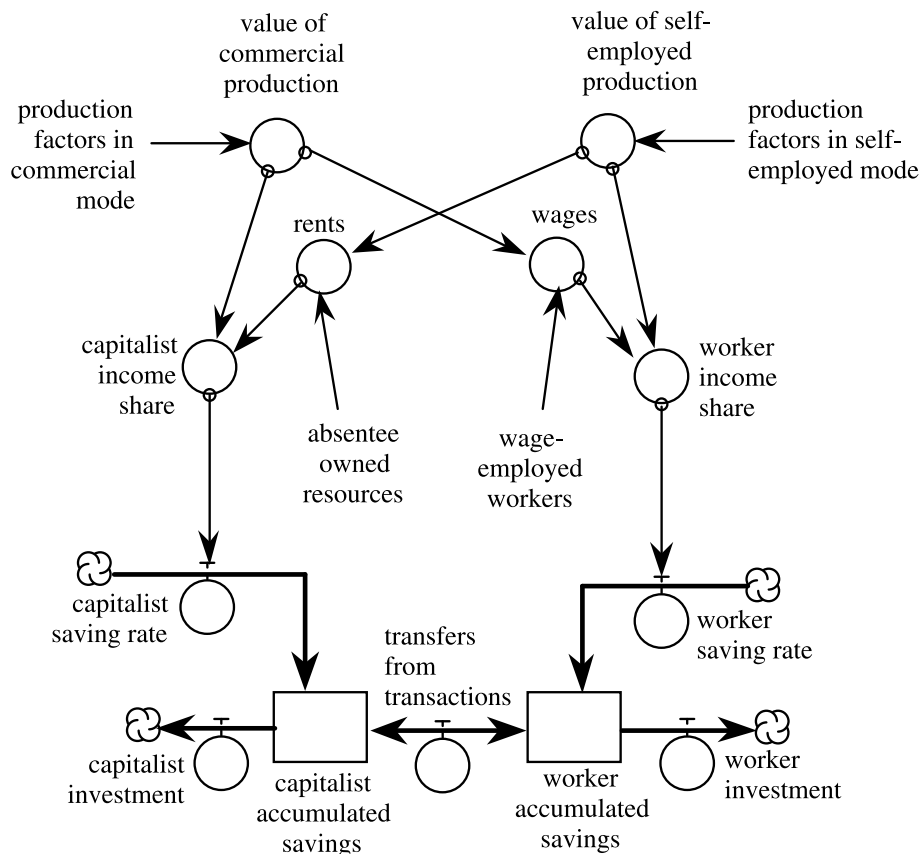
solute level of income compares with their inflexible consumption [23,26,31].

The broad mathematical and behavioral relationships incorporated into the model are given in the Appendix “Model Description”. Technical documentation and a machine-readable listing of the model written in DYNAMO code are available from the author on request.

Replicating Income Distribution Patterns Implicit in Models of Alternative Economic Systems

The model is simulated under different assumptions about wages, rents, financial markets and technology and its behavior analyzed in relation to the various theoretical and empirical premises its information relationships represent.

As an arbitrary initial condition, production factors are equally divided between the two sectors and equilibrium in both product and factor markets is assumed to exist under the conditions of a perfect economic system as



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 2
Income disbursement process

described in neo-classical economics. Thus, the marginal revenue products of land and capital are initially assumed to be equal to their respective marginal factor costs determined by an exogenously specified interest rate which represents the general pattern of preferences of the community for current as against future consumption [22]. The marginal revenue product of workers is initially set equal to wage rate. The market is initially assumed to be clear and there is no surplus of supply or demand.

Replicating the Theoretical Neo-classical System

This experiment is aimed at understanding internal trends of a system representing the neo-classical economic theory. To transform the model to represent this system, it is assumed that the production factors employed by each sector are owned by it and no renting practice exists [5]. The wage rate is assumed to be determined by the marginal revenue product of workers and the availability of labor instead of the opportunity cost to the workers of supplying wage-labor. Financial markets are assumed to be perfect and investment decisions of the two sectors are uncoupled from their respective liquidity positions. It is also assumed that the technology of production is the same in the two sectors and, in terms of the model, only traditional capital is available to both of them. The only difference between the two sectors is that the capitalist sector can vary all production factors, including labor to come to an efficient mix, while the self-employed sector may absorb all labor not hired by the capitalist sector, while it can freely adjust other production factors to achieve an efficient mix.

The model thus modified stays in equilibrium when simulated as postulated in neo-classical economic theory. When this equilibrium is disturbed arbitrarily by transferring a fraction of the workers from the capitalist to the self-employed sector, the model tends to restore its equilibrium in a manner also similar to that described by the neo-classical economic theory. This is shown in Fig. 3.

The transfer raises the marginal revenue product of workers in the capitalist sector, which immediately proceeds to increase its workforce. The transfer also raises the intensity of workers in the self-employed sector as a result of which the marginal revenue products of land and capital in that sector rise. Hence, it proceeds to acquire more land and capital. These activities continue until the marginal revenue products of the factors and their proportions are the same in the two sectors. Note that while the factor proportions and marginal revenue products of the factors are restored by the model to their original values, the absolute amounts of the various factors are different

when new equilibrium is reached. There is, however, no difference in endowments per worker between the capitalist and the self-employed sectors.

Since factor payments are determined purely on the basis of contribution to the production process while the quantities of production factors allocated to each sector depend on economic efficiency, the wages and factor allocations seem to be determined fairly and efficiently, as if by an invisible hand. Ownership in such a situation can either be communal or very widely distributed among households since otherwise the wage bargaining process will not lead to fair wages. Renting of production factors among households is irrelevant since transfer to parties who can efficiently employ them is automatic.

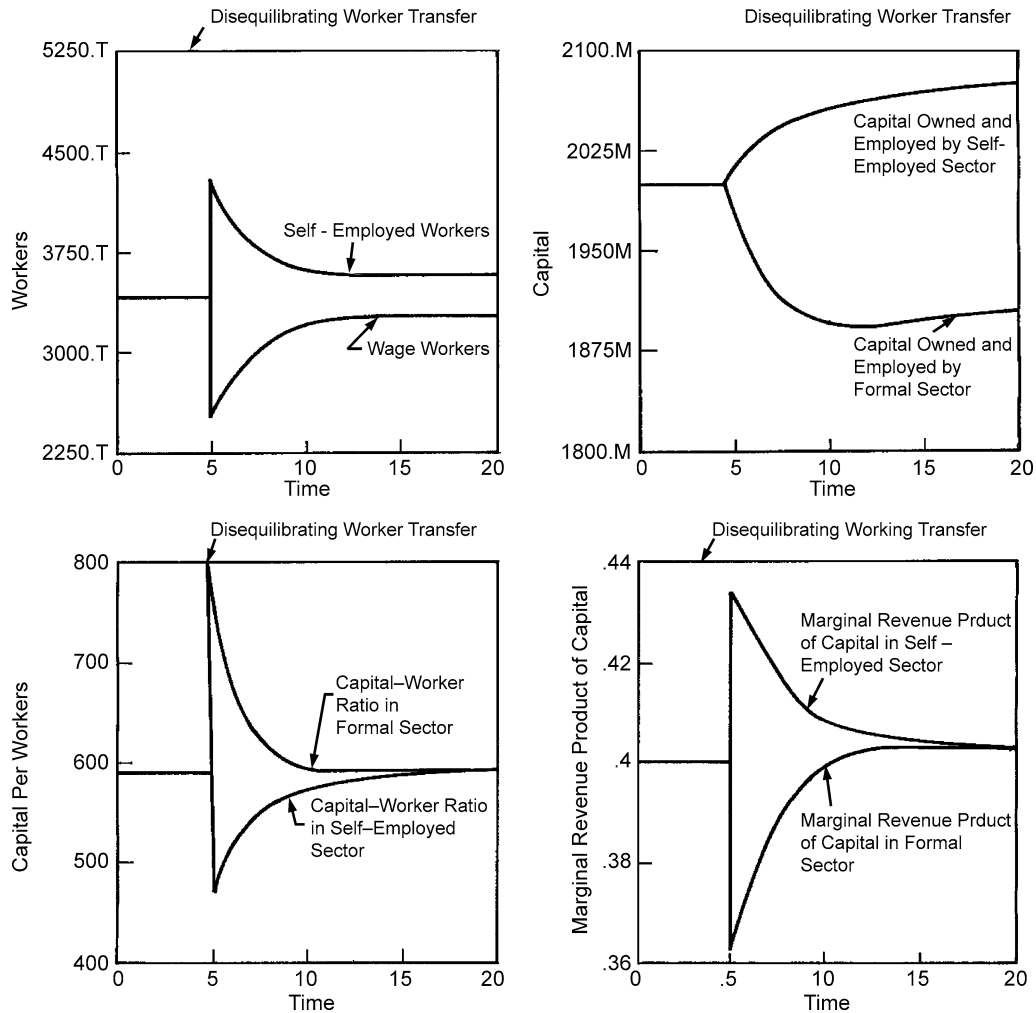
Before anything is said about the empirical validity of the simplifying assumptions made in this model, the historical context of these assumptions must be examined carefully. The simplified model is based on Adam Smith's description of an industrial economy observed at the start of the industrial revolution. This economy was run by artisan-turned capitalists and there were many of these capitalists competing with one another, although, none had the financial muscle to outbid the others except through his/her ability to employ resources efficiently [60].

As far as labor wage rate was concerned, although there were instances of exploitation of workers at a later stage of the industrial revolution, the artisan workers could obtain a wage that was equal to their contribution of labor to the production process, as otherwise they could easily be self-employed since the economy was still quite labor intensive and the tools needed for self-employment may not have cost very much. Also, since ownership of the tools of a trade may have been quite widespread while the contribution of capital resources to the production process was quite small as compared to that of labor, a major part of the income might have accrued to the working households. In such circumstances, the simplifying assumptions of the neo-classical model may appear quite reasonable.

The neo-classical model became irrelevant, however, as the system made progress in the presence of a social or organizational framework that legally protected ownership of all types and freely allowed the renting of assets, thus making possible an absentee mode of owning productive resources while technological changes also made the contribution of capital resources to the production process more significant.

Creating Worker Capitalism

It is not only methodologically expedient but also pedagogically interesting to explore what ownership and wage



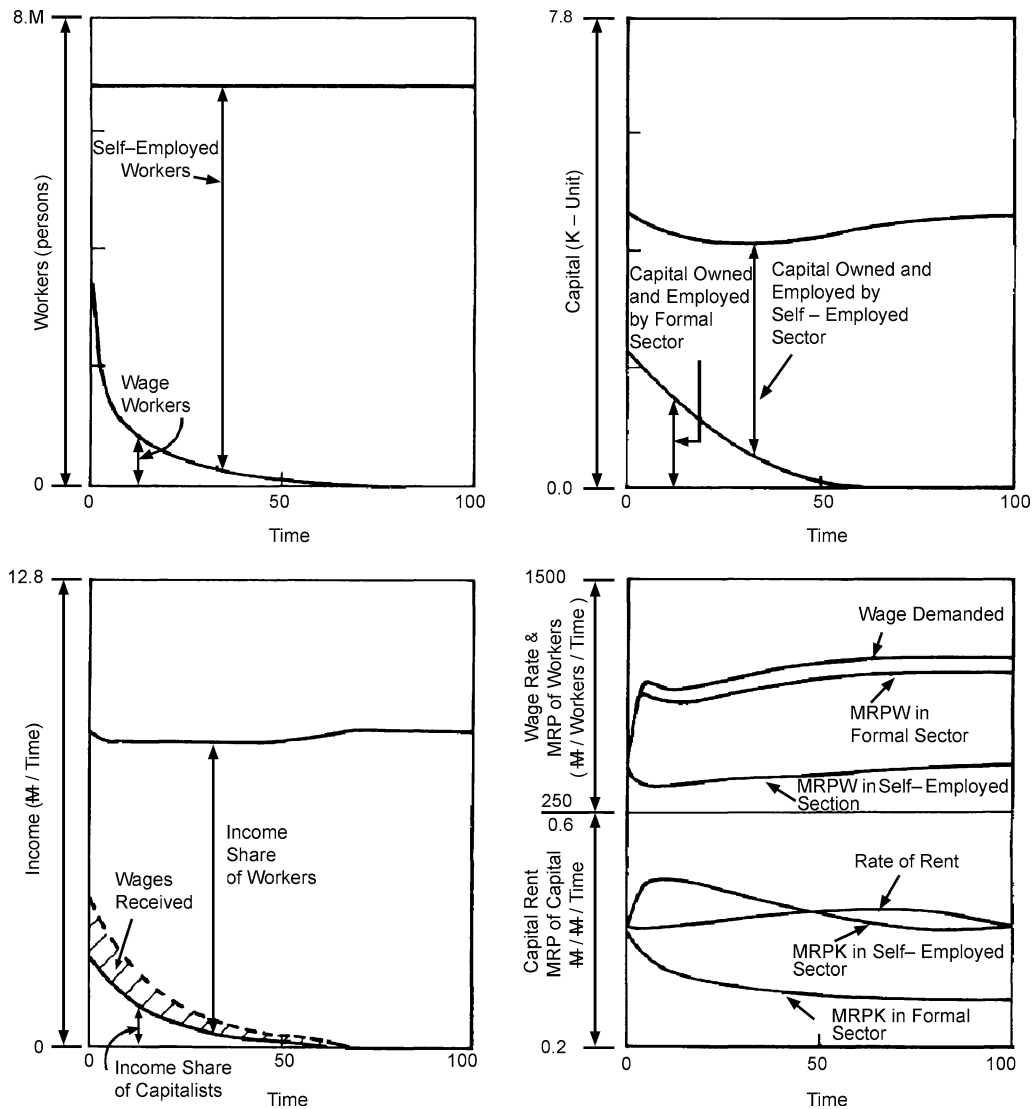
Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 3
 Recovery from dis-equilibrium in a neo-classical system

patterns might have emerged if labor-wages were determined through bargaining mechanisms incorporated into the model instead of fair payment equal to the marginal revenue product of workers, while all other assumptions of a perfect market of the experiment of the last section were maintained.

Figure 4 shows a simulation of the model in which wage rate is determined by the average consumption expenditure per worker (as given in Eqs. (1) and (2) of the model described in the Appendix “[Model Description](#)”) while renting of production factors and financial fragmentation of the households are still not allowed. This change in assumptions disturbs the initial market equilibrium in the model thus activating its internal tendency to seek a new equilibrium. No exogenous disequilibrating changes

are needed to generate the dynamic behavior in this simulation and in those discussed hereafter.

As a result of this change, the compensation demanded for working in the capitalist sector becomes much higher than the marginal revenue product of the workers. Thus, wage-workers are laid off and accommodated in the self-employed sector. Consequently, the marginal revenue product of land and capital in the self-employed sector increases and its bids for these resources rise. On the other hand, the decrease in the workforce of the capitalist sector increases its land and capital intensities and hence lowers their marginal revenue products. The falling productivity of these resources increases the opportunity cost of holding them. Since renting is not allowed, the capitalist sector is persuaded to sell the resources to the self-employed who



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 4

The develop of worker capitalism when wages depend on bargaining position of workers

can easily buy them since investment in the model is not subject to internal self-finance.

As the self-employed sector increases its land and capital holdings, its production rises. When increases in the production of this sector exceed the wage income lost due to decreasing wage disbursements from the capitalist sector, the net revenue of the workers, and hence their average consumption, rises. The wage rate is thus pushed up further, which necessitates further reductions in wage-workers. These processes spiral into a gradual transfer of all resources to the self-employed sector.

The marginal revenue products of land and labor in the two sectors tend to equilibrate at different val-

ues, but the capitalist sector exists only in theory because towards the end of the simulation almost all the resources are owned and managed by the self-employed. Since no part of the income is obtained by absentee owners, and working households may own and manage resources according to the quantity of labor they can supply, the income distribution may appear to be truly egalitarian.

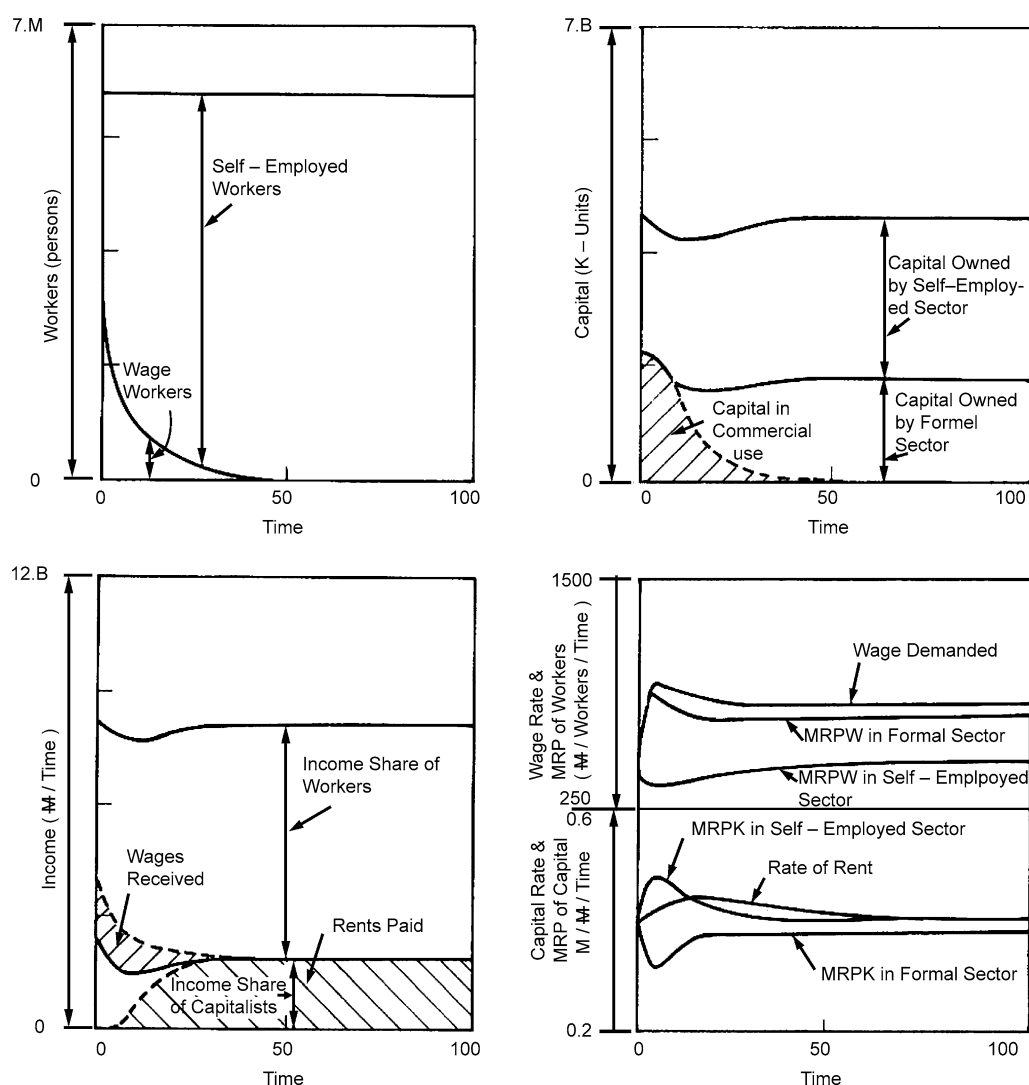
Even though the above simulation is hypothetical, the wage and income distribution pattern shown by it may be experienced when the separation of resources from the households employing them is socially or legally ruled out or the state allocates capital resources and land ac-

according to the quantity and quality of labor supplied by a household. Instances of peasant economies having such characteristics have been recorded in history in tribal cultures and, in a somewhat advanced form, in medieval India [35]. Interestingly, such implicit assumptions are also subsumed in the illusive perfect market the neoclassical economic theory is based on.

Appearance of Absentee Ownership

When ownership of resources is legally protected, whether they are productively employed or owned in absentia, many renting and leasing arrangements may appear which

may allow a household to own resources without having to employ them for production [47]. This is borne out in the simulation of Fig. 5, in which resources are divided by the capitalist sector between commercial production and renting activities depending on the rates of return in each. Rents depend on long-term averages of the marginal revenue products of the respective factors and on the demand for renting as compared with the supply of rentable assets. In the new equilibrium reached by the model, the commercial mode of production and wage-employment gradually disappear but a substantial part of the resources continues to be owned by the capitalist sector, which rents these out to the self-employed sector.



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 5

The appearance of absentee ownership when renting is also allowed

Such a pattern develops because of the combined effect of wage and tenure assumptions incorporated into the model. When workers are laid off by the capitalist sector in response to a high wage rate, the marginal revenue products of land and capital for commercially employing these resources in this sector fall. However, as the laid-off workers are accommodated in the self-employed sector, the marginal revenue products of land and capital, and hence their demand in this sector, rise. Therefore, rents are bid up and the capitalist sector is able to get enough return from renting land and capital to justify maintaining its investment in these.

Again, the marginal revenue products of the production factors in the commercial mode of production are only hypothetical as that mode is not practiced towards the end of the simulation. The renting mechanism allows the self-employed sector to adjust its factor proportions quickly when it is faced with the accommodation of a large number of workers. When the economy reaches equilibrium, the marginal rates of return of the production factors in the self-employed sector are the same as those at the beginning of the simulation. But, the wage demanded equilibrates at a level lower than that for the exclusively self-employed economy described in the simulation of Figure 4, because a part of the income of the economy is now being obtained by the absentee owners of the capitalist sector in the form of rent.

Note that, although the total income of the economy falls a little during the transition, it rises back to the original level towards the end equilibrium since the technology is uniform, irrespective of the mode of production. Also note that the end equilibrium distribution of income depends on initial distribution of factors when modifying assumptions are introduced, and on the volume of transfers occurring over the course of transition. Thus, an unlimited number of income and ownership distribution patterns would be possible depending on initial conditions and the parameters of the model representing the speeds of adjustment of its variables. The common characteristics of these patterns, however, are the presence of absentee ownership, the absence of a commercial mode of production, and a shadow wage that is less than an exclusively self-employed system.

Separation of Ownership from Workers and the Creation of a Marxist System

The ownership of resources becomes separated from the workers and concentrated in the capitalist sector in the model, irrespective of the initial conditions of resource

distribution, when the assumption about the existence of a perfect financial market is also relaxed.

Figure 6 shows the ownership and wage pattern which develops when acquisition of resources by the capitalist and self-employed sectors is made dependent, in addition to their profitability, on the ability to self-finance their purchase. Recall also that the ability to self-finance depends on the unspent balance of savings, and the saving rate of the self-employed sector is sensitive both to the utility of saving in this sector to support investment for self-employment and to the rent burden of this sector compared with the factor contribution to its income from land and capital. The saving rate of the capitalist sector is assumed to be constant.

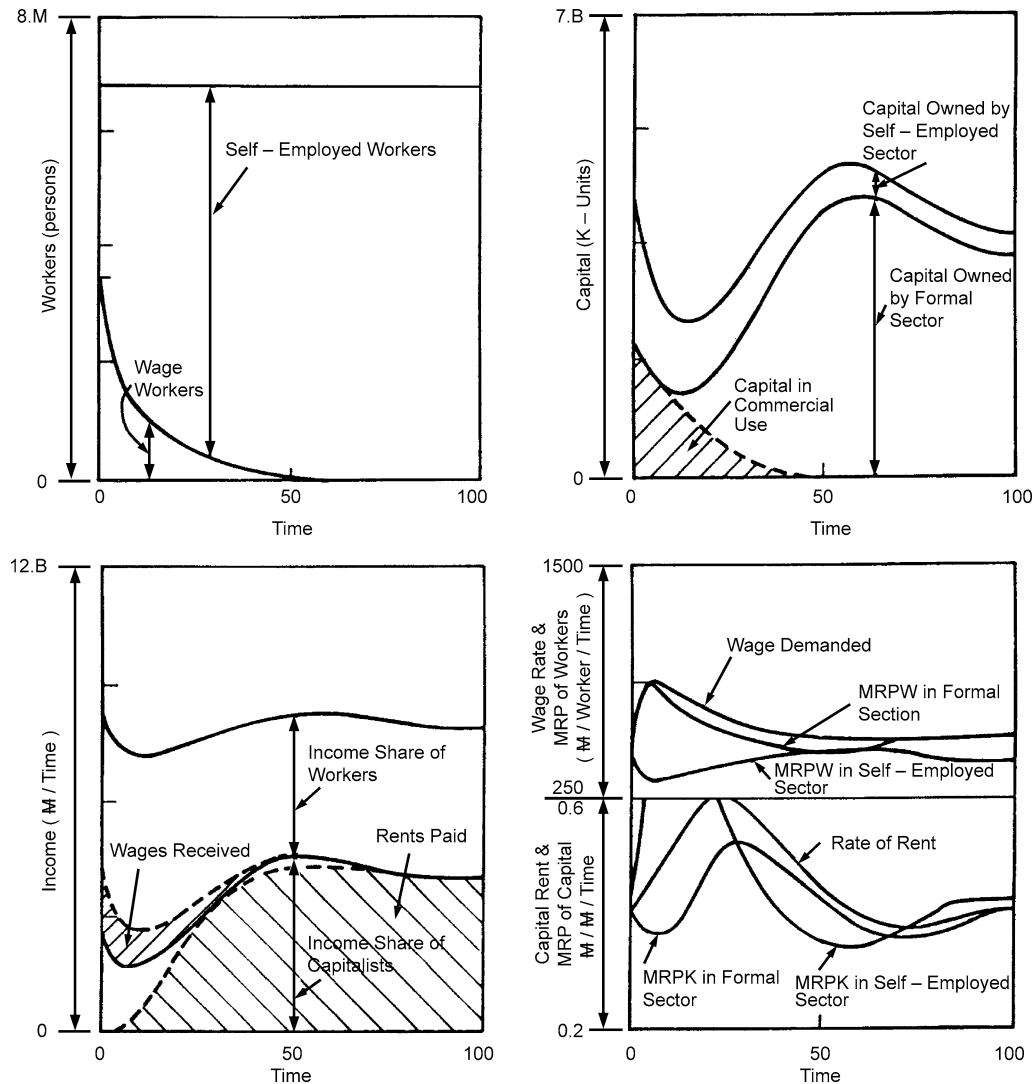
Such a pattern develops because of an internal goal of the system to employ resources in the most efficient way while the ownership of these resources can only be with the households who have adequate financial ability, which is also not independent of ownership.

Creation of a Dualist System

A dualist system is characterized by the side-by-side existence of both commercial and self-employed modes of production. The former appears to be economically efficient and is often also capital-intensive. The latter is seen to be economically inefficient and is also invariably labor-intensive. The side-by-side existence of these two modes of production in many developing countries has often puzzled observers, since according to the neo-classical economic theory, any inefficient production mode must be displaced by the efficient one.

A stable commercially run capital-intensive production sector existing together with a self-employed labor-intensive sector develops in the model if a technological differentiation is created between the capitalist and self-employed sectors. This is shown in the simulation in Fig. 7, in which an exogenous supply of modern capital is made available after end equilibrium of the simulation in Fig. 6 is reached.

Capital differentiation between the two sectors appears since the scale of the self-employed producers does not allow them to adopt modern technologies requiring indivisible capital inputs. The capitalist sector starts meeting its additional and replacement capital needs by acquiring a mixture of modern and traditional capital while the self-employed sector can use only traditional capital. However, the capital demand of the capitalist sector is met by modern capital as much as the fixed supply permits. The balance of its demand is met by acquiring traditional capital.



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 6

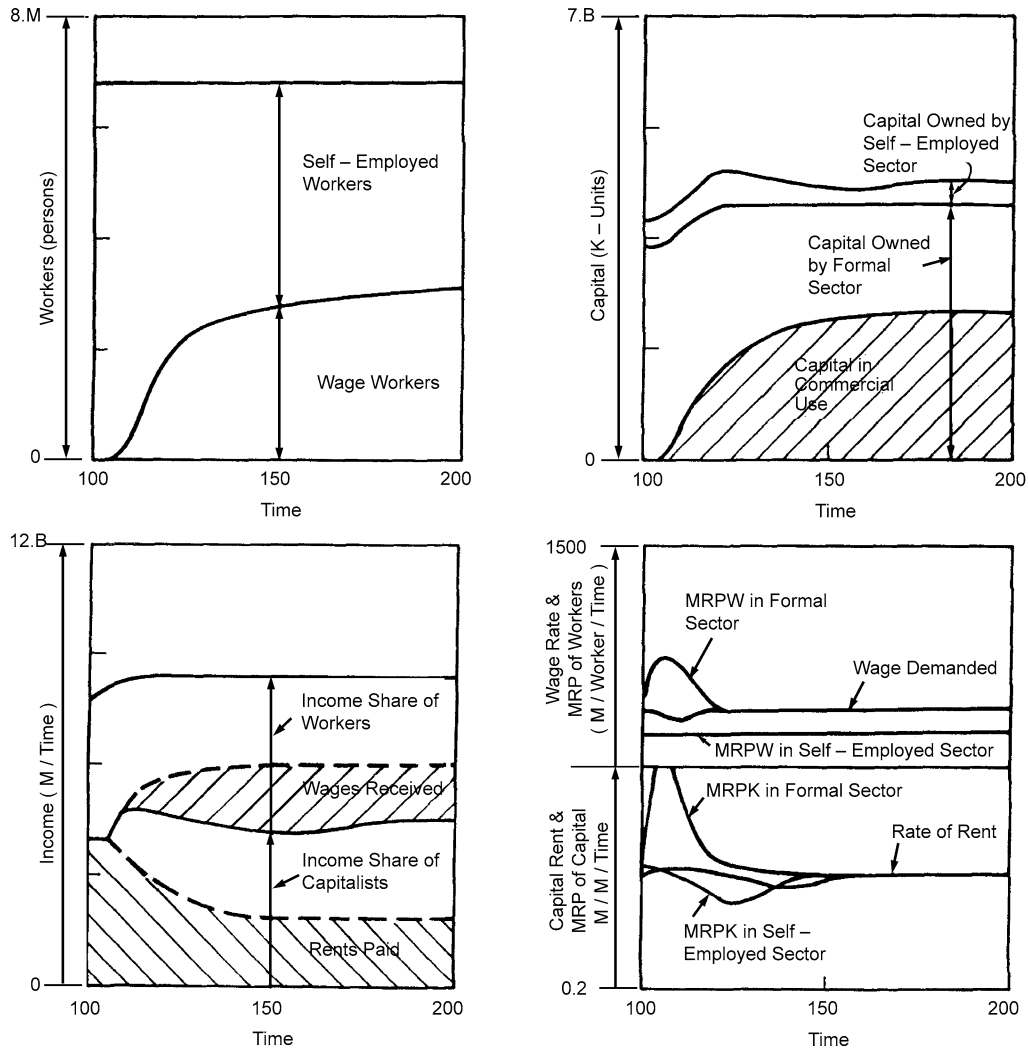
Separation of ownership from workers as postulated by Marx system when investment must also be internally financed

The output elasticity of modern capital is assumed to be higher than that of the traditional capital while the use of the former also allows an autonomous increase in output. The output elasticity of land is assumed to remain constant. The assumption of uniform returns to scale is maintained. Thus, the output elasticity of workers decreases when modern capital is introduced. These assumptions serve to represent the high productivity and labor-saving characteristics of the modern capital.

As its capital becomes gradually more modern and potentially more productive, the capitalist sector is able to employ its productive resources with advantage in the commercial mode of production, instead of renting these

out, and to employ wage-workers at the going wage rate. The increased productivity and income derived from this make it both economically and financially viable for the capitalist sector to invest more. Thus, its share of resources, when a new equilibrium is reached, is further increased.

Since the output elasticity of workers falls with the increase in the fraction of modern capital, the marginal revenue product of workers in the commercial mode may not rise much with the increase in its output. At the same time, since resources are being transferred away by the capitalist sector from renting to commercial employment, the labor intensity and the demand for renting rises in the self-em-



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 7

Creation of dualist system when technological differentiation develops between the capitalist and self-employed sectors

ployed sector. Hence rents are bid up and it again becomes profitable for the capitalist sector to allocate resources to renting. The amount of resources rented out, however, will depend on the degree of technological differentiation that may be created between the two sectors.

The wage rate reaches equilibrium at a lower level and the rents at higher levels than without technological differentiation. Rents, however, equal marginal revenue products of land and capital, which rise in the capitalist sector because of employing superior technology and in the self-employed sector due to increased labor intensity.

Interestingly, dualist patterns appeared in the developing countries, both in the agricultural and industrial sectors, only after modern capital inputs became available

in limited quantities. Labor-intensive peasant agriculture and small-scale industry and services carried out by the self-employed came to exist side-by-side with the commercially run farms and large-scale industry employing wage labor and modern technologies. However, worker income, both in wage-employment and self-employment, remained low [19].

Feedback Loops Underlying Wage and Income Patterns

The internal goal of a dynamic system represented by a set of non-linear ordinary differential equations is created by the circular information paths or feedback loops which are

formed by the causal relations between variables implicit in the model structure. These causal relations exist in the state space independently of time (unless time also represents a state of the system). The existence of such feedback loops is widely recognized in engineering and they are often graphically represented in the so-called block and signal flow diagrams [17,40,65].

While many feedback loops may be implicit in the differential equations describing the structure of a system, only a few of these would actively control the system behavior at any time. The nonlinearities existing in the relationships between the state variables determine which of the feedback loops would actively control the system behavior. A change may occur in the internal goals of a system if its existing controlling feedback loops become inactive while simultaneously other feedback loops present in its structure become active. Such a shift in the controlling feedback loops of a system is sometimes called a structural change in the social sciences and it can result both from the dynamic changes occurring over time in the states of the system and from policy intervention. The realization of a specific wage and income distribution pattern depends not on assumptions about initial conditions but on legal and social norms concerning ownership, renting, financing of investment and the state of technology, determining which feedback loops would be dominant [14,40].

Figure 8 describes the feedback loops, formed by the causal relations implicit in the model structure that appear to polarize income distribution by separating asset ownership from working households and creating a low wage

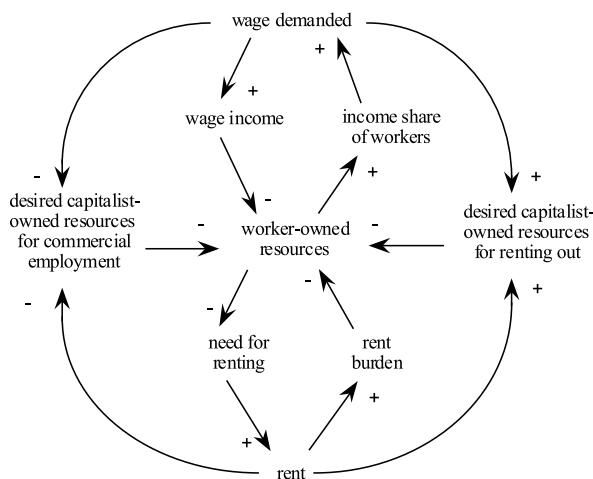
rate, as shown in Fig. 6. An arrow connecting two variables indicates the direction of the causality while a positive or a negative sign shows the slope of the function relating cause to effect. For clarity, only key variables located along each feedback path are shown.

When productive resources can potentially be engaged in commercial or self-employed modes by owners and renters, any autonomous increase in the wage rate would not only decrease the desired capitalist owned resources for commercial employment, it would also concomitantly decrease the utility of investing in resources for self-employment. Thus, while the ownership of resources freed from commercial employment is not transferred to the self-employed sector, the surplus labor released by the commercial sector has to be absorbed in self-employment. As a result, worker income is depressed while the demand for renting rises. Thus, it not only continues to be profitable for the capitalist sector to hold its investments in land and capital, it also gives this sector a financial edge over the self-employed sector, whose savings continue to decline as its rent burden rises. These actions spiral into an expansion of ownership of resources by the capitalist sector even though the commercial mode of production is eliminated due to the high cost of wage labor. This also precipitates a very low wage rate when equilibrium is reached since a low claim to income of the economy creates low opportunity costs for the self-employed workers for accepting wage-employment.

Ironically, the fine distinction between the corporate, artisan and absentee types of ownership is not recognized in the political systems based on the competing neoclassical and Marxist economic paradigms. The former protects all types of ownership; the latter prohibits all. None creates a feasible environment in which a functional form of ownership may help to capture the entrepreneurial energy of the enterprise.

Possibilities for Poverty Alleviation

A functional economic system must incorporate the mechanisms to mobilize the forces of self-interest and entrepreneurship inherent in private ownership of the resources. Yet, it must avoid the conflicts inherent in the inequalities of income and resource ownership that led to the creation of the alternative socialist paradigm, which is devoid of such forces. According to the preceding analysis, the fundamental mechanism which creates the possibility of concentration of resource ownership is the equal protection accorded to the artisan and absentee forms of ownership by the prevailing legal norms. The financial fragmentation of households and the differences in their



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 8

Feedback loops creating dysfunctional income distribution trends in the capitalis system

saving patterns further facilitate the expansion of absentee ownership. Technological differences between the capitalist and self-employed sectors not only make possible the side-by-side existence of the two modes of production, they also exacerbate the dichotomy between ownership of resources and workership. Apparently, the policy agenda for changing resource ownership and income distribution patterns should strive to limit renting and should additionally prevent the development of financial fragmentation and technological differentiation between the commercial and self-employed production modes if the objective is to minimize the conflicts related to income distribution.

Assisting the Poor

Programs to provide technological, organizational, and financial assistance to the poor have been implemented extensively in the developing countries over the past few decades although they have changed neither income distribution nor wage rate as reflected in many learned writings over these decades as well as the data published by UN and World Bank. This occurred because the increased productivity of the self-employed mode first pushed up wage rate, making renting-out resources more attractive for the capitalist sector than commercial production. However, the consequent decrease in wage payments and increase in rent payments pushed down the income share of the workers, which again suppressed the wage rate. Any efforts to facilitate the small-scale sector to increase its productivity through technological development also failed to affect income distribution since the mechanism of renting allowed the gains of the improved productivity to accrue to the absentee owners of the resources [56]. This experience is verified by the simulation of Fig. 9, which incorporates the policies to improve productivity, creating financial institutions and assisting the self-employed to adopt modern technologies. These policies only increase the size of the self-employed sector without increasing worker income, due to the possibility of separation of the mode of production from the ownership of resources. This indicates that influencing the decision to retain resources in absentee mode for renting out should be the key element of a policy framework to improve income distribution that should alleviate poverty.

Influencing Income Distribution

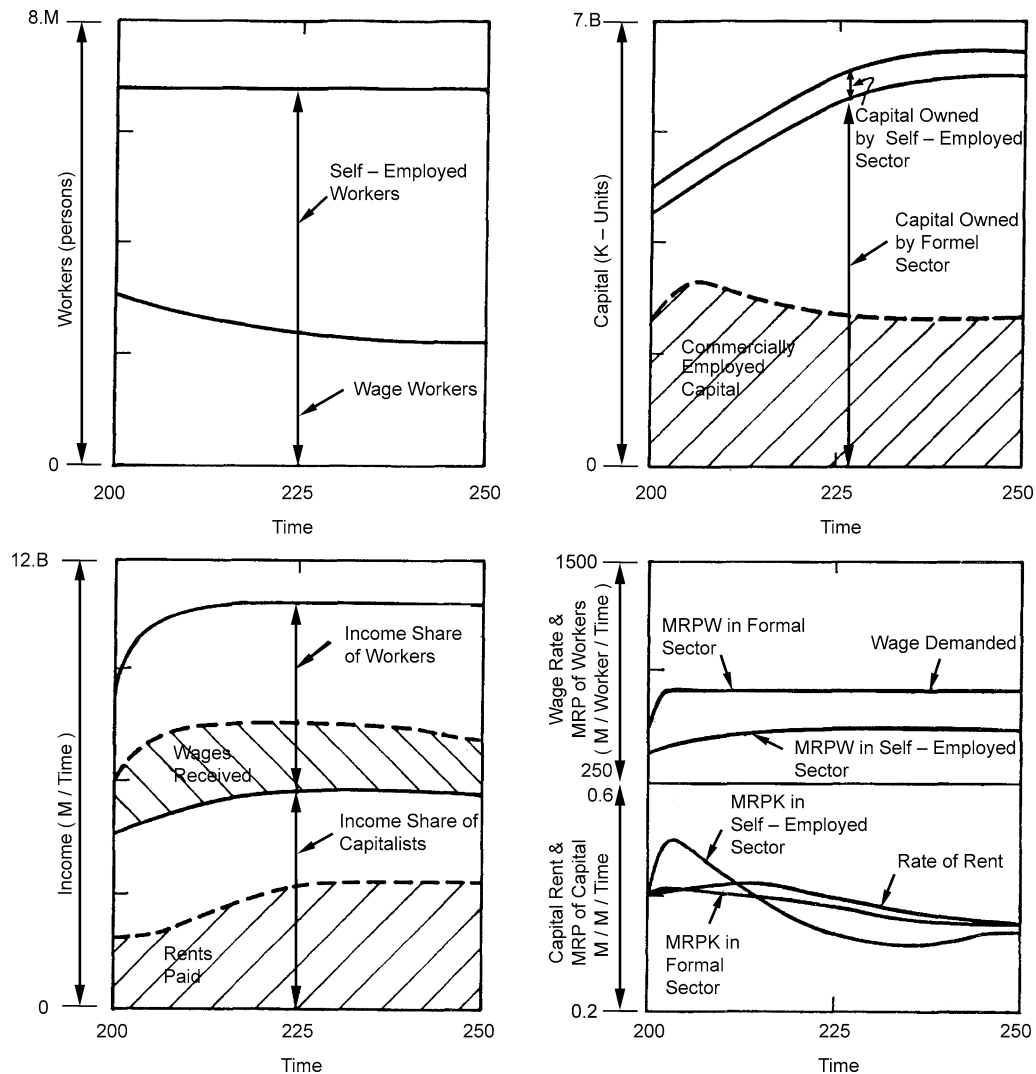
The cost of owning capital resources in absentee form can be increased by imposing a tax on rent income. The results of implementing this policy, together with the policies of Fig. 9 are shown in Fig. 10. In the face of a tax on rent in-

come, resources which cannot be employed efficiently under the commercial system are offered for sale to the self-employed instead of being leased out to them. Purchase of these resources by the self-employed raises the entitlement of the workers to the income of the economy, which increases the opportunity cost of supplying wage-labor to the commercial sector. This raises wage rate, which makes the commercial mode of production even more uneconomical, unless it is able to apply a superior technology. Such changes spiral in the long run into a transfer of a substantial amount of resources to the self-employed sector. Concomitant efforts to decrease the financial fragmentation of households and the technological differentiation between the two modes of production, along with improving productivity, further accelerate these changes.

Facilitation of Innovation and Productivity Improvement

Macroeconomic analyses concerning the industrialized countries show that technological innovation is one of the most important sources of growth [9,61]. Studies conducted at the organizational level in the industrialized countries also show that innovations creating technological progress originate largely from small entrepreneurs or from large companies structured in a way to encourage small independent working groups [38,42]. Thus, entrepreneurial activity is often credited with raising productivity through creation of technical and business-related innovations [6]. The high and rising cost of labor in the developed countries, possibly also forces the wage-employers into finding innovative ways of maintaining high labor productivity and continuously striving to improve it.

On the other hand, economic growth has been dominated in the developing countries by relatively large and often highly centralized and vertically integrated companies. Technologies of production have mostly been replanted from the industrialized countries and indigenous innovation and technological development have had a poor track record [1]. These technologies often do not perform as well as at their respective sources, but due to the availability of cheap labor, their inefficient performance still yields comfortable profits; hence little effort is made to improve productivity. There also exist serious limitations on the number of small players as large cross-section of the households in the developing countries lack the resources to effectively participate in any form of entrepreneurial activity [53,58]. Innovation being a probabilistic process, limited participation drastically limits its scope.



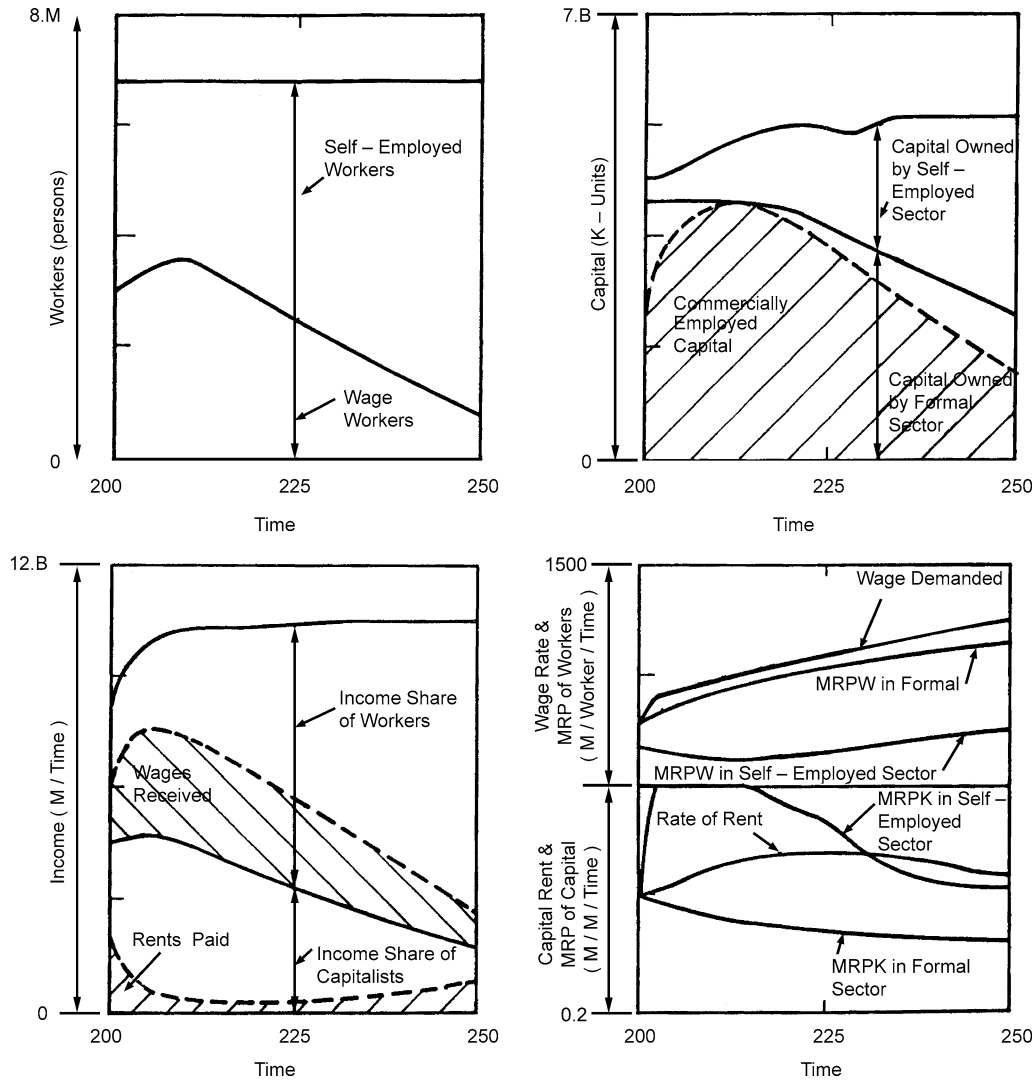
Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 9

Perpetuation of low wage and unequal income distribution resulting from widely used economic development policies

There exists a promising institution in most developing countries, however, which has great potential as a focal point of entrepreneurial activity, which has remained dormant for lack of empowerment. This institution is the small family enterprise in the self-employed sector, which may take the form of a shop-house or an artisan manufacturing firm in the urban sector or a peasant farm in the rural sector. It allows participation from all members of the family while also providing the informal small-group organization considered conducive to innovation in many studies. Its members are highly motivated to work hard and assume the risk of enterprise because of their commitment to support the extended family. This enterprise is somewhat similar to the small manufacturing units

that created the industrial revolution in England in the early nineteenth century. It has also been observed that the small family enterprise tends to maximize consumption; hence its income significantly affects demand, which creates new marketing opportunities [1,4]. Unfortunately, this enterprise has been systematically suppressed and discriminated against in favor of the large-scale capitalist sector. Even its output remains largely unaccounted for in the national accounting systems of most countries [11,20].

The small family enterprise, variously described as the informal, labor-intensive, traditional, peasant, peripheral and sometimes inefficient sector in the developing countries has been stifled in the first instance by a set of social and legal norms through which the wealth has become



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 10

Changes in wage and income distribution resulting from adding taxation of rent income to the policy package

concentrated in an absentee ownership mode. Working households are mostly poor and own few assets [19]. The prosperity of these households will not only provide the much-needed financial resources for entrepreneurial activity, their capacity to spend will also create many marketing opportunities for the potential entrepreneur. Thus, influencing income distribution, through the policy framework proposed in the last section, ranks first on the agenda also for developing entrepreneurship and encouraging innovation. Once a significant cross-section of the populace becomes a potential participant in the economic activity, the development of infrastructure and facilitation to manage risk will also appear to be effective instruments to sup-

port entrepreneurship and innovation [51]. The rise in the wage rates due to the possibility of alternative self-employment opportunities would, at the same time, force the large commercial enterprise to invest in technological innovation for productivity improvement, which should further improve the efficacy of the overall system.

Conclusion

Both neoclassical and Marxist models of economic growth seem to make restricting assumptions about ownership and mechanisms of wage determination, which are linked with specific time- and geography- related historical ev-

idence. These restricting assumptions give internal consistency and a semblance of sustainability to each model, although they remove both from reality. A failure in the free-market system based on the neo-classical model occurs when the invisible hand concentrates ownership of resources in a small minority, suppressing wage rate and creating social conflict due to income inequalities. On the other hand, a failure in the socialist system based on the Marxist model occurs, when the visible hand empowered to act in the public interest stifles entrepreneurial energy while also ignoring public interest in favor of its power interests [48,50].

A behavioral model underlying wage and income distribution has been proposed in this paper, in which the opportunity cost of supplying a unit of labor to the capitalist sector is used as a basis for negotiating a wage. Neither this opportunity cost nor the ownership pattern are taken as given, while the dynamic interaction between the two creates a tendency in the system to generate numerous wage and income distribution patterns, subsuming those postulated in the neo-classical and Marxist theories of economics. The realization of a specific wage and income distribution pattern depends on legal and social norms concerning ownership, renting, the financing of investment and the state of technology.

Private ownership seems to have three forms, commercial, artisan and absentee. Predominance of artisan ownership creates an egalitarian wage and income distribution pattern while a healthy competition between the commercial and artisan firms may release considerable entrepreneurial energy. These functional forms can grow only if the renting of resources can be discouraged. On the other hand, absentee ownership creates a low wage rate and an unequal income distribution, while the growth of this form of ownership is facilitated through the renting mechanism. Potentially, all three ownership forms can exist in an economic system. The problem, therefore, is not to favor or condemn private ownership *per se* as the alternative theories of economics have often advocated, but to understand the reasons behind the development of a particular ownership pattern and identify human motivational factors that would change an existing pattern into a desired one.

The most important reform needed at government level to alleviate poverty is the discouragement of the absentee ownership of capital assets, which would create a wider distribution of wealth. Widespread artisan ownership resulting from this would increase participation in entrepreneurial activity, which would allow adequate performance from the human actors in the system. Such reforms may however not be possible in the authoritarian systems

of government pervasive in the developing countries since they must often limit civil rights and public freedoms to sustain power. Hence, the creation of a democratic political system may be a pre-condition to any interventions aimed at poverty alleviation. This, I have discussed elsewhere [50,53,54].

Future Directions

While the market system has often been blamed by the proponents of central planning for leading to concentration of wealth and income among few, it in fact offers a powerful means for redistributing income if the process of concentration is carefully understood and an intervention designed on the basis of this understanding. In fact, all economic systems can be reformed to alleviate the dysfunctional tendencies they are believed to have, provided the circular relationships creating such dysfunctions can be understood which should be the first objective of policy design for economic development.

Contrary to this position, economic development has often viewed developmental problems as pre-existing conditions, which must be changed through external intervention. Poverty, Food shortage, poor social services and human resources development infrastructure, technological backwardness, low productivity, resource depletion, environmental degradation and poor governance are cases in point. In all such cases, the starting point for a policy search is the acceptance of a snapshot of the existing conditions. A developmental policy is then constructed as a well-intended measure that should improve existing conditions. Experience shows, however, that policies implemented with such a perspective not only give unreliable performance, they also create unintended consequences. This happens because the causes leading to the existing conditions and their future projections are not adequately understood. The well-intentioned policies addressing problem symptoms only create ad hoc changes, which are often overcome by the system's reactions.

Table 1 collects three key developmental problems, poverty, food shortage and social unrest, and the broad policies implemented over the past several decades to address them. These problems have, however, continued to persist or even become worse.

The policy response for overcoming poverty was to foster economic growth so aggregate income could be increased; that for creating food security was intensive agriculture so more food could be produced; and for containing social unrest, the broad prescription is to strengthen internal security and defense infrastructure so public could be protected from social unrest. The unintended

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Table 1
Developmental problems, policies implemented to address them and unintended consequences experienced

Initially perceived problems	Policies implemented	Unintended consequences
Poverty	Economic growth capital formation sectoral development technology transfer external trade	Low productivity indebtedness natural resources depletion environmental degradation continuing/increased poverty
Food shortage	Intensive agriculture land development irrigation fertilizer application use of new seeds	Land degradation depletion of water aquifers vulnerability to crop failure population growth continuing/increased vulnerability to food shortage
Social unrest	Spending on internal security and defense infrastructure limiting civil rights	Poor social services poor economic infrastructure authoritarian governance insurgence continuing/increased social unrest

consequences of these policies are many, but in most instances, they include a continuation or worsening of the existing problems.

Thus, poverty and income differentials between rich and poor have in fact shown a steady rise, which is also accompanied by unprecedented debt burdens and extensive depletion of natural resources and degradation of environment. Food shortages have continued but are now accompanied also by land degradation, depletion of water aquifers, the threat of large-scale crop failure due to a reduction in crop diversity and a tremendous growth in population. Social unrest has often intensified together with appearance of organized insurgence burgeoning expenditures on internal security and defense, which has stifled development of social services and human resources and have created authoritarian governments with little commitment to public welfare.

The unintended consequences are often more complex than the initial problems and have lately drawn concerns at the global level, but whether an outside hand at the global level would alleviate them is questionable. This is evident from the failure to formulate and enforce global public policy in spite of active participation by national governments, global agencies like the UN, the World Bank, the World Trade Organization, and advocacy networks sometimes referred to as the civil society. This failure can largely be attributed to the lack of a clear understanding of the roles of the actors who precipitated those problems and whose motivations must be influenced to turn the tide.

Thus, development planning must adopt a problem solving approach in a mathematical sense if it is to achieve sustainable solutions. In this approach, a problem must

be defined as an internal behavioral tendency and not as a snap shot of existing conditions. It may represent a set of patterns, a series of trends or a set of existing conditions that appear either to characterize a system or to be resilient to policy intervention. In other words, an end condition by itself must not be seen as a problem definition. The complex pattern of change implicit in the time paths preceding this end condition would, on the other hand, represent a problem. The solution to a recognized problem should be a solution in a mathematical sense, which is analogous to creating an understanding of the underlying causes of a delineated pattern. A development policy should then be perceived as a change in the decision rules that would change a problematic pattern to an acceptable one. Such a problem solving approach can be implemented with advantage using system dynamics modeling process that entails building and experimenting with computer models of problems, provided of course a succinct problem definition has first been created.

Appendix

Model Description

Wage rate WR is assumed to adjust over period $WRAT$ towards indicated wage rate IWR .

$$d/dt[WR] = (IWR - WR)/WRAT \quad (1)$$

IWR depends on the wage-bargaining position of the workers, which is determined by their opportunity cost of accepting wage-employment. It is assumed that the opportunity cost of transferring a self-employed worker to wage-work is zero when wage offered is equal to the current con-

sumption expenditure per worker averaged over the whole workforce.

$$IWR = [(R_s * (1 - SP_s) + (AS_s/LAS))/TW], \quad (2)$$

where R_s , SP_s and AS_s are, respectively, income share, saving propensity and accumulated unspent savings of the self-employed sector. LAS and TW are, respectively, life of accumulated unspent savings and total workforce. Subscripts s and f designate, respectively, self-employed and capitalist sectors.

Ownership of land and capital as well as contribution to labor are the bases for claim to income while absentee ownership is possible through leasing arrangements. Thus, R_s is computed by adding together the value of output produced by the self-employed sector VQ_s and the wage payments received by the wage-workers W_f , and subtracting from the sum the rent payments made to the absentee owners. R_f is given by adding together the value of output produced by the capitalist sector VQ_f and the rent payments it receives from the self-employed sector, and subtracting from the sum the wage-payments it makes.

$$R_s = VQ_s + WR * W_f - LR * RL - KR * RK, \quad (3)$$

$$R_f = VQ_f - WR * W_f + LR * RL + KR * RK, \quad (4)$$

where LR , RL , KR , and RK , are, respectively, land rent, rented land, capital rent, and rented capital.

KR and LR depend, respectively, on the long-term averages of the marginal revenue products of capital and land ($AMRPK$ and $AMRPL$) in the economy, and the demand for renting capital and land (RKD and RLD) as compared with the supply of rentable assets (RK and RL). The demand for renting, in turn, depends on the lack of ownership of adequate resources for productively employing the workers in the self-employed sector.

$$KR = AMRPK * f_1[RKD/RK]; \quad f'_1 > 0 \quad (5)$$

$$RKD = DKE_s - KO_s. \quad (6)$$

Where DKE_s is desired capital to be employed in the self-employed sector and KO_s is capital owned by it. Land rent LR and demand for renting land RLD are determined similarly.

The saving propensity of all households is not uniform. Since capitalist households associated with the capitalist sector receive incomes which are much above subsistence, their saving propensity is stable. On the other hand, the saving propensity of the worker households depends on their need to save for supporting investment for self-employment and on how their absolute level of income compares with their inflexible consumption. Thus, SP_s in

the model is determined by the utility of investment in the self-employed sector arising from a comparison of worker productivity in the sector with the wage rate in the capitalist sector, and the rent burden of this sector compared with the factor contribution to its income from land and capital.

$$SP_s = \mu * f_2[MRPW_s/WR] * f_3[(LR * RL + KR * RK)/(VQ_s - MRPW_s * W_s)], \quad (7)$$

$$SP_f = \mu, \quad (8)$$

where $f'_2 > 0$, $f'_3 < 0$, μ is a constant, and $MRPW$ is marginal revenue product of workers.

AS represent the balance of unspent savings, which determine the availability of liquid cash resources for purchase of assets. AS are consumed over their life LAS whether or not any investment expenditure occurs.

$$\begin{aligned} d/dt[AS_i] \\ = R_i * SP_i - AS_i/LAS - LA_i * PL - \sum_j KA_i^j * GPL; \\ i = s, f; \quad j = m, t, \end{aligned} \quad (9)$$

where LA , PL , KA , and GPL are, respectively, land acquisitions, price of land, capital acquisitions, and general price level. Subscript i refers to any of the two sectors, self-employed (s) and capitalist (f), and superscript j to the type of capital, modern (m) or traditional (t).

W_f is assumed to adjust towards indicated workers IW_f given by desired workers DW_f and total workforce TW . TW is assumed to be fixed, although, relaxing this assumption does not alter the conclusions of this paper. All workers who are not wage-employed must be accommodated in self-employment. Thus W_s represents the remaining workers in the economy.

$$d/dt[W_f] = (IW_f - W_f)/WAT \quad (10)$$

$$IW_f = TW * f_4(DW_f/TW) \quad (11)$$

$$W_s = TW - W_f \quad (12)$$

where $1 \geq f_4 \geq 0$, and $f'_4 > 0$. WAT is worker adjustment time.

The desired workers in each sector DW_i is determined by equating wage rate with the marginal revenue product of workers. A modified Cobb–Douglas type production function is used.

$$DW_i = E_i^w * VQ_i/WR, \quad (13)$$

where E_i^w is the elasticity of production of workers in a sector.

Land and capital owned by the capitalist sector (LO_f and KO_f) are allocated to commercial production (KE_f and LE_f) and renting (RK and RL) activities depending on the desired levels of these factors in each activity. Thus,

$$RK = (DRK/(DRK + DKE_f)) * KO_f \quad (14)$$

$$RL = (DRL/(DRL + DLE_f)) * LO_f \quad (15)$$

$$KE_f = KO_f - RK \quad (16)$$

$$LE_f = LO_f - RL \quad (17)$$

Capital and land employed by the self-employed sector consist of these production factors owned by them and those rented from the capitalist sector.

$$KE_s = KO_s + RK \quad (18)$$

$$LE_s = LO_s + RL \quad (19)$$

Desired capital and land to be employed in any sector (DKE_i and DLE_i) are determined on the basis of economic criteria.

$$d/dt(DKE_i)/KE_i = f_6[MRPK_i/MFCK] \quad (20)$$

$$d/dt(DLE_i)/LE_i = f_5[MRPL_i/MFCL] \quad (21)$$

where f'_5 and $f'_6 > 0$. $MRPL_i$ and $MRPK_i$ are respectively marginal revenue products of land and capital in a sector, and $MFCL$ AND $MFCK$ are respectively marginal factor costs of land and capital.

$$MRPL_i = (E_i^l * VQ_i/LE_i) \quad (22)$$

$$MRPK_i = (E_i^k * VQ_i/KE_i) \quad (23)$$

$$MFCL = PL * IR \quad (24)$$

$$MFCK = IR + (1/LK) * GPL \quad (25)$$

where E_i^l and E_i^k are, respectively, elasticities of production of land and capital in a sector. PL is price of land, IR is exogenously defined interest rate, LK is life of capital and GPL is general price level.

Changes in the quantities of capital and land desired to be rented out (DRK and DRL) depend on their respective rents KR and LR compared with their marginal factor costs $MFCK$ and $MFCL$.

$$d/dt[DRK]/RK = f_7[KR/MFCK]; \quad f'_7 > 0 \quad (26)$$

$$d/dt[DRL]/RL = f_8[LR/MFCL]; \quad f'_8 > 0. \quad (27)$$

The value of output produced by each sector is given by the product of the quantity it produces Q_i and the general price level GPL .

$$VQ_i = Q_i * GPL \quad (28)$$

$$Q_i = A_i * K_i^{E_{ki}} * L_i^{E_{li}} * W_i^{E_{wi}}, \quad (29)$$

where K_i , L_i , and W_i represent capital, land and workers employed by a sector. A_i represent technology constants, which increase with the use of modern capital.

$$A_i = \hat{A} * f_9[K_i^m/(K_i^t + K_i^m)], \quad (30)$$

where $f'_9 > 0$ and \hat{A} is a scaling factor based on initial conditions of inputs and output of the production process.

Ownership is legally protected and the financial market is fragmented by households. Thus, purchase of any productive assets must be self-financed by each sector through cash payments. Land ownership LO_i of each sector changes through acquisitions LA_i from each other. Each sector bids for the available land on the basis of economic criteria, its current holdings, and the sector's liquidity.

$$LA_i = d/dt[LO_i] \quad (31)$$

$$LO_i = \left(DLO_i / \sum_i DLO_i \right) * TL, \quad (32)$$

where DLO_i is desired land ownership in a sector and TL is total land which is fixed,

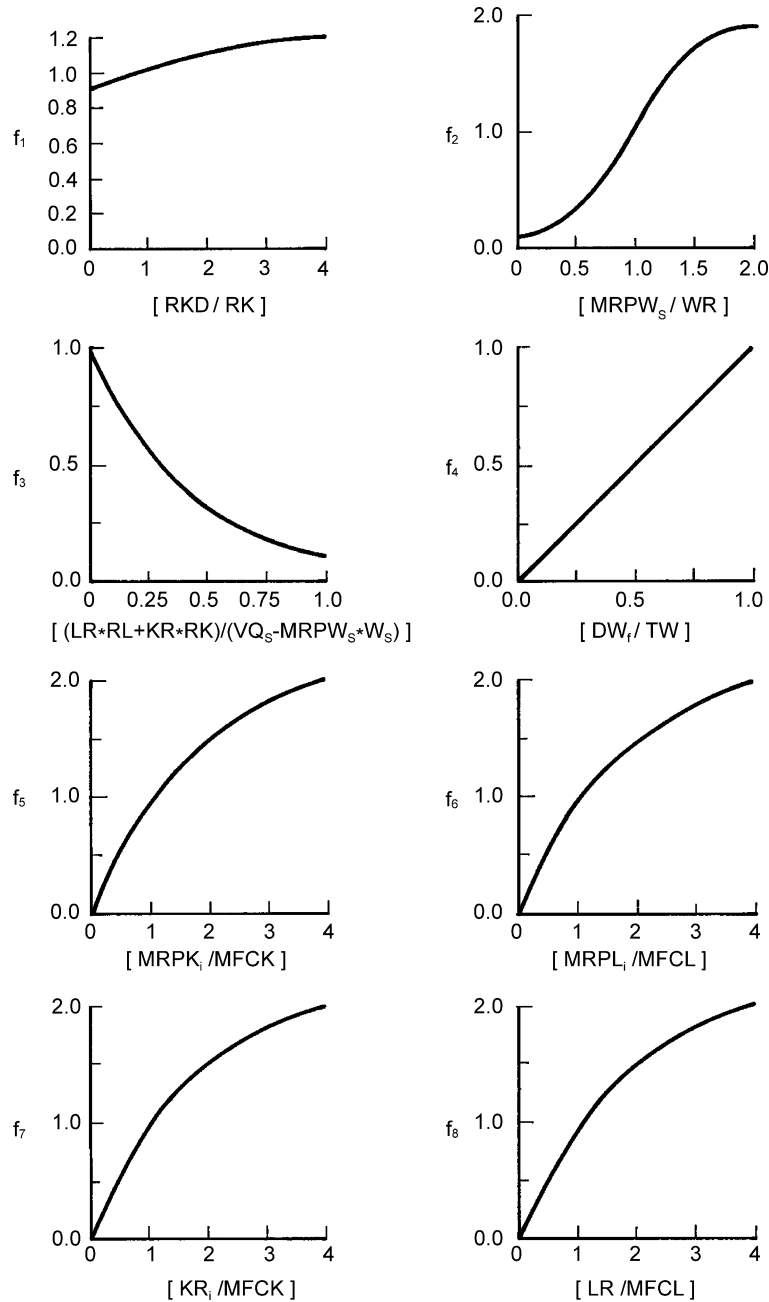
$$DLO_i = LO_i * f_6[MRPL_i/MFCL] * f_{11}[CA_i], \quad (33)$$

where $f'_{11}[CA_i]$ is > 0 , and CA_i is cash adequacy of a sector.

Cash adequacy of a sector CA_i is given by the ratio of its accumulated unspent savings to the desired savings. The latter is computed by multiplying cash needed to finance investment and the traditional rate of consumption of savings in the sector by cash coverage CC .

$$CA_i = AS_i / \left(\left((AS_i/LAS) + (LA_i * PL) + \left(\sum_j KA_{ij} * GPL \right) \right) * CC \right). \quad (34)$$

Capital ownership in a sector $KO_i = KO_i^t + KO_i^m$ changes through acquisitions KA_i^j and decay. Although there is a preference for modern capital, its acquisition KA_i^m de-



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 11
Behavioral relationships f_1 through f_8

depends on the ability to accommodate the technology represented by it. Inventory availability of each type of capital KIA^j also limits its purchases.

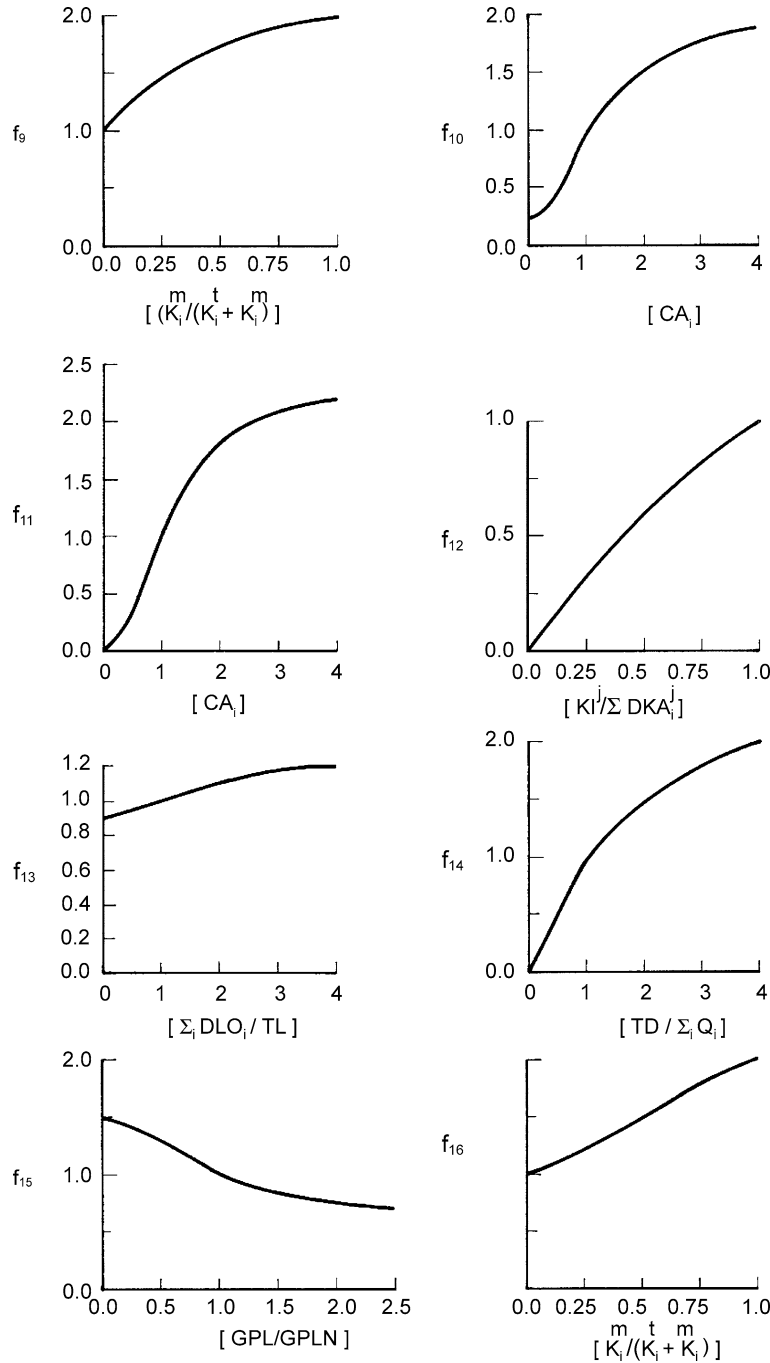
$$d/dt[KO_i] = \sum_j KA_i^j - KO_i/LK, \quad (35)$$

$$KA_i^j = DKA_i^j * KIA^j, \quad (36)$$

$$DKA_i^m = (KO_i/LK) * f_5[MRPK_i/MFCK] * f_{11}[CA_i] * TCF_i, \quad (37)$$

$$DKA_i^t = (KO_i/LK) * f_5[MRPK_i/MFCK] * f_{11}[CA_i] * (1 - TCF_i), \quad (38)$$

where DKA_i are desired capital acquisitions, $f'_{11} \geq 0$, and



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation, Figure 12

Behavioral relationships f_9 through f_{16}

LK is life of capital. TCF_i represent exogenously defined technological capability. $0 < TCF_i < 1$.

where $0 \leq f_{12} \leq 1$, $f'_{12} > 0$, and KIC is capital inventory coverage

$$KIA^j = f_{12} \left[KI^j / \left(\sum_i DKA_i^j \right) * KIC \right], \quad (39)$$

$$d/dt[KI^j] = KQ^j - \sum_i KA_i^j, \quad (40)$$

where KQ^j represent supply of capital. KQ^m is imported, while KQ^t is created within the economy by allocating a part of the capacity to its production.

$$KQ^t = \sum Q_i * \left(\sum_i DKA_i^t / TD \right). \quad (41)$$

The price of land PL is assumed to adjust towards indicated price of land IPL which is given by the economy-wide average of the marginal revenue product of land $AMRPL$, interest rate IR and the desired land ownership in each sector DLO_i

$$d/dt[PL] = (IPL - PL)/LPAT, \quad (42)$$

$$IPL = (AMRPL/IR) * f_{13} \left[\sum_i DLO_i / TL \right]; \quad \text{where } f'_{13} > 0. \quad (43)$$

General price level GPL is determined by supply and demand considerations.

$$d/dt[GPL] = GPLN * f_{14} [TD / \sum_i Q_i] \quad (44)$$

where $f'_{14} > 0$. $GPLN$ is normal value of GPL and TD is total demand for goods and services to be produced within the economy. TD is given by adding up non-food consumption C_i , traditional capital acquisition KA_i^t and production of traditional capital for inventory, food demand FD and government spending G which is equal to taxes, if any, collected.

$$TD = \sum C_i + \sum_i DKA_i^t + \left(\left(KIC * \sum_i DKA_i^t - KI^j \right) / IAT \right) + FD + G, \quad (45)$$

$$d/dt(C_i) = \frac{[(((R_i * (1 - SP_i) + AS_i / LAS) / GPL) * FNFC_i) - C_i]}{CAT}, \quad (46)$$

where IAT is inventory adjustment time, $FNFC_i$ fraction non-food consumption, and CAT is consumption adjustment time. Food demand FD is given by multiplying population P with normal per capita food demand $NFPCD$ and a function f_{15} representing a weak influence of price.

$$FD = P * NFPCD * f_{15}[GPL/GPLN], \quad (47)$$

where $f'_{15} < 0$ and P bears a fixed proportion with total workforce TW .

The elasticity of production of land E_i^l is assumed to be constant as is suggested by empirical evidence concerning agricultural economies [Strout 1978, Heady and Dillon 1961]. Elasticity of production of capital E_i^k depends on the technology of production, which is determined by the proportions of traditional and modern capital employed. Since constant returns to scale are assumed, E_i^w is given by (47).

$$E_i^k = f_{16} [K_i^m / (K_i^t + K_i^m)]; \quad f'_{16} > 0, \quad (48)$$

$$E_i^w = 1 - E_i^k - E_i^l. \quad (49)$$

Behavioral Relationships

Sixteen behavioral relationships $[f_1 \dots f_{16}]$ have been incorporated into the model. The slope characteristics of these relationships have already been described in above equations. The graphical forms of the functions representing these relationships are shown in Figs. 11 and 12 placed below. General considerations for specifying such relationships are discussed in [63].

Bibliography

Primary Literature

1. APO (1985) Improving productivity through macro-micro linkage. Survey and Symposium Report. Tokyo: Asian Productivity Organization
2. Applebaum E (1979) The labor market. In: Eichner A (ed) A Guide to Post-Keynesian Economics. ME Sharpe, White Plains, New York
3. Averitt RT (1968) The dual economy: the dynamics of american industry structure. Norton, New York
4. Bardhan PK (1973) A model of growth in a dual agrarian economy. In: Bhagwati G, Eckus R (eds) Development and planning: essays in honor of Paul Rosenstein-Roden. George Allen and Unwin Ltd, New York
5. Barro RJ (1997) Macroeconomics, 5th edn. MIT Press, Cambridge
6. Baumol WJ (1988) Is entrepreneurship always productive? J Dev Plan 18:85-94
7. Boeke JE (1947) Dualist economics. Oriental economics. Institute of Pacific Relations, New York
8. Cornwall J (1978) Growth and stability in a mature economy. John Wiley, London
9. Denison EF (1974) Accounting for United States economic growth 1929-1969. Brookings Institution, Washington
10. Eichner A, Kregel J (1975) An essay on post-Keynesian theory: a new paradigm in economics. J Econ Lit 13(4):1293-1314
11. Eisner R (1989) Divergences of measurement theory and some implications for economic policy. Am Econ Rev 79(1):1-13

12. Fie JC, Ranis G (1966) Agrarianism, dualism, and economic development. In: Adelman I, Thorbecke E (eds) *The Theory and Design of Economic Development*. Johns Hopkins Press, Baltimore
13. Forrester JW (1979) Macrobehavior from microstructure. In: Karman NM, Day R (eds) *Economic issues of the eighties*. Johns Hopkins University Press, Baltimore
14. Forrester JW (1987) Lessons from system dynamics modelling. *Syst Dyn Rev* 3(2):136–149
15. Galbraith JK (1979) *The nature of mass poverty*. Harvard University Press, Cambridge
16. Gordon DM (1972) *Economic theories of poverty and underemployment*. DC Heath, Lexington
17. Graham AK (1977) *Principles of the relationships between structure and behavior of dynamic systems*. Ph D Thesis. MIT, Cambridge
18. Griffin K, Ghose AK (1979) Growth and impoverishment in rural areas of Asia. *World Dev* 7(4/5):361–384
19. Griffin K, Khan AR (1978) Poverty in the third world: ugly facts and fancy models. *World Dev* 6(3):295–304
20. Hicks J (1940) The valuation of the social income. *Economica* 7(May):163–172
21. Higgins B (1959) *Economic development*. Norton, New York
22. Hirshliefer J (1976) *Price theory and applications*. Prentice Hall, Englewood Cliffs
23. Kaldor N (1966) Marginal productivity and the macro-economic theories of distribution. *Rev Econ Stud* 33:309–319
24. Kaldor N (1969) Alternative theories of distribution. In: Stiglitz J, Ozawa H (eds) *Readings in modern theories of economic growth*. MIT Press, Cambridge
25. Kalecki M (1965) *Theory of economic dynamics*, revised edn. Allen and Unwin, London
26. Kalecki M (1971) *Selected essays on dynamics of capitalist economy*. Cambridge Univ Press, London
27. Kindelberger C, Herrick B (1977) *Economic development*, 3rd edn. McGraw Hill, New York
28. Leontief W (1977) Theoretical assumptions and non-observable facts. In: Leontief W (ed) *Essays in Economics*, vol II. ME Sharpe, White Plains
29. Lewis WA (1958) Economic development with unlimited supply of labor. In: Agarwala I, Singh SP (eds) *The Economics of Underdevelopment*. Oxford University Press, London
30. Lipton M (1977) *Why poor people stay poor*. Harvard University Press, Cambridge
31. Marglin SA (1984) *Growth, distribution and prices*. Harvard Univ Press, Cambridge
32. Marx K (1891) *Capital*. International Publishers, New York (Reprinted)
33. McKinnon RI (1973) *Money and capital in economic development*. The Brookings Institution, New York
34. Minsky H (1975) *John Maynard Keynes*. Columbia University Press, New York
35. Mukhia H (1981) Was there feudalism in Indian history. *J Peasant Stud* 8(3):273–310
36. Myrdal G (1957) *Economic theory and under-developed regions*. Gerald Duckworth Ltd, London
37. Pack SJ (1985) *Reconstructing Marxian economics*. Praeger, New York
38. Quinn JB (1985) *Managing innovation: controlled chaos*. Harvard Bus Rev 85(3):73–84
39. Ricardo D (1817) *Principles of political economy and taxation*, Reprint 1926. Everyman, London
40. Richardson GP (1991) *Feedback thought in social science and systems theory*. University of Pennsylvania Press, Philadelphia
41. Riech M, Gordon D, Edwards R (1973) A theory of labor market segmentation. *Am Econ Rev* 63:359–365
42. Roberts EB (1991) *Entrepreneurs in high technology, lessons from MIT and beyond*. Oxford University Press, New York
43. Robinson J (1955) Marx, Marshal and Keynes: three views of capitalism. Occasional Paper No. 9. Delhi School of Economics, Delhi
44. Robinson J (1969) The theory of value reconsidered. *Aust Econ Pap* June 8:13–19
45. Robinson J (1978) *Contributions to modern economics*. Basil Blackwell, Oxford
46. Robinson J (1979) *Aspects of development and underdevelopment*. Cambridge University Press, London
47. Roulet HM (1976) The Historical context of Pakistan's rural agriculture. In: Stevens RD et al (eds) *Rural development in Bangladesh and Pakistan*. Hawaii University Press, Honolulu
48. Rydenfelt S (1983) *A pattern for failure. Socialist economies in crisis*. Harcourt Brace Jovanovich, New York
49. Saeed K (1988) Wage determination, income distribution and the design of change. *Behav Sci* 33(3):161–186
50. Saeed K (1990) Government support for economic agendas in developing countries. *World Dev* 18(6):758–801
51. Saeed K (1991) Entrepreneurship and innovation in developing countries: basic stimulants, organizational factors and hygies. *Proceedings of Academy of International Business Conference*. Singapore, National University of Singapore
52. Saeed K (1992) Slicing a complex problem for system dynamics modelling. *Syst Dyn Rev* 8(3):251–261
53. Saeed K (1994) Development planning and policy design: a system dynamics approach. Foreword by Meadows DL. Ashgate/Avebury Books, Aldershot
54. Saeed K (2002) A pervasive duality in economic systems: implications for development planning. In: "Systems dynamics: systemic feedback modeling for policy analysis". In: *Encyclopedia of life support systems (EOLSS)*. EOLSS Publishers, Oxford
55. Saeed K (2005) Limits to growth in classical economics. 23rd International Conference of System Dynamics Society, Boston
56. Saeed K, Prankprakma P (1997) Technological development in a dual economy: alternative policy levers for economic development. *World Dev* 25(5):695–712
57. Sen AK (1966) Peasants and dualism with or without surplus labor. *J Polit Econ* 74(5):425–450
58. Sen AK (1999) *Development as freedom*. Oxford University Press, Oxford
59. Simon HA (1982) *Models of bounded rationality*. MIT Press, Cambridge
60. Skinner A (ed) (1974) *Adam Smith: The wealth of nations*. Pelican Books, Baltimore
61. Solow R (1988) Growth theory and after. *Am Econ Rev* 78(3):307–317
62. Sraffa P (1960) *Production of commodities by means of commodities*. Cambridge University Press, Cambridge

63. Sterman J (2000) Business dynamics. McGraw Hill, Irwin
64. Streeten P (1975) The limits of development studies. 32nd Montague Burton Lecture on International Relations. Leeds University Press, Leeds
65. Takahashi Y et al (1970) Control and dynamic systems. Addison-Wesley, Reading
66. Weintraub S (1956) A macro-economic approach to theory of wages. *Am Econ Rev.* 46(Dec):835–856

Books and Reviews

- Atkinson G (2004) Common ground for institutional economics and system dynamics modeling. *Syst Dyn Rev* 20(4):275–286
- Ford A (1999) Modeling the environment. Island Press, Washington DC
- Forrester JW (1989) The system dynamics national model: macrobehavior from microstructure. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International System Dynamics Conference. Springer, Berlin
- Forrester N (1982) A Dynamic synthesis of basic macroeconomic theory: implications for stabilization policy analysis. Ph D Dissertation, MIT
- Forrester N (1982) The life cycle of economic development. Pegasus Communications, Waltham
- Hines J (1987) Essays in behavioral economic modeling. Ph D Dissertation, MIT, Cambridge
- Mass NJ (1976) Economic cycles, an analysis of the underlying causes. Pegasus Communications, Waltham
- Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. *J Econ Issues* 37(1):133–173
- Randers J (1980) Elements of system dynamics method. Pegasus Communications, Waltham
- Richardson GP (1996) Modeling for management: simulation in support of systems thinking. In: Richardson GP (ed) The international library of management. Dartmouth Publishing Company, Aldershot
- Saeed K (1980) Rural development and income distribution, the case of pakistan. Ph D Dissertation. MIT, Cambridge
- Saeed K (1994) Development planning and policy design: a system dynamics approach. Ashgate/Avebury Books, Aldershot
- Saeed K (1998) Towards sustainable development, 2nd edn: Essays on System Analysis of National Policy. Ashgate Publishing Company, Aldershot
- Saeed K (2002) System dynamics: a learning and problem solving approach to development policy. *Glob Bus Econ Rev* 4(1):81–105
- Saeed K (2003) Articulating developmental problems for policy intervention: A system dynamics modeling approach. *Simul Gaming* 34(3):409–436
- Saeed K (2003) Land Use and Food Security – The green revolution and beyond. In: Najam A (ed) Environment, development and human security, perspectives from south asia. University Press of America, Lanham
- Saeed K (2004) Designing an environmental mitigation banking institution for linking the size of economic activity to environmental capacity. *J Econ Issues* 38(4):909–937
- Sterman JD (2000) Business dynamics. Systems thinking and modeling for a complex world. Irwin McGraw Hill, Boston
- Wolstenholme E (1990) System enquiry, a system dynamics approach. John Wiley, Chichester

Dynamics of Language

ROBERT PORT

Linguistics/Cognitive Science, Indiana University, Bloomington, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Speech and Meters](#)

[Further Problems with an Alphabetical Model of Language](#)

[Two Levels of Complex System](#)

[Language as a Social Institution](#)

[Realtime Language Processing](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Phone A ‘minimal’ speech sound whether consonant or vowel, the unit that is represented by a single letters of the International Phonetic Alphabet. A phone is invariant across syllable positions, neighboring context, speaking rate, speaker, etc.

Phoneme An abstract speech sound unit in a particular language, typically consisting of several phone types that are treated as the same (despite any differences) by speakers of the language. Thus, English /t/ (slashes indicate use of a symbol as a phoneme, not a phone) includes both an **allophone** (a particular phoneme variant) that is aspirated (as in the word *take*), another allophone that is a glottal stop (in the usual American pronunciation of *cotton*) and another that is a flap or tap (as in *butter*).

Phonology The branch of linguistics that studies the nature of the speech sound patterns of particular languages, such as the inventory of phonemes, the patterns of syllable construction, stress patterns, etc. Thus for the English word *stops*, the *st-* is the onset of the syllable, the vowel is the nucleus and the coda is the *-ps*. Phonology should be concerned with intonation and speech timing as well, although these are not traditional interests.

Meter An underlying temporal pattern of a fixed number of ‘beats’ within a larger cycle. Thus, there are meters consisting of 2 beats per measure, 3 beats or 4, 5 or 6. Speech often aligns itself with such patterns, e. g., in

chant or song, by locating vowel onsets close in time to the pulse of each cycle.

Orthography The set of conventions about how to write a language. This includes conventions about which letters are used and what sounds they represent in the ideal case. The conventions also include standard spellings for every word, plus conventions of capitalization, punctuation, the definition of a sentence and so forth.

Definition of the Subject

Language has been a topic of great interest at least since the Greek philosophers. Clearly speech is one of the most sophisticated and complex of human capabilities and has remained a major puzzle through the centuries. Complex systems promise to provide analytical tools that will help break through some of the mystery. From this perspective, it appears that there are three complex systems relevant to speech. The first is natural selection, the slow influences on the gene pool of the species that give us whatever innate skills we need to learn at least one language. However, we will have nothing further to say about this issue here since the time scale involved is so long relative to human history. The remaining two complex systems are highly relevant. They are, first, the language itself, viewed here as a particular type of social institution, a set of cultural practices that present the language learner with a broad range of linguistic materials with which to communicate with others. This system clearly evolves as history and the cultural practices of the community change. It sometimes requires only a couple hundred years (a dozen or so generations) for a language to split into mutually incomprehensible dialects.

The final complex system is the one that the individual speaker develops as he/she gradually becomes competent in the use of the language. This system begins before birth and continues to learn and adapt right up until death as the language environment of an individual changes. It involves controlling the many muscle systems of the vocal tract and respiratory system as well as specializations of the auditory system that customize it for hearing and understanding speech in the ambient language or languages.

There is one historical development of central importance that greatly complicates our ability to apply our intuitions to language. This is the 3 thousand-year-old tradition of training our children to read and write using an alphabet. This training, involving hundreds of hours over many years, typically results in our hearing speech as though it consists of a sequence of discrete letter-sized sounds, the consonants and vowels. The lifelong train-

ing and practice interpreting letter strings as continuous speech and interpreting speech as a string of letters has, in this author's opinion, encouraged unwarranted confidence in the reliability of our intuitions about the structure of language. Thus, one goal of this essay is to show that our intuitions about the segmental structure of speech are untrustworthy.

Introduction

What kind of structures are exhibited by a natural spoken language? Lay people think first of *words*. Words are the archetypal units of language. It seems every language displays sequences of words and must have a dictionary, its foundational list of words. Linguists agree with this intuition and add another unit that may be somewhat less obvious to laymen: the speech sound, the *phone* or phoneme, that is, the fragment of speech that is represented by a single *letter* from some alphabetical orthography or from the International Phonetic Alphabet [23]. To begin the study of language by using letters and a list of words seems very natural but turns out to be problematic if our goal is to understand linguistic cognition because letters (and phonemes) provide inadequate models of speech. The main problem is that letters (and phonemes) provide too few bits of information to represent what speakers need to know about speech. One unfortunate consequence of this supposed efficiency is that only the ordering of letters can represent the temporal layout of speech events. Let us begin then by making observations, not about our intuitive description of language structure in terms of an alphabet, but rather by making observations of some specific realtime linguistic behavior.

Demonstration

In order to provide an intuitive example of realtime linguistic behavior, the reader is urged to participate in a tiny experiment on oneself. The reader is encouraged to repeat the following phrase out loud: *Take a pack of cards*. Please repeat this phrase over and over – aloud (a whisper will work fine), between 5 and 10 times. Now. After cycling through the phrase, notice the rhythm pattern that you adopted unconsciously. When a person repeats a short phrase like this over and over, it seems the easiest way to do it is to slip into some rhythmic timing pattern that is the same from repetition to repetition. A remarkable fact is that there is only a small number of ways that most people will arrange the timing of a cycled phrase like this one. To discover which rhythm you used, repeat the phrase again just as you did before but tap a finger in time with the stressed syllables of the phrase. You should find your-

Example 1

- Pattern (1a) has 2 beats per cycle, a beat on *Take* and another on *cards* with the next beat on *Take* again. There is a finger tap for each beat (but none on *pack*). Pattern (1b) has 3 beats (and taps) per cycle: the first on *Take*, the second on *cards* and the third is a beat that has no syllable, just a longer pause than (1a) has – equivalent to a musical ‘rest’. Patterns (a) and (b) must have a long period so that *Take a pack of* can be spoken in a single beat-cycle. Example (1c) also has 3 beats, on *Take*, *pack* and *cards*, and, as in the waltz rhythm, the repetition cycle begins immediately with another *Take*. Most Americans first discover pattern (a) of Example 1, but most people can produce all three without much difficulty. To those familiar with music notation, we could say Example (1a) has a 2-beat measure with *Take* and *cards* beginning at each beat and dividing the measure into equal halves. Both of the others have a 3-beat measure but (1b) has a rest on beat 3. The unstressed syllables seem to be squeezed or stretched to assure that the salient target syllable onsets begin at the proper time, that is, as close as possible to phase zero of one or both of the coupled oscillators generating the metrical time structure.

Example 2 | : Take a pack of cards, [rest] : |
(1 2) (3 4) (5 6)

Take after *cards*. This creates a 5-beat pattern. (If leaving out the rest on beat 6 proves difficult for the reader, then count, 1, 2, 1, 2, 3, 1, 2, 1, 2, 3 a few times, then switch to the test phrase.) English speakers can learn to produce these more exotic 5- and 6-beat meters, but it takes some effort and practice to achieve them. There is some anecdotal evidence, however, that both Japanese and Finnish speakers, for example, are able to effortlessly produce the 5-beat pattern. A few experimental tasks like these have been looked at with speakers of several languages – at least Arabic, Japanese, German and Finnish [53,57]. Everyone tested so far seems to find at least one of the temporal patterns of Example 1 to be very natural and almost unavoidable for many short phrases that are cycled in any language.

Since it is likely that differences between languages in preferred rhythm patterns will continue to hold up, it is likely that rhythmic patterns are an intrinsic aspect of any language, and thus that research on the phonology of a language must begin to rely on continuous time data. These language-specific constraints on speech timing (as well as many subtle phonetic differences between dialects and languages) show that it is reckless to begin the study of language by first encoding all speech data into a small set of uniform letter-like tokens that have only serial order instead of real time. The problem is that, even though our intuitions about language tell us letters are obvious units of language, letters are actually a culturally-transmitted technology that humans have learned to apply to (and, if necessary, impose upon) spoken language. Letters and their psychological counterparts, phones and phonemes, have very weak experimental support as psychological units used for cognitive “spelling” of words [42,43,45,47].

Given the intuitive descriptions of the timing of repeated short phrases, an experimental study is needed to verify that these three patterns are the primary ones that English speakers can reliably produce. But first, for rhythmically produced speech, an important question arises: just what in the acoustics are subjects aligning with the finger taps? That is, what is the physical definition of a “beat” or auditory pulse? Most scientists working in this area of speech agree that it is approximately the onset of a vowel [2,12,30,49]. This is usually located by low-pass filtering of speech (at about 800 Hz) and integrating the remaining audio signal over about a 250 ms time window and looking for positive peaks in the second derivative that indicate a rapid increase in the energy in the low frequen-

cies. That is where people tap their finger in time to a syllable.

In the experiment, speakers were asked to cycle short phrases like *Dig for a dime* and *King for a day* and to put the final noun (e.g., *dime*) at specific phase lags relative to the *Dig-Dig* interval [12]. The target phase angle was indicated by a metronomic pattern of 2 tones repeated 10–15 times per trial. Can speakers begin the word *dime* anywhere we ask within the cycle? It was expected the participants would find it easier to locate the onset at the simple harmonic fractions discovered above, i.e., one half, one third and two-thirds of the cycle. These phase angles might be accurately reproduced and other phase lags would be less accurately reproduced. Subjects were played a metronomic pattern of 2 tones, one High and one Low which they listened to for several cycles. Then they were to repeat the indicated phrase 10 or 12 times in a row aligning the Low tone with *Dig* and the High tone with *dime*. Now in the stimulus patterns, the phase angle φ by which the High tone (i.e., for *dime*) divided the Low-Low interval (i.e., *Dig-Dig*) was varied randomly and uniformly over the range from $\varphi = 0.20$ to 0.70 . The overall period of *Dig-Dig* was also varied inversely with the ratio. The earlier the target phase angle, the slower the tempo of the repetition cycle, so as to leave a constant amount of time for pronouncing the phrase.

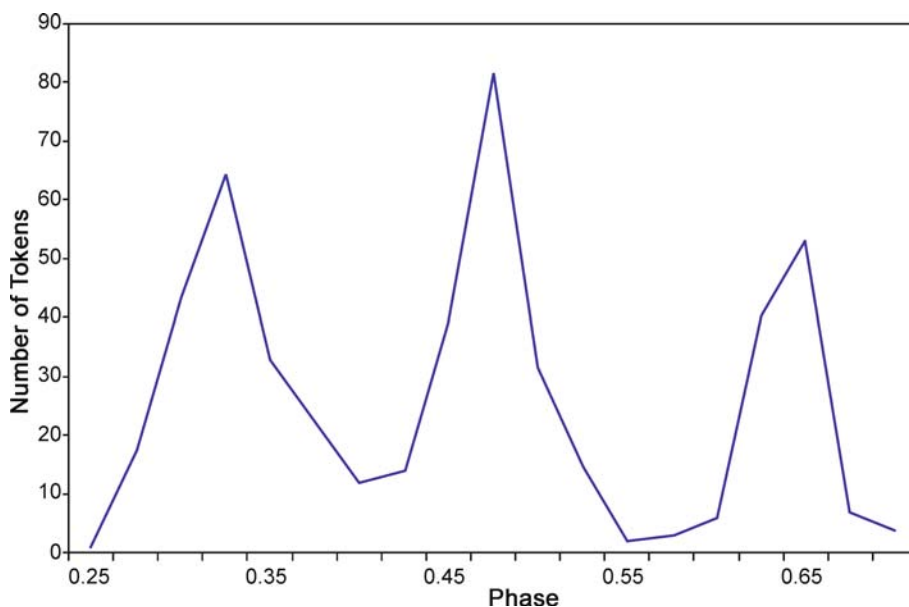
If the American participants were able to imitate those phase angles without bias, as they were asked to do, then

they should produce a flat, uniform distribution of the onset of the vowel of *dime* relative to *Dig* onsets over that range – thereby replicating the pattern of the stimulus metronome. But Fig. 1 shows the observed frequency histogram of the median phase angle of the onset of *dime* for each trial across 8 speakers. The histogram represents about 1400 trials where each trial consisted of 12–15 cycles of a test phrase said to a single metronome pattern.

Clearly the participants exhibited a strong tendency to locate their syllable onsets near just the 3 points: 1/3 of the cycle, 1/2 of the cycle or 2/3 of the cycle corresponding to the readings of Example (1b, 1a) and (1c) respectively. But they do not show evidence of any preference for fifths of the repetition cycle – no hint of peaks at 0.4 or 0.6. The subjects were trying to do as they were told, but when the high tone fell at, e.g., $\varphi = 0.4$, the speakers tended to produce φ closer either to 0.5 or 0.33.

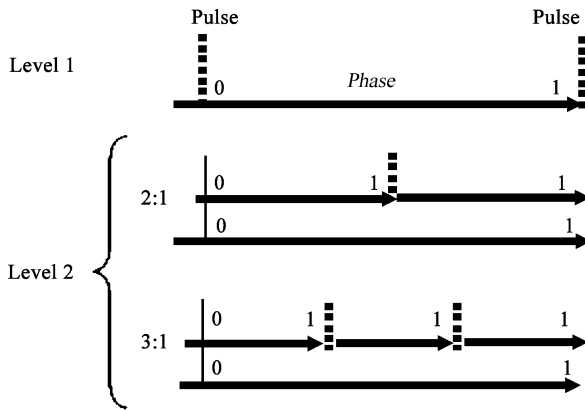
Dynamical Model of Meter

These timing phenomena can be accounted for with dynamical models for neural oscillations correlated with auditory patterns [27,32,47]. A natural interpretation for the three peaks is that they represent attractors of a cyclical metric pattern specified by 2 coupled oscillators locked into one of 2 patterns, either a 2-beats-per-cycle pattern or 3-beats-per-cycle pattern. An important point about meter is that it is audition, the auditory sense, that has the



Dynamics of Language, Figure 1

The distribution of the vowel onsets for the last stressed syllable by 8 speakers saying phrases resembling *Dig for a dime* with target phase angles specified by a metronomic sequence of 2 tones. Figure from [44]



Dynamics of Language, Figure 2

Simple meters represented as circles opened out into a line. Phase 0 and Phase 1 are, of course, the same point on the circle. The upper panel shows a single oscillator with periodic phase zero serving as an attractor for perceived or produced energy onsets. The second and third panels show two ways that a second oscillator can be coupled with the first. Each produces either 2 cycles or 3 cycles before coinciding with the slower cycle. It is assumed that all phase zeros attract energy onsets both in speech and nonspeech sound. The 3 patterns of Example 1 can be accounted for by the 2:1 and 3:1 coupled oscillator or meter systems. Figure from [44]

strongest link to meter. Of course, we cannot tell from these data whether this bias toward the 3 attractor phase angles comes primarily from the participants' perception systems (so, e.g., $\varphi = 0.4$ simply *sounds* like $\varphi = 0.5$) or if the distortion reflects the influence of their motor system which might find it easier to consistently produce the auditorily salient pulse at one of the attractors.

Notice that these patterns involve cycles at two time scales nested within each other as shown in Fig. 2. The slower cycle is the repetition cycle of the whole phrase, represented in the upper panel in the figure. Its phase zero coincides with the onset of *Dig*. For all the phrases in this experiment, there is also a second cycle (thereby implying a state space that is a torus although a torus is inconvenient for display). The second one is faster and tightly coupled with the repetition-cycle oscillator. It cycles either 2 or 3 times before its phase zero aligns with phase zero of the slower oscillator [44]. If we assume that the temporal location of any phase zero is an attractor for a prominent auditory onset, then such a system of 2 coupled oscillators could define several musical meters and account for the 3 powerful patterns of Example 1.

Of course, subject behavior in "speech cycling" experiments is not the only evidence for a preference by humans for periodic speech. Seemingly all human communities have developed traditions of religious, martial or commer-

cial chants and songs as well as other genres of vocal music that build on the metrical principles exhibited by people performing basic speech repetition. So the phenomena reviewed here point toward several important lessons about how to analyze human speech behavior.

The first lesson is that human speech performance, like so many other kinds of dynamical systems, finds periodic cycling easy and natural to do even though there is no mass-bearing object (such as a leg or finger) oscillating at the observed frequencies. The speech periodicities reflect oscillations that must be occurring in various regions of the brain stem, cortex and cerebellum see [38,40,56]. Humans, in just about any language, can warp their speech into singsong or chant very easily. We seem to enjoy entraining our speech to each other see [11] and to external auditory patterns [41]. Just above, the reader was asked simply to repeat a phrase and within a couple repetitions, the performance slipped into one of a small number of periodic attractors that this particular fragment of English can settle into. Once there, the speech production system tends to stay there until you force yourself out of the attractor (e.g., by trying to obey impertinent instructions from the author about assigning stress to bold-face syllables and tapping a finger).

The most important lesson from the phenomena reviewed so far is that we must be very careful about taking alphabetical forms of language as the basic data for a study of human speech. We may do so (as most linguistic research has done for a century) for convenience but must keep in mind that alphabetical representation makes constrained timing relationships like those revealed in the demonstration in the "Introduction" completely invisible, since letters and words in their written form exhibit only serial order, not continuous time. It seems like for this speech cycling task the speech is riding on the dynamics of low frequency oscillations. Examples 1a, 1b and 1c are, of course, identical in terms of alphabetical description, but they are far from identical in timing. Speakers of different languages exhibit different rhythmic effects in speech cycling and in other aspects of timing. Clearly it makes little sense to define language as somehow completely independent of these powerful speech behaviors.

Further Problems with an Alphabetical Model of Language

The invisibility of real time is one reason to reject representations of language using an alphabet. There are many more reasons. For example, the results of the demonstration and the laboratory experiment show that many properties of human speech resemble those of familiar dynamical systems.

ical systems in which oscillatory patterns are frequently observed [1]. But within linguistics and much of psychology, language has been defined for the past century or so as a ‘symbol system’ or a ‘code’ shared by the speaker and hearer [3,7,13,19,37]. These timing patterns cannot be symbolically coded since they are defined by measurements in continuous time. The tendency of people to warp speech timing to fit patterns of integer-ratio time intervals is inexplicable by any model consisting only of serially ordered symbols – like words or the consonant and vowel segments. There are many kinds of temporal patterns in various languages that are inexpressible with any alphabet. So before approaching the study of speech and language in dynamical terms, we need to clarify that we will address a domain that is much broader than ‘language’ as most linguists and psychologists have considered it.

The domain of phenomena for linguistic research needs to be *speech events in continuous time* of people in contextualized linguistic performance. This kind of raw data can easily be recorded, thanks to recent technologies on audio and video or partially displayed as a sound spectrogram (where time is mapped onto one spatial axis). But transcription or reduction of raw speech sound into any alphabet cannot be presumed, whether a hypothetical psychological alphabet such as of (language specific) phonemes, or an orthographic alphabet (such as the form you see printed on this page), or a technical phonetic alphabet (e.g., the International Phonetic Alphabet used by linguists for all languages for over a century now). Letters throw away far too much of what is critical for the dynamics of language as used by speakers.

In addition to speech rhythm, a second major problem with the traditional view is that it predicts invariant acoustic correlates for each segment. But it has been known for almost 50 years, that speech perception relies on formant trajectories and details of spectral shape and spectral change that cannot be captured by any invariant, context-free acoustic cues (e.g., [34,35]). Speech cues for phonetic segments simply cannot be defined so as to do the job of specifying phones or phonemes independently of the identity of neighboring vowels and consonants, speaker identity, speaking rate, etc. despite many attempts to do so see [28,52]. Thus, the shape of the second formant that specifies [d] before an [u] (as in English *do*) shows a falling trajectory whereas the second formant rises for [d] before an [e¹] (as in *day*). Such examples are found throughout the domain of speech cues. The acoustic cues for each phonetic segment tend to be different for each context, different enough so that a general definition can not be formulated that works across a range of different contexts. One consequence of these failures is that engineers working on

speech recognition long ago abandoned attempts to recognize words by first recognizing segments or segmental features as a way of specifying words. Instead, the engineers try to specify whole words and phrases directly in terms of acoustic spectra rather than try to detect the kind of segmental units that linguists continue to insist are essential for word specification. By representing words using a codebook of spectrum slice descriptions, successful artificial speech perception systems effectively store context-specific versions of the phones and phonemes [22,25].

A third difficulty with the view that spoken words have a phonemic spelling for memory is that it makes the wrong prediction about the results of recognition memory experiments, where a subject indicates whether each word in a list is repeated (that is, has occurred earlier in the list) or not. The data show that words repeated by the same voice have a significant advantage in recognition accuracy over words that are repeated in a different voice. This shows that subjects do retain speaker-specific information (often called “indexical information”) in memory. This is difficult to reconcile with the claim that words are stored using only representations like phones or phonemes that are abstracted away from speaker-specific properties indicating the speaker’s voice [14,39,43,45].

Finally, if words are perceived, produced and stored using a discrete list of distinct, abstract phones or phonemes, then linguistic variation of all kinds – dialect variation, historical changes in pronunciation, speaker idiosyncrasies, etc. – should all show discrete jumps in pronunciation between these discrete sound types. Speech should not show continuous variation in parameters like vowel height or backness, degrees of lip rounding, place of articulation, voice-onset time, segment durations, etc. All these parameters should exhibit audible jumps when the transcription changes, e.g., from [t] to [d] or from [i] to [I] (except for possible noisy variation in production). However, generations of phonetics research reveals no evidence of discrete changes in any of these parameters as far as I can tell [6,32,36,45,46]. The apparent target values of vowel quality show no signs of discrete jumps in target in a plot of the first formant against the second formant. And for voice-onset time, the time lag between the release of a stop and the onset of voicing exhibits many apparently continuous values of VOT target depending on the degree of stress, the following vowel or consonant identity, etc. [45]. Nowhere is there evidence that speech sounds are discretely different from each other in the same way that letters are. Nor is there evidence that the gestures are nonoverlapping in time (quite the opposite, they overlap a great deal, [4]). So, the hypothesis that words are cognitively spelled, that is, represented in memory, in terms of

discrete, alphabet-like units simply finds no experimental support. It is extremely unlikely that the standard view of lexical representations used in linguistics will turn out to be correct.

Could there be some reason we have been overlooking for why the nervous system might not rely on these abstract units like phones and phonemes? One problem is probably that a vast amount of essential information about the speech would be lost in the transcription process. In addition to all temporal properties of speech (such as segment durations, rhythmic patterns, etc.), information about coordination details of the speaker motor activity, details of voice quality and speaker identifying properties, etc. are all discarded when speech is reduced to alphabetical form. Many linguists have suggested that an alphabetical transcription is a “more efficient” representation than an audio or video recording [3,10,24]. But more efficient for what? Is it more efficient as a form of word memory that is useful to a human speaker? This is, of course, an empirical question. The answer depends on how memory for words actually works. For a memory system optimized for dealing with a non-discrete world where non-discrete sound patterns are the norm, perhaps storing as much information as possible works better and might actually be “easier”.

Like most modern linguists, Chomsky [7,8] followed his predecessors in the “structural linguistics” movement (e.g., [3,13,21]) in presuming that words have a representation in a psychological alphabet that is small enough that each segment type can be assumed to be distinct from all the others. Every language is expected to exhibit this kind of static ‘representation’. Words are presumed to be distinguished from each other by letter-like segments (unless they happen to be homonyms, like *two*, *too*, *to*, and *steal*, *steel*, *stele*). Yet the evidence reviewed above shows that speakers **do** encode much phonetic detail in their speech perception and in speech productions. Consider all that a good actor encodes into speech: subtle, constantly changing moods, and the region of the speaker’s origin, their social class, etc. as well as the text of the screenplay. Listeners, too, are able to interpret the speaker’s mood, feelings, region, etc. The traditional picture of language as represented in memory in alphabet-like form finds no more support from these familiar phenomena about speech than it found in experimental studies.

Origin of Alphabetical Intuitions

Where might these intuitions come from if, in fact, words are not stored in the abstract way that we supposed? Why are we so comfortable and satisfied with alphabet-like de-

scriptions of speech? The answer should have been obvious. For roughly the past 3k years, the educated minority in the Western cultural tradition have had a powerful technological scaffold for thinking about language [9,10]. They have had an alphabetical orthography for their language. Writing captures a few important and relatively distinctive properties of speech that can be encoded (by someone with appropriate training) into letters and individual words, as in either an orthographic or a phonetic transcription. Please consider that every reader of this page has spent many hours a week since age 4 or 5, practicing the conversion of letters into words and words into letters. These hundreds of hours of training with an alphabetical orthography surely play a major role in shaping the intuitions that all of us share about language, and surely biased the very definition of language, and thus of linguistics. In the linguistic tradition since Saussure, including the Chomskyan tradition, knowledge of a language is assumed to consist in large part, of linguistic representations of words in memory. Since we are talking about ‘knowledge’, each word is stored just once – just as in a published dictionary of a language. In addition to the set of lexical tokens, there are supposed to be rules in the “grammar” that manipulate the linguistic representations as simple tokens (moving segments or segment strings around, deleting or inserting them). This symbol-processing mechanism constitutes a powerful assumption about highly abstract representations of words. But why is it so easy for people to make such elaborate assumptions about language? The traditional linguistic representations for phonemes and words, etc., turn out to get many of their properties by generalization from the technology of conventional orthographic letters and words. Because of our experience using the technology of writing, it has come to seem natural to assume that something like this would be used by speakers mentally.

An orthographic system is a collection of customs about how to write a language. Development of a useable orthography requires deciding exactly what the alphabet will be and its application to sounds spoken in the language, which fragments of speech constitute a ‘word’ and how each word is to be spelled. Other orthographic conventions address what patterns count as a sentence and what words a sentence may or may not begin with. For a learner of English orthography (including linguists and foreign language teachers), one problem is that the relation between letters and sounds is far from consistent [48]. Thus, we no longer pronounce either the **k** or the **gh** in *knight*. The letter **i**, which in most languages sounds like the vowel in *pea*, often sounds like *my* in English, and the vowel [ɛ] can be spelled either **ea** (in *head*), **e** (in *bed*) or

ai (in *said*). Because of these well-known inconsistencies in many languages of Europe, scientists of speech together with second-language pedagogues developed the first version of the IPA phonetic alphabet about 1890. This system employs primarily Roman letters plus many others to describe speech sounds as pronounced with a certain level of detail. But it takes some training to be able to accurately hear some sounds that speakers use. This method pays no attention to ‘intended’ speech sounds or to orthographic spellings. It can be seen that now the 3k year-old technology of an alphabet was employed for two very different purposes: It plays its traditional role of providing the basis for a useful orthography for various languages but also very recently now provides a consistent notation system for scientific recording of speech gestures and speech sounds in many languages.

It seems likely that the development of the IPA alphabet influenced theories of language. It was immediately recognized that one could transcribe to varying levels of detail, so the distinction between *broad* vs. *narrow* transcription was discussed. Within 20 years of the release of the first IPA alphabet, some language scientists (in particular, a Polish linguist named Beaudoin de Courtenay) began speaking of ‘*phonemes*’ (see [53] for the early history of the term). These were to be a minimal set of abstract speech sounds, the smallest set that seems adequate to write a language. So the phoneme appeared as a kind of *idealized letter* that is hypothesized to describe something in the head of speakers. To see how the phoneme is an idealized letter, consider first the properties of letters. Letters in an alphabetical orthography:

1. are *discretely* different from each other.
2. There is a *fixed small set* of them available in any orthography.
3. They are *serially ordered* and graphically *nonoverlapping*.
4. Each letter has an *invariant shape* across different writers. (Although handwriting and font may differ, the letters are still discretely identifiable.)
5. Each word has a *single canonical spelling* in the ‘dictionary’ (that is, in the mental inventory of the linguistic representations of the language).

Notice that because in orthographies words are spelled only from letters, words will always be discretely distinct as well. In fact, even syntax is assured of inheriting the discreteness of the orthographic alphabet. The usefulness and efficiency of letters (plus their intimate familiarity to literate scientists) makes it very plausible to hypothesize psychological counterparts to letters. It seems quite natural to imagine a *phone* or *phoneme* as the psychological ana-

logue of a letter. Thus linguists routinely assume that there are:

1. *discrete* differences between *phones* or *phonemes* which are
2. *drawn from a small set* of phoneme types for each language,
3. and are *serially ordered* and *nonoverlapping in time*.
4. Each phone or phoneme has an *invariant physical form* across contexts, speakers, rates of speech, etc.
5. The words of a language have a *single “canonical” representation* in memory, that is, some prototype version to which incoming tokens are compared.

From this traditional view, linguistic memory is primarily a dictionary with every word having its distinctive spelling in a small, abstract alphabet so it can be differentiated discretely from all non-homophones in the language. A major problem has been Assumption 4 about an invariant physical form for each phoneme. A century of phonetics research on phones and phonemes has shown unambiguously that **there is no way to give these abstract linguistic segments concrete specification**. Successful acoustic definitions for all linguistic segments and features have not been found – in fact, there are no definitions for *any* of the segments or features that are effective across neighboring contexts, speaking rate, etc. Thus there is an incommensurability between letter representations and the actual representations in memory. Very simply, there are no “acoustic letters” for speech. We phoneticians spent a generation or two trying to provide physical acoustic definitions of linguistic units like particular features and segments, but we have not been able to. It almost certainly cannot be done.

So before starting to analyze the dynamics of human speech, we must accomplish some mental self-cleansing. It is important to really understand that letters in the phonetic alphabet are just one kind of model for speech. This model is very useful – especially to those of us with literacy education. It is convenient for a writing system (because it requires learning only a small number of letters). However, such a representation is not, it turns out, closely related to the way speakers, literate or illiterate, store words in memory, recognize them or produce them.

A century ago, the phonetic alphabet provided the best technology available for representing speech for scientific study. But for up to half a century now continuous-time technical displays like audio waveforms, sound spectrograms and smoothed electromyographic plots have been available to show us vividly that letter-like units of speech sound are **not** what people use for representing speech. We have been making excuses for this counter-evidence for

over half a century. It is time to give up the illusion that words have a cognitive spelling into units that resemble letters. There are only two places where they really have such a spelling: on paper when people use orthographic letters to write words down, and in the conscious conceptualization of language by literate people.

The conclusion is that we must reject letter-like phones or phonemes as actual units that define spoken language. But then how can we begin to analyze any particular language or human language in general? We must begin afresh with a clean slate, looking at language without the biases that come from our cultural history. This literate culture demands that we develop great skill at interpreting speech sound as letters and reading letters as speech sound.

Two Levels of Complex System

A very different basic cut of the phenomena surrounding human speech is required, one that distinguishes the two primary complex systems that support human speech communication. It is proposed, following Smith, Brighton and Kirby [50], that we separate:

- (A) the properties of the *language patterns of a community of speakers*, that is, what we might call the social institution of language, on one hand, from
- (B) the skills of *real-time speech processing*, on the other.

The “grammar of a language” should be understood as a description of patterns across a community of speakers and contexts summarized over some historically brief time window (of, say, a generation). Linguistics is (or should be) studying the patterns and the cultural categories that are found to define the unit-like chunks in any language. Linguists should describe the most common patterns and the categories of sounds, and make whatever generalizations they are able to, stated in terms of whatever descriptive vocabulary for timing and articulatory state seem to work. But, if we look into the details, it will be clear that all speakers in a community discover their own pattern definitions.

This implies that approximation across speakers is the best that can be done. There is an unavoidable uncertainty about the precise definitions of the auditory micro-features that “spell” words (or word-like units) for any individual. In addition, each speaker has been exposed to a different subset of the corpus of speech in that language (differing in regional and social dialects, foreign accents, etc.). Thus any given language may have a gross description, the kind linguists and phoneticians are trained to produce using the IPA alphabet or one that is provided by an orthographic representation. But the closer we look, the more

variation will be found. Spoken languages do not have a consistent or uniform finely detailed description. Very simply, there exists no universal vocabulary with which to describe it. The reason for this is, again, that the fine description in each speaker is created independently by developmental processes based on the language-exposure history of each speaker. This is why phonological observations about the sound patterns of any language exist only at the group level and are not necessarily “in” the individual speaker (nor even in *any* speaker). The idea that phonology only exists in the corpus of a community and not in the individual speaker’s representations simply acknowledges that speakers can and will differ from each other to varying degrees.

The acquisition of the skills of real-time speech processing can now be understood in outline. The child, well before birth, begins learning the auditory patterns (human generated and otherwise) [26]. During the first year, the child learns to recognize the typical speech sound-figures of his ambient language and classifies them into categories [31,55]. On this theory, the child increasingly develops his auditory analysis system to perceive speech in his ambient language and then to produce it. Notice that this system has to work without any a priori description of what the speech patterns of the community actually are. The learner has no idea in advance what “the grammar” really is, so the perceptual system learns as best it can to predict the patterns based on very concrete descriptions. It appears the learner makes progress by storing lots of detail about specific utterances that have been heard. Counter-intuitively (at least to literates like us), the real-time speech system makes no use whatever of the abstract segments or letter-like units that are so prominent in our conscious experience of language. Similarly, speech perception takes place with no necessity of recovering any low-dimensional alphabet-like description along the way.

The story told here implies several observations: (1) discrete categories of linguistic patterns are not the same as discrete symbol structures, (2) the abstract and partly discrete structure of phonology exists in a speaker only as a structure of categories. But phonological categories are sets of speech chunks specified in a rich code that are taken to be ‘the same’ as other members of some category. Their discreteness is never guaranteed because they are fairly high dimensional. Plus some events will be found where category assignment is difficult to impossible (or ambiguous). Finally note that (3) speakers can exhibit patterns in their behavior that they have no explicit representation of (and thus the patterns are not appropriately described as ‘knowledge’ of any kind). The child learns to perceive and produce speech that is compatible with the

way others talk in his or her community but each does so using idiosyncratic components (because each is controlling only his own vocal tract and exploiting his own auditory system whose detailed structure reflects the speaker's personal history).

So then, the two complex systems of human speech are, first, the cultural institution, the set of linguistic habits, patterns and category types exhibited by the community of speakers (what we usually call the phonology, grammar and lexicon but also cultural style, etc.), and, second, the individual's working system for speech production (controlling a vast number of articulatory degrees of freedom) and speech perception (employing the speaker's partly idiosyncratic speech sound analysis system). This realtime system seeks sufficient compatibility with the community's usage patterns that the speaker can speak and be understood. The two systems each have their own development processes on very different time scales. The social institution is shaped slowly over time by changes in the statistical patterns of a community of speakers, but the speaker continues to develop new linguistic skills throughout life. We now look closer at each of these systems.

Language as a Social Institution

The first level of complex structure is in the **language as a system of shared speech patterns produced by speakers in a community**. When coming into the world, the child begins to have exposure to a corpus of the language of his community in various social contexts, first from the parents and later from siblings, peers, teachers and pop singers. The system of regularities in the corpus of language a child is exposed to has been self-organized by the speech behavior of the community of speakers of the language over many generations. The result is the existence of patterns that seem to underlie most word shapes, including certain "speech sound" types (typical consonant and vowels of the language), syllable-structure patterns, lexical items and "grammatical patterns." The child learns to perceive and to produce adequate versions of this system of composite patterns as he becomes more skilled at listening to and producing speech [31,55].

There is plenty of evidence of discreteness in these phonological categories. For example, the vowels in the series *bead*, *bid*, *bade*, *bed* seem to be the same as the vowels in *mean*, *min*, *Maine*, *men* and in *peal*, *pill*, *pail*, *Pell* (at least in many English dialects). These similar vowel contrasts suggest a set of categories of words in English (e.g., the category that includes *bead*, *mean* and *peal* versus a different category that includes *bed*, *men* and *Pell*). But the claim that these are best understood psychologically as cat-

egories and not as symbol types implies greatly weakening the predictions that can be made about them. There are many ways to be assigned to a mere category, some rule-based and easily brought to awareness and others quite arbitrary and explicable only historically. Linguistic categories (such as /i/ or /t/, and what pronunciations we recognize as productions of the word *and*, etc.) are just aspects of the culture of a community. We cannot expect that *bead*, *mean* and *peal* will be identical in respect to any particular acoustic or auditory property. Speakers do not necessarily represent these three words as sharing any specific feature. They share only the category that we represent with the IPA symbol /i/. We English speakers (or probably primarily educated English speakers) think of them as "sharing the same vowel." It may be alright to speak this way, but it is important to keep in mind that they are not represented by or spelled in memory with any specific symbol token. The same analysis applies to the variants of English /t/ as found in *teach* [t^h], *stop* [t], *butter* [ɾ] and *cat* [t^ʔ]. They are considered by all of us to be /t/s at the same time that we know they are not the same but differ greatly. Ultimately, the main reason to call them all varieties of /t/ is that orthographically we conventionally spell them all with a *t* (and, of course, they were once pronounced with more similar *t*-like sounds a few centuries ago).

Small changes in statistical distributions by many individuals gradually have the effect of facilitating a continual self-organization process in the community. For reasons that are not yet clear, this process tends to lead eventually to such typical phonological patterns as (a) the use of a similar vowel inventory in different consonantal contexts, (b) the symmetrical patterns of voiced and voiceless stop and nasal consonants (e.g., [b d g, p t k, m n ŋ]), that appear in many languages, (c) the constraints or preferences on the form of syllables in each language, and so on.

Although languages have been known for almost two centuries now to change gradually over time, research has yet to provide a convincing explanation of how or why these changes occur. Since it is very difficult to study such slow changes in real human communities, various attempts have been made to simulate socially defined language-like systems in computational systems employing various kinds of linguistic agents [5,29,50,51]. This work, much of it based on the techniques of artificial life, seems very likely to contribute to our understanding of the process of language evolution.

Realtime Language Processing

The second level of complex system that is central to human speech is the **psychological system for speech pro-**

duction and perception – the neural control system that manages an individual's physiological speech system, the lips, tongue, velum, larynx and respiratory system plus the cognitive mechanisms that accomplish speech perception and understanding in realtime. The individual speaker acquires these skills in developmental time, that is, within the speaker's lifetime. However, the units of the social institution of language are not very useful for representation in memory or linguistic processing of speakers. Speech perceivers apparently rely on detailed episodic representations – much more like an auditory description of a spectrotemporal pattern than like something spelled using a small, abstract alphabet.

It is now clear that dynamical systems play a central role in speech production and perception skills [54]. The individual speaker-hearer must discover ways to coordinate all their muscular systems for the production of speech and also somehow to simulate the auditory-phonetic patterns heard in the speech of others in a real-time perceptual system. These skills require several years to achieve normal competence at basic speaking and listening. The more successful scientific efforts to model the real-time processing of speech production and speech perception have focused in recent years on dynamical system models, most often implemented as “neural networks” [15,16,17]. These models simulate the realtime processing of language but do not rely much on the letter and word based description.

Speech Production Modeling

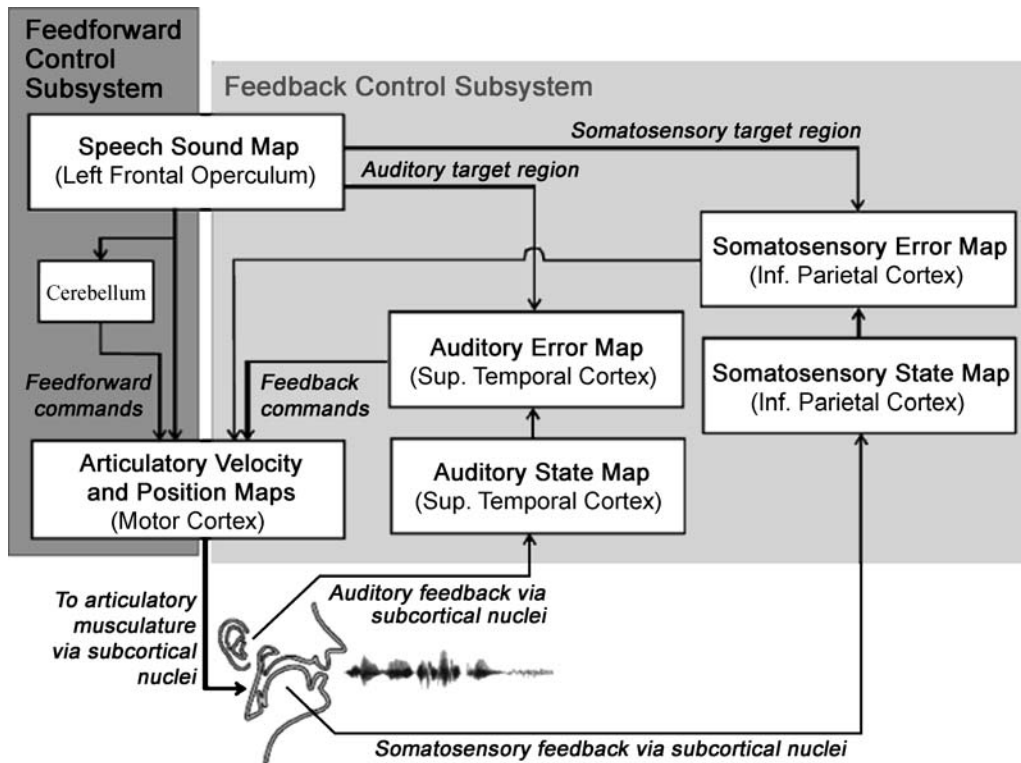
There has been a great deal of research and modeling effort in the study of speech production. The challenge here is that speech production is a formidable motor problem that seems to tax the limits of human motor capabilities. It demands impressive feats of coordination and skilled control of a great many articulatory variables. There are many domains for modeling aspects of human speech performance. But, in order to give some feeling for the kind of dynamical models that appear to be successful, the DIVA model of Frank Guenther will be sketched here.

This model addresses both the problem of skilled speech production as it occurs in realtime, as well as the problem of how such skills could be acquired by the child language learner. In addition, this is one of the first models of speech production for which specific neural sites can be proposed for each functional component of the model (although the neurophysiology will not be discussed here). The model gains inspiration from much previous work in neuroscience demonstrating, for example, that the brain contains many ‘maps’ of cells arranged in two or three-di-

mensional arrays that are topographically arranged, meaning that specific regions of the map are devoted to specific content (or actions) and that neighboring regions are devoted to similar content (or action). Thus, in vision, neighboring regions of visual cortex typically correspond to neighboring regions of the visual field while in audition, neighboring regions of auditory cortex correspond to similar frequencies of sound. Typically some other features (such as frequency rise vs. fall will be represented by the axis perpendicular to the frequency axis).

In the DIVA model, we can begin the overview with the Speech Sound Map (SSM, near the upper left region of Fig. 3) of an adult speaker, where each speech sound type (which might be a specific gesture, syllable or even phrase) is postulated to have a small area whose activation (and its inhibition of competing regions/gestures) sets off the pattern of neural events that give rise to the skilled production of the sound type. There are three output mappings from the Speech Sound Map. Going downward in the figure is the Feedforward Control Subsystem. Neural paths from the SSM to the motor cortex produce the gesture sequence that the speaker learned for producing this sound type. In addition, there are 2 sets of sensory expectations generated: what the sound of this speech gesture should be like in auditory terms and what sensory patterns in the vocal tract should be like. Thus, there is a mapping from the SSM to a somatosensory target region where a pattern of expectations is generated for a specific temporal sequence of sensations from the vocal tract (lips, tongue, palate, glottis, etc.) as the gesture is produced. (These expectations were learned during an earlier babbling stage and from previous speech production experience.) A similar set of predictions is produced about the expected sequence of auditory patterns. Then as the speech gestures are carried out (due to the Feedforward Control Subsystem), they produced actual auditory and somatosensory patterns which are compared to the expectations in the Auditory and Somatosensory Error Maps. If the inhibition from the realtime sensory feedback matches the pattern of excitations from the SSM, then the error maps register zero. If there are deviations (because something expected did not occur or if something unexpected occurs in either stream), then the error is detected and some form of motor correction is applied.

During the acquisition of speech skills, beginning with random babbling, there are many errors and the speaker learns gradually to produce better versions of feedforward gestural control, but during normal, skilled speech, there will be very few errors or deviations. It can be seen that this system gathers whatever rich sensory information it can, both for audition and for proprioception and pre-



Dynamics of Language, Figure 3

Schematic diagram of the DIVA model of speech production (from [18])

dicts the sensory patterning in continuous time. As long as predictions are fulfilled, all proceeds as an approximately ‘open loop’ system. But clearly speech production is not simply a matter of issuing commands. There is much real-time control keeping the gesture on track at every instant.

Speech Perception Models

For audition of speech, the primary problem, rather than the control of many motor variables, it is recognition of the vast number of variants that can arise in ordinary speech. Audition skills begin early since hearing is possible in the womb. Infants at birth are already familiar with much about the auditory environment their mother lives in and even with her voice [26]. Infants gradually learn the auditory patterns of their language and the many variants that speech exhibits. Increasing perception skills lead to recognition of “chunks” of speech that are statistically predominant [16] so speech perception results in a kind of parsing of the speech signal into recognizable chunks of variable size.

Literacy has a dramatic influence on us. Somewhere between about 12 months and 5 years, children in many communities are taught the names of the letters and pro-

ceed over the next few years (often for many years) to become more and more literate and familiar with the conventions of the written language styles of the culture as well. But the written orthography has much that is unknown in the spoken language. The written language presumes discrete words, separated by spaces, enforces fixed conventional spellings, imposes sentence structure on texts, marks proper names and paragraphs, and can even support tabular layouts of text [20]. Literate people become very skilled at interpreting letter strings as speech and interpreting speech as letter sequences. These skills are the reason for our powerful intuitions that speech comes automatically and uniformly in letter-like form. But the description of language offered by alphabetical orthographies, despite its enormous influence on how linguists conceptualize their domain, plays almost no role whatever beyond its role in our conscious experience of language.

Future Directions

The discipline of linguistics has been severely hampered by the inability of linguists to escape from their intuitive understanding of language, an understanding that has been shaped by the literacy training of linguists. Because of our

lifelong experience using alphabetical representations of language, it has been nearly impossible to look beyond our intuitions to see what the experimental data show. The conceptualization of the problem of language presented here says that a language is a social product and not a cognitive one. Speakers cobble together methods that are suitable for talking and hearing within the conventions and customs of their linguistic culture, but this does not imply (as we thought earlier) that the apparent units of language (the sound types, lexical entries, etc.) are themselves cognitive units used and manipulated by speakers.

Simulation work on the social aspects of language, i.e., how a language comes into being in a community and how it evolves over time, is well under way. And the study of “sociolinguistics”, studying language variation in its social context has recently become important. Similarly, there has been much work done on speech production and perception. But the view that intuitive linguistic units, like phones, phonemes, syllables, morphemes, words, etc., must comprise stages in the realtime processing of language has proven to be a red herring that has led us astray for at least a century. It is time to let this go and to focus on the two complex systems, one communal and one personal, that underlie human speech.

Bibliography

Primary Literature

1. Abraham R, Shaw C (1983) *Dynamics: The Geometry of Behavior*, Part 1. Aerial Press, Santa Cruz
2. Allen G (1972) The location of rhythmic stress beats in English: An experimental study I. *Lang Speech* 15:72–100
3. Bloomfield L (1933) *Language*. Holt Reinhart Winston, New York
4. Browman C, Goldstein L (1992) Articulatory phonology: An overview. *Phonetica* 49:155–180
5. Cangelosi A, Parisi D (2002) *Simulating the Evolution of Language*. Springer, New York
6. Cho T, Ladefoged P (1999) Variations and universals in VOT: Evidence from 18 languages. *J Phonetics* 27:207–229
7. Chomsky N (1965) *Aspects of the Theory of Syntax*. MIT Press, Cambridge
8. Chomsky N, Halle M (1968) *The Sound Pattern of English*. Harper and Row, New York
9. Clark A (1997) *Being There: Putting Brain, Body, and World Together Again*. Bradford Books/MIT Press, Cambridge
10. Clark A (2006) Language, embodiment and the cognitive niche. *Trends Cogn Sci* 10:370–374
11. Cummins F (2003) Rhythmic grouping in word lists: Competing roles of syllables, words and stress feet. In: *Proceedings of the 15th International Conference on Spoken Language Processing*, Barcelona, 2003, pp 325–328
12. Cummins F, Port R (1998) Rhythmic constraints on stress timing in English. *J Phonetics* 26:145–171
13. de Saussure F (1916) *Course in General Linguistics*. Philosophical Library, New York
14. Goldinger SD (1996) Words and voices: Episodic traces in spoken word identification and recognition memory. *J Exp Psychol: Learn Mem Cogn* 22:1166–1183
15. Grossberg S (1995) The neurodynamics of motion perception, recognition learning and spatial attention. In: Port R, v Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge
16. Grossberg S (2003) The resonant dynamics of speech perception. *J Phonetics* 31:423–445
17. Guenther F (1995) Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production. *Psychol Rev* 102:594–621
18. Guenther F, Perkell J (2004) A neural model of speech production and supporting experiments. In: *From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*. Cambridge, Massachusetts
19. Harris R (1981) *The Language Myth*. Duckworth, London
20. Harris R (2000) *Rethinking Writing*. Continuum, London
21. Hockett C (1968) *The State of the Art*. Mouton, The Hague
22. Huckvale M (1997) 10 things engineers have discovered about speech recognition. In: *NATO ASI Speech Patterning Conference* Jersey, 1997
23. IPA (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge
24. Jakobson R, Fant G, Halle M (1952) *Preliminaries to Speech Analysis: The Distinctive Features*. MIT, Cambridge
25. Jelinek F (1988) Applying information theoretic methods: Evaluation of grammar quality. In: *Workshop on Evaluation of Natural Language Processing Systems*, 7–9 Dec 1988. Wayne
26. Jusczyk P (1997) *The Discovery of Spoken Language*. MIT Press, Cambridge
27. Kelso JAS (1995) *Dynamic Patterns: Self-organization of Brain and Behavior*. MIT Press, Cambridge
28. Kewley-Port D (1983) Time-varying features as correlates of place of articulation in stop consonants. *J Acoust Soc Am* 73:322–335
29. Kirby S (2001) Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE J Evol Comput* 5:102–110
30. Kochanski G, Orphanidou C (2008) What marks the beat of speech? *J Acoust Soc Am* 123:2780–2791;
31. Kuhl P, Iverson P (1995) Linguistic experience and the perceptual magnet effect. In: Strange W (ed) *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York Press, Timonium, pp 121–154
32. Ladefoged P, Maddieson I (1996) *Sounds of the World's Languages*. Blackwell, Oxford
33. Large E, Jones M (1999) The dynamics of attending: How we track time-varying events. *Psychol Rev* 106:119–159
34. Liberman AM, Harris KS, Hoffman H, Griffith B (1957) The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol* 54:358–368
35. Liberman AM, Delattre P, Gerstman L, Cooper F (1968) Perception of the speech code. *Psychol Rev* 74:431–461
36. Lisker L, Abramson A (1964) A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20:384–422
37. Love N (2004) Cognition and the language myth. *Lang Sci* 26:525–544
38. Mayville J, Jantzen K, Fuchs A, Steinberg F, Kelso JS (2002) Cor-

tical and subcortical networks underlying syncopated an synchronized coordination reveals using fMRI. *Hum Brain Mapp* 17:214–229

39. Palmeri TJ, Goldinger SD, Pisoni DB (1993) Episodic encoding of voice attributes and recognition memory for spoken words. *J Exp Psychol, Learn Mem Cogn* 19:309–328
40. Patel A (2003) Language, music, syntax and the brain. *Nat Neurosci* 6:674–681
41. Patel A, Iverson J, Chen Y, Repp B (2005) The influence of metricality and modality on synchronization with a beat. *Exp Brain Res* 163:226–238
42. Pisoni D, Levi S (2006) Some observations on representation and representational specificity in spoken word processing. In: Gaskell G (ed) *Oxford Encyclopedia of Psycholinguistics*. Oxford University Press, Oxford
43. Pisoni DB (1997) Some thoughts on ‘normalization’ in speech perception, in Talker variability. In: Johnson K, Mullennix J (eds) *Speech processing*. Academic Press, San Diego, pp 9–32
44. Port R (2003) Meter and speech. *J Phonetics* 31:599–611
45. Port R (2007) What are words made of?: Beyond phones and phonemes. *New Ideas in Psychology* 25:143–170
46. Port R (2008) All is prosody: Phones and phonemes are the ghosts of letters. *Prosody* 2008, Campinas
47. Port RF, Leary A (2005) Against formal phonology. *Language* 81:927–964
48. Rayner K, Foorman B, Perfetti C, Pesetsky D, Seidenberg M (2001) How psychological science informs the teaching of reading. *Psychol Sci Public Interest* 2:31–74
49. Scott S (1998) The point of P-centres. *Psychol Res* 61:4–11
50. Smith K, Brighton H, Kirby S (2003) Complex systems in language evolution: The cultural emergence of compositional structure. *Adv Complex Syst* 6:537–558
51. Steels L (2006) Experiments on the emergence of human communication. *Trends Cogn Sci* 10:347–434
52. Stevens K, Blumstein S (1978) Invariant cues for place of articulation in stop consonants. *J Acoust Soc Am* 64:1358–1368
53. Twaddell WF (1935) On defining the phoneme. *Language*, vol *Language Monograph* 16
54. van Gelder T, Port R (1995) Its about time. In: Port R, v Gelder T (eds) *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, pp 1–44
55. Werker J, Tees RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav Dev* 7:49–63
56. Zanto T, Snyder J, Large E (2006) Neural correlates of rhythmic expectancy. *Adv Cogn Psychol* 2:221–231
57. Zawaydeh B, Tajima K, Kitahara M (2002) Discovering Arabic rhythm through a speech cycling task. In: Parkinson D, Benmamoun E (eds) *Perspectives on Arabic Linguistics*. John Benjamins, pp 39–58

Books and Reviews

- Abler W (1989) On the particulate principle of self-diversifying systems. *J Soc Biol Struct* 12:1–13
- Guenther F, Ghosh S, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang* 96:280–301
- Guenther FH, Gjaja M (1996) The perceptual magnet effect as an emergent property of neural map formation. *J Acoust Soc Am* 100:1111–1121

Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transitions in human hand movements. *Biol Cybern* 51:347–356

Patel A, Lofquist A, Naito W (1999) The acoustics and kinematics of regularly timed speech: A database and method for the study of the P-center problem. In: 14th International Congress of Phonetic Sciences, San Francisco, 1–7 Aug

Smith E, Medin D (1981) *Categories and Concepts*. Harvard University Press, Cambridge

Tajima K, Port R (2003) Speech rhythm in English and Japanese. In: Local J, Ogden R, Temple R (eds) *Phonetic Interpretation: Papers in Laboratory Phonology*, vol 6. Cambridge University Press, Cambridge, pp 317–334

Dynamics of Parametric Excitation

ALAN CHAMPNEYS

Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Linear Resonance or Nonlinear Instability?](#)

[Multibody Systems](#)

[Continuous Systems](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Parametric excitation Explicit time-dependent variation of a parameter of a dynamical system.

Parametric resonance An instability that is caused by a rational relationship between the frequency of parametric excitation and the natural frequency of free oscillation in the absence of the excitation. If ω is the excitation frequency, and ω_0 the natural frequency, then parametric resonance occurs when $\omega = (n/2)\omega_0$ for any positive integer n . The case $n = 1$ is usually the most prominent form of parametric resonance, and is sometimes called the principle subharmonic resonance.

Autoparametric resonance A virtual parametric resonance that occurs due to the coupling between two independent degrees of freedom within a system. The output of one degree of freedom acts like the parametric excitation of the other.

Ince–Strutt diagram A two-parameter bifurcation diagram indicating stable and unstable regions, specifically plotting the instability boundaries as the required

amplitude of parametric excitation against the square of the ratio of natural to excitation frequency.

Floquet theory The determination of the eigenvalue spectrum that governs the stability of periodic solutions to systems of ordinary differential equations.

Bifurcation A qualitative change in a system's dynamics as a parameter is varied. One-parameter bifurcation diagrams often depict invariant sets of the dynamics against a single parameter, indicating stability and any bifurcation points. Two-parameter bifurcation diagrams depict curves in a parameter plane on which one-parameter bifurcations occur.

Modal analysis The study of continuum models by decomposing their spatial parts into eigenmodes of the dynamic operator. The projection of the full system onto each mode gives an infinite system of differential equations in time, one for each mode.

Monodromy matrix The matrix used in Floquet theory, whose eigenvalues (also known as Floquet multipliers) determine stability of a periodic solution.

Definition of the Subject

Parametric excitation of a system differs from direct forcing in that fluctuations appear as temporal modulation of a parameter rather than as a direct additive term. A common paradigm is that of a pendulum hanging under gravity whose support is subjected to a vertical sinusoidal displacement. In the absence of any dissipative effects, instabilities occur to the trivial equilibrium whenever the natural frequency is a multiple of *half* the excitation frequency. At amplitude levels beyond the instability, further *bifurcations* (dynamical transitions) can lead to more complex quasi-periodic or chaotic dynamics. In multibody mechanical systems, one mode of vibration can effectively act as the parametric excitation of another mode through the presence of multiplicative nonlinearity. Thus *autoparametric* resonance occurs when one mode's frequency is a multiple of half of the other. Other effects include combination resonance, where the excitation is at a sum or difference of two modal frequencies. Parametric excitation is important in continuous systems too and can lead to counterintuitive effects such as stabilization “upside-down” of slender structures, complex wave patterns on a fluid free surface, stable pulses of light that overcome optical loss, and fluid-induced motion from parallel rather than transverse flow.

Introduction

Have you observed a child on a playground swing? In either a standing or a sitting position, generations of chil-

dren have learned how to “pump” with their legs to set the swing in motion without making contact with the ground, and without external forcing. Note that pumping occurs at *twice* the natural frequency of swing of the pendulum, since the child extends her legs maximally at the two extremities of the swing's cycle, that is, two times per period. What is happening here? The simplest case to analyze is where the child is standing. Then, it has been argued by Curry [17] that the child plus swing is effectively a simple pendulum system, where the child's pumping has the effect of periodic variation of the position of the center of the mass along the arm of the pendulum (see Fig. 1a,b). We shall return to the self-propelled swing problem in Sect. “Future Directions” below, where we shall find out that all is not what it seems with this motivating example, but for the time being let's make the simple hypothesis that the child's pumping causes the center of mass to move up and down.

Using the standard model of a pendulum that consists of a lumped mass suspended by a rigid massless rod of length l from a pivot, the equations of motion for the angular displacement $\theta(t)$ of a pendulum hanging under gravity can be written in the form

$$\ddot{\theta}(t) + \frac{g}{l} \sin[\theta(t)] = 0.$$

Here g is the acceleration due to gravity, a dot denotes differentiation with respect to time t and, for the time being, we have ignored any damping that must inevitably be present. Now, the moving center of mass means that the effective length of the rigid rod oscillates, and let us suppose for simplicity that this oscillation is sinusoidal about some mean length l_0 :

$$l(t) = l_0[1 - \varepsilon \cos(\omega t)].$$

Here $\varepsilon > 0$ assumed to be small and as yet we haven't specified the frequency ω . Then, upon Taylor expansion in ε , we find to leading order

$$\ddot{\theta} + \omega_0^2[1 + \varepsilon \cos(\omega t)] \sin \theta = 0,$$

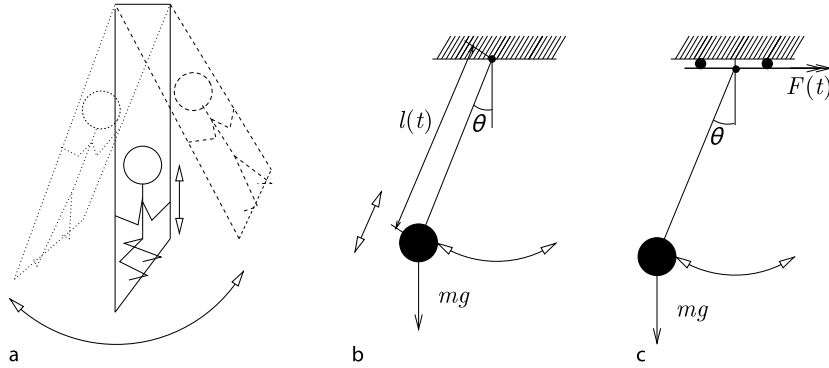
where $\omega_0 = \sqrt{g/l_0}$ is the natural frequency of small amplitude swings in the absence of the pump. Finally, upon rescaling time $\tilde{t} = \omega t$ and dropping the tildes we find

$$\ddot{\theta} + [\alpha + \beta \cos t] \sin \theta = 0, \quad (1)$$

where

$$\alpha = \frac{\omega_0^2}{\omega^2} \quad \text{and} \quad \beta = \varepsilon \frac{\omega_0^2}{\omega^2} \quad (2)$$

are dimensionless parameters that respectively describe



Dynamics of Parametric Excitation, Figure 1

a A child standing on a swing modeled as **b** a parametrically excited pendulum **c** a directly forced pendulum

the square of the ratio of natural frequency to excitation frequency, and the amplitude of the excitation.

Equation (1) is a canonical example of a *parametrically excited* system. Note that (1) also describes, at least to leading order, the case of a pendulum hanging under gravity that is excited by a vertical force of size $\varepsilon\omega_0^2 \cos t$ which we can think of as a periodic modulation of the gravity. This would be in contrast to a *directly forced* simple pendulum where the forcing would act in the direction of swing for small amplitude motion; see Fig. 1c. For the forced pendulum, the equivalent equations of motion would look more like

$$\ddot{\theta} + \alpha \sin \theta = \beta \cos t, \quad (3)$$

where the periodic forcing occurs as an additive input into the differential equation. In contrast, for (1) the periodic excitation occurs as a modulation of a *parameter* (which you can think of as gravity), hence the name ‘parametric’. Another feature of (1) that makes it a good paradigm for this article is that it is nonlinear and, as we shall see, a full appreciation of the dynamics resulting from parametric excitation requires treatment of nonlinear terms.

An interesting contrast between parametrically excited and directly forced systems is in the nature of the trivial response to a non-zero input. In particular, for any α and β , Eq. (1) admits the trivial solution $\theta(t) \equiv 0$, whereas there is no such solution to (3) for non-zero β . The simplest “steady” solution to the directly forced system for small β is a small-amplitude periodic oscillation of period 2π (in our dimensionless time units). So any analysis of a parametrically excited system should start with an analysis of the steady state $\theta = 0$, which corresponds to the pendulum just hanging under gravity. To analyze stability of this equilibrium state let $\theta = 0 + x$ where x is small, then x satisfies linearization (1)

$$\ddot{x} + (\alpha + \beta \cos t)x = 0. \quad (4)$$

The equilibrium position is stable if all solutions $x(t)$ of this equation, remain bounded as $t \rightarrow \infty$ and it is unstable if there are solutions $x(t) \rightarrow \infty$ as $t \rightarrow \infty$.

Equation (4) is known as the *Mathieu equation* [42], and is a specific example of Hill’s equation [29,40]

$$\ddot{x} + [\alpha + V(t)]x = 0, \quad (5)$$

where $V(t)$ is any periodic function with period 2π . Hill’s equation can be solved using classical methods for linear time-dependent ordinary differential equations (ODEs), see e.g. [31,33]. In particular, by writing $y_1 = x$ and $y_2 = \dot{x}$, Hill’s equation (5) is just a two-dimensional example of a linear first-order time-periodic equation

$$\dot{\mathbf{y}} = P(t)\mathbf{y}, \quad \mathbf{y} \in \mathbb{R}^n. \quad (6)$$

The general solutions to (6) for any initial condition can be expressed in terms of a fundamental matrix $\Phi(t)$, with $\Phi(0) = I_n$ (the identity matrix of dimension n):

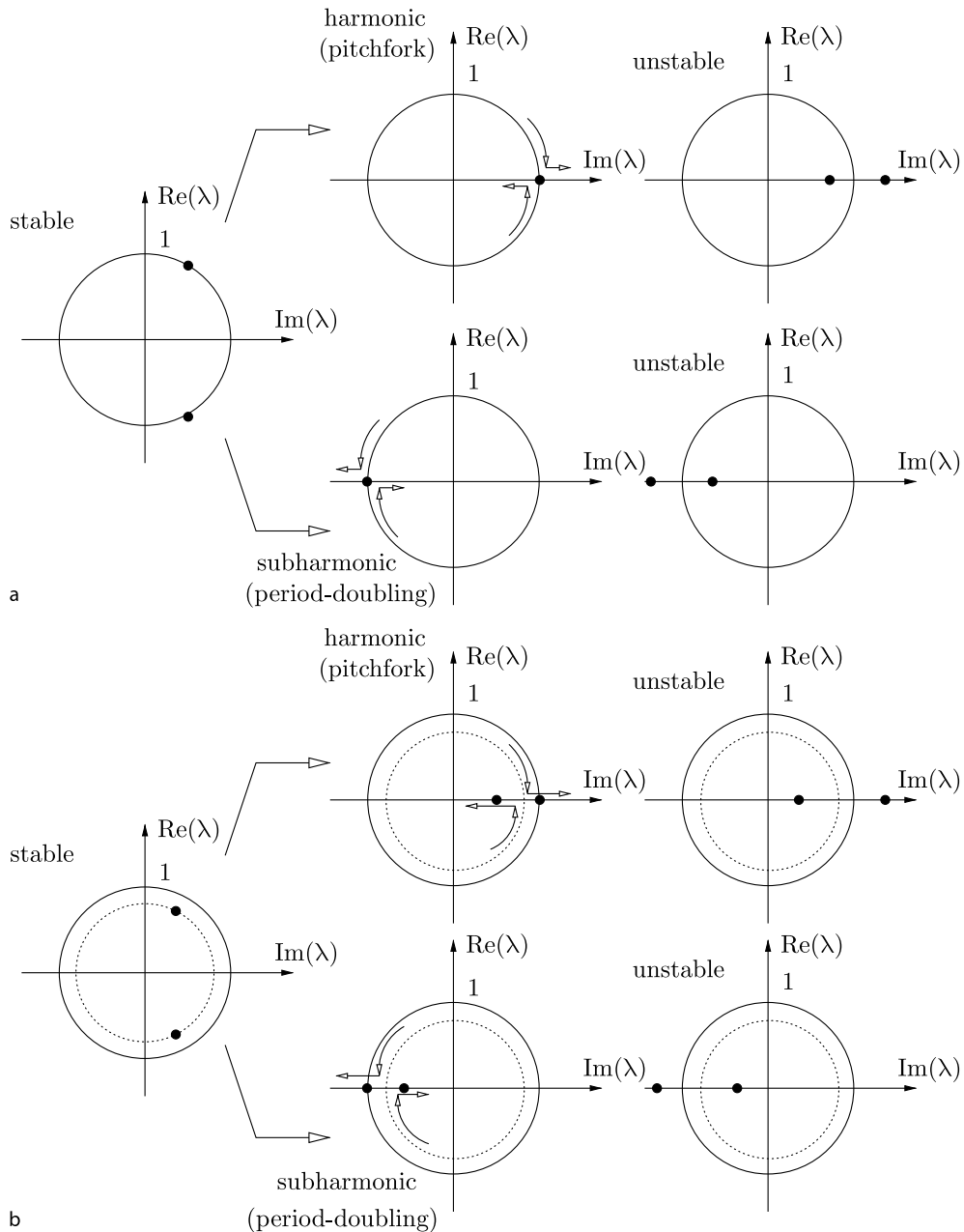
$$\mathbf{y}(t) = \Phi(t)\mathbf{y}(0).$$

Note that just because $P(t)$ has period 2π it does not follow that $\Phi(t)$ does. Also $\Phi(t)$ cannot be computed exactly, except in highly specialized cases, and one has to rely on approximate methods as described in the next Sect. “[Floquet Analysis](#)” and in the companion article to this one [► Perturbation Analysis of Parametric Resonance](#). However, in order to study stability we don’t actually need to construct the full solution $\Phi(t)$, it is sufficient to consider $M = \Phi(2\pi)$, which is called the *monodromy matrix* associated with Hill’s equation. Stability is determined by studying the eigenvalues of M : solutions to (5) with small initial conditions remain small for all time if all eigenvalues of M lie on or within the unit circle in the complex plane; whereas a general initial condition will lead to un-

bounded motion if there are eigenvalues outside this circle. See Fig. 2a. Eigenvalues of the Monodromy matrix are also known as *Floquet multipliers* of the system, and the general study of the stability of periodic systems is called *Floquet theory*.

In the case of the Mathieu equation, where $V(t) = \beta \cos(t)$, we have

$$P(t) = \begin{bmatrix} 0 & 1 \\ 0 & \alpha + \beta \cos(t) \end{bmatrix}. \quad (7)$$



Dynamics of Parametric Excitation, Figure 2

Possible behavior of Floquet multipliers λ (eigenvalues of the monodromy matrix M) for a conservative Hill equation and **b** in the presence of small damping

Figure 3 depicts graphically the behavior of Floquet multipliers for this case. Here, in what is known as an *Ince–Strutt diagram* [31,63], the shaded regions represent the values of parameters at which the origin $x = 0$ is stable (with Floquet multipliers on the unit circle) and the white regions are where the origin is unstable (at least one multiplier outside the unit circle). We describe how to calculate such a diagram in the next section. For the time being we note that there are *resonance tongues* that emanate from $\beta = 0$ whenever α is the square of a half-integer: $\alpha = 1/4, 1, 9/4, 4, 25/4$, etc. The width (in α) of the tongue for small β turns out to be proportional to β to the power of twice the half-integer that is being squared. That is, the $\alpha = 1/4$ tongue has width $O(\beta)$, the $\alpha = 1$ tongue has width $O(\beta)^2$, the $\alpha = 9/4$ tongue has width $O(\beta^3)$ etc.

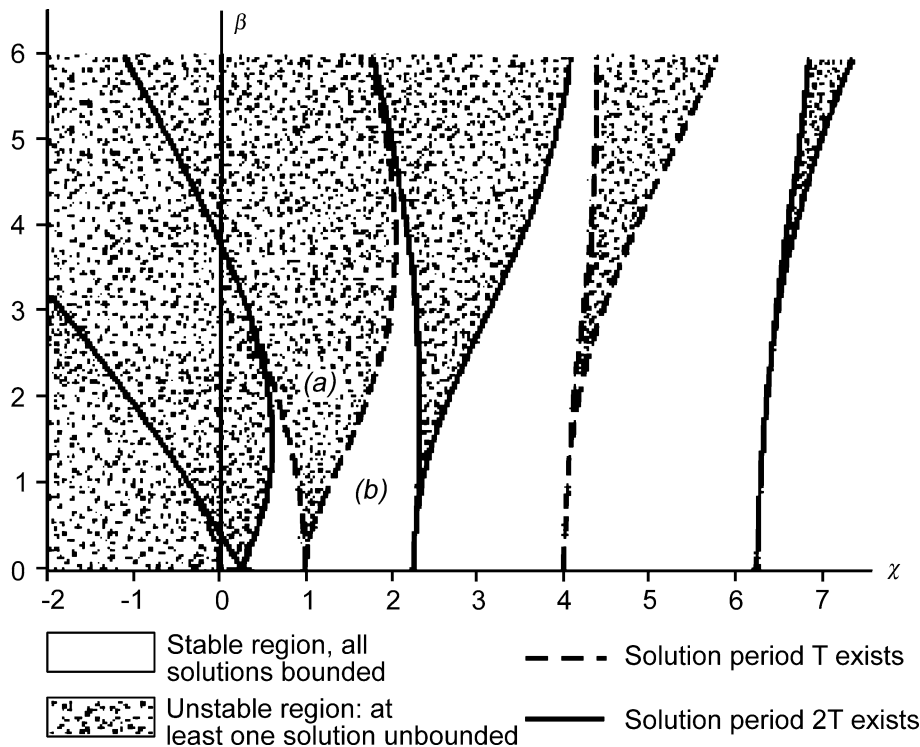
On the tongue boundaries (between the shaded and the unshaded regions of Fig. 3), the Mathieu equation admits non-trivial periodic solutions. For the tongues that are rooted at square integers, $\alpha = 1, 9, 16$, etc., these solutions have period 2π (the *same* period as that of the excitation) and the instability that occurs upon crossing such a tongue boundary is referred to as a *harmonic instability*.

On the other boundaries, those of tongues rooted at $\alpha = 1/4, 9/4, 25/4$ etc., the periodic solution has period 4π (*twice* that of the excitation). Hence such instabilities are characterized by frequencies that are *half* that of the excitation and are hence referred to as *subharmonic instabilities*.

So far, we have ignored any dissipative effects, so that Eq. (5) can be regarded as a conservative or *Hamiltonian* system. Damping can be introduced via a dissipative force that is proportional to velocity, hence (4) becomes

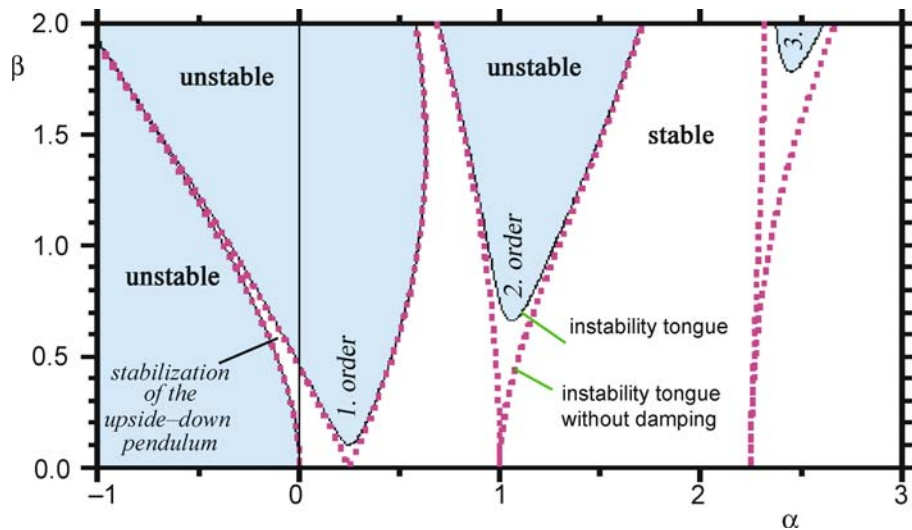
$$\ddot{x} + \delta \dot{x} + (\alpha + \beta \cos t)x = 0, \quad (8)$$

where δ is a dimensionless damping coefficient. One can also apply Floquet theory to this damped Mathieu equation; see Fig. 2b. The results for Eq. (8) are plotted in Fig. 4, where we can see the intuitive effect that damping increases the areas of stability. In particular, in comparison with Fig. 3, note that the resonance tongues have been lifted from the α -axis, with the amount of the raise being inversely proportional to the width of the undamped tongue. Thus, the resonance tongue corresponding to



Dynamics of Parametric Excitation, Figure 3

Ince–Strutt diagram showing resonance tongues of the Mathieu equation (4). Regions of instability of the trivial solution $x = 0$ are shaded and a distinction is drawn between instability curves corresponding to a period- T ($T = 2\pi$) and to a period- $2T$ orbit. After (2nd edn. in [33]), reproduced with permission from Oxford University Press



Dynamics of Parametric Excitation, Figure 4

Similar to Fig. 3 but for the damped Mathieu equation (8) with $\delta = 0.1$. After <http://monet.unibas.ch/~elmer/pendulum/parres.htm>, reproduced with permission

$\alpha = 1/4$ is easily the most prominent. Not only for fixed forcing amplitude β does this tongue occupy for the greatest interval of square frequency ratio α , but it is accessible with the least forcing amplitude β . Thus, practicing engineers often think of this particular *subharmonic instability* close to $\alpha = 1/4$ (i. e. where the forcing frequency is approximately half the natural frequency) as being *the* hallmark of parametric resonance. Such an instability is sometimes called the *principal* parametric resonance. Nevertheless, we shall see that other instability tongues, in particular the *fundamental* or *first harmonic* parametric resonance near to $\alpha = 1$ can also sometimes be of significance.

Returning briefly to the problem of how children swing, it would seem that the choice $\alpha \approx 1/4$ is in some sense preferable. This is the largest instability tongue for small β . That is, the smallest effort is required in order to cause an exponential instability of the downwards hanging solution, and set the swing in motion. And, according to (2), $\alpha = 1/4$ corresponds to $\omega = 2\omega_0$. Thus generations of children seem to have hit upon an optimal strategy [48], by pumping their legs at precisely twice the frequency of the swing.

Finally, looking at Fig. 4, note the curious feature that we have continued the diagrams into negative α , despite the fact that α was defined as a square term. In the case of a pendulum though, $\alpha = l/(g\omega^2)$, and so we can think of negative α as representing negative g . That is, a pendulum with negative α is defying gravity by ‘hanging upwards’. Clearly without any excitation, such a situation would be violently unstable. The big surprise then is that there is

a thin wedge of stability (which is present either with or without damping) for $\alpha < 0$. Note though that the wedge is thickest for small β and small $|\alpha|$, which limit corresponds to small amplitude, high-frequency oscillation. This remarkable stabilization of an upside-down pendulum by tiny, rapid vertical vibration was first described by Stephenson [53] in 1908, re-discovered by Kapitza [34] in 1951 (and hence often called the Kapitza pendulum problem) and has been experimentally verified many times; see e. g. [3]. We shall return to this upside down stability in Sect. “Upside-Down Stability”.

The rest of this article is outlined as follows. In Sect. “Linear Resonance or Nonlinear Instability?” we shall examine how the Ince–Strutt diagram is constructed and see that this is just an example of the linearized analysis around any periodic state of a nonlinear system. Further quasi-periodic or chaotic motion can ensue as we push beyond this initial instability. Section “Multibody Systems” concerns mechanical systems with multiple degrees of freedom. We consider *autoparametric resonance* where the motion of one degree of freedom acts as the parametric excitation of another. We also look at conditions in multibody systems where a combination of two natural frequencies can be excited parametrically, focusing in particular on the somewhat illusive phenomenon of *difference* combination resonances. Section “Continuous Systems” then looks at parametric excitation in continuous systems, looking at phenomena as diverse as pattern formation on the surface of a liquid, flow-induced oscillation of pipes, a clever way of overcoming loss in opti-

cal fibers and the oscillation-induced stiffening of structures. Finally, Sect. “Future Directions” draws conclusions by looking at future and emerging ideas.

Linear Resonance or Nonlinear Instability?

Let us now consider a more detailed analysis of parametric resonance of the Mathieu equation and of the implications for a nonlinear system, using the parametrically excited pendulum as a canonical example.

Floquet Analysis

We begin by giving a reasoning as to why the Ince–Strutt diagram Fig. 3 for the Mathieu equation (ignoring damping for the time being) looks the way it does, by making a more careful consideration of how Floquet multipliers behave as parameters are varied. Essentially, we explain the behavior of Fig. 2a. Our treatment follows the elementary account in [33], to which the interested reader is referred for more details. For a general linear system of the form (6), a lot of insight can be gained by considering the so-called *Wronskian* $W(t) = \det(\Phi(t))$. From the elementary theory of differential equations (e. g. [33]) we have

$$W(t) = W(t_0) \exp \left(\int_{t_0}^t \text{trace}[P(s)] ds \right), \quad (9)$$

where the *trace* of a matrix is the sum of its diagonal elements. Note that the specific form of $P(t)$ for the Mathieu equation (7) has $\text{trace}[P(t)] = 0$, hence, from (9) that $\det(M) = \det[W(2\pi)] = 1$. But from elementary algebra we know that the trace of the matrix is the product of its eigenvalues. So we know that the two Floquet multipliers $\lambda_{1,2}$ satisfy $\lambda_1 \lambda_2 = 1$ and must therefore solve a characteristic equation of the form

$$\lambda^2 - 2\phi(\alpha, \beta)\lambda + 1 = 0,$$

for some unknown real function ϕ . There are two generic cases $|\phi| < 1$ and $|\phi| > 1$. If $\phi > 1$, then the Floquet multipliers are complex and lie on the unit circle. This corresponds to stability of the zero-solution of the Mathieu equation. Conversely if $|\phi| < 1$, then the Floquet multipliers are real and satisfy $|\lambda_1| < 1$ and $|\lambda_2| = 1/|\lambda_1| > 1$. The larger multiplier represents an exponentially growing solution and hence corresponds to an instability. The two boundary cases are: $\phi = 1$, in which case we have a double Floquet multiplier at $\lambda = 1$; or $\phi = -1$, which corresponds to a double multiplier at $\lambda = -1$. These represent respectively the harmonic and the subharmonic stability boundaries in the Ince–Strutt diagram.

But, how do we compute ϕ ? The answer is that we don't. We simply recognize that $\phi = 1$ corresponds to the existence of a pure 2π -periodic solution of the fundamental matrix $M = \Phi(2\pi)$, and that $\phi = -1$ represents a periodic solution of $\Phi(2\pi)^2 = \Phi(4\pi)$ and hence a 4π -periodic solution of the fundamental matrix. To look for a 2π -periodic solution of the Mathieu equation, it is convenient to use Fourier series. That is, we seek a solution (4) in the form

$$x(t) = \sum_{n=-\infty}^{n=+\infty} c_n e^{int}.$$

Substitution into (4), using $\cos(t) = (1/2)(e^{it} + e^{-it})$ leads to infinitely many equations of the form

$$(1/2)\beta c_{n+1} + (\alpha - n^2)c_n + (1/2)\beta c_{n-1} = 0, \\ n = -\infty, \dots, \infty, \quad \text{with } c_{-n} = c_n^*, \quad (10)$$

where an asterisk represents complex conjugation. Note that the final condition implies that x is real. Clearly, for $\beta = 0$, a nontrivial solution can be found for any integer p which has $c_n = 0$ for all $n \neq p$, provided $\alpha^2 = p^2$. That is, solution curves bifurcate from $\beta = 0$ for each α -value that is the square of a whole integer. Furthermore, nonlinear analysis can be used to show that for small positive β , there are *two* solution branches that emerge from each such point (provided $p \neq 0$); one corresponding to c_p being real, which corresponds to motion $x(t) \sim \cos(pt)$ in phase with the excitation; and another corresponding to c_p imaginary, which corresponds to motion $x(t) \sim \sin(pt)$ out of phase with the excitation. Similarly, if we look for solutions which are 4π -periodic by making the substitution

$$x(t) = \sum_{m=-\infty}^{m=+\infty} d_m e^{imt/2}, \quad (11)$$

we arrive at an infinite system of equations

$$(1/2)\beta d_{m+2} + (\alpha - (1/4)m^2)d_m + (1/2)\beta d_{m-2} = 0, \\ m = -\infty, \dots, \infty, \quad \text{with } d_{-m} = d_m^*. \quad (12)$$

Note that the equations for m even and m odd in (12) are decoupled. The equations for even m become, on setting $m = 2n$, precisely the system (10) with $d_{2m} = c_n$. For $m = 2n + 1$, however, we arrive at new points of bifurcation, by exactly the same logic as we used to analyze (10). Namely whenever $\alpha^2 = (p/2)^2$ for any odd integer p we have the bifurcation from $\beta = 0$ of two 4π -periodic solutions (one like $\cos(pt/2)$ and one like $\sin(pt/2)$).

Perturbation analysis can be used to obtain asymptotic expressions for the transition curves that bifurcate from

$\beta = 0$ at these special α -values; see the companion article ► [Perturbation Analysis of Parametric Resonance](#) for the details. In particular it can be shown that the width of the stability tongue (separation in α -values from the sine and the cosine instability curves) originating at $\alpha = (p/2)^2$ (for p being odd or even) scales like

$$\text{tongue width}_p \sim \beta^p + \text{h.o.t.}$$

Hence the principal subharmonic instability arising from $\alpha = 1/4$ has width $O(\beta)$ (actually the transition curves are given by $\alpha = 1/4 \pm (1/2)\beta + O(\beta^2)$), and the fundamental harmonic instability arising from $\alpha = 1$ has width $O(\beta^2)$, etc. Hence the principal subharmonic instability is by far the most prevalent as it occupies the greatest frequency width for small β , and is often thought to be the hallmark of parametric resonance in applications.

Note that we have not as yet mentioned the transition curve emerging from $\alpha = \beta = 0$, which corresponds to $n = 0$ in the above analysis. This case can also be analyzed asymptotically, and the single transition curve can be shown to scale like $\alpha = -(1/2)\beta^2 + O(\beta^3)$ for small β . It is this bending back into $\alpha < 0$ that leads to the Kapitza pendulum effect, which we shall return to in Sect. “[Up-side-Down Stability](#)” below.

Finally, the addition of small damping can be shown to narrow the tongues and lift them off from the $\beta = 0$ axis as shown in Fig. 4. The amount of lift is in proportion to the width of the tongue. This property further enhances the pre-eminence of the principal subharmonic instability for parametrically excited systems in practice, as they thus require the least amplitude of vibration in order to overcome the damping forces.

Bifurcation Theory

We have just analyzed the Mathieu equation, which is purely linear. But, for example as shown in the Introduction, such an equation is often derived in the course of analyzing the stability of a trivial solution to a nonlinear parametrically excited system. It is worthwhile therefore to consider the nonlinear implications of the harmonic and subharmonic instabilities. The time-periodic Mathieu equation can be written as an example of a three-dimensional autonomous system of ODEs

$$\dot{x} = y, \quad \dot{y} = f(x, y, s), \quad \dot{s} = 1 \pmod{2\pi},$$

for $(x, y, s) \in \mathbb{R}^2 \times S^1$, where $f(x, y, s) = -\delta y - (\alpha + \beta \cos(s))x$. As such, the trivial solution $x = 0$ should actually be seen as the periodic solution $x = y = 0, s = t \pmod{2\pi}$. Hence we can understand instability of the

trivial solution as points of bifurcation of periodic solutions, for which there is a mature theory, see for example [27,37,65,67]. In particular, the subharmonic instability with Floquet multiplier -1 is a *period doubling bifurcation*, and the harmonic bifurcation corresponding to multiplier $+1$ is a *pitchfork* bifurcation. Note that both these bifurcations come in super- and sub-critical forms. Hence, depending on the nature of the nonlinearity we may find either stable bounded motion close to the instability inside the resonant tongue close to the boundary edge, or motion that is not of small amplitude. See Fig. 5.

The case of the Mathieu or Hill equations without damping have received special attention in the literature. This is because of the extra mathematical structure they possess, namely that they can be expressed as *reversible* and *Hamiltonian* systems. Specifically, Hill’s equation of the form (5) can be shown to conserve the energy function

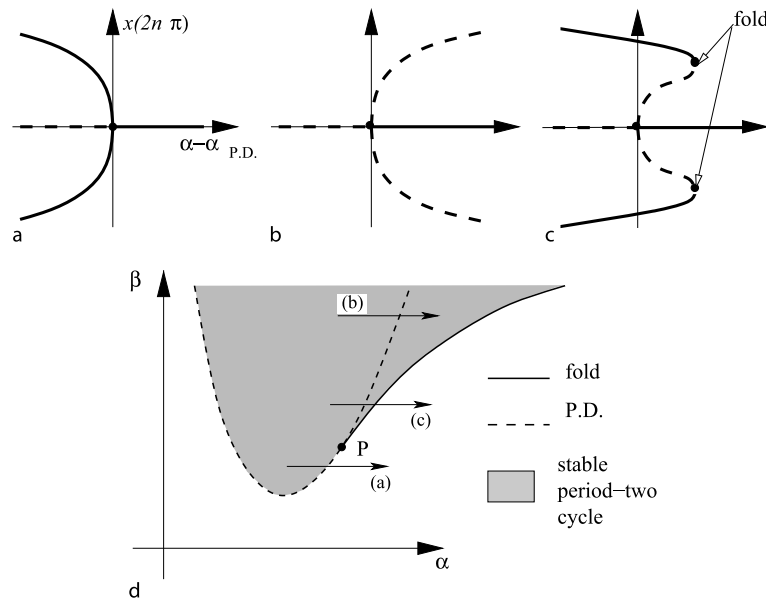
$$H(x, y, t) = (1/2)y^2 + (1/2)(\alpha + V(t))x^2,$$

$y = \dot{x}$, and to be expressible in Hamiltonian form

$$\dot{x} = \frac{\partial H}{\partial y}, \quad \dot{y} = -\frac{\partial H}{\partial x}.$$

Reversibility of Hill’s equation is expressed in the fact that the model is invariant under the transformation $t \rightarrow -t$, $y \rightarrow -y$. Broer and Levi [7] used Hamiltonian and reversible theory to look at Hill’s equations for general periodic functions $V(t)$. They showed that under certain conditions, the resonance tongues may become of zero width for nonzero β . Effectively, the left and right-hand boundaries of the tongue pass through each other. This is a generic phenomenon, in that a small perturbation within the class of Hill’s equations will lead not destroy such an intersection. However, the Mathieu equation for which $V(t) = \cos(t)$ does not possess this property.

Many nonlinear models that undergo parametric resonance also preserve the properties of being Hamiltonian and reversible. Special tools from reversible or Hamiltonian systems theory can sometimes be used to analyze the bifurcations that occur under parametric resonance in nonlinear systems, see for example [8,9]. Points outside the resonance tongues correspond to where the trivial solution is an *elliptic* (weakly stable) fixed point of the time- 2π map, and those inside the resonance tongue to where it is a *hyperbolic* (unstable) fixed point. One therefore does not generally see asymptotically stable non-trivial solutions being born at a resonance tongue boundary, rather one sees elliptic nontrivial dynamics. In fact, for a general periodic solution in a Hamiltonian system, pitchfork and period-doubling bifurcations are not the only kind of bifurcations that produce non-trivial elliptic periodic or-



Dynamics of Parametric Excitation, Figure 5

Contrasting a super-critical and **b** sub-critical bifurcations close to the resonance tongue boundary of a subharmonic (period-doubling) instability, here represented by the abbreviation P.D. *Solid lines* represent stable solutions, *dashed lines* represent unstable solutions. **c** A one-parameter bifurcation diagram close to the super/sub-critical boundary. **d** A two parameter unfolding of the codimension-two point P where the super/sub-critical transition occurs. *Shading* is used to represent where a stable period-two cycle exists. Note in this case that for higher β -values the right-hand boundary of this existence region is therefore the fold curve and not the period-doubling point. Between the period-doubling and the fold there is hysteresis (coexistence) between the stable trivial solution and the stable period-two cycle

bits. If the Floquet multipliers pass through any n th root of unity, there is a *period multiplying* (subharmonic) bifurcation at which a periodic orbit of period $2n\pi$ is born (see, e.g. [62]). A period-doubling is just the case $n = 2$. Such bifurcations can also occur for the non-trivial elliptic periodic orbits within a resonance tongue, but they are all destroyed by the presence of even a small amount of damping.

Beyond the Initial Instability

For a nonlinear system, the pitchfork or period-doubling bifurcation that occurs upon crossing a resonance tongue might just be the first in a sequence of bifurcations that is encountered as a parameter is varied. As a general rule, the more β is increased for fixed α , the more irregular the stable dynamics becomes. For example, the period-doubling bifurcation that is encountered upon increasing β for α close to $1/4$ is typically just the first in a sequence of period-doublings, that accumulate in the creation of a chaotic attractor, obeying the famous Feigenbaum scaling of this route to chaos (see e.g. [18]).

Van Noort [64] produced a comprehensive numerical and analytical study into the nonlinear dynamics of

the parametrically excited pendulum (1) without damping. Of interest in that study was to complete the nonlinear dynamics of one of the simplest non-integrable Hamiltonian and reversible dynamical systems. The parameter β acts like a perturbation to integrability. Certain structures that can be argued to exist in the integrable system, like closed curves of quasiperiodic motion in a Poincaré section around an elliptic fixed point, serve as approximate organizers of the dynamics for non-zero β . However, KAM theory (see e.g. [55]) shows that these closed curves become invariant tori that typically breakup as β is increased. At the heart of the regions of chaotic dynamics lie the resonance tongues associated with a period-multiplying bifurcation. Inside the resonance tongues are non-trivial fixed points, whose surrounding invariant tori break up into heteroclinic chains, which necessarily imply chaos.

Acheson [2] studied the dynamics of the parametrically forced pendulum in the presence of damping (i.e. Eq. (1) with an additional term $\Delta\dot{x}$ on the right-hand side) via computer simulation. He found parameter regions inside the principal resonance tongue of the pendulum where it undergoes what he referred to as *multiple nodding* oscillations. The subharmonic motion inside the tongue should be such that during one cycle of the excita-

tion, the pendulum swings to the left and back, and during the next cycle it swings to the right and back. That is, one complete oscillation per two cycles. Multiple nodding motion by contrast is asymmetric and involves two swings to the left, say, (one swing per cycle of the excitation), followed by one swing to the right, with the motion repeating not every two but every three cycles. He also found similar four-cycle and five-cycle asymmetric motion. This behavior can be explained as pockets of stable behavior that survives from the period-multiplying bifurcations that are only present in the case of zero damping [21].

Clearly, beyond the initial instability the universality of the Ince–Strutt diagram for the Mathieu equation, or for any other Hill equation for that matter, disappears and the precise features of the nonlinear dynamics depends crucially on the details of the particular nonlinear system being investigated.

Multibody Systems

Of course, there is little in the real world that can be accurately described by deterministic, parametrically excited single degree-of-freedom ODEs. In this section we shall consider slightly more realistic situations that are governed by higher-dimensional systems, albeit still deterministic, and still described by finitely many dynamic variables. For simplicity, we shall restrict attention to rigid-body mechanisms that can be described by a system of (finitely many) coupled modes of vibration. Furthermore, we shall adopt a Lagrangian framework where the dynamics of mode i can be represented by a generalized coordinate q_i and its associated generalized velocity \dot{q}_i , such that we arrive at a system of (generally nonlinear) coupled second-order ODEs; one equation for each acceleration \ddot{q}_i .

Autoparametric Resonance

Autoparametric resonance occurs when there is asymmetric nonlinear multiplicative coupling between modes of a multibody system. Specifically, suppose that a mode (q_1, \dot{q}_1) is coupled to a mode (q_2, \dot{q}_2) through a term proportional to $q_1 q_2$ in the equation of motion for the 2nd mode (i. e. the \ddot{q}_2 equation), yet there is no corresponding $q_2 q_1$ term in the \ddot{q}_1 equation. For example, consider

$$\ddot{q}_1 + \delta_1 \dot{q}_1 + \alpha_1 q_1 = f_1(q_1, t) \quad (13)$$

$$\ddot{q}_2 + \delta_2 \dot{q}_2 + \alpha_2 q_2 + \beta q_1 q_2 = f_2(q_2), \quad (14)$$

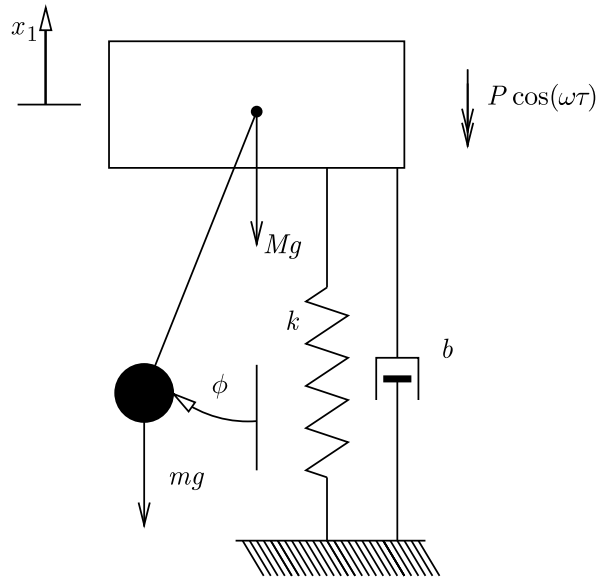
where f_1 and f_2 are nonlinear functions of their arguments and the dependence of f_1 on t is periodic. Furthermore suppose that mode 1 undergoes a periodic motion (either as a result of direct forcing, its own parametric

resonance, or self-excitation). For example, if $f_2 = 0$ and f_1 were simply $\cos(\omega t)$ for some ω^2 sufficiently detuned from α_1 , then the solution $q_1(t)$ to (13) would, after allowing sufficient time for transients to decay, be simply proportional to $\cos(\omega t + \phi)$ for some phase ϕ . Then, the $q_1 q_2$ term in (14) would be effectively act like $\beta \cos(\omega t + \phi) q_2$ and so (14) would be precisely the Mathieu equation. Values of $\omega^2 = 4\alpha_2/n$ for any positive integer n correspond to the root points of the resonance tongues of the Ince–Strutt diagram. Such points are referred to as *internal* or *autoparametric* resonances.

A canonical example of an autoparametrically resonant system is the simple spring-mass-pendulum device depicted in Fig. 6, which we shall use to discuss the some key generic features of autoparametric resonance; see the book by Tondl et al. [58] for more details. In the figure, a mass M is mounted on a linear spring with stiffness k , is forced by an applied displacement $P \cos(\omega \tau)$ and is attached to the pivot of a pendulum of length l and lumped mass m that hangs under gravity. The motion of the mass and the pendulum are assumed to be subject to proportional viscous damping with coefficients b and c respectively. Using standard techniques, the equations of motion can of such a device can be written in the dimensionless form

$$\ddot{x} + \delta_1 \dot{x} + \alpha_1 + \mu(\ddot{\phi} \sin(\phi) + \dot{\phi}^2 \cos(\phi)) = A \cos t \quad (15)$$

$$\ddot{\phi} + \delta_2 \dot{\phi} + \alpha_2 \sin \phi + \ddot{x} \sin \phi = 0, \quad (16)$$



Dynamics of Parametric Excitation, Figure 6
A mass-spring-pendulum system

where $x = y/l$, time t has been scaled so that $t = \omega\tau$, and the dimensionless parameters are

$$\delta_1 = \frac{b}{\omega(M+m)}, \quad \alpha_1 = \frac{k}{\omega^2(M+m)}, \quad \mu = \frac{m}{M+m},$$

$$A = \frac{P}{l\omega^2(M+m)}, \quad \delta_2 = \frac{c}{\omega m l^2}, \quad \alpha_2 = \frac{g}{l\omega^2}.$$

The simplest form of output of such a model is the so-called *semi-trivial* solution in which the pendulum is at rest, $\dot{\phi} = \phi \equiv 0$, and the mass-spring system alone satisfies the simple equation

$$\ddot{x} + \delta_1 \dot{x} + \kappa_1 x = A \cos t.$$

After transients have died out, this single equation has the attracting solution

$$x(t) = R_0 \cos(t + \psi_0),$$

$$\text{where } R_0 = \frac{A}{\sqrt{\delta_1^2 + (\alpha_1 - 1)^2}}$$

$$\text{and } \tan \psi_0 = \frac{\delta_1}{\alpha_1 - 1}. \quad (17)$$

Note that the solution (17) exists for all values of the physical parameters, in particular for all positive α_1 provided $\delta_1 > 0$.

Substitution of the semi-trivial solution (17) into the pendulum Eq. (16), leads to

$$\ddot{\phi} + \delta_2 \dot{\phi} + [\alpha_2 + R_0 \cos(t + \psi_0)] \sin \phi = 0,$$

which, after a time shift $t \rightarrow t - \psi_0$ is precisely the Eq. (1) for a parametrically excited pendulum in the presence of damping. Thus, the Ince–Strutt diagram of the damped Mathieu equation, with $\beta = A/(\sqrt{\delta_1^2 + (\alpha_1 - 1)^2})$, $\delta = \delta_2$ and $\alpha = \alpha_2$ tell us precisely the parameter values for which the semi-trivial solution is stable. Inside the resonance tongues we have an instability of the semi-trivial solution, which leads to stable non-trivial motion with the pendulum swinging. To determine precisely what happens away from the boundaries of the stability tongue, one again has to perform fully nonlinear analysis, as outlined in Sect. “[Linear Resonance or Nonlinear Instability?](#)”. Note now though that for nonzero pendulum motion, $\phi \neq 0$, the Eqs. (15) and (16) become a fully coupled two-degree-of-freedom system, whose dynamics can be even more complex than that of the one-degree-of-freedom parametrically excited pendulum. Indeed in [58], a further destabilizing secondary-Hopf (Neimark–Sacker) bifurcation is identified within the principal ($\alpha_2 = 1/4$)

subharmonic resonance tongue, and numerical evidence is found for several different kinds of quasiperiodic motion.

There is one difference though between autoparametric resonance and pure parametric resonance. The paradigm for parametric resonance is that the external parametric excitation is held constant (its amplitude and frequency are *parameters* of the system). Here, however, for nonzero ϕ , the system is fully coupled, so that the parametric term (17) can no longer be thought of as being independent of ϕ , effectively the excitation becomes a *dynamic variable* itself. Thus, motion of the pendulum can affect the motion of the mass-spring component of the system. In particular if there is an autoparametric instability where the pendulum is set in motion, then energy must be transferred from the mass-spring component to the pendulum, a phenomenon known as *quenching*. But the pendulum is only set in motion close to certain resonant frequency ratios α_2 . Thus, we have a *tuned damper* that is designed to take energy out of the motion of the mass at certain input frequencies ω frequencies.

To see how such a damper might work in practice, suppose we build a civil engineering structure such as a bridge or a building, that has a natural frequency ω_1 in some mode, whose amplitude we call $x(t)$. Suppose the system is subjected to an external dynamic loading (perhaps from traffic, wind or earthquakes) that is rich in a frequency ω which is close to ω_1 . Then, if the damping is small (as is typical in large structures) using (15) with $\mu = 0$, we find that the response given by (17) is large since α_1 is close to 1 and δ_1 is small. This is the fundamental *resonant response* of the structure close to its natural frequency. Now suppose we add the pendulum as a tuned damper. In particular we design the pendulum so that $\alpha_2 \approx \alpha_1$. The pendulum now sees a large amplitude parametric excitation within its main subharmonic resonance tongue. It becomes violently unstable, thus sucking energy from the structure.

To see just how effective this quenching effect can be, consider the following asymptotic calculation [58]. Suppose the external forcing is very small, $A = \varepsilon^2 \hat{A}$, for some small parameter ε , and that the damping and frequency detunings scale with ε : $\delta_1 = \varepsilon \hat{\delta}_1$, $\delta_2 = \varepsilon \hat{\delta}_2$, $\alpha_1 = 1 + \varepsilon \hat{\alpha}_1$, $\alpha_2 = (1/4) + \varepsilon \hat{\alpha}_2$, where the hatted quantities are all assumed to be $O(1)$ as $\varepsilon \rightarrow 0$. Upon making these substitutions, and dropping all hats, Eqs. (15) and (16) become

$$\ddot{x} + x + \varepsilon[\delta_1 \dot{x} + \sigma_1 - (1/4)\mu \phi^2 + \mu \dot{\phi}^2] = \varepsilon A \cos(t) + O(\varepsilon^2), \quad (18)$$

$$\ddot{\phi} + \frac{1}{4}\phi + \varepsilon[\delta_2 \dot{\phi} + (1/2)\sigma_2 \phi - x\phi] = O(\varepsilon^2). \quad (19)$$

We then look for solutions of the form

$$x = R_1(t) \cos[t + \psi_1(t)], \quad \phi = R_2(t) \cos[(t/2) + \psi_2(t)].$$

Making this substitution into (18), (19), and keeping only $O(\varepsilon)$ terms we find

$$\begin{aligned} 2\dot{R}_1 &= \varepsilon[-\delta_1 R_1 - (1/4)R_2^2 \sin(\psi_1 - 2\psi_2) - A \sin(\psi_1)], \\ 2R_1\dot{\psi}_1 &= \varepsilon[\sigma_1 R_1 - (1/4)\mu R_2^2 \cos(\psi_1 - 2\psi_2) - A \cos \psi_1], \\ 2\dot{R}_2 &= \varepsilon[-\delta_2 R_2 + R_1 R_2 \sin(\psi_1 - 2\psi_2)], \\ 2R_2\dot{\psi}_2 &= \varepsilon[\sigma_2 R_2 - R_1 R_2 \cos(\psi_1 - 2\psi_2)]. \end{aligned}$$

Equilibrium solutions of these equations correspond to 4π -periodic solutions that are valid close to the instability tongue. Upon setting the left-hand sides to zero, a straightforward combination of the final two equations yields

$$R_1 = \sqrt{\sigma_2^2 + \delta_2^2}, \quad (20)$$

and a slightly longer calculation reveals that this leads to a unique (stable) solution for R_2 provided $A^2 > (\delta_1^2 + \sigma_1^2)(\delta_2^2 + \sigma_2^2)$.

Equation (20) gives the amplitude of the quenched solution for the structure. Compare this with the amplitude R_0 of the semi-trivial solution (17), which is the forced response of the structure without the pendulum-tuned-damper added. Written in the rescaled coordinates, this solution is

$$R_0 = \frac{A^2}{\sigma_1^2 + \delta_1^2}.$$

Here we make the important observation that, unlike the simple forced response R_0 , the quenched amplitude R_1 is independent of the forcing amplitude A , and does not blow up at the fundamental resonance $\sigma_1 \rightarrow 0$. In fact, at least to leading order in this asymptotic approximation, the amplitude of the quenched solution is *independent of the frequency detuning and damping constant of the structure, and depends only on the frequency detuning and proportional damping of the pendulum!*

Tondl et al. [58] give a number of applications of autoparametric resonance. First, the above quenching effect means that analogues of the mass-spring-pendulum system can be used as tuned dampers, that are capable of suppressing certain undesirable response frequencies from structures or machines. Another important example of parametric resonance they cite is in flow-induced oscillation, where periodic flow features such as vortex shedding can cause a parametric response of a structure. The galloping of cables in strong winds is an example of just

such an effect. Cable-stayed bridges are engineering structures that are particularly susceptible to autoparametric effects. Not only is there potential for resonant interaction between the cables and the wind, but, since the cables are all of different lengths, there is a good chance that there is at least one cable whose natural frequency is either half or twice that of a global vibration mode of the entire bridge. Models of deck-cable interaction have shown the propensity for undesirable effects, see e.g. [23,24]. For example, vibrations caused by traffic could cause particular cables to undergo large-amplitude vibration leading to potential problems with wear. Alternatively, large-amplitude cable vibrations (caused for example by fluid-structure interaction) could cause significant deck vibration leading to an uncomfortable ride for those crossing the bridge.

Combination Parametric Resonances

Mailybayev and Seyranian [41] consider general systems of second-order equations of the form

$$M\ddot{y} + \delta D\dot{y} + (A + \beta B(\Omega t))y = 0, \quad y \in \mathbb{R}^n, \quad (21)$$

where M , D and A are positive definite symmetric matrices, and B is a periodic matrix with period $2\pi/\Omega$. They derive conditions under which parametric resonance may occur. Suppose that when $\delta = 0 = \beta = 0$, the linear system has non-degenerate natural frequencies, $\omega_1, \omega_2, \dots, \omega_n$. Then they showed that the only kinds of parametric resonance that are possible are when

$$\Omega = \frac{2\omega_j}{k}, \quad j = 1, \dots, m, \\ k = 1, 2, \dots, \quad (\text{simple resonance}),$$

$$\text{or } \Omega = \frac{\omega_i \pm \omega_j}{k}, \quad i, j = 1, \dots, m, \quad i > j, \\ k = 1, 2, \dots, \quad (\text{combination resonance}). \quad (22)$$

The sign '+' or '-' in (22) gives what are called *sum* or *difference* combination resonances respectively.

There are many examples of such combination resonances in the literature. For example, the book by Nayfeh [45] considers many such cases, especially of mechanical systems where the two modes are derived from a Galerkin reduction of a continuum system such as a plate or beam. In all these examples, though, it is a sum combination resonance that is excited. That is, by forcing the system at the sum of the two individual frequencies $\Omega = \omega_1 + \omega_2$, individual responses are found at frequency ω_1 or ω_2 .

However, there do not seem to be any concrete examples of *difference* resonances in the literature. In fact, such a mechanical device might be somewhat strange. If we had $\omega_1 \approx \omega_2$, then $\Omega = \omega_1 - \omega_2$ would be small, perhaps several orders of magnitude smaller. So a difference combination resonance would give a response at a high frequency from low frequency excitation. This would be like making a drum vibrate by sending it up and down in an elevator!

In fact, for the case that $B = B_0\phi(t)$ is a constant matrix times a function of time, then it is shown in (Theorem 1 in [41]) that a given system can only excite *either* sum *or* difference combination resonances, but not both. To explain this result, it is sufficient to consider a simplification of (21) to the case $n = 2$. Consider (21) in the case $\delta = 0$ and where $B(t)$ is a constant matrix times a pure function of time with period Ω . Without loss of generality, let us suppose this function to be a pure cosine $B = B_0 \cos(\Omega t)$. Suppose that we change coordinates so that the system with $\beta = 0$ is written in diagonal form, and finally that time has been rescaled so that $\Omega = 1$. Hence we obtain a system of the form

$$\begin{pmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{pmatrix} + \left[\begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} + \beta \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \cos(t) \right] \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0, \quad (23)$$

where $\alpha_1 = \omega_1^2/\Omega^2$, $\alpha_2 = \omega_2^2/\Omega^2$ and the matrix $\hat{B} = \{b_{ij}\}$ is the original constant matrix within B_0 written in the transformed coordinates. Recalling, how in Subject. “Floquet Analysis” we found the transition curves for the Mathieu equation using Floquet theory, we can look for solutions to (23) in the form

$$x_1 = c_n \sum_{n=-\infty}^{\infty} e^{int/\sqrt{2}}, \quad x_2 = \pm c_n \sum_{n=-\infty}^{\infty} e^{sint/\sqrt{2}}, \quad (24)$$

where $s = \pm 1$. We find an infinite system of algebraic equations analogous to (10). It is straightforward to see that in the case $\beta = 0$, there is a non-trivial solution with $c_n = 0$ for all $n \neq k$ and $c_k \neq 0$, whenever $\alpha_1 + s\alpha_2 = k^2$. This would suggest that both sum and difference combination resonances are possible. However, when looking at the conditions for the bifurcation equations to have a real solution for nonzero β , one finds the following condition

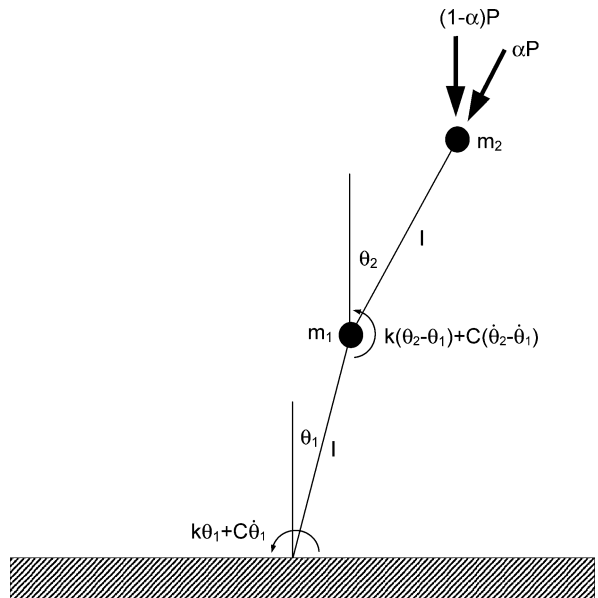
$$\text{sign}(b_{12}b_{21}) = s. \quad (25)$$

That is, to excite a sum resonance, the diagonal entries of B_0 must be of the same sign, and to excite a difference res-

onance these diagonal entries must be of opposite sign. In particular, if B_0 is a symmetric matrix (as is the case in many mechanical applications) then only sum combinations can be excited. In fact, it is argued in [68] that pure Hamiltonian systems cannot ever excite difference combination resonance; in other words, if difference parametric resonance is possible at all, then the matrix $B(\Omega t)$ must contain dissipative terms.

One might wonder whether difference combination resonance can ever be excited in any physically derived system. To show that they can, consider the following example with a non-conservative parametric excitation force, a detailed analysis of which will appear elsewhere [59].

Figure 7 depicts a double pendulum (which we think of as lying in a horizontal plane so that gravity does not affect it) with stiff, damped joints and is loaded by an end force – a so-called *follower load* – that is maintained in the direction of the axis of the second pendulum. Note that such a load is by its nature non-conservative because work (either positive or negative) has to be done to maintain the direction of the follower. In practice, such a force might be produced by a jet of fluid emerging from the end of the outer pendulum, for example if the pendula were actually pipes, although the equations of motion (and the consequent nonlinear dynamics) are rather different for



Dynamics of Parametric Excitation, Figure 7

The partially follower-loaded elastic double pendulum, depicting a definition of the various physical quantities. In this study we take $m_1 = m_2 = m$ and $\alpha = 1$

that case [5,12,13]. Non-dimensional equations of motion for this system are derived in [56], where it was shown that the system can undergo quite violent chaotic dynamics for large deflections. We shall restrict ourselves here to small deflection theory and allow the follower load to be a harmonic function of time $P = A \cos \Omega t$. After nondimensionalization using $\beta = |A|l/k$, $\delta = C/\sqrt{km}l^2$, and rescaling time according to $t_{\text{new}} = \sqrt{k/m_2 l^2} t_{\text{old}}$, we obtain the equations

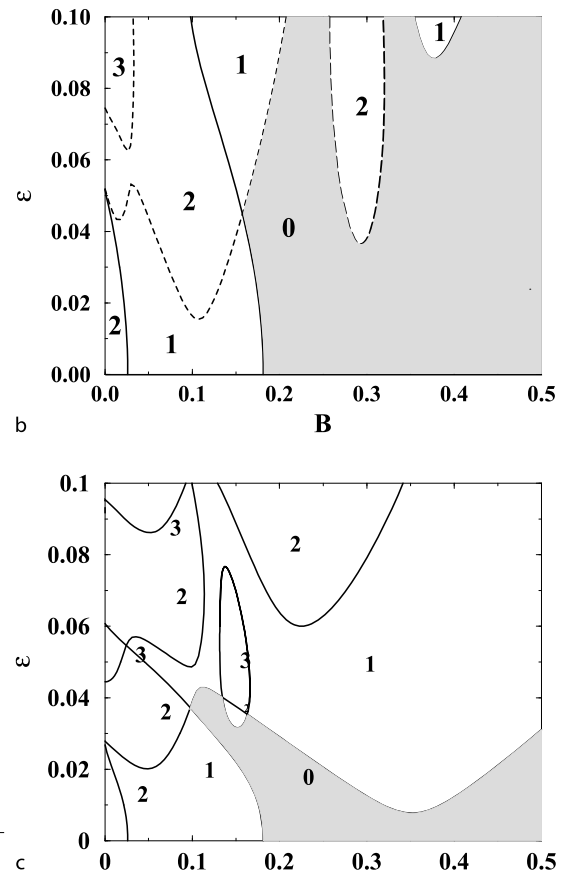
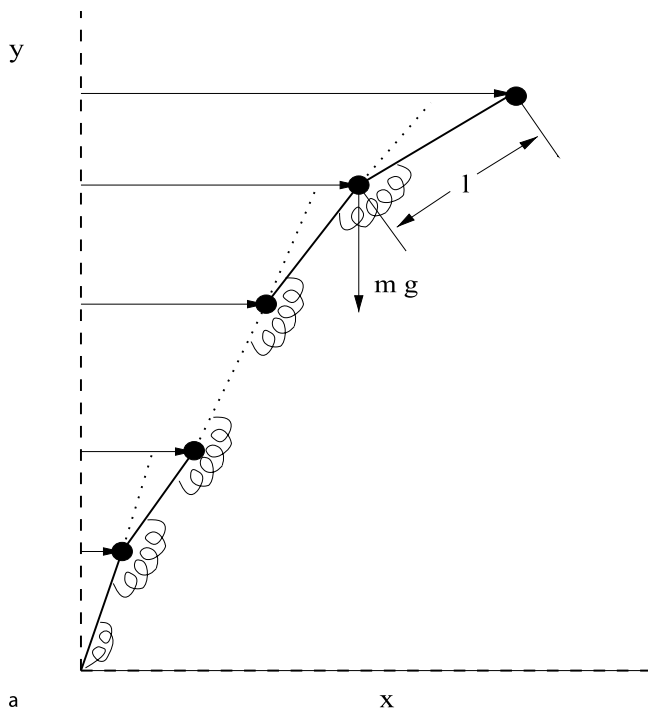
$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{pmatrix} + \delta \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + p \cos \Omega t \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = 0. \quad (26)$$

Now, (26) is not in diagonal form, but a simple coordinate transformation can put it in the form (21) with M the identity matrix,

$$D = A = \begin{pmatrix} 0.1492 & 0 \\ 0 & 3.3508 \end{pmatrix}$$

$$\text{and } B_0 = \begin{pmatrix} -0.0466 & 0.3788 \\ -0.1288 & 1.0466 \end{pmatrix}.$$

Hence the system has two natural frequencies $\omega_1 = \sqrt{0.1492} = 0.3863$ and $\omega_2 = \sqrt{3.3508} = 1.8305$, and the quantities $b_{12}b_{21} = 0.3788 \times -0.1288 = -0.0488 < 0$. Hence, according to (25), a difference combination resonance should be possible by exciting the system at $\Omega = \omega_2 - \omega_1 = 1.4442$, as verified in [59].



Dynamics of Parametric Excitation, Figure 8

a A multiple pendulum with stiff joints. **b,c** Ince-Strutt diagrams of dimensionless amplitude of vibration ε against scaled elastic stiffness B for eight identical jointed pendulums with weak damping. **b** For dimensionless applied frequency $\omega = 10$; **c** $\omega = 20$. Shading represents the stability region of the upright equilibrium and digits in other regions give the number of unstable Floquet multipliers. After [22], to which the reader is referred to for the precise details; reproduced with permission from Elsevier

Upside-Down Stability

As mentioned in the Introduction, it has been known for 100 years [53] that the ‘negative gravity’ ($\alpha < 0$) region of the Ince–Strutt diagram for the Mathieu equation implies that a simple pendulum can be stabilized in the inverted position by application of parametric excitation. More recently Acheson and Mullin [3] demonstrated experimentally that a double and even a triple pendulum can be similarly stabilized when started in the inverted position. Moreover, the stabilized equilibrium is quite robust, such that the system relaxes asymptotically back to its inverted state when given a moderate tap.

At first sight, these observations seem remarkable. A multiple pendulum has several independent normal modes of oscillation and naively one might expect that single frequency excitation could at best stabilize only one of these modes at a time. However, Acheson [1] (see also the earlier analysis of Otterbein [46]) provided a simple one-line proof that, in fact, for sufficiently high frequency and sufficiently small amplitude sinusoidal excitation, in theory *any* ideal finite chain of N pendulum can be stabilized in the inverted position. To see why this is true, consider the multiple pendulum system represented in Fig. 8 but without the springs at the joints. Then, upon performing a modal analysis, we find that the equations of motion reduce to the study of N uncoupled Mathieu equations each with different parameters $\alpha_i > 0$ and $\beta_i > 0$, for $i = 1, \dots, N$. Hence the frequency can be chosen to be sufficiently high and the amplitude chosen to be sufficiently small for each α_i, β_i to lie in the thin wedge of stability for $\alpha < 0$ in Fig. 4. Thus, each mode is stable, and

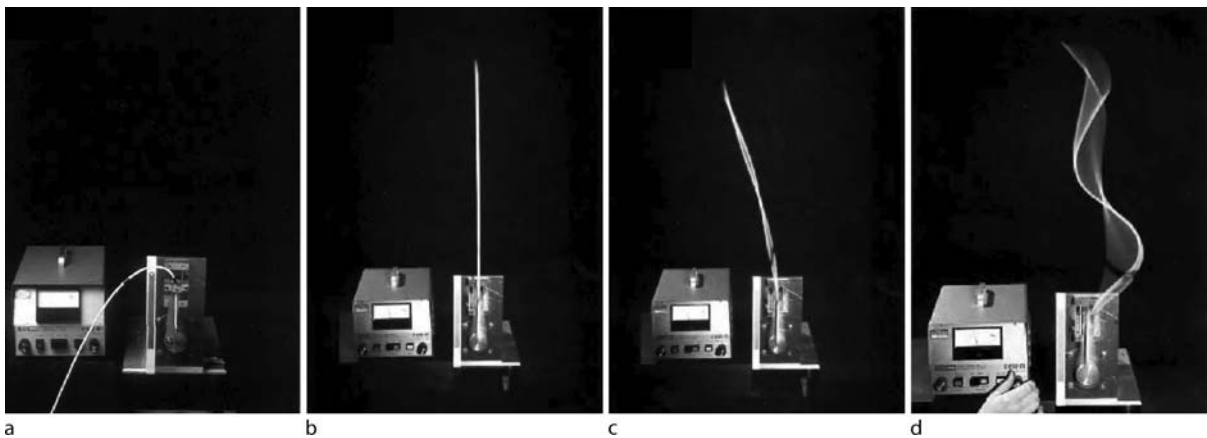
so the whole finite chain has been stabilized by parametric excitation.

It has been suggested that the limit of such a multiple pendulum, in which the total length and mass of the system stays constant while the number of pendulums in the chain becomes infinite, could be used as a possible explanation of the so called ‘*Indian rope trick*’, a classic conjuring trick in which a rope is ‘charmed’ by a magician to appear vertically out of a basket like a snake. Unfortunately, as noted [1], in this limit, which is that of a piece of string, the stability region becomes vanishingly small and so this potential scientific explanation of the magic trick fails.

However, a further experiment by Mullin announced in [4], demonstrated that a piece of ‘bendy curtain wire’, clamped at the bottom and free at the top, that is just too long to support its own weight can be stabilized by parametric excitation (see Fig. 9). In order to explain this result, Galán et al. [22] introduced the model depicted in Fig. 8 and studied the stability of the vertical equilibrium position in the presence of both damping and stiffness in each of the joints. The normal modes of this problem lead to fully coupled equations and so Acheson’s argument does not apply. Using numerical bifurcation analysis Galán et al. were able to find stability boundaries and also to find an appropriate continuum limit; providing a good motivation for us to turn our attention to parametric resonance in continuous systems.

Continuous Systems

Parametric resonance is also of importance in continuum systems. Canonical effects we shall discuss here include the



Dynamics of Parametric Excitation, Figure 9

Experimental results on a piece of ‘bendy curtain wire’ which is just long enough to buckle under its own weight (a). For tiny vertical oscillations between about 15 and 35 Hz, the wire can be made to stand upright (b), and to be stable against even quite large perturbations (c). Beyond the upper frequency threshold, a violent dynamic stability sets in which involves the third vibration mode excited harmonically (d). After [44], reproduced with permission from the Royal Society

wire stiffening we have just introduced, the excitation of patterns at fluid interfaces, more general theories of flow-induced oscillations of structures, and the stabilizing effects of periodic modulation of optical fibers.

Structural Stiffening; the ‘Indian Rod Trick’

Thomsen [57] has shown that parametric excitation generally speaking has a stiffening effect on a structure. This is essentially because of the same effect that causes the upside down stability of the simple pendulum for high-enough frequencies and small enough amplitude. However, detailed experiments by Mullin and co-workers [44], reproduced in Fig. 9, showed that there can also be a destabilizing effect of parametric resonance. In particular, a piece of curtain wire (a thin helical spring surrounded by plastic) that was just too long (about 50 cm) to support its own weight, was clamped vertically upright and subjected to rapid, small-amplitude oscillations of about 1 cm peak-to-peak. At around 15 Hz the rod was found to be stabilized, such that when released it remains erect (see Fig. 9b). This effect continues for higher frequencies, until about 30 Hz at which point a violent dynamic instability sets in that is commensurate with the driving frequency (harmonic, rather than subharmonic). The experimentally determined parameter values in which the upside-down sta-

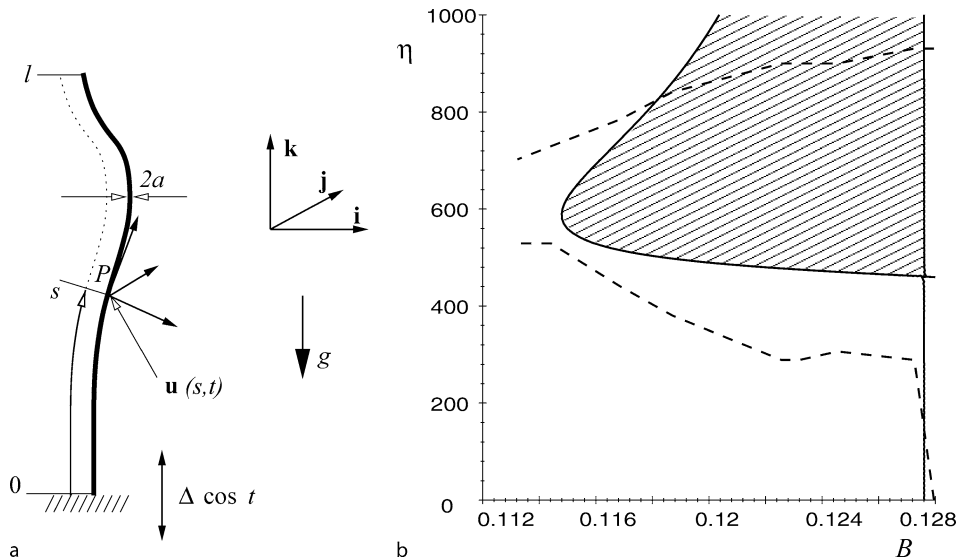
bilization can be achieved are represented in terms of dimensionless parameters in Fig. 10b.

Champneys and Fraser [14,15,20], introduced a continuum theory to explain these experimental results, which produced the shaded region in Fig. 10b, that we now explain. Consider an intrinsically straight column of length ℓ , with a uniform circular cross-section of radius $a \ll \ell$ and mass linear density m per unit length (see Fig. 10a). The column is assumed to be inextensible, un-shearable and linearly elastic with bending stiffness \hat{B} and material damping coefficient $\hat{\gamma}$ (where hatted variables are dimensional). The lower end is clamped upright and is to vertical displacement $\Delta \cos \omega \hat{t}$, whereas the upper end is free. In [15], it is shown that the stability of the vertical equilibrium position can be studied by analyzing solutions to the PDE

$$Bu_{ssss} + [(1-s)u_s]_s + \eta(u_{tt} + \gamma u_{ssst} - \varepsilon[(1-s)u_s]_s \cos t) = 0, \quad (27)$$

for the scaled lateral displacement $u(s, t)$ assumed without loss of generality to be in a fixed coordinate direction, where the scaled arc length $s \in (0, 1)$, the scaled time variable $t = \omega \hat{t}$, and the boundary conditions are

$$u = u_s = 0, \text{ at } s = 0; \quad u_{ss} = u_{ssss} = 0 \text{ at } s = 1. \quad (28)$$



Dynamics of Parametric Excitation, Figure 10

a Definition sketch of the parametrically excited flexible rod model. **b** Comparison between theory (solid line and shading) and the experimental results of [44] (dashed line), for the stability region as a function of the dimensionless parameters B and η using the experimental value $\varepsilon = 0.02$, and the representative damping value $\gamma_1 = 0.01$ (which was not measured in the experiment). The theory is based on calculations close to the resonance tongue interaction between when $B_{0,1} = B_{1,3}$ when $\eta = \eta_c = 460.7$. After [15], reproduced with permission from SIAM

The dimensionless parameters appearing in (27) are

$$\gamma = \frac{\Gamma \ell \omega}{mg}, \quad B = \frac{\hat{B}}{mg\ell^3}, \quad \eta = \frac{\omega^2 \ell}{g}, \quad \varepsilon = \frac{\Delta}{\ell}, \quad (29)$$

which represent respectively material damping, bending stiffness, the ratio of forcing frequency to the equivalent pendulum natural frequency, and forcing amplitude.

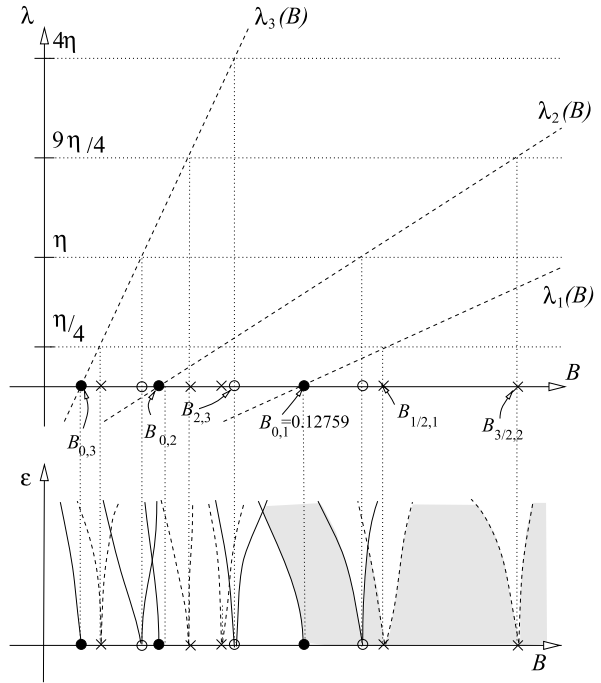
If we momentarily consider the problem with no forcing or damping, $\varepsilon = \gamma = 0$, then the eigenmodes ϕ_n and corresponding natural frequencies $\sqrt{\lambda_n/\eta}$ of free vibration of the rod satisfy

$$B(\phi_n)_{ssss} + [(1-s)(\phi_n)_s]_s \phi_n - \lambda_n \phi_n = 0,$$

and form a complete basis of functions that satisfy the boundary conditions (28). Now the eigenmodes $\phi_n(s; B)$ are in general not expressible in closed form except at the special values of B at which $\lambda_n(B) = 0$, in which case [20] there is a solution in terms of the Bessel function $J_{-1/3}$. The same analysis shows that there are infinitely many such B -values, $B_{0,n}$, for $n = 1, 2, 3, 4, \dots$, accumulating at $B = 0$, the first few values of which are $B_{0,1} = 0.127594$, $B_{0,2} = 0.017864$, $B_{0,3} = 0.0067336$ and $B_{0,4} = 0.0003503$. These correspond respectively to where each eigenvalue locus $\lambda_n(B)$, for $n = 1, 2, 3, 4$, crosses the B -axis, with the corresponding eigenmodes there having $n-1$ internal zeros. The lowering of B through each $B_{0,n}$ -value implies that the n th mode becomes linearly unstable. Hence for $B > B_c := B_{0,1}$ the unforced rod is stable to self-weight buckling (a result known to Greenhill [26]). Moreover, it can be shown [14,15] that each eigenmode $\phi_n(s)$ retains a qualitatively similar mode shape for $B > B_{0,n}$ with the n th mode having $n = 1$ internal zeros, and that the corresponding loci $\lambda_n(B)$ are approximately straight lines as shown in the upper part of Fig. 11.

The lower part of Fig. 11 shows the results of a parametric analysis of the Eq. (27) for $\gamma = 0$. Basically, each time one of the eigenvalue loci passes through η times the square of a half-integer $(j/2)^2$ we reach a B -value for which there is the root point of an instability tongue. Essentially we have a Mathieu-equation-like Ince–Strutt diagram in the (B, ε) plane for each mode n . These diagrams are overlaid on top of each other to produce the overall stability plot. However, note from Fig. 11 that the slope of each eigenvalue locus varies with the frequency ratio η . Hence as η increases, the values of all the $B_{j/2,n}$, $j = 0, \dots, \infty$ for a given n increase with the same fixed speed. But this speed varies with the mode number n . Hence the instability tongues slide over each other as the frequency varies.

Note the schematic diagram in the lower panel of Fig. 11 indicates a region of the stability for $B < B_{0,1}$ for



Dynamics of Parametric Excitation, Figure 11

Summarizing schematically the results of the parametric resonance analysis of the parametric excited vertical column in the absence of damping. The upper part shows eigenvalue loci $\lambda_n(B)$ and the definition of the points $B_{j/2,n}$, $j = 0, 1, 2, 3, \dots$. In the lower part, instability tongues in the (B, ε) -plane are shown to occur with root points $B_{j/2,n}$ and to have width ε^j . The shaded regions correspond to the where the vertical solution is stable. Solid lines represent neutral stability curves with Floquet multiplier $+1$ (where α is an integer) and dashed lines to multipliers -1 (where α is half an odd integer). Reproduced from [14] with permission from the Royal Society

small positive ε . Such a region would indicate the existence of rod of (marginally) lower bending stiffness than that required for static stability, that can be stabilized by the introduction of the parametric excitation. Essentially, the critical bending stiffness for the buckling instability is reduced (or, equivalently since $B \propto \ell^{-3}$, the critical length for buckling is increased). This effect has been dubbed the ‘Indian rod trick’ because the need for bending stiffness in the wire means that mathematically this is a rod, not a rope (or string); see, e.g. the account in [4].

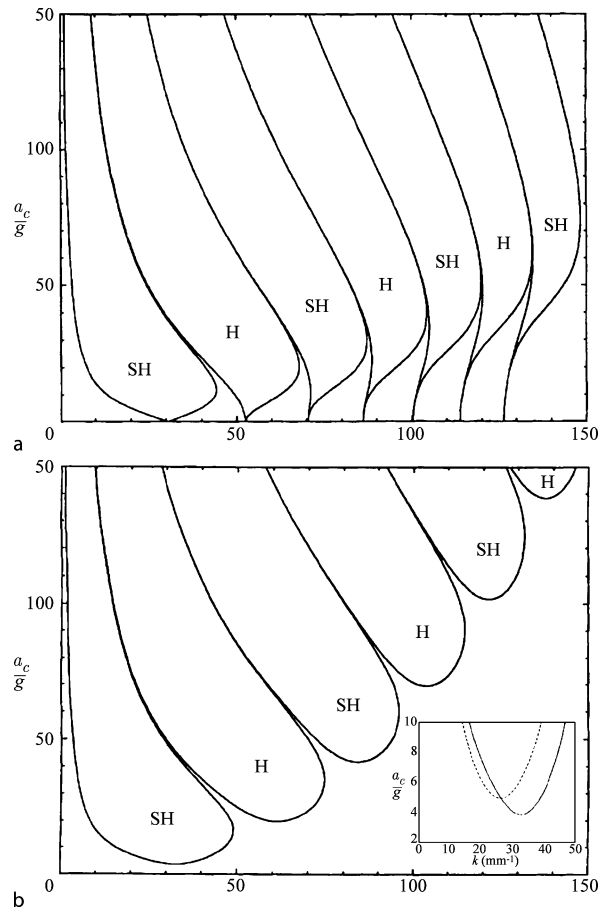
To determine in theory the precise parameter values at which the rod trick works turns out to be an involved process. That the instability boundary emerging from $B_{0,1}$ in the lower panel of Fig. 11 should usually bend back to the left can be argued geometrically [7]. However, for η -values at which $B_{0,1} \approx B_{1,n}$ for some n , it has been shown that there is actually a singularity in the coefficient of the quadratic part of this buckling curve causing the instabil-

ity boundary to bend to the right [20]. A complete unfolding of this so-called *resonance tongue interaction* was performed in [15], also allowing for the presence of material damping. The resulting asymptotic analysis needs to be carried out in the presence of *four* small parameters; ε , γ , $B - B_{0,1}$ and $\eta - \eta_c$ where η_c is the frequency-ratio value at which the resonance-tongue interaction occurs. The results produce the shaded region shown in Fig. 10b, which can be seen to agree well with the experimental results.

Faraday Waves and Pattern Formation

When a vessel containing a liquid is vibrated vertically, it is well known that certain standing wave patterns form on the free surface. It was Faraday in 1831 [19] who first noted that the frequency of the liquid motion is half that of the vessel. These results were confirmed by Rayleigh [50] who suggested that what we now know as parametric resonance was the cause. This explanation was shown to be correct by Benjamin and Ursell [6] who derived a system of uncoupled Mathieu equations from the Navier–Stokes equations describing the motion of an ideal (non-viscous) fluid. There is one Mathieu equation (4), for each wave number i that fits into the domain, each with its own parameters α_i and β_i . They argue that viscous effects would effectively introduce a linear damping term, as in (8) and they found good agreement between the broad principal (subharmonic) instability tongue and the results of a simple experiment. Later, Kumer and Tuckerman [36] produced a more accurate Ince–Strutt diagram (also known as a dispersion relation in this context) for the instabilities of free surfaces between two viscous fluids. Through numerical solution of appropriate *coupled* Mathieu equations that correctly take into account the boundary conditions, they showed that viscosity acts in a more complicated way than the simple velocity proportional damping term in (8). They produced the diagram shown in Fig. 12b, which showed a good match with the experimental results. Note that, compared with the idealized theory, Fig. 12a, the correct treatment of viscosity causes a small shift in the critical wavelength of each instability, a broadening of the higher-harmonic resonance tongues and a lowering of the amplitude required for the onset of the principal subharmonic resonance if damping is introduced to the idealized theory (see the inset to Fig. 12b).

More interest comes when Faraday waves are excited in a vessel where two different wave modes have their subharmonic resonance at nearly the same applied frequency, or when the excitation contains two (or more) frequencies each being close to twice the natural frequency of a standing wave. See for example the book by Hoyle [30] and ref-



Dynamics of Parametric Excitation, Figure 12

Stability boundaries for Faraday waves from: a Benjamin and Ursell's theory for ideal fluids; b the results of full hydrodynamic stability analysis by Kumar and Tuckerman. The horizontal axis is the frequency and the vertical scale the amplitude of the excitation. The inset to panel b shows the difference between the true instability boundary for the principal subharmonic resonance (lower curve) and the idealized boundary in the presence of simple proportional damping. After [36], to which the reader is referred for the relevant parameter values; reproduced with permission from Cambridge University Press

erences therein, for a glimpse at the amazing complexity of the patterns of vibration that can result and for their explanation in terms of dynamical systems with symmetry. Recently a new phenomenon has also been established, namely that of localized spots of pattern that can form under parametric resonance, so-called *oscillons* see e. g. [61] for a review and [39] for the first steps at an explanation.

Fluid-Structure Interaction

Faraday waves are parametric effects induced by the gross mechanical excitation of a fluid. In Subject. “Autopara-

metric Resonance" we briefly mentioned that fluid-structure interaction can go the other way, namely that external fluid flow effects such as vortex shedding can excite (auto)parametric resonances in simple mechanical systems. Such flow-induced parametric resonance can also occur for continuous structures, in which waves can be excited in the structure due to external parametric loading of the fluid; see for example the extensive review by Païdousis and Li [47]. They also treat cases where the flow is internal to the structure: we are all familiar with the noisy ducts and pipes, for example in central heating or rumbling digestive systems, and it could well be that many of these effects are a result of parametric resonance.

Semler and Païdousis [52] consider in detail parametric resonance effects in the simple paradigm of a cantilevered (free at one end, clamped at the other) flexible pipe conveying fluid. Think of this as a hosepipe resting on a garden lawn. Many authors have considered such an apparatus as a paradigm for internal flow-induced oscillation problems (see the extensive references in [47,52]), starting from Benjamin's [5] pioneering derivation of equations of motion from the limit of a segmented rigid pipe. The equations of motion for such a situation are similar to those of the rod (27), but with significant nonlinear terms arising from inertial and fluid-structure interaction effects. In the absence of periodic fluctuations in the flow, then it is well known that there is a critical flow rate beyond which the pipe will become oscillatory, through a Hopf bifurcation (flutter instability). This effect can be seen in the way that hosepipes writhe around on a lawn, and indeed forms the basis for a popular garden toy. Another form of instability can occur when the mean flow rate is beneath the critical value for flutter, but the flow rate is subject to a periodic variation, see [52] and references therein. There, a parametric instability can occur when the pulse rate is twice the natural frequency of small amplitude vibrations of the pipe. In [52] this effect was observed in theory, in numerical simulation and in a careful experimental study, with good agreement between the three. Parametric resonance could also be found for super-critical flow rates. In this case, the effect was that the periodic oscillation born at the Hopf bifurcation was excited into a quasi-periodic response.

Another form of fluid structure interaction can occur when the properties of an elastic body immersed in a fluid are subject to a temporal periodic variation. With a view to potential biological applications such as active hearing processes and heart muscles immersed in blood, Cortez et al. [16] consider a two-dimensional patch of viscous fluid containing a ring filament that is excited via periodic variation of its elastic modulus. Using a mixture of

immersed boundary numerical methods and analytic reductions, they are able to find an effective Ince-Strutt diagram for this system and demonstrate the existence of parametric instability in which the excitation of the structure causes oscillations in the fluid. Taking a fixed wave number p of the ring in space, they find a sequence of harmonic and subharmonic instabilities as a parameter effectively like α in the Mathieu equation is increased. Interestingly though, even for low viscosity, the principal subharmonic resonance tongue (corresponding to $\alpha = 1/4$) does not occur for low amplitude forcing, and the most easily excited instability is actually the fundamental, harmonic resonance (corresponding to $\alpha = 1$).

Dispersion Managed Optical Solitons

The stabilizing effect of high-frequency parametric excitation has in recent years seen a completely different application, in an area of high technological importance, namely optical communications. Pulses of light sent along optical fibers are a crucial part of global communication systems. A pulse represents a packet of linear waves, with a continuum of different frequencies. However, optical fibers are fundamentally dispersive, which means that waves with different wavelengths travel at different group velocities. In addition, nonlinearity, such as the Kerr effect of intensity-dependent refractive index change, and linear losses mean that there is a limit to how far a single optical signal can be sent without the need for some kind of amplifier. Depending on the construction of the fiber, dispersion can be either *normal* which means that lower frequencies travel faster, or *anomalous* in which case higher frequencies travel faster.

One idea to overcome dispersive effects, to enable pulses to be travel over longer distances without breaking up, is to periodically vary the medium with which the fiber is constructed so that the dispersion is alternately normally and anomalously dispersive [38]. Another idea, first proposed by Hasegawa and Tappert [28], is to balance dispersion with nonlinearity. This works because the complex wave envelope A of the electric field of the pulse can be shown to leading order to satisfy the nonlinear Schrödinger (NLS) equation

$$i \frac{dA}{dz} + \frac{1}{2} \frac{d^2 A}{dt^2} + |A|^2 A = 0, \quad (30)$$

in which time and space have been rescaled so that the anomalous dispersion coefficient (multiplying the second-derivative term) and the coefficient of the Kerr nonlinearity are both unity. Note the peculiarity in optics compared to other wave-bearing system in that time t plays the role

of the spatial coordinate, and the propagation distance z is the evolution variable. Thus we think of pulses as being time-traces that propagate along the fiber. Now, the NLS equation is completely integrable and has the well-known soliton solution

$$A(z, t) = \eta \operatorname{sech}(\eta t - cz) e^{i\eta^2 z}, \quad (31)$$

which represents a smooth pulse with amplitude η , which propagates at ‘speed’ c . Moreover, Eq. (30) is completely integrable [69] and so arbitrary initial conditions break up into solitons of different speed and amplitude. Optical solitons were first realized experimentally by Mollenauer et al. [43]. However, true optical fibers still suffer linear losses, and the necessary periodic amplification of the optical solitons to enable them to survive over long distances can cause them to *jitter* [25] and to interact with each other as they move about in time.

The idea of combining nonlinearity with dispersion management in order to stabilize soliton propagation was first introduced by Knox, Forysiak and Doran [35], was demonstrated experimentally by Suzuki et al. [54] and has shown surprisingly advantageous performance that is now being exploited technologically (see [60] for a review). Essentially, dispersion management introduces a periodic variation of the dispersion coefficient $d(z)$, that can be used to compensate the effects of periodic amplification, which with the advent of modern erbium-doped optical amplifiers can be modeled as a periodic variation of the nonlinearity. As an envelope equation we arrive at the modified form of (30),

$$i \frac{dA}{dz} + \frac{d(z)}{2} \frac{d^2 A}{dz^2} + \sigma(z) |A|^2 A = 0, \quad (32)$$

where the period L of the dispersion map $d(z)$ and the amplifier spacing (the period of $\sigma(z)$) Z may not necessarily be the same. Two particular regimes are of interest. One, of application to long-haul, sub-sea transmission systems where the cable is already laid with fixed amplifiers, is to have $L \gg Z$. Here, one modifies an existing cable by applying bits of extra fiber with large dispersion of the opposite sign only occasionally. In this case, one can effectively look only at the long scale and average out the effect of the variation of σ . Then the parametric term only multiplies the second derivative. Another interesting regime, of relevance perhaps when designing new terrestrial communication lines, is to have $L = Z$ so that one adds loops of extra opposite-dispersion fiber at each amplification site. In both of these limits we have an evolution equation whose coefficients are periodic functions of the evolution variable z with period L .

In either regime, the periodic (in z) parameter variation has a profound stabilizing influence, such that the effects of linear losses and jitter are substantially reduced. A particular observation is that soliton solutions appear to (32) *breathe* during each period; that is, their amplitude and phase undergo significant periodic modulation. Also, with increase in the amplitude of the oscillatory part of d , so the average shape of the breathing soliton becomes less like a sech-pulse and more like a Gaussian. An advantage is that the Gaussian has much greater energy the regular NLS soliton (31) for the same average values of d and σ , hence there is less of a requirement for amplification. Surprisingly, it has been found that the dispersion managed solitons can propagate stably even if the average of d is slightly positive, i. e. in the normal dispersion region where there is no corresponding NLS soliton. This has a hugely advantageous consequence. The fact that the local dispersion can be chosen to be high, but the average dispersion can be chosen to be close to zero, is the key to enabling the jitter introduced by the amplifiers to be greatly suppressed compared with the regular solitons. Finally the large local modulation of phase that is a consequence of the breathing, means that neighboring solitons are less likely to interact with each other. A mathematical analysis of why these surprising effects arise is beyond the scope of this article, see e.g. [32,60] for reviews, but broadly-speaking the stable propagation of dispersion-managed solitons is due to the stabilizing effects of high frequency parametric excitation.

Future Directions

This article has represented a whistle-stop tour of a wide areas of science. One of the main simplifications we have taken is that we have assumed the applied excitation to the system to be periodic; in fact, in most examples it has been pure sinusoidal. There is a growing literature on quasiperiodic forcing, see e.g. [51], and it would be interesting to look at those effects that are unique to parametrically excited systems that contain two or more frequencies. The inclusion of noise in the excitation also needs to be investigated. What effect does small additive or multiplicative noise have on the shape of the resonance tongues? And what is the connection with the phenomenon of stochastic resonance [66]?

There is a mature theory of nonlinear resonance in systems with two degrees of freedom which, when in an integrable limit, undergoes quasiperiodic motion with two independent frequencies. Delicate results, such as the KAM theorem, see e.g. [55], determine precisely which quasiperiodic motions (invariant tori) survive as a func-

tion of two parameters, the size of the perturbation from integrability and the frequency detuning between the two uncoupled periodic motions. For the tori that break up (inside so-called Arnol'd tongues) we get heteroclinic tangles and bands of chaotic motion. Parametric resonance tends to be slightly different though, since in the uncoupled limit, one of the two modes typically has zero amplitude (what we referred to earlier as a semi-trivial solution). It would seem that there is scope for a deeper connection to be established between the Arnol'd tongues of KAM theory and the resonance tongues that occur under parametric excitation.

In terms of applications, a large area that remains to be explored further is the potential exploitation of parametric effects in fluid-structure interaction problems. As I write, the heating pipes in my office are buzzing at an extremely annoying frequency that seems to arise on warm days when many radiators have been switched off. Is this a parametric effect? It would also seem likely that nature uses parametric resonance in motion. Muscles contract longitudinally, but motion can be in a transverse direction. Could phenomena such as the swishing of bacterial flagella, fish swimming mechanisms and the pumping of fluids through vessels, be understood in terms of exploitation of natural parametric resonances? How about motion of vocal chords, the Basilar membrane in the inner ear, or the global scale pumping of the heart? Does nature naturally try to tune tissue to find the primary subharmonic resonance tongue? Perhaps other effects of parametric excitation that we have uncovered here, such as the quenching of resonant vibrations, structural stiffening, and frequency tuned motion, are already being usefully exploited in nature. If so, there could be interesting consequences for nature inspired design in the engineered world.

Finally, let us return to our opening paradigm of how children effectively excite a parametric instability in order to induce motion in a playground swing. In fact, it was carefully shown by Case and Swanson [11] that the mechanism the child uses when in the sitting position is far more accurately described as an example of direct forcing rather than parametric excitation. By pushing on the ropes of the swing during the backwards flight, and pulling on them during the forward flight, the child is predominately shifting his center of gravity to and fro, rather than up and down; effectively as in Fig. 1c rather than Fig. 1b. Later, Case [10] argued that, even in the standing position, the energy gained by direct forcing from leaning backwards and forwards is likely to greatly outweigh any gained from parametric up-and-down motion of the child. The popular myth that children swinging is a suitable playground science demonstrator of parametric resonance was finally

put to bed by Post et al. [49]. They carried out experiments on human patients and analyzed the mechanisms they use to move, reaching the conclusion that even in the standing position, typically 95% of the energy comes from direct forcing, although parametric effects do make a slightly more significant contribution for larger amplitudes of swing.

Bibliography

1. Acheson D (1993) A pendulum theorem. *Proc Roy Soc Lond A* 443:239–245
2. Acheson D (1995) Multiple-nodding oscillations of a driven inverted pendulum. *Proc Roy Soc Lond A* 448:89–95
3. Acheson D, Mullin T (1993) Upside-down pendulums. *Nature* 366:215–216
4. Acheson D, Mullin T (1998) Ropy magic. *New Scientist* February 157:32–33
5. Benjamin T (1961) Dynamics of a system of articulated pipes conveying fluid. 1. Theory. *Proc Roy Soc Lond A* 261:457–486
6. Benjamin T, Ursell F (1954) The stability of a plane free surface of a liquid in vertical periodic motion. *Proc Roy Soc Lond A* 255:505–515
7. Broer H, Levi M (1995) Geometrical aspects of stability theory for Hills equations. *Arch Ration Mech Anal* 131:225–240
8. Broer H, Vegter G (1992) Bifurcation aspects of parametric resonance. *Dynamics Reported* (new series) 1:1–53
9. Broer H, Hoveijn I, van Noort M (1998) A reversible bifurcation analysis of the inverted pendulum. *Physica D* 112:50–63
10. Case W (1996) The pumping of a swing from the standing position. *Am J Phys* 64:215–220
11. Case W, Swanson M (1990) The pumping of a swing from the seated position. *Am J Phys* 58:463–467
12. Champneys A (1991) Homoclinic orbits in the dynamics of articulated pipes conveying fluid. *Nonlinearity* 4:747–774
13. Champneys A (1993) Homoclinic tangencies in the dynamics of articulated pipes conveying fluid. *Physica D* 62:347–359
14. Champneys A, Fraser W (2000) The 'Indian rope trick' for a continuously flexible rod; linearized analysis. *Proc Roy Soc Lond A* 456:553–570
15. Champneys A, Fraser W (2004) Resonance tongue interaction in the parametrically excited column. *SIAM J Appl Math* 65:267–298
16. Cortez R, Peskin C, Stockie J, Varela D (2004) Parametric resonance in immersed elastic boundaries. *SIAM J Appl Math* 65:494–520
17. Curry S (1976) How children swing. *Am J Phys* 44:924–926
18. Cvitanovic P (1984) *Universality in Chaos*. Adam Hilger, Bristol
19. Faraday M (1831) On the forms of states of fluids on vibrating elastic surfaces. *Phil Trans Roy Soc Lond* 52:319–340
20. Fraser W, Champneys A (2002) The 'Indian rope trick' for a parametrically excited flexible rod: nonlinear and subharmonic analysis. *Proc Roy Soc Lond A* 458:1353–1373
21. Galán J (2002) Personal communication
22. Galán J, Fraser W, Acheson D, Champneys A (2005) The parametrically excited upside-down rod: an elastic jointed pendulum model. *J Sound Vibration* 280:359–377
23. Gattulli V, Lepidi M (2003) Nonlinear interactions in the

- planar dynamics of cable-stayed beam. *Int J Solids Struct* 40:4729–4748
24. Gattulli V, Lepidi M, Macdonald J, Taylor C (2005) One-to-two global-local interaction in a cable-stayed beam observed through analytical, finite element and experimental models. *Int J Nonlin Mech* 40:571–588
 25. Gordon J, Haus H (1986) Random walk of coherently amplified solitons in optical fiber transmission. *Opt Lett* 11:665–666
 26. Greenhill A (1881) Determination of the greatest height consistent with stability that a pole or mast can be made. *Proceedings of the Cambridge Philosophical Society IV*, Oct 1880 – May 1883, pp 65–73
 27. Guckenheimer J, Holmes P (1983) *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer, New York
 28. Hasegawa A, Tappert F (1991) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibres. i. anomalous dispersion. *Appl Phys Lett* 23:142–144
 29. Hill G (1886) On the part of the motion of lunar perigee which is a function of the mean motions of the sun and moon. *Acta Math* 8:1–36
 30. Hoyle R (2005) *Pattern Formation; An Introduction to Methods*. CUP, Cambridge
 31. Ince E (1956) *Ordinary Differential Equations*. Dover, New York
 32. Jones C (2003) Creating stability from instability. In: *Nonlinear dynamics and chaos; where do we go from here?* IOP Publishing, Bristol, chap 4, pp 73–90
 33. Jordan DW, Smith P (1998) *Nonlinear Ordinary Differential Equations*, 3rd edn. Oxford University Press, Oxford
 34. Kapitza P (1951) Dinamicheskaya ustoychivost mayatnika pri koleblyushchetsya tochke podvesa. *Z Eksperimentalnoi i Teoreticheskoi Fiziki* 21:588–597, in Russian
 35. Knoz F, Forsyia W, Doran N (1995) 10 gbit/s soliton communication systems over standard fiber at $1.55\mu\text{m}$ and the use of dispersion compensation. *IEEE J Lightwave Technol* 13:1960–1995
 36. Kumar K, Tuckerman L (1994) Parametric instability of the interface between two fluids. *J Fluid Mech* 279:49–68
 37. Kuznetsov YA (2004) *Elements of Applied Bifurcation Theory*, 3rd edn. Springer, New York
 38. Lin C, Kogelnik H, Cohen L (1980) Optical-pulse equilization of low-dispersion transmission in single fibers in the $1.3\text{--}1.7\mu\text{m}$ spectral region. *Opt Lett* 5:476–480
 39. Lloyd D, Sandstede B, Avitabile D, Champneys A (2008) Localized hexagon patterns in the planar swift-hohenberg equation. To appear in *SIAM Appl Dyn Sys*
 40. Magnus W, Winkler S (1979) *Hill's Equation*. Dover, New York
 41. Mailybayev A, Seyranian AP (2001) Parametric resonance in systems with small dissipation. *PMM J App Math Mech* 65:755–767
 42. Mathieu E (1868) Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique. *J Math Pure Appl* 13:137–203
 43. Mollenauer L, Stolen R, Gordon J (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. *Phys Rev Lett* 45:1095–1097
 44. Mullin A, Champneys T, Fraser W, Galan J, Acheson D (2003) The 'Indian wire trick' via parametric excitation: a comparison between theory and experiment. *Proc Roy Soc Lond A* 459:539–546
 45. Nayfeh A (2000) *Nonlinear Interactions: Analytical, Computational and Experimental Methods*. Wiley Interscience, New York
 46. Otterbein S (1982) Stabilization of the *N*-pendulum and the Indian link trick. *Arch Rat Mech Anal* 78:381–393
 47. Paidoussis MP, Li G (1993) Pipes conveying fluid: a model dynamical problem. *J Fluids Structures* 7:137–204
 48. Piccoli B, Kulkarni J (2005) Pumping a swing by standing and squatting: Do children pump time optimally? *IEEE Control Systems Magazine* 25:48–56
 49. Post A, de Groot G, Daffertshofer A, Beek P (2007) Pumping a playground swing. *Motor Control* 11:136–150
 50. Rayleigh L (1883) On the crispations of fluid resting upon a vibrating support. *Phil Mag* 16:50
 51. Romeiras F, Romeiras F, Bondeson A, Ott E, Antonsen TM, Grebogi C (1989) Quasiperiodic forcing and the observability of strange nonchaotic attractors. *Phys Scr* 40:442–444
 52. Semler C, Paidoussis M (1996) Nonlinear analysis of the parametric resonances of a planar fluid-conveying cantilevered pipe. *J Fluids Structures* 10:787–825
 53. Stephenson A (1908) On a new type of dynamical stability. *Mem Proc Manch Lit Phil Soc* 52:1–10
 54. Suzuki M, Morita N, Edagawa I, Yamamoto S, Taga H, Akiba S (1995) Reduction of gordon-haas timing jitter by periodic dispersion compensation in soliton transmission. *Electron Lett* 31:2027–2035
 55. Tabor M (1989) *Chaos and Integrability in Nonlinear Dynamics: An Introduction*. Wiley, New York
 56. Thomsen J (1995) Chaotic dynamics of the partially follower-loaded elastic double pendulum. *J Sound Vibr* 188:385–405
 57. Thomsen J (2003) *Vibrations and stability: advanced theory, analysis, and tools*, 2nd edn. Springer, New York
 58. Tondl A, Ruijgrok T, Verhulst F, Nabergoj R (2000) *Autoparametric Resonance in Mechanical Systems*. Cambridge University Press, Cambridge
 59. Truman M, Galán J, Champneys A (2008) An example of difference combination resonance. In preparation
 60. Turytsyn S, Shapiro E, Medvedev S, Fedoruk M, Mezentssev V (2003) Physics and mathematics of dispersion-managed optical solitons. *CR Physique* 4:145–161
 61. Umbanhowar PB, Melo F, Swinney HL (1996) Localized excitations in a vertically vibrated granular layer. *Nature* 382:793–796
 62. Vanderbauwhede A (1990) Subharmonic branching in reversible-systems. *SIAM J Math Anal* 21:954–979
 63. van der Pol B, Strutt M (1928) On the stability of the solutions of mathieu's equation. *Phil Mag* 5:18–38, sp. Iss. 7th Series
 64. van Noort M (2001) The parametrically forced pendulum. a case study in $1\frac{1}{2}$ degree of freedom. Ph D thesis, RU Groningen
 65. Verhulst F (2000) *Nonlinear Differential Equations and Dynamical Systems*. Springer, New York
 66. Wiesenfeld K, Moss F (1995) Stochastic resonance and the benefits of noise: from ice ages to crayfish and squids. *Nature* pp 33–36
 67. Wiggins S (2003) *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, 2nd edn. Springer, New York
 68. Yakubovich VA, Starzhinskii VM (1987) *Parametric Resonance in Linear Systems*. Nauka, Moscow, in Russian
 69. Zakharov V, Shabat A (1971) Exact theory of two-dimensional self focussing and one-dimensional modulation of waves in nonlinear media. *Sov Phys JETP* 33:77–83