O

Observability (Deterministic Systems) and Realization Theory

Jean-Paul André Gauthier
Department of Electrical Engineering,
University of Toulon, Toulon, France

Article Outline

Glossary

Observability An observed (and eventually controlled) dynamical system is observable if two distinct initial conditions can be distinguished (via the observations) by choosing the control function.

Universal inputs A universal input is a control function allowing to distinguish between all initial conditions.

Observer An observer system is a device, given in general under the guise of a differential equation (or a differences equation in the discrete case), allowing to track asymptotically the state trajectory of the system, using only the controls and the observations.

Input-output map An input-output map is a mapping (for fixed initial condition) which to "control functions" associates "output functions". It is in general assumed to be "causal" in some sense.

Realization A realization of an input-output map is a (controlled) nonlinear system realizing the given input-output map. A realization (system) is said to be minimal if it is controllable and observable.

Definition of the Subject

Observability analysis, design of nonlinear observers and realization of input-output maps are subjects of central interest in control theory and systems analysis. Related to the synthesis of observer systems is the very important question of "dynamic output stabilization": usually in practice a stabilizing feedback law is applied to the system via the estimation of the state provided by some observer device. Also, the topic is strongly connected with filtering theory, including the standard linear Kalman filter but also nonlinear filtering theory. Realization of some input-output behavior covers the practical idea of modeling systems by differential equations on the basis of input-output experiments (identification).

Introduction

In this article, we discuss the basic concepts and methods in observability, observation and realization theories. The area is so large that there are thousands of contributions. We provide a nonexhaustive limited list of references which is certainly far from complete, but corresponds to our taste: an entirely subjective selection. We focus on the continuous finite dimensional case, but there are very important developments for systems governed by PDE's, and for discrete time systems.

In this continuous, finite dimensional context, we chose the geometric setting, however there are other possibilities (algebraic setting, formal power series, Volterra series, ...).

For more details, we provide a list of books of significant interest dealing with the topics.

After setting the general definitions, we consider briefly linear systems for which the theory has been well established for a long time, the pioneers being Kalman and Luenberger.

Then we state some important results from the geometric nonlinear observability theory, the most significant contributions being undoubtedly those of Hermann and Krener [12] and Sussmann [25,26,27]. Also, contrarily

to the case of linear systems, the observability of a system depends on the control applied to it. The existence of universal controls is a very important point, clarified by Sussmann [28]. We state the main result. Concerning observability in an analytic-geometry setting, there are also interesting and important results by Bartoziewicz.

The next part of the paper is devoted to realization theory, where mostly two problems may be considered:

1. Given a nonlinear system, find a minimal realization;
2. Given some input-output mapping, find a realization of it (it will be minimal by construction).

The most important contribution in this setting is that of Jakubczyk [14,15]. In fact, it follows a basic idea of Kalman, first for finite automata and second for linear systems. We like Jakubczyk's approach since in particular, it contains very naturally the linear case. To our knowledge, this natural approach has not been used (in the nonlinear framework) for practical identification of nonlinear systems. However, it is rather clear that interesting developments are possible. Moreover, it is not so hard to show complete equivalence between this geometric approach and the formal power series approach.

The contribution of Crouch [4] about realization of finite Volterra series is also important, original and involves a lot of geometric considerations. We just refer to the original paper.

Realizing or approximating a system by a bilinear or state linear one is an important question in view of the observer synthesis problem. We state some results on the subject. In particular, there is an important geometric representation theorem (by bilinear systems) due to Fliess and Kupka [10], that we explain.

After these theoretical considerations, we go to a more practical topic: observers. Besides the linear case, there are several contributions on nonlinear observers synthesis (sliding modes, high gain, ...). Here, we focus on two natural generalizations of the linear results:

1. The output injection method (the equivalent for observability of feedback linearization) due mostly to Isidori, Krener, Respondek [19,20];
2. The use of the deterministic version of the linear Kalman's filter: it applies to bilinear systems, that are popular also by several approximation results (Fliess and Jacob in particular [13]).

Preliminaries

Surprisingly in the nonlinear case controllability plays a role in the observability properties of a system. It is the reason for the title of the next section.

Nonlinear Systems Under Consideration and Controllability

We consider nonlinear systems (Σ) of the usual form:

$$(\Sigma)\quad \begin{cases} \dot{x} = f(x,u)\,, \quad u \in U\,, \\ y = h(x)\,. \end{cases} \tag{1}$$

Here, the state x lives either in $\mathbb{R}^n$ or more generally in some n-dimensional differentiable manifold X. The set U of values of control u is some arbitrary set (for simplicity, we assume a closed subset of $\mathbb{R}^l$, may be finite). The observation function h takes values in $\mathbb{R}^p$. To simplify, we will consider the analytic case only, i. e. f and h are real-analytic w.r.t. x. In the special cases where U has some analytic structure (i. e. $U = R^l$ for instance) we assume joint real analyticity w.r.t. (x,u).

If W is an open subset of X, we denote by $\Sigma|_W$ the system Σ restricted to W.

Some initial condition $x_0 \in X$ being fixed, such a system Σ defines (via Cauchy existence and uniqueness Theorem) an input-output mapping $P_\Sigma : L^\infty[U] \to AC[\mathbb{R}^p]$, $u(\cdot) \to y(\cdot)$, where $L^\infty[U]$ is the set of functions defined on semi-open intervals $[0, T_u[$ (depending on the control $u(\cdot)$). Possibly $T_u = +\infty$. Here $AC[\mathbb{R}^p]$ denotes the set of absolutely continuous functions over some interval $[0, T_y[$ possibly depending on the output function $y(\cdot)$. Moreover, $T_y = \inf\{T_u, e(u,x_0)\}$, where $e(u,x_0)$ is the explosion time of the solution of (1) associated with the initial condition x_0, and the control $u(\cdot)$.

Particular cases of systems under consideration are the usual linear systems (L), bilinear systems (B) or state-linear systems (LX):

$$\begin{cases} (L) & \dot{x} = Ax + Bu,\ y = Cx;\ X = \mathbb{R}^n,\ U = \mathbb{R}^l\,, \\ (B) & \dot{x} = Ax + Bx \otimes u,\ y = Cx;\ X = \mathbb{R}^n,\ U = \mathbb{R}^l, \\ (LX) & \dot{x} = A(u)x,\ y = Cx;\ X = \mathbb{R}^n\,. \end{cases} \tag{2}$$

In these formulas, $A, B, C, A(u)$ are linear. Of course, in the case of a linear system (L), with initial condition $x_0 = 0$, the input-output mapping P_L is a linear mapping.

Our system Σ is said to be "**controllable**" if the Lie algebra $Lie(\Sigma)$ of smooth vector fields on X generated by the vector fields f_u, $u \in U$ (where $f_u(x) = f(x,u)$) has dimension n at each point of X.

Also, we say that a system Σ is **symmetric** if $\forall\, u \in U$, $\exists\, v \in U$ s. t. $f_v = -f_u$, and Σ is **complete** if all the vector fields f_u, $u \in U$, are complete.

The following fact is standard, for analytic systems. A system is controllable iff:

1. The accessibility set $A(x_0)$ of $x_0 \in X$, i. e. the set of points that can be reached from x_0 by some trajectory of Σ, **in positive time**, has open interior in X, whatever $x_0 \in X$.
2. The orbit $O(x_0)$ of $x_0 \in X$, i. e. the set of points that can be joined to x_0 by some continuous curve which is a concatenation of trajectories of Σ **in positive or negative time**, is equal to X, whatever $x_0 \in X$.

Moreover in 1, 2 above, it is enough to restrict to piecewise constant control functions. Also, if Σ is symmetric, $O(x_0) = A(x_0)$, $\forall\, x_0 \in X$.

Definition and Characterization of Observability, Minimal Systems

Here, $C^\omega(X)$ denotes the vector space of real analytic functions over X. First, let $\Theta \subset C^\omega(X)$ denote the "**observation space of** Σ", i. e. the smallest vector subspace of $C^\omega(X)$ containing the p components $h_i(\cdot)$ of the output function h and closed under Lie derivation L_{f_u} in the direction of the vector fields f_u, $u \in U$. Then, Θ is also closed under Lie derivation in the direction of vector fields in $Lie(\Sigma)$ and Θ is generated as a real vector space by the functions $(L_{f_{u_r}})^{k_r}(L_{f_{u_{r-1}}})^{k_{r-1}} \cdots (L_{f_{u_1}})^{k_1} h_i$.

Definition 1 The observability distribution Δ of Σ is the distribution $\ker(\mathrm{d}\Theta)$ formed by the kernel of the one-forms $\mathrm{d}\theta, \theta \in \Theta$. The system Σ is said to be rank-observable if the distribution Δ is trivial. This fact is also called the "observability rank condition".

The important fact relating the observability and controllability properties is that the observability distribution Δ **has no singularities as soon as** Σ **is controllable**: the rank of Δ is preserved along trajectories of vector fields f_u. Moreover, it is clear that Δ is involutive, hence integrable by Frobenius's Theorem. Leaves of Δ are **levels** of Θ.

Definition 2 (Indistinguishability and weak indistinguishability relations) Let I be the binary relation over X defined by $x_0^1 I x_0^2$ if for any (piecewise constant) control $u(\cdot)\colon [0, T_u[\to U$ such that $e(u, x_0^1) = e(u, x_0^2) = T_u$, then the corresponding output functions $y_1(t), y_2(t)$ from both initial conditions x_0^1, x_0^2 are equal, $t \in [0, T_u[$. The relation I is called the indistinguishability relation for Σ. If V is an open subset of X, we denote by I_V (V-indistinguishability relation) the indistinguishability relation for the restriction $\Sigma|_V$. The weak-indistinguishability relation, denoted by I^w is the equivalence relation associated with the foliation of X generated by Δ.

The indistinguishability relation is an equivalence relation as soon as Σ is complete. It is not an equivalence relation in general. Hence in general, V-indistinguishability also is not equivalence over V.

Definition 3 The system Σ is said to be observable if the relation I is the trivial relation. It is said to be weakly observable if for all $x_0 \in X$, there is a neighborhood W of x_0 such that for each neighborhood V of x_0, $V \subset W$, $I_V(x_0) = x_0$.

Then weak observability means that **locally**, we can find inputs such that the initial conditions are distinguished by the observations, in arbitrarily short time. Observability means just that distinct initial conditions can be distinguished by observations. The system Σ being observable, analytic, this can be done in arbitrary short time.

In view of realization theory, we say that Σ is **minimal** if it is both controllable and observable. We say that it is **weakly minimal** if it is controllable and weakly observable.

Definition 4 A universal input for Σ is an input $u(\cdot)$, that distinguishes among any pair of distinct states in arbitrarily short time.

Observers

For a system Σ of the form (1) (that we assume to be observable) an observer is a system of the form:

$$\begin{cases} \dot{z} = F(z, y, u)\,, \\ \hat{x} = H(z, u)\,, \end{cases} \tag{3}$$

where $z \in Z$, some manifold. The observer system is fed by $y(t)$ and $u(t)$, the output and input of Σ. The mapping $H\colon Z \times U \to X$, and we require that, for a large set of initial condition z_0 for z, the output $\hat{x}(t)$ tracks asymptotically the state $x(t)$ of the system, i. e. at least,

$$\lim_{t \to +\infty} d(\hat{x}(t), x(t)) = 0\,, \tag{4}$$

where d is some (Riemannian) metric over X. In general, there are additional requirements on the rate of convergence to zero of the **estimation error** $\varepsilon(t) = d(\hat{x}(t), x(t))$ (such as exponential convergence, with arbitrary exponential rate).

Of course even without such additional requirements, this definition is very vague and not serious at all. It has to be made more precise, depending on the context. There are mostly two types of problems:

- This definition depends on the metric d. It may happen that $\varepsilon(t)$ goes to zero for some Riemannian metric d, although it goes to infinity for some other metric d'.

Also, the state variables z or x may explode in finite time. Therefore, in general it is reasonable to require (4) only for trajectories of Σ that remain in a given compact subset of X for all positive times. In that case, the usual convergence requirements becomes independent of the Riemannian metric d.

- One cannot expect to observe unobservable systems. Therefore, one has to require convergence for "good" inputs only.

Abstract Definition of an Input-Output Map

We define the topological group G (resp. the topological semi group S) of extended (resp. positive time) **piecewise constant** controls as follows: typical elements of G and S are words of the form:

$$\check{u}(\check{t}) = (u_k, t_k)\cdots(u_1, t_1)\,, \tag{5}$$

where $u_i \in U$ and $t_i \in \mathbb{R}$ (resp. $\mathbb{R}_+$). The operation over G and S is the concatenation of words. We consider also the neutral element ε: $\check{u}(\check{t})\varepsilon = \varepsilon\check{u}(\check{t}) = \check{u}(\check{t})$. We define the equivalence relation $\sim$ over G and S as being generated by the relations:

$$\begin{cases} (u, 0) \sim \varepsilon\,, \\ (u, s)(u, \theta) \sim (u, s + \theta)\,. \end{cases} \tag{6}$$

We consider the quotient spaces $G := G/\!\sim$, $S := S/\!\sim$. Both are embedded with the topology co-induced by the maps:

$$\check{u}(\cdot)\colon \mathbb{R}^k \to G \quad (\text{resp. } (\mathbb{R}_+)^k \to S)\,.$$

For $\theta \in \mathbb{R}_+$ and $\check{u}(\check{t}) = (u_k, t_k)\cdots(u_1, t_1) \in S$, we define

$$\begin{aligned} \theta * \check{u}(\check{t}) &= (u_{r+1}, \theta - \eta_r)(u_r, t_r)\cdots(u_1, t_1) \\ &\qquad\qquad \text{for } \theta \in [\eta_r, \eta_{r+1}[\,, \\ \eta_r &= t_1 + \cdots + t_r\,, \\ \theta * \check{u}(\check{t}) &= \check{u}(\check{t}) \quad \text{for } \theta \geq \eta_k\,. \end{aligned}$$

A real mapping: $P\colon D \subseteq G \to \mathbb{R}$ (resp. $S \to \mathbb{R}$) with open domain D is said to be analytic if, for all $\check{u}(\check{t}) \in D$, the mapping $\check{t} \to P(\check{u}(\check{t}))$ is analytic at $\check{t}$ as a mapping $\mathbb{R}^k \to \mathbb{R}$.

The domain D of P: $D \subseteq S \to \mathbb{R}$ is said to be "star-shaped" if $\theta * a \in D$ for all $\theta \in \mathbb{R}_+$ and $a \in D$.

Denote $\check{B}(\check{s}) = ((\check{b}_m(\check{s}_m), \ldots, \check{b}_1(\check{s}_1)) \in G^m$ (resp. S^m), with

$$\check{b}_i(\check{s}_i) = \left(b_{i_{n_i}}, s_{i_{n_i}}\right)\cdots\left(b_{i_1}, s_{i_1}\right), \quad b_{i_j} \in U,$$

and set:

$$\begin{aligned} \Psi^{\check{B}(\check{s})}_{\check{u}(\check{t})} &= \left(\Psi^{\check{b}_m(\check{s}_m)}_{\check{u}(\check{t})}, \ldots, \Psi^{\check{b}_1(\check{s}_1)}_{\check{u}(\check{t})}\right)\,, \\ \Psi^{\check{b}_i(\check{s}_i)}_{\check{u}(\check{t})} &= P(\check{b}_i(\check{s}_i)\check{u}(\check{t}))\,. \end{aligned}$$

The rank of P is defined as

$$\operatorname{rank}(P) = \sup_{k, \check{B}(\check{s}), \check{u}(\check{t})} \operatorname{rank} D_{\check{t}} \Psi^{\check{B}(\check{s})}_{\check{u}(\check{t})}\,,$$

where $D_{\check{t}}$ means the differential w.r.t. $\check{t} \in \mathbb{R}^k$, and all the arguments belong to the possible domain defined by the domain D of P.

Definition 5 An (abstract) input-output mapping P is an analytic mapping, from some open and star-shaped subset $D \subset S$, with finite rank.

An extension P^+ of an analytic mapping P is an analytic mapping such that $\operatorname{dom}(P) \subset \operatorname{dom}(P^+) \subset S$ and $P = P^+|_{\operatorname{dom}(P)}$ (restriction of P^+ to $\operatorname{dom}(P)$).

Remark 6 Given a pointed nonlinear system (Σ, x_0) where Σ is of the form (1) and $x_0 \in X$, it is clear that the associated input-output mapping defines an abstract input-output mapping, the rank of which is the dimension n of the state space.

Linear Systems

The simplest case for observability, design of observers and realization theory is the linear case.

Given a linear system (L) from (2) the following results are standard and more or less obvious:

- The observability property is independent of the control $u(\cdot)$ applied to the system, i.e (L) is observable iff it is observable for some fixed arbitrary control $u(\cdot)$.
- The observability distribution Δ is a field of constant planes, given by $\Delta = \cap_{i=1}^{n-1} \ker(CA^i)$. Then Σ is observable iff $\operatorname{rank}(\Delta) = 0$. This condition is known as the **observability rank condition**.
- If (L) is observable the following device (**Luenberger observer**):

$$\begin{cases} \dot{z} = (A - \Omega C)z + \Omega y + Bu\,, \\ \Omega\colon \mathbb{R}^n \to \mathbb{R}^p\,, \quad z \in \mathbb{R}^n\,, \end{cases} \tag{7}$$

is an **arbitrary exponential rate observer**, i. e. the matrix Ω can be chosen in such a way that the matrix $A - \Omega C$ has arbitrary spectrum, which implies:

$$\begin{aligned} \|\varepsilon(t)\| &= \|z(t) - x(t)\| \\ &\leq k(\alpha)e^{-\alpha t}\|z_0 - x_0\| = k(\alpha)e^{-\alpha t}\|\varepsilon_0\|\,, \end{aligned} \tag{8}$$

where $\alpha > 0$ is arbitrary, and k is some polynomial in α, independent of Ω.

- Any linear system restricts to a controllable one on some subspace, and is mapped to an observable one, by the canonical projection $\Pi : \mathbb{R}^n \to \mathbb{R}^n/I$ (where I is the indistinguishability relation from Definition 2).
- Let $Y(t)$ denote an "impulse response" ($Y(t) : \mathbb{R}^l \to \mathbb{R}^p, t \geq 0$). The input-output map is the causal linear mapping $P : u(\cdot) \to y(\cdot) = Y * u$, where $*$ denotes the convolution of (positive time) signals. Then assume that (as a formal power series) $Y(t) = \sum_{k=1}^{\infty} G_k \frac{t^k}{k!}$, and let $\mathcal{H}$ denote the infinite block-Hankel matrix constructed from the sequence of blocks $G_1, G_2, \ldots, G_k, \ldots$
  Then, $Y(t)$ is the impulse response of a linear system (L) iff $\mathcal{H}$ has finite rank n.

Observability of Nonlinear Systems

What is clear is that if a system is rank-observable, then it is weakly observable. This is due to a Baker–Campbell–Hausdorf like formula, valid for piecewise constant controls $\check{u}(\check{t})$:

$$y(t) = \sum (L_{f_{u_k}})^{r_k} \cdots (L f_{u_1})^{r_1} h(x_0) \frac{t_k^{r_k} \cdots t_1^{r_1}}{r_k! \cdots r_1!} . \qquad (9)$$

Indeed by real analyticity, if $y_1(t) = y_2(t)$, all the terms $(L_{f_{u_k}})^{r_k} \cdots (L f_{u_1})^{r_1} h(x_0^1)$ and $(L_{f_{u_k}})^{r_k} \cdots (L f_{u_1})^{r_1} h(x_0^2)$ are equal, which contradicts the rank assumption, for x_0^1, x_0^2 close enough.

Conversely, assume that Σ is controllable and not rank-observable. Then, the observability distribution Δ is constant rank, integrable, nontrivial. Leaves of Δ are levels of Θ. By the same formula (9) points of such leaves are indistinguishable. Therefore Σ is not weakly observable. Then, the following theorem holds:

Theorem 7 *A controllable system Σ is weakly observable iff it is rank-observable.*

The other important result (Sussmann [28]) is:

Theorem 8 *If Σ is observable, there is a universal input. Moreover, the set of universal inputs is generic.*

Realization Theory

Minimal Realizations Given a Realization

We are given a realization i. e. a pointed system (Σ, x_0), $x_0 \in X$. In fact, the results follow from the Sussmann's theorem on quotient manifolds: a closed equivalence relation $\mathcal{R}$ differentiably passes to the quotient (i. e. quotient is a manifold and canonical projection is submersive) if there are enough complete vector fields that respect $\mathcal{R}$. We apply this theorem to the indistinguishability relation $\mathcal{R} = I$ in the case of a complete and controllable system. Then all vector fields of $Lie(\Sigma)$ respect I. This is exactly Sussmann's requirement, so that not only there is a quotient manifold and canonical mapping is submersive, but moreover vector fields of $Lie(\Sigma)$ pass to the quotient. Also, the elements of Θ obviously pass to the quotient.

If Σ is not controllable, then, as a first step, we can use the standard Hermann–Nagano Theorem to restrict to a (controllable) leaf of the distribution $Lie(\Sigma)$. Then, we have a similar theorem to the one of the linear case.

Theorem 9 *If Σ is complete, then we can restrict to the leaf of $Lie(\Sigma)$ containing $x_0 \in X$ to get a controllable system. Passing to the quotient manifold by the indistinguishability relation I, we get a minimal realization. Moreover, two minimal realizations are unique up to a diffeomorphism of the state spaces.*

For complete systems, there is an interesting refinement of this theorem. A realization is said to be **weakly-minimal** if it is controllable, weakly observable. It turns out that the equivalence relation I^w associated to Δ meets also Sussmann's conditions. It follows that the system goes to the quotient, and we get a weakly minimal realization $\hat{\Sigma}$ with state space $\hat{X}$. We can apply the previous Theorem 9 to $\hat{\Sigma}$ to get again the (unique) minimal realization Σ_m of Σ, with state space X_m. The following theorem is almost obvious.

Theorem 10 *$\hat{X}$ is a covering space of X_m. Moreover, any covering space of X_m determines a weakly-minimal realization of Σ, by a trivial lifting procedure.*

In particular, there is (up to diffeomorphisms) a single simply-connected weakly-minimal realization.

Note that in fact the relation I^w is the same relation as: $x_0^1 I^w x_0^2$ if there is a continuous curve $\gamma : [0, 1] \to X$ connecting x_0^1 to x_0^2 and for $r, s \in [0, 1]$, $\gamma(r) I \gamma(s)$.

Minimal Realizations Given an Abstract Input-Output Map

The set of controls U being given, we consider an abstract input-output map defined over the whole group G (domain $D = G$). Note that this is the case in particular for the input-output mappings determined by a complete symmetric system.

In that case we have the following theorem, due to Jakubczyk [14].

Theorem 11 *An abstract input-output mapping with domain G has a unique minimal realization, which is complete.*

Remark 12 The finite rank assumption for the input-output mapping is a generalization of the finite rank assumption of the Hankel matrix of the linear case. It is also the analog of certain finite rank assumptions appearing in the formal power series approach of Fliess, or in the Volterra-kernels approach.

Remark 13 There is one ugly detail in this theory: in general, we do not get a paracompact manifold as the state space X.

The idea for the proof of the theorem is very simple: we consider the subgroup H of G defined by $H = \{a \in G | P(ca) = P(c), \forall c \in G\}$. Then, the state space will just be $X = G/H$. The finite rank condition implies that X has the structure of a Hausdorff analytic manifold. The output function h is defined by $h(gH) = P(g)$. The vector-field f_u is defined via its one parameter group: $\exp(tf_u)(gH) = \Pi((u,t)g)$, where $\Pi : G \to G/H$ is the canonical projection.

A more practical result is the following: if we assume that the set U of values of the control **is a finite set**, then the following global result (containing a local one) can be proven.

Theorem 14 *Assume U is finite, then, a necessary and sufficient condition for P to have a realization (weakly-minimal) is that P has an extension P^+ with star-shaped domain D^+. The state space X of this realization is Hausdorff, paracompact.*

In the general analytic case with infinite U, existence of certain **local** realizations only, can be proved.

Bilinear or State-Linear Realization

This point will be extremely important for the problem of constructing observer systems (Sect. "Observers"). A system is said to be control affine if the vector fields f_u form an affine family w.r.t. u. The single control case ($l = 1$) is just the case $f(x,u) = f(x) + g(x)u$ where f and g are two vector fields on X. Note that a bilinear system is just a state linear system, which is moreover affine in the controls.

A state linear realization (LX, x_0) from Eq. (2) is said to be minimal if it is observable and controllable in the following sense: the orbit of x_0 is not contained in a strict subspace of $\mathbb{R}^n$ (the smallest such subspace would be automatically invariant under all the operators $A(u), u \in U$). First, it is rather simple to show that any pointed state-linear system (LX, x_0) has a minimal state-linear realization. Of course, the additional property to be bilinear is hereditary.

Note that for state-linear systems, the observation space is a (finite-dimensional) vector space of linear forms over X. It turns out that this finite dimensionality condition is in fact a necessary and sufficient condition. This is a very important result from Fliess and Kupka [10]:

Theorem 15 *Assume that Σ has a finite dimensional observation space Θ. Then, Σ is embeddable in a state-linear system. In other terms (Σ, x_0) has a state linear (minimal) realization.*

The proof is very easy. It is enough to take:

- $X = \Theta^*$ (dual space of Θ),
- For $\varphi \in \Theta^*$, $C_i\varphi = \varphi(h_i)$, $i = 1, \ldots, p$,
- $A(u) = (L_{f_u})^*$ (transpose of L_{f_u}),
- The initial state $\hat{x}_0$ meets $\hat{x}_0(\varphi) = \varphi(x_0)$ for $\varphi \in \Theta$.

Besides the fact that this result allows one to solve the observer problem for such systems, "truncating" in some manner the observation space is a way to approximate systems by state-linear ones, and to get approximate observers.

An interesting particular case where this theorem applies is the case of systems with polynomial observation h and state-linear dynamics:

$$\begin{cases} \dot{x} = A(u)x\,, \\ y = P(x)\,, \end{cases}$$

where P is some polynomial mapping. It is clear that Θ is finite-dimensional. More generally, if we start with a system with state-linear dynamics, we can uniformly approximate h on compact sets by a polynomial mapping to get a state-linear realization (and later on, an approximate observer device).

State-Linear Skew-Adjoint Realization

Here, for the sake of simplicity in the exposition we limit ourselves to the single output case $p = 1$.

This section describes some particular cases and some generalizations of the results of the previous section, in view of synthesis of observers with a method presented in Subsect. "Observers for Skew-Adjoint State Linear Systems".

For some reason that will be made clear in the Subsect. "Observers for Skew-Adjoint State Linear Systems" we would like to know when it is possible to embed a system (or to have a realization of a system) into a skew-symmetric, or more generally skew-adjoint, state-linear one.

This means that all the matrices $A(u)$ are skew-symmetric w.r.t. the usual scalar product over the state space $\mathbb{R}^n$ of the realization. By Theorem 15, a necessary condition for the nonlinear system Σ (minimal and complete) to be embeddable is that Θ be finite dimensional and hence the group of diffeomorphisms of X generated by the vector fields f_u be a Lie group G. One could think that a necessary condition is that G be a compact Lie group. This is not the case as the following example shows:

$$\begin{cases} \dot{x} = u\,, & x, u \in \mathbb{R}\,, \\ y = \cos(x) + \cos(\alpha x)\,, & \text{where } \alpha \text{ is irrational}\,, \end{cases}$$

since the group of diffeomorphisms is $\mathbb{R}$, while a skew symmetric embedding does exist.

The proper condition is given by the following theorem:

Theorem 16 *The system Σ (complete, minimal) can be embedded into a state-linear skew symmetric system iff:*

1. $\dim(\Theta) < \infty$ *(from which it follows that G is a Lie group),*
2. *The observation function $h(x)$ lifts over G into $\tilde{h}$ (in a natural way), an almost periodic function over G.*

Recall that an almost periodic function over G is a function that prolongs into a continuous function over the Bohr compactification G^{b} of G [5]. The two conditions of Theorem 16 are equivalent to the fact that G is a Lie group and $\tilde{h}$ is a finite linear combination of coefficients[1] of unitary irreducible finite dimensional representations of G.

If G is "embeddable in a compact group", i. e. if G is the semi-direct product of a compact group by a finite dimensional real vector space then, any h can be approximated in some sense by an almost periodic one.

Actually, a special interesting case is the following: the system Σ is such that $X = G$, a compact Lie group, and the vector fields f_u are right invariant vector fields over G. We can take h as any continuous function $h\colon G \to \mathbb{R}$, and consider the abstract Fourier transform $\hat{h}$ of h. In fact, by Peter–Weyl's Theorem [3], h is a uniform limit over G of **finite** linear combinations of the form

$$h(g) = \sum_i \alpha_i \Phi_i(g)\,,$$

where $\Phi_i(g)$ is a coefficient of an irreducible (hence finite dimensional) unitary representation of G. This means that h has approximations h_m that converge uniformly to h over G, such that each system

$$(\Sigma_m) \begin{cases} \dot{g} = A(u)g\,, \\ y = h_m(g)\,, \end{cases}$$

has a state-linear minimal realization of the form:

$$\begin{cases} \dot{x} = A_m(u)x\,, \quad x \in \mathbb{C}^n\,, \quad A_m(u) \text{ is skew-adjoint}\,, \\ y = C_m x\,. \end{cases}$$

Hence **the input-output mapping of any right invariant system over a compact group can be approximated by the one of a skew-adjoint state-linear one**.

Now, let us consider again a (complete, minimal) system Σ, with finite dimensional Lie algebra, but the group G is not compact. In that case $\tilde{h}$ (a lift of h over G) can be approximated uniformly on any compact subset of G by a function h_m, which is a finite linear combination of "positive type" functions over G. This approximation result is known as the Gelfand–Raikov Theorem [5]. As a consequence we have the theorem:

Theorem 17 *The system*

$$(\Sigma_n) \begin{cases} \dot{g} = A(u)g\,, \\ y = h_m(g)\,, \end{cases}$$

has a (infinite dimensional) skew-adjoint state linear realization on a separable complex Hilbert space $\mathcal{H}$, i.e:

$$\begin{cases} \dot{\Psi} = A(u)\Psi\,, \\ y = \langle \Psi, \xi \rangle\,. \end{cases}$$

Here $\xi, \Psi \in \mathbb{H}$ and $\langle \cdot, \cdot \rangle$ is the scalar product over $\mathcal{H}$. All the operators $A(u)$ are densely defined, essentially skew-adjoint operators, infinitesimal generators of strongly continuous one parameter groups of unitary operators over $\mathcal{H}$.

With this result, in Subsect. "Observers for Skew-Adjoint State Linear Systems", we will be able to construct reasonable approximate observers for Σ.

Observers

Kalman's Observer for State-Linear Systems

This is just the deterministic version of the linear time-dependant Kalman filter. Therefore, inputs being known, it applies to state-linear systems (LX) from (2). Contrarily to linear systems, observability for those systems is not a property independent of the inputs: for some input $u(\cdot)$

[1] A coefficient of a representation is a coefficient of the matrix representing the representation operator in certain orthonormal basis.

it might be observable, for others it might not be. Clearly, if we want the observer to have some asymptotic property of convergence of the estimation error, it is reasonable to require that the input under consideration keeps a certain **minimum level of observability** when the time grows to infinity. It is natural to consider inputs living in the space $\mathcal{U} = L^\infty_{[0,\infty[,\mathbb{R}^p}$ of measurable U-valued bounded functions. For an input $u \in \mathcal{U}$ and for a real $a \geq 0$, set $u_a(t) = u(t+a)$. We denote by $\Phi_u(t)$ the matrix resolvent of the linear equation $\dot{\Phi}_u(t) = A(u(t))\Phi_u(t)$. Then for $T > 0$, the Gramm-observability matrix:

$$G_{u,T} = \int_0^T \Phi_u(t)^* C^* C \Phi_u(t)\, \mathrm{d}t\,, \tag{10}$$

where * stands for adjoint operator, measures observability in the following sense: the system is observable for $u\colon [0, T] \to U$ iff $G_{u,T}$ is positive definite. Hence there are several types of assumptions that are possible to express that $u\colon [0, +\infty[\to U$ keeps a certain level of observability when time passes. The most simple one is the following:

There are $\alpha, T, T_0 > 0$ such that for all $\theta \geq T_0$, $G_{u_\theta,T} \geq \alpha.Id_n$, where Id_n is the identity matrix. This condition means intuitively that, from time T_0 on, the input u has minimum observability level α on all time intervals of length T. Such an input could be called a "**persistent-excitation**" for Σ.

Then, the following theorem is just a restatement of the classical results about the deterministic version of the linear time-dependant Kalman's filter:

Theorem 18 ([18]) *The matrices Q and R being positive definite symmetric matrices with adequate dimensions, the Riccati system:*

$$\begin{cases} (1) \quad \dot{S} = -A(u(t))'S(t) - S(t)A(u(t)) + C^* R^{-1} C - SQS, \\ (2) \quad \dot{z} = A(u(t))z - S^{-1}C^* R^{-1}(Cz - y(t))\,, \end{cases} \tag{11}$$

is an asymptotic observer for persistent-excitations $u(\cdot)$. Convergence of the estimation error is exponential. The matrices $S(t)$ (as soon as the same holds for the initial condition S_0) live in the open cone of positive definite symmetric matrices.

Observers for Systems that are Injectable in a State-Linear One

Of course, the technique of the previous section applies stricto-sensu to such systems from Subsect. "Bilinear or State-Linear Realization".

The Output-Injection Idea

It turns out that both the Luenberger observer (7) for linear systems and the Kalman observer (11) for state-linear systems can be applied in more general nonlinear situations.

Assume that Σ is **linear "up to output injection"**, i. e.

$$(\Sigma) \left\{ \begin{array}{c} \dot{x} = Ax + \varphi(y,u) \\ y = Cx \end{array} \right\}\,, \tag{12}$$

or respectively that Σ is **state-linear (or bilinear) up to output injection**, i. e.

$$(\Sigma) \left\{ \begin{array}{c} \dot{x} = A(u)x + \varphi(y,u) \\ y = Cx \end{array} \right\}\,, \tag{13}$$

where φ (the output injection) is some nonlinear term depending on the output and input only. Then there are easy modifications of the Luenberger observer (resp. the Kalman's observer) that provide exactly the same results of convergence of the estimation error as for the corresponding systems without the output-injection term.

For case (12) we take the observer under the Luenberger-modified form:

$$\dot{z} = (A - \Omega C)z + \varphi(y,u) + \Omega(y - Cz)\,,$$

while for case (13) we take:

$$\begin{cases} \dot{S} = -A(u(t))'S(t) - S(t)A(u(t)) + C^* R^{-1} C - SQS \\ \dot{z} = A(u(t))z + \varphi(y,u) - S^{-1}C^* R^{-1}(Cz - y(t))\,. \end{cases}$$

To check the result it is enough to write the estimation error equation and to see that it is exactly the same as in the situation without output-injection.

For that reason, **it is important to characterize systems that can be put under the form of a linear or state-linear system up to output-injection**.

There is an industry around this question. It starts with the works of Isidori, Krener, Respondek [19,20]. The first result of this type is in the uncontrolled case. For an uncontrolled system

$$(\Sigma) \begin{cases} \dot{x} = f(x) \\ y = h(x)\,, \end{cases}$$

with single output ($p = 1$), consider the vector fields X_i defined by

$$L_{X_1}(L_f)^{i-1}h = \delta_{i,n}\,, \quad i = 1, \ldots, n\,,$$
$$\text{where } \delta \text{ is the Kronecker symbol}\,,$$
$$X_j = -[f, X_{j-1}]\,, \quad j = 2, \ldots, n\,,$$

The system Σ can be linearized up to a diffeomorphism

and an output injection iff the two following conditions are met [19]:

1. The family $\{\mathrm{d}h, \mathrm{d}L_f h, \ldots, \mathrm{d}(L_f)^{n-1}h\}$ has full rank n at all points of X.
2. $[X_k, X_m] = 0$ for $1 \leq k, m \leq n$.

Of course, this is a "local almost everywhere result" only.

There is also a lot of results on the problem of characterizing systems that are diffeomorphic to or **embeddable in state-linear systems up to output injection**. A significant result to the problem of embedding up to output injection is the one of Jouan [16].

Observers for Skew-Adjoint State Linear Systems

Again, to simplify the exposition we consider the single output case $p = 1$ only.

In this case we have a (minimal) state-linear realization which is also skew-adjoint, there is a construction of an observer which is **much simpler** than Kalman's one [no Riccati equation besides the prediction-correction Eq. (11), (2)]. Moreover this construction **extends to infinite-dimensional realizations**, a fact which allows it to treat any (complete minimal) system with finite dimensional Lie algebra.

To start, consider some skew-symmetric state linear system:

$$(LX)\begin{cases}\dot{x} = A(u)x\,, \quad A(u) \text{ skew-symmetric } \forall\, u \in U\,,\\ y = Cx\,,\end{cases} \tag{14}$$

We consider the following candidate observer system:

$$\dot{z} = A(u)z - rC^*(Cz - y)\,, \tag{15}$$

in which $r > 0$ is a parameter. The estimation error is $\varepsilon = z - x$:

$$\dot{\varepsilon} = (A(u) - rC^*C)\varepsilon\,.$$

Then it is not so hard to show that, if $u\colon [0, \infty[\to U$ is a "persistent excitation" of Σ in some sense (for instance in the sense of Subsect. "Kalman's Observer for State-Linear Systems", then we have:

$$\lim_{t\to+\infty} \|\varepsilon(t)\| = 0\,.$$

As a consequence, the systems with compact group G of diffeomorphisms (or with G semidirect product of compact group by vector space), admit also approximate observers, using the results of Subsect. "State-Linear Skew-Adjoint Realization".

It turns out that this method can be extended in a reasonable way to systems with (infinite dimensional) skew-adjoint state-linear realization. In particular, it is possible to construct approximate observers for all (complete minimal) systems with finite dimensional Lie algebra.

Consider a skew adjoint realization from Subsect. "State-Linear Skew-Adjoint Realization":

$$\begin{cases}\dot{\Psi} = A(u)\Psi\,,\\ y = \langle \Psi, \xi\rangle\end{cases}$$

on the (separable) Hilbert space $\mathcal{H}$. Then, the candidate observer device is:

$$\dot{\Lambda} = A(u)\Lambda - r\xi(\langle \Lambda, \xi\rangle - y(t))\,. \tag{16}$$

In fact, the persistency assumption cannot be of the same type as in the finite dimensional case. The reason is that the Gramm observability matrix $G_{u,T}$ is a compact operator in that case. Hence it cannot satisfy an inequality of the type $G_{u,T} \geq \alpha Id_{\mathbb{H}}$ since $\mathcal{H}$ is infinite dimensional.

Hence, the definition of a persistent excitation has to be replaced by one of the following type: there is a time $T > 0$ and a real sequence θ_n, $\theta_n \to +\infty$, with $\theta_{n+1} - \theta_n$ bounded, such that the translated inputs $u_{\theta_n}\colon [0, T] \to \mathbb{R}^l$ converge to u^* in the weak–* topology of $L^\infty_{[0,T],\mathbb{R}^l}$ (which topology is precompact over bounded sets) and u^* is a universal input for Σ on $[0, T]$.

This means also that a certain level of observability is preserved, on regularly spaced time intervals, while the time increases.

In that case, of course the result is weaker than in the finite dimensional case. We have only:

$$\text{weak–}\lim_{t\to+\infty} \varepsilon(t) = 0\,.$$

Future Directions

For observability and synthesis of observers, besides the improvement of the current methods (including sliding modes, high gain, ...) several directions have to be investigated more deeply, namely infinite dimensional systems, delay and hybrid systems.

For realization theory, and as a consequence identification theory, almost no "**practical result**" is known in the nonlinear context. However, we think interesting and consistent developments are possible, even starting from the apparently abstract theory outlined there. This is clearly the challenge for the future.

Bibliography

Primary Literature

1. Bartosiewicz Z (1995) Local observability of nonlinear systems. Syst Control Lett 25(4,1):295–298
2. Bartosiewicz Z (1998) Real analytic geometry and local observability. Proc Sympos Pure Math. Am Math Soc 64:65–72
3. Barut AO, Raczca R (1986) Theory of Group Representations and Applications. World Scientific, Singapore
4. Crouch PE (1981) Dynamical realizations of finite volterra series. SIAM J Control Opt 19:177–202
5. Dixmier J (1964) Les $\mathbb{C}^*$ algébres et leurs représentations. Cahiers Scientifiques, fasc. XXIX. Gauthier-Villars, Paris
6. Fliess M (1981) Fonctionnelles causales non linéaires et indéterminées non commutatives. Bull Soc Math France 109: 3–40
7. Fliess M (1983) Réalisation locale des systèmes non linéaires, algèbres de Lie filtrés transitives et séries génératrices non commutatives. Inventiones Mathematicae 71:521–537. Springer
8. Fliess M (1973) Sur la realization des systemes dynamiques bilineaires. CR Acad Sc Paris A 277:923–926
9. Fliess M (1980) Nonlinear realization theory and abstract transitive Lie algebras. Bull Amer Math Soc (NS) 2:444–446
10. Fliess M, Kupka I (1983) A finiteness criterion for nonlinear input-output differential systems. SIAM J Control Optim 21:721–728
11. Grizzle JW, Moraal PE (1995) Observer design for nonlinear systems with discrete-timemeasurements. IEEE Trans Autom Control 40(3):395–404
12. Hermann R, Krener AJ (1977) Nonlinear controllability and observability. IEEE Trans Autom Control 22(5):728–740
13. Jacob G, Hespel C (1991) Approximation of nonlinear dynamic systems by rational series. Theor Comp Sci Arch 79(1):151–162
14. Jakubczyk B (1986) Local realizations of nonlinear causal operators. SIAM J Control Optim 24(2):230–242
15. Jakubczyk B (1984) Réalisations locales des opérateurs causals non linéaires. Comptes rendus des séances de l'Académie des sciences. Série A 299(15):787–793
16. Jouan P (2003) Immersion of nonlinear systems into linear systems modulo output injection. SIAM J Control Optim 41(6):1756–1778
17. Kalman RE (1963) Mathematical description of linear dynamical systems. SIAM J Control Optim 1(2):152–192
18. Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. J Basic Eng Trans ASME 83:95–108
19. Krener AJ, Isidori A (1983) Linearization by output injection and nonlinear observers. Syst Control Lett 3(1):47–52
20. Krener AJ, Respondek W (1985) Nonlinear observers with linearizable error dynamics. SIAM J Control Optim 23:197–216
21. Luenberger D (1966) Observers for multivariable systems. IEEE Trans Autom Control 11(2):190–197
22. Luenberger D (1971) An introduction to observers. IEEE Trans Autom Control 16(6):596–602
23. Sontag E (1979) On the observability of polynomial systems, I: Finite-time problems. SIAM J Control Optim 17(1):139–151
24. Sussmann HJ (1973) Orbits of families of vector fields and integrability of distributions. Trans Amer Math Soc 180:171–188
25. Sussmann HJ (1975) A generalization of the closed subgroup theorem to quotients of arbitrary manifolds. J Diff Geom 10:151–166
26. Sussmann HJ (1974) On quotients of manifolds: a generalization of the closed subgroup theorem. Bull Amer Math Soc 80:573–575
27. Sussmann HJ (1976) Existence and uniqueness of minimal realizations of nonlinear systems. J Theor Comput Syst 10(1):263–284, 371–393
28. Sussmann HJ (1978) Single-input observability of continuous-time systems. J Theor Comput Syst 12(1):371–393
29. Yamamoto Y (1981) Realization theory of infinite-dimensional linear systems, Part I. Math Syst Theor 15:55–77

Books and Reviews

Isidori A (1995) Nonlinear Control Systems. Springer Communications and Control Engineering Series. Springer, New York

Kalman RE, Falb PL, Arbib MA (1969) Topics in Mathematical System Theory. Mc Graw Hill, New York

Khalil H (2002) Nonlinear systems, 2nd edn. Prentice Hall, Upper Slade River

Nijmeijer H, van der Schaft A (1990) Nonlinear Dynamical Control Systems. Springer

Rugh WJ (1981) Nonlinear System Theory, The Volterra/Wiener Approach. The Johns Hopkins University Press, Baltimore

Sontag ED (1998) Mathematical Control Theory: Deterministic Finite Dimensional Systems, 2nd edn. Springer, New York

Utkin VI (1992) Sliding Modes in Control and Optimization. Springer

Opinion Dynamics and Sociophysics

DIETRICH STAUFFER

Institute for Theoretical Physics, Cologne University, Cologne, Germany

Article Outline

Glossary

Cluster Clusters are sets of neighboring sites of the same type.

Ising model Each site carries a magnetic dipole which points up or down; neighboring dipoles "want" to be parallel.

Opinion dynamics How do people change opinions? Simulations usually ignore all details of the brain and represent the opinion by one or several numbers which can be changed due to contact with others.

Schelling model People belonging to different groups may produce segregated neighborhoods just by their personal preferences, not by outside force.

Sociophysics Application of methods from (mostly statistical) physics to human relations can be traced centuries backwards.

Definition of the Subject

Sociophysics is the study of social questions by physicists using their physics methods. In contrast to biophysics, it is a field which is not yet very well established. Opinion dynamics is one of the most widespread topics of sociophysics.

Introduction

The application of concepts from the natural sciences to social sciences, partly to be reviewed here, is at least 25 centuries old. Then the Greek philosopher Empedokles stated (according to J. Mimkes) that humans are like liquids: Some mix easily like wine and water, and others like oil and water refuse to mix. We start with the Schelling model of 1971, which implemented this idea, and its criticism (see ► Social Processes, Simulation Models of). Then we will review opinion dynamics in large populations, summarizing only shortly other aspects like self-organization of hierarchies or competition between human languages.

Humans do not like to be treated like a number, and indeed the human brain is much more complex than a binary variable (called "spin" by physicists) which is either $+1$ or -1. We do not deal here with the psychological processes of an individual but with mass psychology, and this author learned half a century ago in school that mass psychology is different from individual psychology: The law of large numbers averages out over individual fluctuations and makes general trends more clearly visible. Thus what we call today statistical physics plays a useful rule, and social scientists [15,36] have applied it, without knowing then that they dealt with an Ising model of ferromagnets.

The astronomer Halley, best known through his comet, tried to establish mortality tables already three centuries ago. Of course, the time of death of one given individual is usually difficult to predict but averaged over millions of people the statistical offices of many countries prepare regularly life tables which tell us how probable it is for a newborn child to live up to x years, provided there are no changes of the mortalities in the coming decades. Insurance for automobiles is another example: We do not want to produce accidents, but we know that they happen, and take precautions against their financial consequences. Thus the whole insurance industry is based on treating humans like numbers, ignoring their individuality.

Finally, human opinions are often fluctuating and ill-defined, but nevertheless in elections people cast one choice, out of a limited number of choices. And election results belong to those social data for which we have lots of accurate numbers, based on large populations.

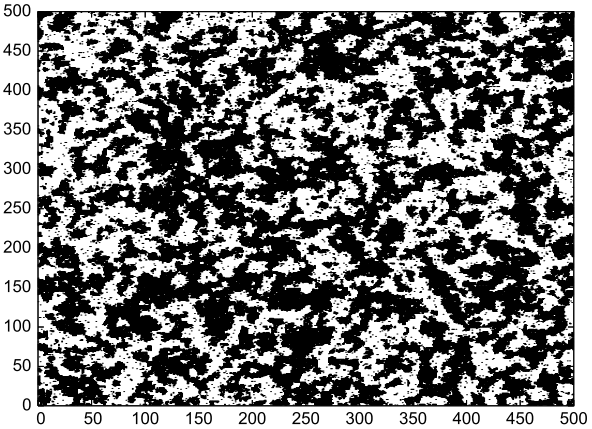
Thus it is not at all the merit (or ignorance) of physicists which treats humans like numbers; this method has a very long tradition and is an indispensable part of modern life.

Schelling Model

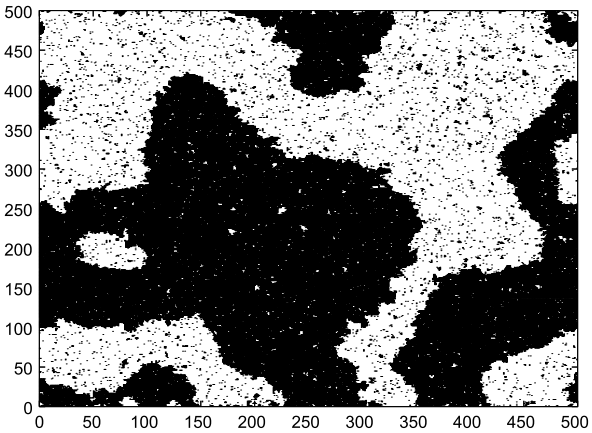
Ising Simulations

Following (but not citing) Empedokles, the later economics Nobel laureate Schelling [36] asked whether the racial segregation in American cities can emerge from intrinsic behavior of the individual people, instead of or in addition to extrinsic reasons like discrimination, rent differences, etc. In particular, can "black" ghettos in the predominantly "white" USA arise just because people prefer to have neighbors of their own group over neighbors from the other group? In many other countries we find many other types of residential segregation, based on religion, ethnicity, ... In physics, such a process is easily simulated through the two-dimensional Ising model, as shown in Figs. 1 and 2.

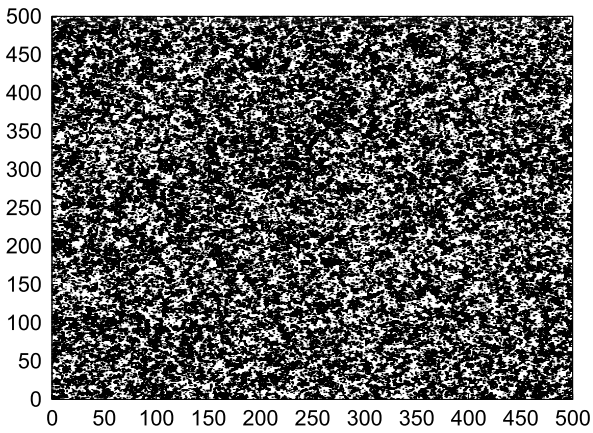
In this Ising model, each site on a square lattice carries a variable $S_i = \pm 1$, and each pair $\langle i, k \rangle$ of nearest neighbors produces an "energy" $-JS_iS_k$ with some proportionality constant J. The total energy E (= total unhap-



Opinion Dynamics and Sociophysics, Figure 1
Ising model after 20 Glauber kinetic steps per site on a 500 × 500 square lattice at $k_BT/J = 2$. We start from a random distribution of equally many black and white sites



Opinion Dynamics and Sociophysics, Figure 2
As Fig. 1 but after 2000 instead of 20 iterations



Opinion Dynamics and Sociophysics, Figure 3
As Fig. 2 but at $k_B T/J = 3$ instead of 2. Only small clusters and no large domains are formed. After 200 and 20,000 iterations the pictures look similar to this one made after 2000 iterations

piness) is the sum of this pair energy over all neighbor pairs of the lattice. In physics, different distributions of the "spins" S_i are realized with a probability proportional to $\exp(-E/k_B T)$ where T is the absolute temperature and k_B the Boltzmann constant. There is no need to worry about values for T, k_B, J since the only relevant quantity is the ratio $k_B T/J$, taken as 2 in these pictures. The "Glauber" kinetics is simulated on the computer by flipping a spin if and only if a random number between 0 and 1 is smaller than the probability $\exp(-\Delta E/k_B T)/[1+\exp(-\Delta E/k_B T)]$. The Fortran program of Algorithm 1 contains less than 40 lines and takes a few seconds.

Such models and programs are taught in courses on computational or theoretical physics all over the world; the model was published in 1925. If in the above flipping probability the denominator is omitted one gets the Metropolis kinetics. If instead of flipping one spin, we exchange two opposite spins, we get the Kawasaki dynamics. For Glauber or Metropolis, after very long times (measured by the number of sweeps through the lattice) one of the two possibilities dominates at the end, if T is not larger than the critical temperature T_c, with $2J/k_B T_c = \ln(1+\sqrt{2}) \simeq 0.88$ known since 1940. For Kawasaki dynamics the fraction of black sites remains constant, and we get two large domains. For higher temperatures above T_c only small clusters and no large domains are formed, Fig. 3.

In this Ising model, two neighboring spins have due to their interaction $-JS_i S_k$ a higher probability to belong to the same group than to belong to the two different groups. If the difference between these two probabilities is large enough, $T < T_c$, domain sizes can grow to infinity in an infinite lattice, Figs. 1 and 2, while only small clusters are formed in Fig. 3 for smaller differences in the probabilities, $T > T_c$. That these probabilities, controlled through $-J/k_B T$, lead to these different regimes, separated by a sharp phase transition at $T = T_c$, is not obvious from the definition of the interaction $JS_i S_k$, took physicists many years to find, and is typical of complex systems.

The social meaning of temperature T is not what we hear in the weather reports but an overall approximation for all the more or less random events which influence our decisions but are not explicitly included in the model. For residential segregation the model only counts how many neighbors of which group one has. But not all people of one group are alike, housing in different parts of a city costs different amounts of money, some parts are more beautiful then others, and job hunting may force us into a temporary residence of a new city which does not conform to our wishes. In this way, a positive temperature allows for rare moves which increase the energy, i. e. we move to a new residence where the neighborhood composition along makes us less happy. At zero temperature, the Ising model does not properly order into one or two "infinite" domains.

Schelling's Version and Later Improvements

Schelling [36] avoided probabilistic rules and thus counted neighbors $S_k = \pm 1$ at zero temperature. Then it does not matter if all neighbors or only a majority of them belong to the own group. Thus people are defined as happy if at least half of the neighbors belong to the own group, and as unhappy otherwise (i. e. if the majority belong to the opposite group). Unhappy people move to the nearest place

```
      parameter(L=500,Lmax=(L+2)*L)
      dimension is(Lmax),iex(9)
      byte is
      data t,max,ibm/2.00,2000,1/
      print *, L,max,ibm,t
      Lp1=L+1
      L2pL=L*L+L
      do 1 i=1,Lmax
         is(i)=-1
         ibm=ibm*16807
1        if(ibm.gt.0) is(i)=1
      do 2 ie=1,9
         ex=exp(-2*(ie-5)/t)
2        iex(ie)=(2.0*ex/(1.0+ex) - 1.0)*2147483647
      ibm=2*ibm+1
      do 3 mc=1,max
         do 4 i=Lp1,L2pL
            ie=5+is(i)*(is(i-1)+is(i+1)+is(i-L)+is(i+L))
            ibm=ibm*16807
4           if(ibm.lt.iex(ie)) is(i)=-is(i)
         mag=0
         do 6 i=Lp1,L2pL
6           mag=mag+is(i)
3        if(mc.eq.(mc/100000)*100000) print *, mc,mag
      do 5 i=Lp1,L2pL
         if(is(i).ne.1) goto 5
         iy = (i-1)/L
         ix=i-L*iy
         print *, ix, iy+1
5     continue
      stop
      end
```

Opinion Dynamics and Sociophysics, Algorithm 1
Simple Fortran program to produce pictures like Figs. 1 to 3

where they are happy. Since Schelling moved only one person (or family) at a time, and made no exchange of two people simultaneously as in Kawasaki kinetics, he introduced a large fraction of empty residences. Thus at each step, one unhappy person or family moves into the closest vacancy where life would be happy.

This model, and also many variants [16,36], fail to give large domains; only small clusters are seen. In reality, Harlem in Manhattan (New York), is not a cluster of a few houses but extends over many square kilometers. Thus the original version does not give the desired results. Large domains are formed if people also change residences if this brings no improvement [46] (hardly a realistic assumption), or at a finite temperature [41]. The latter paper also gives some alternatives to the Schelling model which also allow for large domains, and a simple example of a finite cluster where everybody is "happy" and which therefore never grows or dissolves on its own "will". More quantitative analyses of domain growth are given by [13, 27,39].

Much earlier and simpler is the zero-temperature version of Jones [15] who at each iteration removes a random fraction of the people and fills the vacancies with people who are there happy in the Schelling sense. This randomness, just like the finite temperature, leads to large domains as desired. Neither physicists nor social scientists have taken much note of [15]. The history of the Schelling model is an example how the lack of communication between disciplines has hampered progress in research, even very recently [41,46]. Only computational statistical physicists know everything. ([15] also mention a probabilistic version closer to the Ising model.)

For finite temperatures, [41] follow the above Glauber dynamics, but instead of an energy E uses a variable which is 0 or 1 depending on the happiness of the residents. Moving from one place to the other then depends exponentially

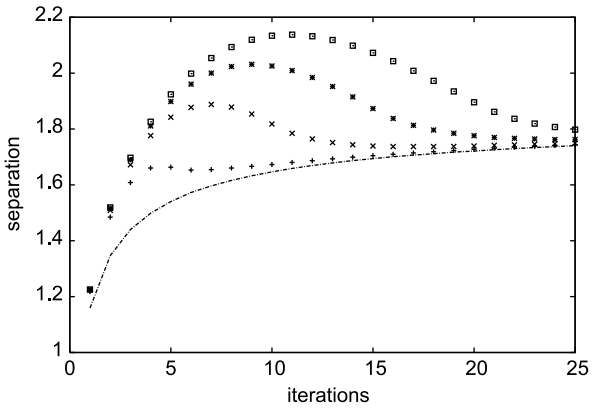
on the ratio of this variable to $k_B T$, instead of on the ratio $\Delta E/k_B T$. Many variants are possible, e. g. in the treatment of neutral cases [30,46] where the number of neighbors of both groups is exactly the same.

But we are on safer grounds and can use decades of physics research if we use the normal Ising model, or its generalization to Q different groups, the Q-state Potts model. Then Refs. [28,37] implemented a suggestion of Weidlich [47] that people slowly learn to live together with neighbors from the other group. Thus T not only takes into account the various accidents from outside the model, but also measures the tolerance: The higher T is the more are people willing to live in neighborhoods of the other group. In the limit $T = \infty$ the neighbors would not matter at all, for intermediate T, Fig. 3 showed small clusters but no large domains, and for low T the domains grow to infinite sizes on an infinite lattice. The learning suggested by Weidlich thus means that this parameter T (= temperature or tolerance) no longer is kept constant but slowly increases.

For an Ising model, [28] showed how an initial large domain dissolves if the temperature is slowly increased from below to above T_c. More realistically, for five (instead of only two) different groups in a modified five-state Potts model, [37] increased T from low to high values and showed that with a slow increase one has appreciable domain formation during intermediate times, while with a fast increase this segregation is mostly avoided, Fig. 4.

Instead of imposing a fixed temperature or tolerance T to everybody, one can also let it self-organize according to the neighborhood [30]. Or one may introduce two different T, one for tolerance against people of the other group, and the other for the random noise from events outside the model, like marriages, job losses, deaths [32].

Poor people cannot afford expensive housing. If we assume one of the two groups to be poor and the other to be rich, and if we assume that each residence is randomly either expensive or cheap, then we have a random-field Ising model [43]. This "field" gives the probability for the poor to select only cheap housing and for the rich to live in expensive residences. For intermediate lattice sizes and intermediate times, the field prevents the growth of infinite domains, and the clusters are the smaller the larger the field is [43].



Opinion Dynamics and Sociophysics, Figure 4
Amount of neighbors of the same type in a Potts model of five groups, normalized to unity for the initial random distribution. The temperature or tolerance increases from low to high values, slowly in the *top curves*, and fast in the *lower curves*; the latter mostly avoid the segregation into different group. (The *lowest line* holds for a constant high temperature.). From [37]

Opinion Dynamics

The following section describes several rules for simulated people to change opinions; each of these rules is applied again and again to these agents until some stationary or static state has been achieved.

Ising Model

Also for human opinions, one could use the Ising model of the previous section [9,19]; see also [47]. People can vote for or against the government or a new constitution, for one of two presidential candidates, or (using generalized models) for one out of $Q > 2$ different parties. Their neighbors on a lattice influence them in their vote, and in addition mass media may influence everybody in one direction. The latter effect can be modeled through an external "magnetic" field, Eq. (1b) in "Phase transitions ..." by this author in this encyclopedia. No motion of people needs to be taken into account, and the complications of Kawasaki kinetics (exchange of two people with opposing opinions) are not needed. Thus the Glauber program of the previous section still can be used, and we only refer here to the old generalization into the social impact model [24,25] and to a recent financial application [50].

Voter Model

Also quite old is the voter model [26]: Each person chooses between two opinions, by taking over the one of a randomly selected neighbor. One may rewrite this rule as stating that each person selects the opinion of the neighborhood, with a probability proportional to the number n of neighbors having that opinion. Thus in contrast to the Ising model where the probabilities depend exponentially on n, now they depend linearly on n. A final equilibrium (absorbing fixed point) is reached if everybody shares the

same opinion. The deviations from that final state can be measured by the magnetization (difference between the numbers for the two opinions) or energy (average number of neighbors having the opposite opinion). The time needed to reach the consensus increases with a power of the lattice size, and the exponent depends on the dimensionality. A nice and short review of the voter model, also on various networks, is given by the Majorca group [35].

Axelrod Model

Axelrod [1] wondered how different opinions or cultures may coexists even if people tend to become more alike in their beliefs. Looking at the above Ising Figs. 1, 2, 3, we see that due to finite time and/or finite temperature such coexistence of two opposing opinions is possible. But Axelrod generalized it not only to $Q > 2$ different possible opinions as in the Potts model of the above "Schelling Model" section, but also to F different questions. People may have one set of opinions on which political party they want to vote for, another set about what is the best football team, a third about recent cinema films, etc. This allows for Q^F different opinion sets on all F questions ("features"). Of course, one could generalize this model to the case where the number Q_f of possible choices is different for the different features f, allowing then for $\Pi_{f=1}^{F} Q_f$ instead of simply Q^F different sets of opinions.

Another aspect of the model takes into account that people prefer to talk to, or to make political coalitions with, others with whom they share many opinions. Thus the probability of one person to take over the different opinion of a neighbor is proportional to the number of features on which their opinions already agree. In the next subsection we will use a similar concept under the name of bounded confidence.

Whether a total consensus ("globalization") is reached or multiculturality persists depends on parameters: Small Q lead to consensus. Again, the Majorca group [35] reviewed the many follow-up papers on this Axelrod model.

Sznajd, Krause–Hegselmann and Deffuant Models

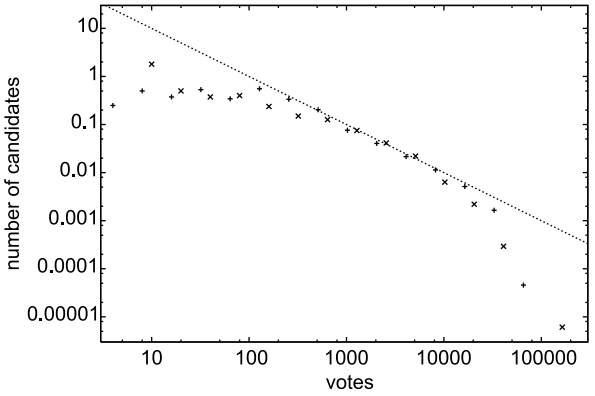
Much of the opinion dynamics research since 2000 centered on three different models S, KH and D, originally invented independently around that year: Sznajd [44] (S), Krause–Hegselmann [23] (KH) and Deffuant et al. [14] (D). They were also called missionaries, opportunists and negotiators by some computational physicists [42].

The S model is closest to the earlier models since it allows for Q discrete opinions, while KH and D use real opinions, e. g. between zero and one. S happens on a lattice or network while for KH and D everybody may interact with everybody. In the most widespread S version a pair of neighboring sites on a square lattice convinces its six neighbors of its opinion, if and only if the two opinions of the pair agree [29]; governments and parties usually lose support if their internal opinion differences make it to the headlines. For KH, the new opinion of a person is the arithmetic average over the opinion of the whole population. For D, each person selects randomly another person and then both move in their opinion towards each other by an amount proportional to their opinion difference.

In all three cases, "bounded confidence" applies: The KH agents average only over those people who differ from their own opinion by less than ϵ, and the D agents only select negotiation partners differing by less than ϵ from their own opinions. In both models $0 < \epsilon < 1$ is a fixed parameter. For S agents with $Q = 2$ such a rule makes no difference, but for $Q > 2$ one can modify the convincing rule such that only neighbors differing by at most ± 1 from the pair opinion adopt the pair opinion. Thus $1/Q$ for S plays the role of ϵ for D and KH. A rule similar to this bounded confidence was mentioned above for the Axelrod model [1].

In spite of the differences in their definitions, the results are quite similar for S, KH and D. For large ϵ or $Q \leq 3$ a complete consensus is usually reached; for small ϵ or $Q \geq 4$ different opinions may coexist forever. In addition to computer simulations, also analytical calculations were made [2,40] which agree with many aspects of the simulations. More results, also for opinions on more than one feature and agents sitting on scale-free networks [8], are summarized in [42].

One particular application is shown in Fig. 5: Various Brazilian election results for candidates in city councils



Opinion Dynamics and Sociophysics, Figure 5
Brazilian elections (x) and simulations of Sznajd model on Barabási–Albert networks (+); from [4]

showed great similarity if the number of candidates getting a given number v of votes is plotted against this v. Putting the S model with $Q = 1000$ candidates onto a scale-free network instead of a square lattice, excellent agreement of simulation and reality was found after the numbers were scaled by suitable factors. Also Indian elections were simulated this way [20], while proportional elections are different [11]. It would be nice to apply other opinion dynamics models to the same election problem. As usual in statistical physics, these studies can predict and simulate the shape of the distributions but not the winner in a specific election, just as we can predict the pressure of the air molecules around us but not where which molecule will be one minute from now.

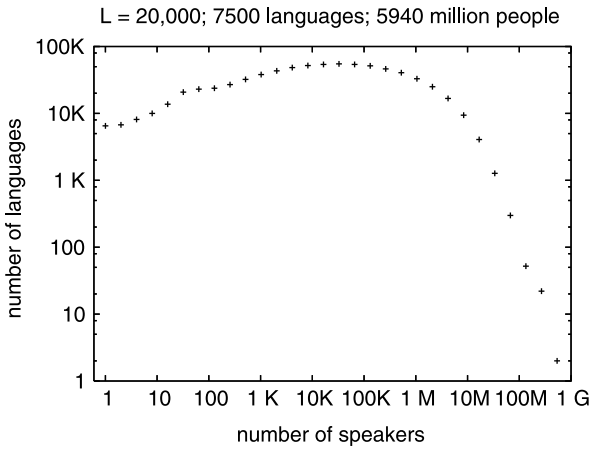
Galam Conservatism

Galam has published since many years theoretical models which may explain why reforms are very difficult and why a minority can stay in power. Usually these models are solvable analytically and assume that the population is divided into small groups of people which to the outside are represented by one person who follows the majority wish of the group. Several of these representatives form a supergroup, and this supergroup again decides according to the majority of the representatives in it. In this way an "infinite" hierarchy of people, groups, supergroups etc can be built. In the case of equally many voting for one choice as for the opposite choice, within one unit, that unit votes for the status quo. Starting with everybody having opinion -1, a very large majority of people must switch to opinion $+1$ before the top of the hierarchy finally also changes opinion [17]. We refer to [42] for a summary of more recent Galam papers, and to [18] for a more complete review.

Languages, Hierarchies and Football

Language Competition

Darwinian survival of the fittest is established biology, but similar concepts can be applied to human languages, bridging the gap to opinion dynamics. There are now thousands of different languages, and their "size" is the number of native speakers of that language. The size distribution extends from 1 (on the verge of extinction) to 10^9 (Mandarin Chinese). The grammar of a language [22] can be characterized by F features each of which can have Q different values, just as in Axelrod's model explained above. Features can change spontaneously or be taken over from a (neighboring) language; speakers of a small language give it up and learn a widespread language (as done with physics research publications since 1945); people migrate to other places and bring their language with them. All these processes can lead to the extinction of existing languages and the creation of new ones (by the branching of one language into several sub-languages.) The present language size distribution is roughly log-normal, with an enhancement at small sizes [21]. Similar languages form families, and the size distribution of families is a power law at intermediate sizes [49] (where the size is now the number of different languages belonging to that family).



Opinion Dynamics and Sociophysics, Figure 6
Simulated language size distribution on a 20,000 × 20,000 square lattice using a modified Viviane model [33]

Various computer simulations of this language competition have been made, mostly since 2003 and reviewed recently [38]; see also [10,12]. We only mention Fig. 6 from a modified Viviane model, which agrees well with the real language size distribution. For language families, the empirical statistics is worse [49] and one model also works well [38]. Good distributions were also obtained in a model which avoids dealing with individual speakers [45].

Self-Organization of Social Hierarchies

The elites of all countries and all times always had excellent reasons why they should be on top and others on the bottom. This holds even when the United Nations criticize the school system as violating human rights. In contrast, the Bonabeau model [7] explains social hierarchies as purely accidental, without any merit. People are put on a lattice, occupying a fraction p of all lattice sites and having an initial score of zero. Then they move randomly to neighboring sites, and whenever one person wants to move into the site occupied by another person, a fight erupts. The winner takes the contested site, the loser moves into (or stays at) the other site. Also, the winner adds one point and the

loser subtracts one point in its score, and in the future the agents with a positive score have a higher probability to win, those with a negative score have a lower probability to win. Slowly the history is forgotten, by reducing the score at each time step by, say, ten percent.

With some suitable feedback between the distribution of scores and the probability to win, a phase transition was simulated such that for p above some critical concentration, the standard deviation in the scores becomes positive for long times and large populations. For $p < p_c$ it fluctuates near zero, which means that everybody has close to a 50 percent chance to win. So, just by accident at a high population density some people rise to the top, and others fall to the bottom. However, the people on top (bottom) are not always the same; only the differences between top and bottom, not the people, remain the same. [3,31,48] are some of the more recent references in this field.

Football

Football (= soccer) is the world's most popular spectator sport, though in the author's city it is more a frustration. Randomness surely plays a role and makes it attractive. Can we explain all results just by chance, in the spirit of Bonabeau hierarchies? Assuming a constant probability to make a goal within one minute, the distribution of goals and victories is more narrow than in reality. If instead we assume that this probability varies from team to team, still no good agreement is found. Good agreement with reality is obtained only if correlations are taken into account [6], in the sense that a goal makes the scoring team happy, shocks the opposing team, and thus with an enhanced probability leads to another goal for the scoring team. Thus if we lose it is not just bad luck; it is also the referee's fault.

Future Directions

The Schelling model of Sect. "Schelling Model" is not the only case of missed opportunities because of a lack of cooperation between social sciences on the one side and physics, mathematics or computer science on the other side. The two books [5,42] were written without the authors of one book knowing of the preparation of the other book. One group of authors works in physics departments; none of the other group lists physics as institutional address. Nevertheless the two books show strong overlap in fields and methods covered, but little overlap in the literature cited. More interdisciplinary cooperation would help.

Bibliography

1. Axelrod R (1997) J Confl Resolut 41:203
2. Ben-Naim E, Krapivsky P, Redner S (2003) Physica D 183:190
3. Ben-Naim E, Vazquez F, Redner S (2006) Eur Phys J B 49:531
4. Bernardes AT, Stauffer D, Kertész J (2002) Eur Phys J B 25:123
5. Billari FC, Fent T, Prskawetz A, Scheffran J (2006) Agent-based computational modelling. Physica, Heidelberg
6. Bittner E, Nussbaumer A, Janke W, Weigel M (2007) Europhys Lett 78:58002 (2007)
7. Bonabeau E, Theraulaz G, Deneubourg JL (1995) Physica A 217:373
8. Caldarelli G (2007) Scale-free networks: complex webs in nature and technology. Oxford University Press, Oxford
9. Callen E, Shapero D (1974) Phys Today 27:23
10. Cangelosi A, Parisi D (eds) (2002) Simulating the evolution of language. Springer, New York
11. Castellano C, Fortunato S, Loreto V (2008) Statistical physics of social dynamics. arXiv:0710:3256, Rev Mod Phys (preprint)
12. Culicover P, Nowak A (2003) Dynamical grammar. Oxford University Press, Oxford
13. Dall'Asta L, Castellano C, Marsili M (2008) J Stat Mech L07002
14. Deffuant G, Amblard F, Weisbuch G, Faure T (2002) J Artif Soc Soc Simul 5(4):1
15. Dethlefsen E, Moody C (1982) Byte 7:178; Jones FL (1985) Aust NZ J Sociol 21:431
16. Fossett M (2006) J Math Sociol 30:185
17. Galam S (1990) J Stat Phys 61:943
18. Galam S (2008) Int J Mod Phys C 19:409
19. Galam S, Gefen Y, Shapir Y (1982) J Math Sociol 9:1
20. González MC, Sousa AO, Herrmann HJ (2004) Int J Mod Phys C 15:45
21. Grimes BF (2000) Ethnologue: Languages of the world, 14th edn. Summer Institute of Linguistics, Dallas. Available at www.ethnologue.org
22. Haspelmath M, Dryer M, Gil D, Comrie C (eds) (2005) The world atlas of language structures. Oxford University Press, Oxford
23. Hegselmann R, Krause U (2002) J Artif Soc Soc Simul 5(3):2
24. Hołyst JA, Kacperski K, Schweitzer F (2001) Annual Reviews of Computational Physics IX. World Scientific, Singapore, p 275
25. Latané B (1981) Am Psychol 36:343
26. Liggett TM (1985) Interacting particle systems. Springer, New York
27. Lim M, Metzler R, Bar-Yam Y (2007) Science 317:1540
28. Meyer-Ortmanns H (2003) Int J Mod Phys C 14:311
29. Milgram S, Bickman L, Berkowitz L (1969) J Pers Soc Psych 13:79
30. Müller K, Schulze C, Stauffer D (2008) Int J Mod Phys C 19:385
31. Naumis GG, del Castillo-Mussot M, Perez LA, Vazquez GJ (2006) Physica A 369:789
32. Ódor G (2008) Int J Mod Phys C 19:393
33. de Oliveira PMC, Stauffer D, Lima FWS, Sousa AO, Schulze C, Moss de Oliveira S (2007) Physica A 376:609
34. de Oliveira PMC, Stauffer D, Wichmann S, Moss de Oliveira S (2008) J Linguist 44:659
35. San Miguel M, Eguíluz VM, Toral R (2005) Comput Sci Engin 7:67
36. Schelling TC (1971) J Math Sociol 1:143
37. Schulze C (2005) Int J Mod Phys C 16:351
38. Schulze S, Stauffer D, Wichmann S (2007) Comm Comput Phys 3:271
39. Singh A, Vainchtein D, Weiss H. Schelling's segregation model: Parameters, scaling, and aggregation. Preprint
40. Slanina F, Lavička H (2003) Eur Phys J B 35:279
41. Stauffer D, Solomon S (2007) Eur Phys J B 57:473

42. Stauffer D, Moss de Oliveira S, de Oliveira PMC, Sá Martins JS (2006) Biology, sociology, geology by computational physicists. Elsevier, Amsterdam
43. Sumour MA, El-Astal AH, Radwan MA, Shabat MM (2008) Int J Mod Phys C 19:637; see also Emboloni F (preprint)
44. Sznajd-Weron K, Sznajd J (2000) Int J Mod Phys C 11:1157
45. Tuncay Ç (2007) Int J Mod Phys C 18:1641
46. Vinkovic D, Kirman A (2006) Proc Natl Acad Sci USA 103:19261
47. Weidlich W (2000) Sociodynamics; A systematic approach to mathematical modelling in the social sciences. Harwood Academic Publishers, 2006 reprint, Dover, Mineola
48. Weisbuch G, Stauffer D (2007) Physica A 384:542
49. Wichmann S (2005) J Linguist 41:117
50. Zhou WX, Sornette D (2007) Eur Phys J B 55:175

Optical Computing

Thomas J. Naughton[2,3], Damien Woods[1,4]
[1] Department of Computer Science, University College Cork, Cork, Ireland
[2] Department of Computer Science, National University of Ireland, Maynooth County Kildare, Ireland
[3] Oulu Southern Institute, University of Oulu, RFMedia Laboratory, Ylivieska, Finland
[4] Department of Computer Sience and Artificial Intelligence, University if Seville, Seville, Spain

Article Outline

Glossary

Coherent light Light of a narrow band of wavelengths (temporally coherent), and a light beam whose phase is approximately constant over its cross sectional area (spatial coherence). For example, coherent light can be produced by a laser.

Incoherent light Light which is not spatially coherent and not temporally coherent. For example, incoherent light is produced by a conventional light bulb.

Source A device for generating light.

Spatial light modulator (SLM) A device that imposes some form of spatially-varying modulation on a beam of light. An SLM may modulate the intensity, phase, or both, of the light.

Detector A device for sensing light.

Continuous space machine (CSM) A general optical model of computation that is defined in Sect. "Continuous Space Machine (CSM)".

Parallel computation thesis This thesis states that parallel time corresponds, within a polynomial, to sequential space, for reasonable parallel and sequential machines [29,52,74,98,126].

P, NP, PSPACE, NC Complexity classes, these classes are respectively defined as the set of problems solvable on polynomial time deterministic Turing machines; polynomial time nondeterministic Turing machines; polynomial space Turing machines; and parallel computers that use polylogarithmic time and polynomial hardware [97].

Definition of the Subject

An optical computer is a physical information processing device that uses photons to transport data from one memory location to another, and processes the data while it is in this form. In contrast, a conventional digital electronic computer uses electric fields (traveling along conductive paths) for this task. The optical data paths in an optical computer are effected by refraction (such as the action of a lens) or reflection (such as the action of a mirror). A principal advantage of an optical data path over an electrical data path is that optical data paths can intersect and even completely overlap without corrupting the data in either path. Optical computers make use of this property to efficiently transform the optically-encoded data from one representation to another, for example, to shuffle or reverse the order of an array of parallel paths, or to convolve the data in several arrays of parallel paths. Other advantages of optical computers include inherent parallelism and the ability to encode a two-dimensional spatial function in the cross-section of a single beam of light, higher bandwidths (in contrast to the free transmission of photons, electric fields generate noise in parallel conductors as they are pushed down their conductor), lower energy consumption (an argument deriving from the fact that optical computers in principle generate very little heat), easier circuit design, and lower latency in comparison to electrical transmission.

However, the property of non-interference of intersecting data paths means that it is not straightforward to

effect a switch or branch instruction in optics since in a vacuum the presence or absence of light in one data path cannot affect another path. In order to perform a conventional computation (e. g. solve a decision problem) optical computers invariably need to be equipped with an electronic interface, which would sense the presence or absence of light at some stage of the computation, and set the optical computer on a new course. Although this limitation has been addressed with varying levels of success through the development of nonlinear optical materials for all-optical switching and storage devices, it is generally accepted that if optical computers become mainstream, it will be through a symbiotic relationship with their extremely flexible digital electronic counterparts. Furthermore, currently there is no convincing alternative to using digital electronics for optical computer data input devices (liquid-crystal display panels, for example) and data output devices (digital cameras, for example).

Optical computing is an inherently multidisciplinary subject whose study routinely involves a spectrum of expertise that threads optical physics, materials science, optical engineering, electrical engineering, computer architecture, computer programming, and computer theory. Applying ideas from theoretical computer science, such as analysis of algorithms and computational complexity, enables us to place optical computing in a framework where we can try to answer a number of important questions. For example, which problems are optical computers suitable for solving? Also, how does the resource usage on optical computers compare with more standard (e. g. digital electronic) architectures? Furthermore, optical computing gives one an opportunity to apply computer theory on a completely new suite of machine models. In contrast to a number of other nature-inspired models of computation, optical computers have very real and immediate realization possibilities.

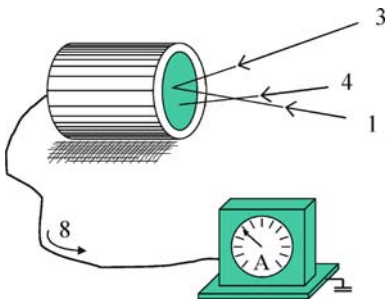
Traditionally, in optical information processing a distinction was made between signal/image processing through optics and numerical processing through optics, with only the latter (and often only the digital version of the latter) being called optical computing [44,73,85,142]. However, it was always difficult to clearly delineate between the two, since it was largely a question of the interpretation the programmer attached to the output optical signals. The most important argument for referring to the latter only as optical computing had to do with the fact that the perceived limits (or at least, ambitions) of the former was simply for special-purpose signal/image processing devices while the ambitions for the latter was general-purpose computation. Given recent results on the computational power of optical image processing architectures [91,131,136], it is not the case that such architectures are limited to special-purpose tasks. Furthermore, as the field become increasingly multidisciplinary, and in particular as computer scientists play a more prominent role, it is necessary to bring the definition of optical computing in line with the broad definition of computing. In particular, this facilitates analysis from the theoretical computer science point of view. The distinction between analog optical computing and digital optical computing is similarly blurred given the prevalence of digital multiplication schemes effected through analog convolution [73]. Our broad interpretation of the term optical computing has been espoused before [25].

Introduction

The three most basic hardware components of an optical information processing system are a source, a modulator, and a detector. A source generates the light, a modulator multiplies the light by a (usually, spatially varying) function, and a detector senses the resulting light. The simplest example of a modulator encoding a spatially varying function is a transparency (a sheet of clear plastic or photographic film) with an opaque pattern handwritten or printed onto it. When placed in the path of an advancing wavefront, which we define simply as being a wide beam of light, the modulator encodes its pattern onto this wavefront. The common liquid-crystal display projector is a programmable example of the same principle. Keeping this kind of system in mind, we now highlight some attributes of optical information processing systems.

Time Efficiency

Consider a light detector that converts incident light into an electrical current. Consider also an encoding scheme whereby the intensity in a beam of light represented a particular nonnegative integer. Further, assume there are no fluctuations in the light source output, that the encoding scheme is linear, and that the detector's response is linear. Then, the sum of two such nonnegative integers incident on the detector could be determined by measuring the detector's current. In fact, several nonnegative integers could be summed in this way, with a single measurement (see Fig. 1). (This concept is not unknown to designers of analog electrical ANNs. However, the important difference is that since the medium is free space, the practical fan-in limitations of Kirchoff Law summation [86] in analog electronics do not apply here.) Such an optical arrangement can find the sum of n nonnegative integers in $O(1)$ addition steps. On a model of a sequential digital electronic computer this would require $n - 1$ addition



Optical Computing, Figure 1
A light detector apparatus converts incident light into an electrical current. Multiple nonnegative integers, encoded in beams of light, can be summed in unit time

operations and even a parallel digital electronic machine with n or more processors requires $O(\log n)$ timesteps. Tasks that rely on scalar summation operations (such as matrix multiplication) would benefit greatly from an optical implementation of the scalar sum operation. Similarly, $O(1)$ multiplication and $O(1)$ convolution operations can be realized optically. In Sect. "Optical Computing and Computational Complexity" we formally describe the time efficiency of a broad class of optical computers. Very recently, an optics-based digital signal processing platform has been marketed that claims digital processing speeds of tera (10^{12}) operations per second [79].

Efficiency in Interconnection Complexity

As optical pathways can cross in free space without measurable effect on the information in either channel, high interconnection densities are possible with optics [20,27]. Architectures with highly parallel many-to-many interconnections between parallel surfaces have already been proposed for common tasks such as sorting [8,37,83,115]. Currently, intra-chip, inter-chip, and inter-board connections are being investigated for manufacturing feasibility [87].

Energy Efficiency

Electrical wires suffer from induced noise and heat, which increases dramatically whenever wires are made thinner or placed closer together, or whenever the data throughput is increased [87]. As a direct consequence of their resistance-free pathways and noise-reduced environments, optical systems have the potential to generate less waste heat and so consume less energy per computation step than electronic systems [21]. This has been demonstrated experimentally with general-purpose digital optical processors [59,116,117].

Coherence

Mutually spatially coherent optical wavefronts (such as from a laser) interfere with each other just as waves in a water tank do. In the theory of physical optics, coherent wavefronts can be described by, and thus can represent, a complex-valued function (both positive and negative values in each of the real and imaginary axes). Using the language of this theory to interpret optical phenomena permits the definition of (at least) three important information processing constant-time operations: spatial modulation, Fourier transformation, and signal squaring [55]. The ability to perform such operations has resulted in many constant-time optical implementations of standard convolution-based digital image processing tasks.

While incoherent light (such as from an ordinary lightbulb) has some advantages in terms of tolerances in misalignment of optical components and less susceptibility to certain types of noise, coherent light is more general (in that its mathematics can be used to describe both coherent and incoherent wavefronts). Incoherent wavefronts can be modeled as being nonnegative everywhere, and so only admit possibilities to directly represent nonnegative spatially-varying and temporally-varying functions.

Optical Image Processing

It has long been appreciated that spatial optical signals are the most natural means of representing continuous tone 2D signals. There are many positive aspects to processing information using these (sometimes unwieldy and always inaccurate) physical signals instead of the more accurate digital electronic representations of 2D signals. These include the ability to concurrently modify all parts of an image (spatial light modulation), the capability to substitute space computational complexity for time computational complexity when performing certain transformations [22,82,91,107,131] (such as constant-time Fourier transformation with coherent light), the potential significant energy savings [21] (in both creating the signal and effecting the computation), and the ease with which analog signals can be digitized or resampled at an arbitrary frequency for subsequent digital electronic handling. The most common applications of optical image processing are pattern recognition and numerical matrix computations (see Sect. "History" for elaboration).

Pattern recognition is one of the most commonly implemented signal, image, and information processing tasks. A significant number of the algorithms for these tasks involve a convolution operation, either as the principal operation (e. g. comparing two images on a pixel-by-pixel basis using correlation) or as part of necessary pre-

processing (e. g. edge detection prior to applying a Hough transform [64]). The inherent parallel nature of optical systems can be used to facilitate low time and space complexity implementations of the convolution operation, either by multiplication in the Fourier domain or by systolic action [26].

Overview of the Chapter

In this chapter, we focus on optical image processing as it is an optical computing paradigm that makes full use of the degrees of freedom afforded by optics. Other optical computing architectures that seek to emulate perfectly in step-by-step fashion the operations of digital electronic architectures (for example, architectures built upon all-optical flip-flops [34] and all-optical network routers [39]) occupy an important place in the taxonomy of optical computing. However, in terms of computational complexity, the analyzes of these digital optical computers are in many respects identical to that of the digital electronic counterparts they emulate.

In Sect. "History", we give a brief overview of the history of optical computing, commonly referred to as optical information processing. Section "History" also includes an overview of existing optical models of computation. This is followed in Sect. "Selected Elements of Optical Computing Systems" by a summary of the most important elements of optical computing that could be used to define the functionality of an optical model of computation. In Sect. "Continuous Space Machine (CSM)", we take a detailed look at a particular model of optical computing (the CSM) that encompasses most of the functionality that coherent optical information processing has to offer. We begin by defining the CSM and a total of seven complexity measures that are inspired by real-world (optical) resources. We go on to discuss how the CSM's operations could be carried out physically. Section "Example CSM Datastructures and Algorithms" contains some example datastructures and algorithms for the CSM. In Sect. "C_2-CSM" we motivate and introduce an important restriction of the model called the C_2-CSM, and in Sect. "Optical Computing and Computational Complexity" we describe a number of C_2-CSM computational complexity results, and their implications. We conclude with some future directions in Sect. "Future Directions".

History

It could be argued that the field of optical information processing began in earnest with the realization that spatially coherent laser light could be used to conveniently Fourier transform an image, allow one to modify the complex-valued spatial frequency components, and then inverse Fourier transform back to the spatial domain. This concept is called spatial filtering [35,36,78,95,121,123,124], it is a generalization that encompasses convolution and correlation operations, and it could be performed over two-dimensional (2D) images in constant time while limited in speed only by the refresh rates of the input and output devices. It first found application in the 1950s for parallel processing and analysis of the huge amounts of radar data produced at the time. The initial special-purpose spatial filtering systems performed optical Fourier transforms, performed image processing (for example, noise reduction and edge enhancement), and recognized patterns through correlation. The fundamentals of optical spatial filtering were formulated in that decade, and built upon previous work on optimum linear filtering in communication theory. Achieving the full potential of optical spatial filtering theory requires filters that are complex-valued, and a technique to obtain such filters was first proposed by VanderLugt [123,125]. The technique allows one to physically encode a complex-valued image on an amplitude-modulating SLM such as an LCD panel.

Research continued into this form of image-based computation. Many important image processing tasks were demonstrated at that time, from character recognition [4], to real-time tracking of moving objects [48,122], to telescope/microscope image deblurring [118]. Two important strands of this research at the time were the development of sophisticated pattern recognition algorithms and numerical computation using values encoded in the complex amplitude or intensity of the light.

Optical Pattern Recognition

In pattern recognition [6,13,16,17,24,31,66,72], effort focused on achieving systems invariant to scaling, rotation, out-of-plane rotation, deformation, and signal dependent noise, while retaining the existing invariance to translating, adding noise to, and obscuring parts of the input. Effort also went into injecting nonlinearities into these inherently linear systems to achieve wider functionality [70,71]. Improvements were made to the fundamental limitations of the VanderLugt filter, most notably the joint transform correlator architecture [130].

Optical correlators that use incoherent sources of illumination (both spatially and temporally) rather than lasers are also possible [14,15,42,100,143]. The simplest incoherent correlator would have the same basic architecture as that used for matched filtering. While coherent systems in principle are more capable than incoherent systems (principally because the former naturally represents complex

functions while the latter naturally represents real functions), incoherent systems require less precise positioning when it comes to system construction and are less susceptible to noise.

A common example of an optical correlator's use in practical systems involved it being used as a front end to a generalized hybrid object recognition system. The optical processing component would quickly and efficiently identify regions of interest in a cluttered scene and pass these on to the slower but more accurate digital electronic components for false-alarm reduction, feature extraction, and classification. Today, the matched filter and the joint transform correlator are the two most widely used optical pattern recognition techniques [45,63,84,140].

Trade-offs between space and time were proposed and demonstrated. These included time integrating correlators [125] (in contrast to the space integrating correlators mentioned thus far) and systolic architectures [18, 26,50] where, for example, the propagation of an amplitude-modulated pressure wave through an acousto-optic device naturally effects the required correlation lags [108]. In addition to pattern recognition, a common application for these classes of architectures was numerical calculation.

Analog Optical Numerical Computation

An important strand of image-based optical computation involved numerical calculations: analog computation as well as multi-level discrete computation. Matrix-vector and matrix-matrix multiplication systems were proposed and demonstrated [5,44,63,73,76,85,125]. The capability to expand a beam of light and to focus many beams of light to a common point directly corresponded to high fan-out and fan-in capabilities, respectively. The limitations of encoding a number simply as an intensity value (finite dynamic range and finite intensity resolution in the modulators and detectors) could be overcome by representing the numbers in some base. Significant effort went into dealing with carry operations so that in additions, subtractions, and multiplications each digit could be processed in parallel. Algorithms based on convolution to multiply numbers in this representation were demonstrated [73], with a single post-processing step to combine the sub-calculations and deal with the carry operations. Residue arithmetic was demonstrated as a viable alternative in which carry operations did not arise at all, and for which a matrix-vector multiplier was proposed [68], but of course conversion to and from residue format is necessary.

An application that benefited greatly from the tightly-coupled parallelism afforded by optics was the solving of sets of simultaneous equations and matrix inversion [1,23]. An application that, further, was tolerant to the inherent inaccuracies and noise of analog optics was optical neural networks [27,43,65,104] including online neural learning in the presence of noise [90].

Digital Optical Computing

The next major advances came in the form of optical equivalents of digital computers [67]. The flexibility of digital systems over analog systems in general was a major factor behind the interest in this form of optical computation [109]. Specific drawbacks of the analog computing paradigm in optics that this new paradigm addressed included no perceived ability to perform general purpose computation, accumulation of noise from one computation step to another, and systematic errors introduced by imperfect analog components. The aim was to design digital optical computers that followed the same principles as conventional electronic processors but which could perform many binary operations in parallel. These systems were designed from logic gates using nonlinear optical elements: semitransparent materials whose transmitted intensity has a nonlinear dependence on the input intensity. Almost always, the coherence of the light was not used in the computation. All-optical bistable devices acting as flip-flops were demonstrated. The field drew on many decades of research into fast on-off optical switching speeds which was heralding an explosion in optical fiber communications. The difficulties that the digital optical paradigm experienced included how to fully exploit the theoretical parallelism of optics within an optical logic framework, how to efficiently manufacture very large systems of cascaded nonlinear optical elements (for which integrated optics holds promise [51]), and the more fundamental mathematical problem of how to parallelize arbitrary algorithms in the first place to exploit the parallelism afforded by digital optics.

Digital optical computing was also proposed as an application of architectures designed originally for image-based processing, for example logic effected through symbolic substitution [11,12]. At the confluence of computing and communication, optical techniques were proposed for the routing of signals in long-haul networks [44,142]. This is a promising application given that most long-haul communications already use light in optical fibers, and the conversion from optical to electronic in order to switch, and them back to optical to retransmit, can be costly. Initial implementations followed the concept of an optoelectronic crossbar switch with n inputs and n outputs [109], while latterly more effort is now going into

all-optical packet switching in a single channel configuration [34,39,49].

Optical Models of Computation

As already discussed, optical computers were designed and built to emulate conventional microprocessors (digital optical computing), and for image processing over continuous wavefronts (analog optical computing and pattern recognition). Here we are interested in the latter class: optical computers that store data as images. Numerous physical implementations exist and example applications include fast pattern recognition and matrix-vector algebra [56,125]. There have been much resources devoted to designs, implementations and algorithms for such optical information processing architectures (for example see [5,22,40,44,56,77,82,85,90,107,125,142] and their references).

However the computational complexity theory of optical computers (that is, finding lower and upper bounds on computational power in terms of known complexity classes) had received relatively little attention when compared with other nature-insired computing paradigms. Some authors have even complained about the lack of suitable models [44,82]. Many other areas of natural computing (e. g. [2,58,62,80,88,89,99,112,139]) have not suffered from this problem. Even so, we discuss some optical computation research that is close to the goals of the theoretical computer scientist.

Reif and Tyagi [107] study two optically inspired models. The first model is a 3D VLSI model augmented with a 2D discrete Fourier transform (DFT) primitive and parallel optical interconnections. The second model is a DFT circuit with operations (multiplication, addition, comparison of two inputs, DFT) that compute over an ordered ring. Parallel time complexity is defined for both models in the obvious way. For the first model, volume complexity is defined as the volume of the smallest convex box enclosing an instance of the model. For the DFT circuit, size is defined as the number of edges plus gates. Constant time, polynomial size/volume, algorithms for a number of problems are reported including 1D DFT, matrix multiplication, sorting and string matching [107].

Feitelson [44] gives a call to theoretical computer scientists to apply their knowledge and techniques to optical computing. He then goes on to generalize the concurrent read, concurrent write parallel random access machine, by augmenting it with two optically inspired operations. The first is the ability to write the same piece of data to many global memory locations at once. Secondly, if many values are concurrently written to a single memory location then a summation of those values is computed in a single timestep. Essentially Feitelson is using 'unbounded fan-in with summation' and 'unbounded fan-out'. His architecture mixes a well known discrete model with some optical capabilities.

A symbolic substitution model of computation has been proposed by Huang and Brenner, and a proof sketched of its universality [11]. This model of digital computation operates over discrete binary images and derives its efficiency by performing logical operations on each pixel in the image in parallel. It has the functionality to copy, invert, and shift laterally individual images, and OR and AND pairs of images. Suggested techniques for its optical implementation are outlined.

In computer science there are two famous classes of problems called P and NP [97]. P contains those problems that are solvable in polynomial time on a standard sequential computer, while NP is the class of problems that are solvable in polynomial time on a nondeterministic computer. NP contains P, and it is widely conjectured that they are not equal. A direct consequence of this conjecture is that there are (NP-hard) problems for which we strongly believe there is no polynomial time algorithm on a standard sequential computer.

It is known that it is possible to solve any NP (and even any PSPACE) problem in polynomial time on optical computers, albeit with exponential use of some other, space-like, resources [131,133,135]. In Sect. "C_2-CSM and Parallel Complexity Theory", we describe how parallel optical algorithms can solve such problems.

Along with these rather general results, there are a number of specific examples of algorithms with related resource usage for NP-hard problems. Shaked et al. [110,111] design an optical system for solving the NP-hard traveling salesman problem in polynomial time. Basically they use an optical matrix-vector multiplier to multiply the (exponentially large) matrix of tours by the vector of intercity weights. They give both optical experiments and simulations. Dolev and Fitoussi [38] give optical algorithms that make use of (exponentially large) masks to solve a number of NP-hard problems. Oltean [94], and Haist and Osten [60], give polynomial time algorithms for Hamiltonian path, and traveling salesman problem, respectively, via light travelling through optical cables. As is to be expected, both suffer from exponential (space-like) resource use. Nature-inspired systems that apparently solve NP-hard problems in polynomial time, while using an exponential amount of some other resource(s), have been around for many years. So the existence of massively parallel optical systems for NP-hard problems should not really surprise the reader. Nevertheless, it is interesting

to know the computational abilities, limitations, and resource trade-offs of such optical architectures, as well as to find particular (tractable or intractable) problems which are particularly suited to optical algorithms.

Reif, Tygar and Yoshida [106] examined the computational complexity of ray tracing problems. In such problems we are concerned about the geometry of the an optical system where diffraction is ignored and we wish to predict the position of light rays after passing through some system of mirrors and lenses. They gave undecidability and PSPACE hardness results, which gives an indication of the power of these systems as computational models.

Selected Elements of Optical Computing Systems

If one is designing an optical model of computation, one will incorporate the functionality of a subset of the following selected elements (devices and functionality) of optical computing systems.

Sources

Lasers are a common source of illumination because at some levels they are mathematically simpler to understand, but incoherent sources such as light-emitting diodes are also used frequently for increased tolerance to noise and when nonnegative functions are sufficient for the computation. Usually, the source is monochromatic to avoid the problem of color dispersion as the light passes through refracting optical components, unless this dispersion is itself the basis for the computation.

Spatial Light Modulators

It is possible to encode a spatial function (a 2D image) in an optical wavefront. A page of text when illuminated with sunlight, for example, does this job perfectly. This would be called an amplitude-modulating reflective SLM. Modulators can also act on phase and polarization, and can be transmissive rather than reflective. They include photographic film, and electro-optic, magneto-optic, and acousto-optic devices [5,44,56,73,85]. One class of note are the optically-addressed SLMs, in which, typically, a 2D light pattern falling on a photosensitive layer on one side of the SLM spatially varies (with an identical pattern) the reflective properties of the other side of the SLM. A beam splitter then allows one to read out this spatially-varying reflectance pattern. The liquid-crystal light valve [41,57,69,128,129] is one instance of this class. Other classes of SLMs such as liquid-crystal display panels and acousto-optic modulators allow one to dynamically alter the pattern using electronics. It is possible for a single device (such as an electronically programmed array of individual sources) to act as both source and modulator.

Detectors and Nature's Square Law

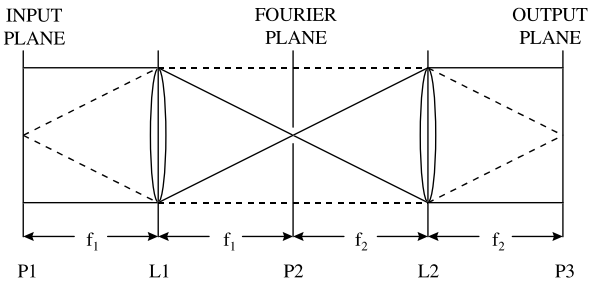
Optical signals can be regarded as having both an amplitude and phase. However, detectors will measure only the square of the amplitude of the signal (referred to as its intensity). This phenomenon is known as Nature's detector square law and applies to detectors from photographic film to digital cameras to the human eye. Detectors that obey this law are referred to as square-law detectors. This law is evident in many physical theories of light. In quantum theory, the measurement of a complex probability function is formalized as a projection onto the set of real numbers through a squaring operation. Square-law detectors need to be augmented with a interferometric or holographic arrangement to measure both amplitude and phase rather than intensity [19,47], or need to be used for multiple captures in different domains to heuristically infer the phase.

Since it squares the absolute value of a complex function, this square law can be used for some useful computation (for example, in the joint transform correlator [130]). Detectors most commonly used include high range point (single pixel) detectors such as photodiodes, highly sensitive photon detectors such as photomultiplier tubes, and 1D and 2D array detectors such as CCD- or CMOS-digital cameras. Intensity values outside the range of a detector (outside the lowest and highest intensities that the detector can record) are thresholded accordingly. The integration time of some detectors can be adjusted to sum all of the light intensity falling on them over a period of time. Other detectors can have quite large light sensitive areas and can sum all of the light intensity falling in a region of space.

Lenses

Lenses can be used to effect high fan-in and fan-out interconnections, to rescale images linearly in either one or two dimensions, and for taking Fourier transforms. In fact, a coherent optical wavefront naturally evolves into its Fresnel transform, and subsequently into its Fourier transform at infinity, and the lens simply images those frequency components at a finite fixed distance.

VanderLugt [125] has derived an expression for the coherent optical Fourier transform, given here in 1D for convenience, using the Fresnel transform of a complex-valued function $f(x)$ positioned in the front focal plane of a convex lens (plane P1 in Fig. 2), which is illuminated from the back by a plane wave of constant amplitude and phase. In terms of the physical coordinate ξ, the signal at the back



Optical Computing, Figure 2

Optical spatial frequency filtering using two convex lenses with a plane wave illuminating the input from the left. Lenses L1 and L2 have focal lengths of f_1 and f_2, respectively. If we assume that the physical dimensions of the lenses allow all diffracted orders to pass, the output is a transposed (and rescaled if $f_1 \neq f_2$) but otherwise identical version of the input. Any modification to the light field in the Fourier plane results in a spatial frequency filtered output

focal plane P2 can be written as

$$F(\xi) = \sqrt{\frac{\mathrm{i}}{\lambda L}} \int_{-\infty}^{\infty} f(x) \exp\left(\mathrm{i}2\pi x\xi/\lambda L\right) \mathrm{d}x \,, \tag{1}$$

where λ is the wavelength of the illumination and L is the focal length of the lens. Rewriting so it is a function of the spatial frequency variable α (α is a measurement of radians per unit distance) gives the common equation

$$F(\alpha) = \sqrt{\frac{\mathrm{i}}{\lambda L}} \int_{-\infty}^{\infty} f(x) \exp\left(\mathrm{i}2\pi\alpha x\right) \mathrm{d}x \,, \tag{2}$$

where $\xi = \lambda L\alpha$, and where we ignore the architecture-specific scaling constant. The Fourier transform in optics can be formed under a wide variety of conditions, and not just with a plane wave and not just in the focal plane of the lens [125]. This formalism adopts the paraxial approximation: that the distance between planes P1 and P2 is very much greater than the extent of the information in P1, thus avoiding the need for curved opposing surfaces in planes P1 and P3. When a Fourier transform is detected directly, only its square, called the power spectrum, is recorded. As mentioned, holography [19,47] overcomes this.

Interference and Complex Addition

Although photons, being bosons, do not interact with each other, coherent wavefronts can be made to interact at a detector. The addition (called superposition) of complex-valued wavefronts at a measurement is termed interference. Optically, interference is the same phenomenon as diffraction and is responsible for the formation of optical Fourier transforms. The linearity property is a useful property when analyzing coherent optical phenomena. If several images are coplanar then the optical field at their common Fourier plane is the superposition of their frequency spectra. This superposition, or interference, of complex-valued signals can be regarded as a pointwise addition of the complex amplitudes of the images.

Incoherent wavefronts are nonnegative everywhere. The addition of several incoherent wavefronts is linear in the intensity of each of those wavefronts. The addition of coherent wavefronts is linear in their complex amplitudes.

Image Copying and Combining

Images can be copied using optically-addressed SLMs or by dividing the optical energy along two paths using a beam splitter [33,125,130]. Beam splitters can also be used to combine several images to make them coplanar.

Multiplication of Images

When a signal's Fourier spectrum $F(\alpha, \beta)$ is coplanar with a transparency that encodes a second Fourier spectrum $H(\alpha, \beta)$, and if their centers are coincident, the complex signal in the plane immediately behind the transparency can be described as the pointwise multiplication of the two spectra

$$G(\alpha, \beta) = F(\alpha, \beta)H(\alpha, \beta) \,. \tag{3}$$

The signal $G(\alpha, \beta)$, which is also a frequency spectrum, could be inverse Fourier transformed to reveal a suitably correlated, convolved, or spatial frequency filtered original signal. A significant proportion of analog optics' role in the area of computation through filtering concerns convolution and those signal processing operations derived from it. Convolution filters are used extensively in the digital signal processing world to perform such tasks as deblurring, restoration, enhancement, and recognition [73]. The possibility of performing constant time convolution operations using coherent light is a promising concept.

To an approximation, optical systems can be regarded as both linear and shift-invariant. This is the basis for their convolution capabilities. Referring to Eq. (3), one can see that $H(\alpha, \beta)$ acts as a spatial frequency filter, altering the frequency content of the signal F. It could be used as a simple band-pass filter, damping the high-frequency components (noise suppression) or low frequency components (edge enhancement), or used to modulate the frequency spectrum with a more sophisticated spatial function. Mathematically, convolution with a mask A is equivalent to a frequency-domain multiplication with the Fourier transform of A.

Phase shifts can be introduced to a coherent wavefront by adding by a constant phase value. These could be constant shifts to effect numerical subtraction [44], or time-varying using a piezoelectric transducer mirror [138].

Other Elements of Optical Computation

A mirror changes the direction of the wavefront and simultaneously reflects it along some axis. A phase conjugate mirror [33] returns an image along the same path at which it approached the mirror.

In-plane rotation of an image by 180° can be accomplished using the apparatus in Fig. 2, a single lens, or even a pinhole camera. An out-of-plane tilt can be accomplished using a prism. In-plane flipping of an image (mirror image) can be accomplished using a prism, or using a beam splitter and some mirrors. Arbitrary in-plane rotation (with some tilting and translation, if required) can be achieved by combining several flip operations using a Dove prism or Pechan prism [96,119] or by combining several shearing operations [81]. Image rescaling can be accomplished using a combination of lenses (rescaling both dimensions identically), or using cylindrical lenses or an anamorphic prism (rescaling in one dimension only).

A prism or diffraction grating can be used to separate by wavelength the components of a multi-wavelength optical channel. For optical fiber communications applications, more practical (robust, economical, and scalable) alternatives exist to achieve the same effect [142].

Polarization is an important property of wavefronts, in particular in coherent optical computing, and is the basis for how liquid crystal displays work. At each point, an optical wavefront has an independent polarization value dependent on the angle, in the range $[0, 2\pi)$, of its electrical field. This can be independent of its successor (in the case of randomly polarized wavefronts), or dependent (as in the case of linear polarization), or dependent and time varying (as in the case of circular or elliptical polarization). Mathematically, a polarization state, and the transition from one polarization state to another, can be described using the Mueller calculus or the Jones calculus.

Photons can also be used for quantum computation, and quantum computers using linear optical elements (such as mirrors, polarizers, beam splitters, and phase shifters) have been proposed and demonstrated [28,75,101].

Continuous Space Machine (CSM)

For the remainder of this chapter we focus on an optical model of computation called the CSM. The model was originally proposed by Naughton [91,92]. The CSM is inspired by analog Fourier optical computing architectures, specifically pattern recognition and matrix algebra processors [56,90]. For example, these architectures have the ability to do unit time Fourier transformation using coherent (laser) light and lenses. The CSM computes in discrete timesteps over a number of two-dimensional images of fixed size and arbitrary spatial resolution. The data and program are stored as images. The (constant time) operations on images include Fourier transformation, multiplication, addition, thresholding, copying and scaling. The model is designed to capture much of the important features of optical computers, while at the same time be amenable to analysis from a computer theory point of view. Towards these goals we give an overview of how the model relates to optics as well as giving a number of computational complexity results for the model.

Section "CSM Definition" begins by defining the model. We analyze the model in terms of seven complexity measures inspired by real-world resources, these are described in Section "Complexity Measures". In Sect. "Optical Realization" we discuss possible optical implementations for the model. We then go on to give example algorithms and datastructures in Sect. "Example CSM Datastructures and Algorithms". The CSM definition is rather general, and so in Sect. "CSM Definition" we define a more restricted model called the C_2-CSM. Compared to the CSM, the C_2-CSM is somewhat closer to optical computing as it happens in the laboratory. Finally, in Sect. "Optical Computing and Computational Complexity" we show the power and limitations of optical computing, as embodied by the C_2-CSM, in terms computational complexity theory. Optical information processing is a highly parallel form of computing and we make this intuition more concrete by relating the C_2-CSM to parallel complexity theory by characterizing the parallel complexity class NC. For example, this shows the kind of worst case resource usage one would expect when applying CSM algorithms to problems that are known to be suited to parallel solutions.

CSM Definition

We begin this section by describing the CSM model in its most general setting, this brief overview is not intended to be complete and more details are to be found in [131].

A complex-valued image (or simply, image) is a function $f : [0, 1) \times [0, 1) \to \mathbb{C}$, where $[0, 1)$ is the half-open real unit interval. We let $\mathcal{I}$ denote the set of complex-valued images. Let $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$, $\mathbb{N} = \mathbb{N}^+ \cup \{0\}$, and for a given CSM M let $\mathcal{N}$ be a countable set of images that encode M's addresses. An address is an element of $\mathbb{N} \times \mathbb{N}$.

$h(i_1; i_2)$	: replace image at i_2 with horizontal 1D Fourier transform of i_1.
$v(i_1; i_2)$	: replace image at i_2 with vertical 1D Fourier transform of image at i_1.
$*(i_1; i_2)$	: replace image at i_2 with the complex conjugate of image at i_1.
$\cdot i_1, i_2; i_3$	: pointwise multiply the two images at i_1 and i_2. Store result at i_3.
$+(i_1, i_2; i_3)$	: pointwise addition of the two images at i_1 and i_2. Store result at i_3.
$\rho(i_1, z_l, z_u; i_2)$	: filter the image at i_1 by amplitude using z_l and z_u as lower and upper amplitude threshold images, respectively. Place result at i_2.
$[\xi\prime_1, \xi\prime_2, \eta\prime_1, \eta\prime_2] \leftarrow [\xi_1, \xi_2, \eta_1, \eta_2]$	: copy the rectangle of images whose bottom left-hand address is (ξ_1, η_1) and whose top right-hand address is (ξ_2, η_2) to the rectangle of images whose bottom left-hand address is $(\xi_1\prime, \eta_1\prime)$ and whose top right-hand address is $(\xi_2\prime, \eta_2\prime)$. See illustration in Fig. 4.

Optical Computing, Figure 3
CSM high-level programming language instructions. In these instructions $i, z_l, z_u \in \mathbb{N} \times \mathbb{N}$ are image addresses and $\xi, \eta \in \mathbb{N}$. The control flow instructions are described in the main text

Additionally, for a given M there is an *address encoding function* $\mathfrak{E} : \mathbb{N} \to \mathcal{N}$ such that $\mathfrak{E}$ is Turing machine decidable, under some *reasonable* representation of images as words.

Definition 1 (CSM) A CSM is a quintuple $M = (\mathfrak{E}, L, I, P, O)$, where

- $\mathfrak{E} : \mathbb{N} \to \mathcal{N}$ is the address encoding function,
- $L = ((s_\xi, s_\eta), (a_\xi, a_\eta), (b_\xi, b_\eta))$ are the addresses: *sta*, *a* and *b*, where $a \neq b$,
- I and O are finite sets of input and output addresses, respectively,
- $P = \{(\zeta_1, p_{1_\xi}, p_{1_\eta}), \ldots, (\zeta_r, p_{r_\xi}, p_{r_\eta})\}$ are the r programming symbols ζ_j and their addresses (p_{j_ξ}, p_{j_η}) where $\zeta_j \in (\{h, v, *, \cdot, +, \rho, st, ld, br, hlt\} \cup \mathcal{N}) \subset \mathcal{I}$.

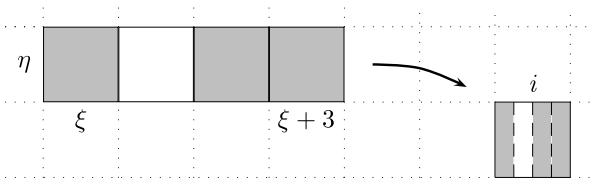
Each address is an element from $\{0, \ldots, \Xi - 1\} \times \{0, \ldots, \Upsilon - 1\}$, where $\Xi, \Upsilon \in \mathbb{N}^+$.

Addresses whose contents are not specified by P in a CSM definition are assumed to contain the constant image $f(x, y) = 0$. We interpret this definition to mean that M is (initially) defined on a grid of images bounded by the constants Ξ and Υ, in the horizontal and vertical directions respectively. The grid of images may grow in size as the computation progresses.

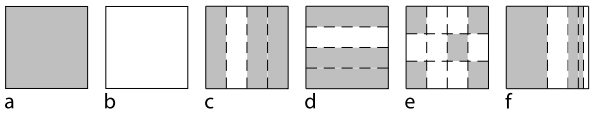
In our grid notation the first and second elements of an address tuple refer to the horizontal and vertical axes of the grid respectively, and image $(0, 0)$ is located at the lower left-hand corner of the grid. The images have the same orientation as the grid. For example the value $f(0, 0)$ is located at the lower left-hand corner of the image f.

In Definition 1 the tuple P specifies the CSM program using programming symbol images ζ_j that are from the (low-level) CSM programing language [131,136]. We refrain from giving a description of this programming language and instead describe a less cumbersome high-level language [131]. Figure 3 gives the basic instructions of this high-level language. The copy instruction is illustrated in Fig. 4. There are also **if/else** and **while** control flow instructions with conditional equality tests of the form $(f_\psi == f_\phi)$ where f_ψ and f_ϕ are *binary symbol images* (see Fig. 5a and b).

Address *sta* is the start location for the program so the programmer should write the first program instruction at *sta*. Addresses *a* and *b* define special images that are frequently used by some program instructions. The function $\mathfrak{E}$ is specified by the programmer and is used to map addresses to image pairs. This enables the programmer to choose her own address encoding scheme. We typically



Optical Computing, Figure 4
Illustration of the instruction $i \leftarrow [\xi, \xi + 3, \eta, \eta]$ that copies four images to a single image that is denoted *i*



Optical Computing, Figure 5
Representing binary data. The *shaded areas* denote value 1 and the *white areas* denote value 0. a Binary symbol image representation of 1 and b of 0, c list (or row) image representation of the word 1011, d column image representation of 1011, e 3 × 4 matrix image, f binary stack image representation of 1101. *Dashed lines* are for illustration purposes only

don't want $\mathfrak{C}$ to hide complicated behavior thus the computational power of this function should be somewhat restricted. For example we put such a restriction on $\mathfrak{C}$ in Definition 7. At any given timestep, a configuration is defined in a straightforward way as a tuple $\langle c, e\rangle$ where c is an address called the control and e represents the grid contents.

Complexity Measures

In this section we define a number of CSM complexity measures. As is standard, all resource bounding functions map from $\mathbb{N}$ into $\mathbb{N}$ and are assumed to have the usual properties [7]. We begin by defining CSM TIME complexity in a manner that is standard among parallel models of computation.

Definition 2 The TIME complexity of a CSM M is the number of configurations in the computation sequence of M, beginning with the initial configuration and ending with the first final configuration.

The first of our six space-like resources is called GRID.

Definition 3 The GRID complexity of a CSM M is the minimum number of images, arranged in a rectangular grid, for M to compute correctly on all inputs.

Let $S : \mathcal{I} \times (\mathbb{N} \times \mathbb{N}) \to \mathcal{I}$, where $S(f(x, y), (\Phi, \Psi))$ is a raster image, with $\Phi\Psi$ constant-valued pixels arranged in Φ columns and Ψ rows, that approximates $f(x, y)$. If we choose a reasonable and realistic S then the details of S are not important.

Definition 4 The SPATIALRES complexity of a CSM M is the minimum $\Phi\Psi$ such that if each image $f(x, y)$ in the computation of M is replaced with $S(f(x, y), (\Phi, \Psi))$ then M computes correctly on all inputs.

One can think of SPATIALRES as a measure of the number of pixels needed during a computation. In optical image processing terms, and given the fixed size of our images, SPATIALRES corresponds to the space-bandwidth product of a detector or SLM.

Definition 5 The DYRANGE complexity of a CSM M is the ceiling of the maximum of all the amplitude values stored in all of M's images during M's computation.

In optical processing terms DYRANGE corresponds to the dynamic range of a signal.

We also use complexity measures called AMPLRES, PHASERES sand FREQ [131,136]. Roughly speaking, the AMPLRES of a CSM M is the number of discrete, evenly spaced, amplitude values per unit amplitude of the complex numbers in M's images, and so AMPLRES corresponds to the amplitude quantization of a signal. The PHASERES of M is the total number (per 2π) of discrete evenly spaced phase values in M's images, and so PHASERES corresponds to the phase quantization of a signal. Finally, the FREQ complexity of a CSM M is the minimum optical frequency necessary for M to compute correctly, this concept is explained further in [136].

Often we wish to make analogies between space on some well-known model and CSM 'space-like' resources. Thus we define the following convenient term.

Definition 6 The SPACE complexity of a CSM M is the product of all of M's complexity measures except TIME.

Optical Realization

In this section, we outline how some of the elementary operations of the CSM could be carried out physically. We do not intend to specify the definitive realization of any of the operations, but simply convince the reader that the model's operations have physical interpretations. Furthermore, although we concentrate on implementations employing visible light (optical frequencies detectable to the human eye) the CSM definition does not preclude employing other portion(s) of the electromagnetic spectrum.

A complex-valued image could be represented physically by a spatially coherent optical wavefront. Spatially coherent illumination (light of a single wavelength and emitted with the same phase angle) can be produced by a laser. SLM could be used to encode the image onto the expanded and collimated laser beam. One could write to a SLM offline (expose photographic film, or laser print or relief etch a transparency) or online (in the case of a liquid-crystal display [90,129,141] or holographic material [32,105]). The functions h and v could be effected using two convex cylindrical lenses, oriented horizontally and vertically, respectively [55,56,90,125]. As mentioned, a coherent optical wavefront will naturally evolve into its own Fourier spectrum as it propagates to infinity. What we do with a convex lens is simply image, at a finite distance, this spectrum at infinity. This finite distance is called the focal length of the lens. The constant θ used in the definitions of h and v could be effected using Fourier spectrum size reduction techniques [56,125] such as varying the focal length of the lens, varying the separation of the lens and SLM, employing cascaded Fourier transformation, increasing the dimensions/reducing the spatial resolution of the SLM, or using light with a shorter wavelength. The function $*$ could be implemented using a phase conjugate mirror [33]. The function $\cdot$ could be realized by placing a SLM encoding an image f in the path of a wavefront encoding another image g [56,123,125]. The wave-

front immediately behind the SLM would then be $\cdot(f, g)$. The function + describes the superposition of two optical wavefronts. This could be achieved using a 50:50 beam splitter [33,125,130]. The function ρ could be implemented using an electronic camera or a liquid-crystal light valve [129]. The parameters z_l and z_u would then be physical characteristics of the particular camera/light valve used. z_l corresponds to the minimum intensity value that the device responds to, known as the dark current signal, and z_u corresponds to the maximum intensity (the saturation level).

A note will be made about the possibility of automating these operations. If suitable SLMs can be prepared with the appropriate 2D pattern(s), each of the operations h, v, $*$, $\cdot$, and + could be effected autonomously and without user intervention using appropriately positioned lenses and free space propagation. The time to effect these operations would be the sum of the flight time of the image (distance divided by velocity of light) and the response time of the analog 2D detector; both of which are constants independent of the size or resolution of the images if an appropriate 2D detector is chosen. Examples of appropriate detectors would be holographic material [32,105] and a liquid-crystal light valve with a continuous (not pixellated) area [129]. Since these analog detectors are also optically-addressed SLMs, we can very easily arrange for the output of one function to act as the input to another, again in constant time independent of the size or resolution of the image. A set of angled mirrors will allow the optical image to be fed back to the first SLM in the sequence, also in constant time. It is not known, however, if ρ can be carried out completely autonomously for arbitrary parameters. Setting arbitrary parameters might fundamentally require offline user intervention (adjusting the gain of the camera, and so on), but at least for a small range of values this can be simulated online using a pair of liquid-crystal intensity filters.

We have outlined some optics principles that could be employed to implement the operations of the model. The simplicity of the implementations hides some imperfections in our suggested realizations. For example, the implementation of the + operation outlined above results in an output image that has been unnecessarily multiplied by the constant factor 0.5 due to the operation of the beam splitter. Also, in our suggested technique, the output of the ρ function is squared unnecessarily. However, all of these effects can be compensated for with a more elaborate optical setup and/or at the algorithm design stage.

A more important issue concerns the quantum nature of light. According to our current understanding, light exists as individual packets called photons. As such, in order to physically realize the CSM one would have to modify it such that images would have discrete, instead of continuous, amplitudes. The atomic operations outlined above, in particular the Fourier transform, are not affected by the restriction to quantized amplitudes, as the many experiments with electron interference patterns indicate. We would still assume, however, that in the physical world space is continuous.

A final issue concerns how a theoretically infinite Fourier spectrum could be represented by an image (or encoded by a SLM) of finite extent. This difficulty is addressed with the FREQ complexity measure [136].

Example CSM Datastructures and Algorithms

In this section we give some example data representations. We then to go on to give an example CSM algorithm that efficiently squares a binary matrix.

Representing Data as Images

There are many ways to represent data as images and interesting new algorithms sometimes depend on a new data representation. Data representations should be in some sense reasonable, for example it is unreasonable that the input to an algorithm could (non-uniformly) encode solutions to NP-hard or even undecidable problems. From Sect. "C_2-CSM", the CSM address encoding function gives the programmer room to be creative, so long as the representation is logspace computable (assuming a reasonable representation of images as words).

Here we mention some data representations that are commonly used. Figures 5a and 5b are the binary symbol image representations of 1 and 0 respectively. These images have an everywhere constant value of 1 and 0 respectively, and both have SPATIALRES of 1. The row and column image representations of the word 1011 are respectively given in Figs. 5c and 5d. These row and column images both have SPATIALRES of 4. In the matrix image representation in Fig. 5e, the first matrix element is represented at the top left corner and elements are ordered in the usual matrix way. This 3×4 matrix image has SPATIALRES of 12. Finally the binary stack image representation, which has exponential SPATIALRES of 16, is given in Fig. 5f.

Figure 4 shows how we might form a list image by copying four images to one in a single timestep. All of the above mentioned images have DYRANGE, AMPLRES and PHASERES of 1.

Another useful representation is where the value of a pixel directly encodes a number, in this case DYRANGE

becomes crucial. We can also encode values as phase values, and naturally PHASERES becomes a useful measure of the resources needed to store such values.

A Matrix Squaring Algorithm

Here we give an example CSM algorithm (taken from [133]) that makes use of the data representations described above. The algorithm squares a $n \times n$ matrix in $O(\log n)$ TIME and $O(n^3)$ SPATIALRES (number of pixels), while all other CSM resources are constant.

Lemma 1 *Let n be a power of 2 and let A be a $n \times n$ binary matrix. The matrix A^2 is computed by a C_2-CSM, using the matrix image representation, in* TIME $O(\log n)$, SPATIALRES $O(n^3)$, GRID $O(1)$, DYRANGE $O(1)$, AMPLRES *1 and* PHASERES *1.*

Proof In this proof the matrix and its matrix image representation (see Fig. 5e) are both denoted A. We begin with some precomputation, then one parallel pointwise multiplication step, followed by $\log n$ additions to complete the algorithm.

We generate the matrix image A_1 that consists of n vertically juxtaposed copies of A. This is computed by placing one copy of A above the other, scaling to one image, and repeating to give a total of $\log n$ iterations. The image A_1 is constructed in TIME $O(\log n)$, GRID $O(1)$ and SPATIALRES $O(n^3)$.

Next we transpose A to the column image A_2. The first n elements of A_2 are row 1 of A, the second n elements of A_2 are row 2 of A, etc. This is computed in TIME $O(\log n)$, GRID $O(1)$ and SPATIALRES $O(n^2)$ as follows.

Let $A' = A$ and $i = n$. We horizontally split A' into a left image A'_L and a right image A'_R. Then A'_L is pointwise multiplied (or masked) by the column image that represents $(10)^i$, in TIME $O(1)$. Similarly A'_R is pointwise multiplied (or masked) by the column image that represents $(01)^i$. The masked images are added. The resulting image has half the number of columns as A' and double the number of rows, and for example: row 1 consists of the first half of the elements of row 1 of A' and row 2 consists of the latter half of the elements of row 1 of A'. We call the result A' and we double the value of i. We repeat the process to give a total of $\log n$ iterations. After these iterations the resulting column image is denoted A_2.

We pointwise multiply A_1 and A_2 to give A_3 in TIME $O(1)$, GRID $O(1)$ and SPATIALRES $O(n^3)$.

To facilitate a straightforward addition we first transpose A_3 in the following way: A_3 is vertically split into a bottom and a top image, the top image is placed to the left of the bottom and the two are scaled to a single image, this splitting and scaling is repeated to give a total of $\log n$ iterations and we call the result A_4. Then to perform the addition, we vertically split A_4 into a bottom and a top image. The top image is pointwise added to the bottom image and the result is thresholded between 0 and 1. This splitting, adding and thresholding is repeated a total of $\log n$ iterations to create A_5. We 'reverse' the transposition that created A_4: image A_5 is horizontally split into a left and a right image, the left image is placed above the right and the two are scaled to a single image, this splitting and scaling is repeated a total of $\log n$ iterations to give A^2.

The algorithm highlights a few points of interest about the CSM. The CSM has quite a number of space-like resources, and it is possible to have trade-offs between them. For example in the algorithm above, if we allow GRID to increase from $O(1)$ to $O(n)$ then the SPATIALRES can be reduced from $O(n^3)$ to $O(n^2)$. In terms of optical architectures modeled by the CSM this phenomenon could be potentially very useful as certain resources may well be more economically viable than others. The algorithm is used in the proof that that polynomial TIME CSMs (and C_2-CSMs, see below) compute problems that are in the PSPACE class of languages. PSPACE includes the famous NP class. Such computational complexity results are discussed further in Sect. "Optical Computing and Computational Complexity" below.

There are a number of existing CSM algorithms, for these we point the reader to the literature [91,92,93,131,133,135,136].

C_2-CSM

In this section we define the C_2-CSM. One of the motivations for this model is the need to put reasonable upper bounds on the power of reasonable optical computers. As discussed below, it turns out that CSMs can very quickly use massive amounts of resources, and the C_2-CSM definition is an attempt to rein in this power.

Worst Case CSM Resource Usage

For the case of sequential computation it is usually obvious how the execution of a single operation will affect resource usage. In parallel models, execution of a single operation can lead to large growth in a single timestep. Characterizing resource growth is useful for proving upper bounds on power and choosing reasonable model restrictions.

We investigated the growth of complexity resources over TIME, with respect to CSM operations [131,134]. As expected, under certain operations some measures do not grow at all. Others grow at rates comparable to massively parallel models. By allowing operations like the

Optical Computing, Table 1

CSM resource usage after one timestep. For a given operation and complexity measure pair, the relevant table entry defines the worst case CSM resource usage at TIME $T+1$, in terms of the resources used at TIME T. At TIME T we have GRID $= G_T$, SPATIALRES $= R_{S,T}$, AMPLRES $= R_{A,T}$, DYRANGE $= R_{D,T}$, PHASERES $= R_{P,T}$ and FREQ $= \nu_T$

	GRID	SPATIALRES	AMPLRES	DYRANGE	PHASERES	FREQ
h	G_T	∞	∞	∞	∞	∞
v	G_T	∞	∞	∞	∞	∞
$*$	G_T	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
$\cdot$	G_T	$R_{S,T}$	$(R_{A,T})^2$	$(R_{D,T})^2$	$R_{P,T}$	ν_T
$+$	G_T	$R_{S,T}$	∞	$2R_{D,T}$	∞	ν_T
ρ	unbounded	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
st	unbounded	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
ld	unbounded	unbounded	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	unbounded
br	G_T	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
hlt	G_T	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T

Fourier transform we are mixing the continuous and discrete worlds, hence some measures grow to infinity in one timestep. This gave strong motivation for CSM restrictions.

Table 1 summarizes these results; the table defines the value of a complexity measure after execution of an operation (at TIME $T+1$). The complexity of a configuration at TIME $T+1$ is at least the value it was at TIME T, since complexity functions are nondecreasing. Our definition of TIME assigns unit time cost to each operation, hence we do not have a TIME column. Some entries are immediate from the complexity measure definitions, for others proofs are given in the references [131,134].

C_2-CSM

Motivated by a desire to apply standard complexity theory tools to the model, we defined [131,134] the C_2-CSM, a restricted class of CSM.

Definition 7 (C_2-CSM) A C_2-CSM is a CSM whose computation TIME is defined for $t \in \{1, 2, \ldots, T(n)\}$ and has the following restrictions:

- For all TIME t both AMPLRES and PHASERES have constant value of 2.
- For all TIME t each of GRID, SPATIALRES and DYRANGE is $2^{O(t)}$ and SPACE is redefined to be the product of all complexity measures except TIME and FREQ.
- Operations h and v compute the discrete Fourier transform in the horizontal and vertical directions respectively.
- Given some *reasonable* binary word representation of the set of addresses $\mathcal{N}$, the address encoding function $\mathfrak{E}: \mathbb{N} \to \mathcal{N}$ is decidable by a logspace Turing machine.

Let us discuss these restrictions. The restrictions on AMPLRES and PHASERES imply that C_2-CSM images are of the form $f : [0, 1) \times [0, 1) \to \{0, \pm 1/2, \pm 1, \pm 3/2, \ldots\}$. We have replaced the Fourier transform with the discrete Fourier transform [10], this essentially means that FREQ is now solely dependent on SPATIALRES; hence FREQ is not an interesting complexity measure for C_2-CSMs and we do not analyze C_2-CSMs in terms of FREQ complexity [131,134]. Restricting the growth of SPACE is not unique to our model, such restrictions are to be found elsewhere [54,98,102].

In Sect. "CSM Definition" we stated that the address encoding function $\mathfrak{E}$ should be Turing machine decidable, here we strengthen this condition. At first glance sequential logspace computability may perhaps seem like a strong restriction, but in fact it is quite weak. From an optical implementation point of view it should be the case that $\mathfrak{E}$ is not complicated, otherwise we cannot assume fast addressing. Other (sequential/parallel) models usually have a very restricted 'addressing function': in most cases it is simply the identity function on $\mathbb{N}$. Without an explicit or implicit restriction on the computational complexity of $\mathfrak{E}$, finding non-trivial upper bounds on the power of C_2-CSMs is impossible as $\mathfrak{E}$ could encode an arbitrarily complex Turing machine. As a weaker restriction we could give a specific $\mathfrak{E}$. However, this restricts the generality of the model and prohibits the programmer from developing novel, reasonable, addressing schemes.

Optical Computing and Computational Complexity

There have been a number of optical algorithms given that use the inherent parallelism of optics to speed up the solutions to certain problems. An alternative approach is to ask

the following question: How does a given optical model relate to standard sequential and parallel models? Establishing a relationship with computational complexity theory, by describing both upper and lower bounds on the model, gives immediate access to a large collection of useful algorithms and proof techniques.

The parallel computation thesis [29,52,74,98,126] states that parallel time (polynomially) corresponds to sequential space, for reasonable parallel and sequential models. An example would be the fact that the class of problems solvable in polynomial space on a number of parallel models is equivalent to PSPACE, the class of problems solvable on Turing machines that use at most polynomial space [3,9,30,46,53,54,61,113,114,120,127].

Of course the thesis can never be proved, it relates the intuitive notion of reasonable parallelism to the precise notion of a Turing machine. When results of this type were first shown researchers were suitably impressed; their parallel models truly had great power. For example if model M verifies the thesis then M decides PSPACE (including NP) languages in polynomial time. However there is another side to this coin. It is straightforward to verify that given our current best algorithms, M will use at least a superpolynomial amount of some other resource (like space or number of processors) to decide a PSPACE-complete or NP-complete language. Since the composition of polynomials is itself a polynomial, it follows that if we restrict the parallel computer to use at most polynomial time and polynomial other resources, then it can at most solve problems in P.

Nevertheless, asking if M verifies the thesis is an important question. Certain problems, such as those in the class NC, are efficiently parallelisable. NC can be defined as the class of problems that are solvable in polylogarithmic time on a parallel computer that uses a polynomial amount of hardware. So one can think of NC as those problems in P which are solved exponentially faster on parallel computation thesis models than on sequential models. If M verifies the thesis then we know it will be useful to apply M to these problems. We also know that if M verifies the thesis then there are (P-complete) problems for which it is widely believed that we will not find exponential speed up using M.

C_2-CSM and Parallel Complexity Theory

Here we summarize some characterizations of the computing power of optical computers. Such characterizations enable the algorithm designer to know what kinds of problems are solvable with resource bounded optical algorithms.

Theorem 1 below gives lower bounds on the computational power of the C_2-CSM by showing that it is at least as powerful as models that verify the parallel computation thesis.

Theorem 1 ([133,135])
$\text{NSPACE}(S(n)) \subseteq C_2\text{-}CSM\text{-TIME}(O(S^2(n)))$

In particular, polynomial TIME C_2-CSMs accept the PSPACE languages. PSPACE is the class of problems solvable by Turing machines that use polynomial space, which includes the famous class NP, and so NP-complete problems can be solved by C_2-CSMs in polynomial TIME. However, any C_2-CSM algorithm that we could presently write to solve PSPACE or NP problems would require exponential SPACE.

Theorem 1 is established by giving a C_2-CSM algorithm that efficiently generates, and squares, the transition matrix of a $S(n) = \Omega(\log n)$ space bounded Turing machine. This transition matrix represents all possible computations of the Turing machine and is of size $O(2^S) \times O(2^S)$. The matrix squaring part was already given as an example (Lemma 1), and the remainder of the algorithm is given in [133]. The algorithm uses SPACE that is cubic in one of the matrix dimensions. In particular the algorithm uses cubic SPATIALRES, $O(2^{3S})$, and all other space-like resources are constant. This theorem improves upon the time overhead of a previous, but similar, result [131,135] that was established via C_2-CSM simulation of the vector machines [102,103] of Pratt, Rabin, and Stockmeyer.

From the resource usage point of view, it is interesting to see that the older of these two algorithms uses GRID, DYRANGE, and SPATIALRES that are each $O(2^S)$, while the newer algorithm shows that if we allow more SPATIALRES we can in fact use only constant GRID and DYRANGE. It would be interesting to find other such resource trade-offs within the model.

Since NP is contained in PSPACE, Theorem 1 and the corresponding earlier results in [131,135], show that this optical model solves NP-complete problems in polynomial TIME. As described in Sect. "Optical Models of Computation", this has also been shown experimentally, for example Shaked et al. [110] have recently given a polynomial time, exponential space, optical algorithm to solve the NP-complete travelling salesperson problem. Their optical setup can be implemented on the CSM.

The other of the two inclusions that are necessary in order to verify the parallel computation thesis have also been shown: C_2-CSMs computing in TIME $T(n)$ are no more powerful than $T^{O(1)}(n)$ space bounded deterministic Turing machines. More precisely, we have:

Theorem 2 ([131,132])
C_2-CSM-TIME$(T(n)) \subseteq$ DSPACE$(O(T^2(n)))$

This result gives an upper bound on the power of C_2-CSMs and was established via C_2-CSM simulation by logspace uniform circuits of size and depth polynomial in SPACE and TIME respectively [132].

Via the proofs of Theorems 1 and 2 we get another (stronger) result: C_2-CSMs that simultaneously use polynomial SPACE and polylogarithmic TIME solve exactly those problems in the class NC.

Corollary 1
C_2-CSM-SPACE, TIME$(n^{O(1)}, \log^{O(1)} n) =$ NC

Problems in NC highlight the power of parallelism, as these problems can be solved exponentially faster on a polynomial amount of parallel resources than on polynomial time sequential machines. As further work in this area one could try to find alternate characterizations of NC in terms of the C_2-CSM. In particular, one could try to find further interesting trade-offs between the various space-like resources of the model. In the real world this would correspond to computing over various different CSM resources. Also, it might be interesting for optical algorithm designers to try to design optical algorithms for NC problems in an effort to find problems that are well suited to optical solutions. See [137] for details on this argument, and also fpr other CSM characterisations of complexity classes and an implementation of a fast optical search algorithm.

Future Directions

As already noted, optical computing is an inherently multidisciplinary subject whose study routinely involves a spectrum of expertise that threads optical physics, materials science, optical engineering, electrical engineering, computer architecture, computer programming, and computer theory. From the point of view of each of these fields there are various directions for future work. Also, it is generally accepted that if optical computers become mainstream, it will be through a symbiotic relationship with their extremely flexible digital electronic counterparts. At the confluence of computing and communication there is room for optical techniques such as for the routing of signals in long-haul networks via all-optical packet switching in a single channel configuration. So it seems that whether or not optical computers will be adopted in a widespread manner is both a technological and economic issue.

From the algorithmic point of view there is plenty of scope for future work. There are a number of questions related to trade-offs between resources and we believe the CSM gives a good framework to answer such questions. For example can we give useful parallel algorithms that exploit CSM resources such as PHASERES while at the same time using small SPATIALRES and GRID? In a similar vein, one can explore the computing power of restrictions and generalizations of the CSM with the goal of finding new algorithms and characterizations of complexity classes. This has immediate applications in finding new and efficient implementations of optical solutions to computational problems.

Acknowledgments

DW thanks J. Paul Gibson and Cris Moore for interesting discussions. DW acknowledges Junta de Andalucia grant TIC-581, Science Foundation Ireland grant number 04/IN3/1524, and Irish Research Council for Science Engineering and Technology grant number PD/2004/33. TN acknowledges support from the European Commission through a Marie Curie Intra-European Fellowship.

Bibliography

1. Abushagur MAG, Caulfield HJ (1987) Speed and convergence of bimodal optical computers. Opt Eng 26(1):22–27
2. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266:1021–1024
3. Alhazov A, de Jesús Pérez–Jiménez M (2007) Uniform solution to QSAT using polarizationless active membranes. In: Durand–Lose J, Margenstern M (eds) Machines, Computations and Universality (MCU). LNCS, vol 4664. Springer, Orléans, pp 122–133
4. Armitage JD, Lohmann AW (1965) Character recognition by incoherent spatial filtering. Appl Opt 4(4):461–467
5. Arsenault HH, Sheng Y (1992) An Introduction to Optics in Computers. Tutorial Texts in Optical Engineering, vol TT8. SPIE Press, Bellingham, Washington
6. Arsenault HH, Hsu YN, Chalasinska–Macukow K (1984) Rotation-invariant pattern recognition. Opt Eng 23(6):705–709
7. Balcázar JL, Díaz J, Gabarró J (1988) Structural complexity, vols I and II. EATCS Monographs on Theoretical Computer Science. Springer, Berlin
8. Beyette Jr FR, Mitkas PA, Feld SA, Wilmsen CW (1994) Bitonic sorting using an optoelectronic recirculating architecture. Appl Opt 33(35):8164–8172
9. Borodin A (1977) On relating time and space to size and depth. SIAM J Comput 6(4):733–744
10. Bracewell RN (1978) The Fourier transform and its applications, 2nd edn. Electrical and electronic engineering series, McGraw–Hill, New York
11. Brenner KH, Huang A, Streibl N (1986) Digital optical computing with symbolic substitution. Appl Opt 25(18):3054–3060
12. Brenner KH, Kufner M, Kufner S (1990) Highly parallel arithmetic algorithms for a digital optical processor using symbolic substitution logic. Appl Opt 29(11):1610–1618
13. Casasent DP (1984) Unified synthetic discriminant function computational formulation. Appl Opt 23:1620–1627
14. Casasent DP, House GP (1994) Comparison of coherent and noncoherent optical correlators. In: Optical Pattern Recog-

nition V. Proceedings of SPIE, vol 2237. SPIE, Bellingham, pp 170–178
15. Casasent DP, House GP (1994) Implementation issues for a noncoherent optical correlator. In: Optical Pattern Recognition V. Proceedings of SPIE, vol 2237. SPIE, Bellingham, pp 179–188
16. Casasent DP, Psaltis D (1976) Position, rotation, and scale invariant optical correlation. Appl Opt 15(7):1795–1799
17. Casasent DP, Psaltis D (1976) Scale invariant optical transforms. Opt Eng 15(3):258–261
18. Casasent DP, Jackson J, Neuman CP (1983) Frequency-multiplexed and pipelined iterative optical systolic array processors. Appl Opt 22:115–124
19. Caulfield HJ (ed) (1979) Handbook of Optical Holography. Academic Press, New York
20. Caulfield HJ (1989) Computing with light. Byte 14:231–237
21. Caulfield HJ (1989) The energetic advantage of analog over digital computing. In: OSA Optical Computing Technical Digest Series, vol 9. Optical Society of America, Washington, pp 180–183
22. Caulfield HJ (1990) Space-time complexity in optical computing. In: Javidi B (ed) Optical information-processing systems and architectures II. SPIE, vol 1347. SPIE Press, Bellingham, pp 566–572
23. Caulfield HJ, Abushagur MAG (1986) Hybrid analog-digital algebra processors. In: Optical and Hybrid Computing II. Proceedings of SPIE, vol 634. SPIE Press, Bellingham, pp 86–95
24. Caulfield HJ, Haimes R (1980) Generalized matched filtering. Appl Opt 19(2):181–183
25. Caulfield HJ, Horvitz S, Winkle WAV (1977) Introduction to the special issue on optical computing. Proceedings of the IEEE 65(1):4–5
26. Caulfield HJ, Rhodes WT, Foster MJ, Horvitz S (1981) Optical implementation of systolic array processing. Opt Commun 40:86–90
27. Caulfield HJ, Kinser JM, Rogers SK (1989) Optical neural networks. Proc IEEE 77:1573–1582
28. Cerf NJ, Adami C, Kwiat PG (1998) Optical simulation of quantum logic. Phys Rev A 57(3):R1477–R1480
29. Chandra AK, Stockmeyer LJ (1976) Alternation. In: 17th annual symposium on Foundations of Computer Science, IEEE, Houston, Texas, pp 98–108
30. Chandra AK, Kozen DC, Stockmeyer LJ (1981) Alternation. J ACM 28(1):114–133
31. Chang S, Arsenault HH, Garcia–Martinez P, Grover CP (2000) Invariant pattern recognition based on centroids. Appl Opt 39(35):6641–6648
32. Chen FS, LaMacchia JT, Fraser DB (1968) Holographic storage in lithium niobate. Appl Phys Lett 13(7):223–225
33. Chiou AE (1999) Photorefractive phase-conjugate optics for image processing, trapping, and manipulation of microscopic objects. Proc IEEE 87(12):2074–2085
34. Clavero R, Ramos F, Martí J (2005) All-optical flip-flop based on an active Mach–Zehnder interferometer with a feedback loop. Opt Lett 30(21):2861–2863
35. Cutrona LJ, Leith EN, Palermo CJ, Porcello LJ (1960) Optical data processing and filtering systems. IRE Trans Inf Theory 6(3):386–400
36. Cutrona LJ, Leith EN, Porcello LJ, Vivian WE (1966) On the application of coherent optical processing techniques to synthetic-aperture radar. Proc IEEE 54(8):1026–1032
37. Desmulliez MPY, Wherrett BS, Waddie AJ, Snowdon JF, Dines JAB (1996) Performance analysis of self-electro-optic-effect-device-based (seed-based) smart-pixel arrays used in data sorting. Appl Opt 35(32):6397–6416
38. Dolev S, Fitoussi H (2007) The traveling beam: optical solution for bounded NP-complete problems. In: Crescenzi P, Prencipe G, Pucci G (eds) The fourth international conference on fun with algorithms (FUN). Springer, Heidelberg, pp 120–134
39. Dorren HJS, Hill MT, Liu Y, Calabretta N, Srivatsa A, Huijskens FM, de Waardt H, Khoe GD (2003) Optical packet switching and buffering by using all-optical signal processing methods. J Lightwave Technol 21(1):2–12
40. Durand–Lose J (2006) Reversible conservative rational abstract geometrical computation is Turing-universal. In: Logical Approaches to Computational Barriers, Second Conference on Computability in Europe, (CiE). Lecture Notes in Computer Science. Springer, Swansea, vol 3988. pp 163–172
41. Efron U, Grinberg J, Braatz PO, Little MJ, Reif PG, Schwartz RN (1985) The silicon liquid crystal light valve. J Appl Phys 57(4):1356+
42. Esteve–Taboada JJ, García J, Ferreira C (2000) Extended scale-invariant pattern recognition with white-light illumination. Appl Opt 39(8):1268–1271
43. Farhat NH, Psaltis D (1984) New approach to optical information processing based on the Hopfield model. J Opt Soc Am A 1:1296
44. Feitelson DG (1988) Optical Computing: A survey for computer scientists. MIT Press, Cambridge, Massachusetts
45. Feng JH, Chin GF, Wu MX, Yan SH, Yan YB (1995) Multiobject recognition in a multichannel joint-transform correlator. Opt Lett 20(1):82–84
46. Fortune S, Wyllie J (1978) Parallelism in random access machines. In: Proc. 10th Annual ACM Symposium on Theory of Computing, ACM, New York, pp 114–118
47. Gabor D (1948) A new microscopic principle. Nature 161(4098):777–778
48. Gara AD (1979) Real time tracking of moving objects by optical correlation. Appl Opt 18(2):172–174
49. Geldenhuys R, Liu Y, Calabretta N, Hill MT, Huijskens FM, Khoe GD, Dorren HJS (2004) All-optical signal processing for optical packet switching. J Opt Netw 3(12):854–865
50. Ghosh AK, Casasent DP, Neuman CP (1985) Performance of direct and iterative algorithms on an optical systolic processor. Appl Opt 24(22):3883–3892
51. Goldberg L, Lee SH (1979) Integrated optical half adder circuit. Appl Opt 18:2045–2051
52. Goldschlager LM (1977) Synchronous parallel computation. Ph D thesis, University of Toronto, Computer Science Department
53. Goldschlager LM (1978) A unified approach to models of synchronous parallel machines. In: Proc. 10th Annual ACM Symposium on Theory of Computing. ACM, New York, pp 89–94
54. Goldschlager LM (1982) A universal interconnection pattern for parallel computers. J ACM 29(4):1073–1086
55. Goodman JW (1977) Operations achievable with coherent optical information processing systems. Proc IEEE 65(1):29–38
56. Goodman JW (2005) Introduction to Fourier Optics, 3rd edn. Roberts & Company, Englewood
57. Grinberg J, Jacobson AD, Bleha WP, Miller L, Fraas L, Boswell D, Myer G (1975) A new real-time noncoherent to coherent

light image converter: The hybrid field effect liquid crystal light valve. Opt Eng 14(3):217–225
58. Grover LK (1996) A fast quantum mechanical algorithm for database search. In: Proc. 28th Annual ACM Symposium on Theory of Computing. ACM, New York, pp 212–219
59. Guilfoyle PS, Hessenbruch JM, Stone RV (1998) Free-space interconnects for high-performance optoelectronic switching. IEEE Comp 31(2):69–75
60. Haist T, Osten W (2007) An optical solution for the travelling salesman problem. Opt Expr 15(16):10473–10482
61. Hartmanis J, Simon J (1974) On the power of multiplication in random access machines. In: Proceedings of the 15th annual symposium on switching and automata theory. IEEE, The University of New Orleans, pp 13–23
62. Head T (1987) Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. Bull Math Biol 49(6):737–759
63. Horner JL (ed) (1987) Optical Signal Processing. Academic Press, San Diego
64. Hough PVC (1962) Methods and measures for recognising complex patterns. U.S. Patent No. 3069654
65. Hsu KY, Li HY, Psaltis D (1990) Holographic implementation of a fully connected neural network. Proc IEEE 78(10):1637–1645
66. Hsu YN, Arsenault HH (1982) Optical pattern recognition using circular harmonic expansion. Appl Opt 21(22):4016–4019
67. Huang A (1984) Architectural considerations involved in the design of an optical digital computer. Proc IEEE 72(7):780–786
68. Huang A, Tsunoda Y, Goodman JW, Ishihara S (1979) Optical computation using residue arithmetic. Appl Opt 18(2):149–162
69. Jacobson AD, Beard TD, Bleha WP, Morgerum JD, Wong SY (1972) The liquid crystal light value, an optical-to-optical interface device. In: Proceedings of the Conference on Parallel Image Processing, document X-711-72-308, Goddard Space Flight Center. NASA, Washington, pp 288–299
70. Javidi B (1989) Nonlinear joint power spectrum based optical correlation. Appl Opt 28(12):2358–2367
71. Javidi B (1990) Generalization of the linear matched filter concept to nonlinear matched filters. Appl Opt 29(8):1215–1217
72. Javidi B, Wang J (1995) Optimum distortion-invariant filter for detecting a noisy distorted target in nonoverlapping background noise. J Opt Soc Am A 12(12):2604–2614
73. Karim MA, Awwal AAS (1992) Optical Computing: An Introduction. Wiley, New York
74. Karp RM, Ramachandran V (1990) Parallel algorithms for shared memory machines. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science, vol A. Elsevier, Amsterdam, pp 869–941
75. Knill E, LaFlamme R, Milburn GJ (2001) A scheme for efficient quantum computation with linear optics. Nature 409:46–52
76. Lee JN (ed) (1995) Design Issues in Optical Processing. Cambridge Studies in Modern Optics. Cambridge University Press, Cambridge
77. Lee JN (ed) (1995) Design issues in optical processing. Cambridge studies in modern optics. Cambridge University Press, Cambridge
78. Leith EN (1977) Complex spatial filters for image deconvolution. Proc IEEE 65(1):18–28
79. Lenslet Labs (2001) Optical digital signal processing engine. white paper report, Lenslet Ltd., 12 Hachilazon St., Ramat–Gan, Israel 52522
80. Lipton RJ (1995) Using DNA to solve NP-complete problems. Science 268:542–545
81. Lohmann AW (1993) Image rotation, Wigner rotation, and the fractional Fourier transform. J Opt Soc Am A 10(10):2181–2186
82. Louri A, Post A (1992) Complexity analysis of optical-computing paradigms. Appl Opt 31(26):5568–5583
83. Louri A, Hatch Jr JA, Na J (1995) Constant-time parallel sorting algorithm and its optical implementation using smart pixels. Appl Opt 34(17):3087–3097
84. Lu XJ, Yu FTS, Gregory DA (1990) Comparison of Vander lugt and joint transform correlators. Appl Phys B 51:153–164
85. McAulay AD (1991) Optical Computer Architectures: The Application of Optical Concepts to Next Generation Computers. Wiley, New York
86. Mead C (1989) Analog VLSI and Neural Systems. Addison–Wesley, Reading
87. Miller DA (2000) Rationale and challenges for optical interconnects to electronic chips. Proc IEEE 88(6):728–749
88. Moore C (1991) Generalized shifts: undecidability and unpredictability in dynamical systems. Nonlinearity 4:199–230
89. Moore C (1997) Majority-vote cellular automata, Ising dynamics and P-completeness. J Stat Phys 88(3/4):795–805
90. Naughton T, Javadpour Z, Keating J, Klíma M, Rott J (1999) General-purpose acousto-optic connectionist processor. Opt Eng 38(7):1170–1177
91. Naughton TJ (2000) Continuous-space model of computation is Turing universal. In: Bains S, Irakliotis LJ (eds) Critical Technologies for the Future of Computing. Proc SPIE, vol 4109. San Diego, pp 121–128
92. Naughton TJ (2000) A model of computation for Fourier optical processors. In: Lessard RA, Galstian T (eds) Optics in Computing 2000. Proc SPIE, vol 4089. Quebec, pp 24–34
93. Naughton TJ, Woods D (2001) On the computational power of a continuous-space optical model of computation. In: Margenstern M, Rogozhin Y (eds) Machines, Computations and Universality: Third International Conference (MCU'01). LNCS, vol 2055. Springer, Heidelberg, pp 288–299
94. Oltean M (2006) A light-based device for solving the Hamiltonian path problem. In: Fifth International Conference on Unconventional Computation (UC'06). LNCS, vol 4135. Springer, York, pp 217–227
95. O'Neill EL (1956) Spatial filtering in optics. IRE Trans Inf Theory 2:56–65
96. Paek EG, Choe JY, Oh TK, Hong JH, Chang TY (1997) Nonmechanical image rotation with an acousto-optic dove prism. Opt Lett 22(15):1195–1197
97. Papadimitriou CH (1995) Computational complexity. Addison–Wesley, Reading
98. Parberry I (1987) Parallel complexity theory. Wiley, New York
99. Păun G (2002) Membrane computing: an introduction. Springer, Heidelberg
100. Pe'er A, Wang D, Lohmann AW, Friesem AA (1999) Optical correlation with totally incoherent light. Opt Lett 24(21):1469–1471
101. Pittman TB, Fitch MJ, Jacobs BC, Franson JD (2003) Experimental controlled-NOT logic gate for single photons in the coincidence basis. Phys Rev A 68:032316–3

102. Pratt VR, Stockmeyer LJ (1976) A characterisation of the power of vector machines. J Comput Syst Sci 12:198–221
103. Pratt VR, Rabin MO, Stockmeyer LJ (1974) A characterisation of the power of vector machines. In: Proc 6th annual ACM symposium on theory of computing. ACM, New York, pp 122–134
104. Psaltis D, Farhat NH (1985) Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. Opt Lett 10(2):98–100
105. Pu A, Denkewalter RF, Psaltis D (1997) Real-time vehicle navigation using a holographic memory. Opt Eng 36(10):2737–2746
106. Reif J, Tygar D, Yoshida A (1990) The computability and complexity of optical beam tracing. In: 31st Annual IEEE Symposium on Foundations of Computer Science (FOCS). IEEE, St. Louis, pp 106–114
107. Reif JH, Tyagi A (1997) Efficient parallel algorithms for optical computing with the discrete Fourier transform (DFT) primitive. Appl Opt 36(29):7327–7340
108. Rhodes WT (1981) Acousto-optic signal processing: convolution and correlation. Proc IEEE 69(1):65–79
109. Sawchuk AA, Strand TC (1984) Digital optical computing. Proc IEEE 72(7):758–779
110. Shaked NT, Simon G, Tabib T, Mesika S, Dolev S, Rosen J (2006) Optical processor for solving the traveling salesman problem (TSP). In: Javidi B, Psaltis D, Caulfield HJ (eds) Proc SPIE, Optical Information Systems IV, vol 63110G. SPIE, Bellingham
111. Shaked NT, Messika S, Dolev S, Rosen J (2007) Optical solution for bounded NP-complete problems. Appl Opt 46(5):711–724
112. Shor P (1994) Algorithms for quantum computation: Discrete logarithms ande factoring. In: Proceedings 35th Annual Symposium on Foundations Computer Science. ACM, New York, pp 124–134
113. Sosík P (2003) The computational power of cell division in P systems: Beating down parallel computers? Nat Comput 2(3):287–298
114. Sosík P, Rodríguez–Patón A (2007) Membrane computing and complexity theory: A characterization of PSPACE. J Comput Syst Sci 73(1):137–152
115. Stirk CW, Athale RA (1988) Sorting with optical compare-and-exchange modules. Appl Opt 27(9):1721–1726
116. Stone RV (1994) Optoelectronic processor is programmable and flexible. Laser Focus World 30(8):77–79
117. Stone RV, Zeise FF, Guilfoyle PS (1991) DOC II 32-bit digital optical computer: optoelectronic hardware and software. In: Optical Enhancements to Computing Technology, Proc SPIE, vol 1563. SPIE, Bellingham, pp 267–278
118. Stroke GW, Halioua M, Thon F, Willasch DH (1974) Image improvement in high-resolution electron microscopy using holographic image deconvolution. Optik 41(3):319–343
119. Sullivan DL (1972) Alignment of rotational prisms. Appl Opt 11(9):2028–2032
120. Tromp J, van Emde Boas P (1993) Associative storage modification machines. In: Ambos–Spies K, Homer S, Schöning U (eds) Complexity theory: current research. Cambridge University Press, pp 291–313
121. Turin GL (1960) An introduction to matched filters. IRE Trans Inf Theory 6(3):311–329
122. Upatnieks J (1983) Portable real-time coherent correlator. Appl Opt 22(18):2798–2803
123. VanderLugt A (1964) Signal detection by complex spatial filtering. IEEE Trans Inf Theory 10(2):139–145
124. VanderLugt A (1974) Coherent optical processing. Proc IEEE 62(10):1300–1319
125. VanderLugt A (1992) Optical Signal Processing. Wiley, New York
126. van Emde Boas P (1990) Machine models and simulations. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science, vol A. Elsevier, Amsterdam, chap 1
127. van Leeuwen J, Wiedermann J (1987) Array processing machines. BIT 27:25–43
128. Wang CH, Jenkins BK (1990) Subtracting incoherent optical neuron model – Analysis, experiment and applications. Appl Opt 29(14):2171–2186
129. Wang PY, Saffman M (1999) Selecting optical patterns with spatial phase modulation. Opt Lett 24(16):1118–1120
130. Weaver CS, Goodman JW (1966) A technique for optically convolving two functions. Appl Opt 5(7):1248–1249
131. Woods D (2005) Computational complexity of an optical model of computation. Ph D thesis, National University of Ireland, Maynooth
132. Woods D (2005) Upper bounds on the computational power of an optical model of computation. In: Deng X, Du D (eds) 16th International Symposium on Algorithms and Computation (ISAAC 2005). LNCS, vol 3827. Springer, Heidelberg, pp 777–788
133. Woods D (2006) Optical computing and computational complexity. In: Fifth International Conference on Unconventional Computation (UC'06). LNCS, vol 4135. Springer, pp 27–40
134. Woods D, Gibson JP (2005) Complexity of continuous space machine operations. In: Cooper SB, Löewe B, Torenvliet L (eds) New Computational Paradigms, First Conference on Computability in Europe (CiE 2005). LNCS, vol 3526. Springer, Amsterdam, pp 540–551
135. Woods D, Gibson JP (2005) Lower bounds on the computational power of an optical model of computation. In: Calude CS, Dinneen MJ, Păun G, Pérez–Jiménez MJ, Rozenberg G (eds) Fourth International Conference on Unconventional Computation (UC'05). LNCS, vol 3699. Springer, Heidelberg, pp 237–250
136. Woods D, Naughton TJ (2005) An optical model of computation. Theor Comput Sci 334(1–3):227–258
137. Woods D, Naughton TJ (2008) Parallel and sequential optical computing. In: International Workshop on Optical Super-Computing. LNCS. Springer, Heidelberg
138. Yamaguchi I, Zhang T (1997) Phase-shifting digital holography. Opt Lett 22(16):1268–1270
139. Yokomori T (2002) Molecular computing paradigm – toward freedom from Turing's charm. Nat Comput 1(4):333–390
140. Yu FTS (1996) Garden of joint transform correlators: an account of recent advances. In: Second International Conference on Optical Information Processing. Proc SPIE, vol 2969. SPIE, Bellingham, pp 396–401
141. Yu FTS, Lu T, Yang X, Gregory DA (1990) Optical neural network with pocket-sized liquid-crystal televisions. Opt Lett 15(15):863–865
142. Yu FTS, Jutamulia S, Yin S (eds) (2001) Introduction to information optics. Academic Press, San Diego
143. Zhai H, Mu G, Sun J, Zhu X, Liu F, Kang H, Zhan Y (1999) Color pattern recognition in white-light joint transform correlation. Appl Opt 38(35):7238–7244

Optimization Problems and Algorithms from Computer Science

Heiko Rieger
Theoretical Physics, Universität des Saarlandes, Saarbrücken, Germany

Article Outline

Glossary

Combinatorial optimization The search for an optimal configuration in terms of a cost function on a discrete set of allowed configurations.

Ground state The configuration of a model for a physical system of many interacting degrees of freedom described by a Hamiltonian or energy function that has the lowest energy. Also dented as the global minimum of the energy of the system.

Disordered system A physical system with frozen in or quenched inhomogeneities, usually modeled by an energy function containing parameters that are random numbers obeying in prescribed probability distribution.

Universal properties Properties that do not depend on microscopic details of a physical system, like the critical exponents at a continuous phase transition or fractal dimensions.

Network flows A function defined on the edges of a graph that obeys mass balance constraints at each node. A number of polynomial optimization problems relevant for disordered systems can be formulated as network flow models.

Definition of the Subject

Optimization problems in statistical physics occur whenever the ground state of a classical model for a complex condensed matter system has to be determined, which is necessary for understanding its low temperature properties. In some cases calculating the ground state is an easy task as for instance for the paradigmatic model for a ferromagnet: The configuration of all magnetic moments or spins with the lowest energy is the one, where all spins point in the same direction. But usually the situation is much more complex and the problem of calculating the state with the lowest energy is highly non-trivial. This occurs typically in systems with quenched disorder and/or frustration, which means that their Hamiltonian or energy function consists of competing terms that cannot be satisfied simultaneously. Powerful algorithms from computer science have been devised to find the optimum of complex cost-functions and in some cases this can even be achieved in polynomial time. In recent years many of these algorithms could be successfully applied to physically relevant model systems: to polymers in random media, interface problems in random ferromagnets, magnetic flux-lines in disordered environments, spin glasses, and many more.

Introduction

Solid materials which contain a substantial degree of quenched disorder, so called disordered systems, have been an experimental and a theoretical challenge for physicists for many decades. The different thermodynamic phases emerging in random magnets, the aging properties and memory effects of spin glasses, the disorder induced conductor-to-insulator transition in electronic or bosonic systems, the collective behavior of magnetic flux lines in amorphous high temperature superconductors, and the roughening transition of a disordered charge density wave systems are only a few examples for these fascinating phenomena that occur due to the presence of quenched disorder.

Analytic studies of models for these systems are usually based on perturbation theories valid for weak disorder, on phenomenological scaling pictures or on mean-field approximations. Therefore the demand for efficient numerical techniques that allow the investigation of the model Hamiltonians of disordered systems has always been high. Three facts make life difficult here: 1) The regime, where disorder effects are most clearly seen, are at low temperatures – and are even best visible at zero temperature; 2) the presence of disorder slows the dynamics of theses systems down, they become *glassy*, such that for instance conventional Monte-Carlo or molecular dynamics simulations encounter enormous equilibration problems; 3) any numerical computation of disordered systems has to incorporate an extensive disorder average.

In recent years more and more model systems with quenched disorder were found that can be investigated nu-

merically 1) at zero temperature, 2) without equilibration problems, 3) extremely fast, in polynomial time (for reviews see [1,2,3]). This *is* indeed progress, which became possible by the application of *exact* combinatorial optimization algorithms developed by mathematicians and computer scientists over the last few decades. This gift is not for free: first a mapping of the problem of finding the *exact* ground state of the model Hamiltonian under consideration onto a standard combinatorial optimization problem has to be found. If one is lucky, this problem falls into the class of *P*-problems, for which polynomial algorithms exist. If not, the intellectual challenge for the theoretical physicist remains to reformulate the model Hamiltonian in such a way that its universality class is not changed but a mapping on a *P*-problem becomes feasible.

An optimization problem can be described mathematically in the following way: let $\underline{\sigma} = (\sigma_1, \ldots, \sigma_n)$ be a vector with n elements which can take values from a domain $X^n : \sigma_i \in X$. The domain X can be either discrete, for instance $X = \{0, 1\}$ or $X = Z$ the set of all integers (in which case it is an integer optimization problem) or X can be continuous, for instance $X = R$ the real numbers. Moreover, let $\mathcal{H}$ be a real valued function, the cost function or objective, or in physics usually the Hamiltonian or the energy of the system. The *minimization problem* is then:

$$\text{Find } \underline{\sigma} \in X^n, \text{ which minimizes } \mathcal{H} \text{ !}$$

A maximization problem is defined in an analogous way. It is sufficient to consider only minimization problems, since maximizing a function H is equivalent to minimizing $-H$. Minimization problems in which the set X is *countable* are called *combinatorial* [4,5,6]. Optimization methods for real valued variables are treated mainly in mathematical literature and in books on numerical methods, see e. g. [8].

Constraints, must hold for the solution, may be expressed by additional equations or inequalities. An arbitrary value of $\underline{\sigma}$, which fulfills all constraints, is called *feasible*. Usually constraints can be expressed more conveniently without giving equations or inequalities. A famous example is the Traveling Salesman Problem (TSP) [7].

The TSP has attracted the interest of physicist several times. For an introduction, see [9]. The model is briefly presented here. Consider n cities distributed randomly in a plane. Without loss of generality the plane is considered to be the unit square. The minimization task is to find the shortest round-tour through all cities which visits each city only once. The tour stops at the city where it started. The problem is described by

$$X = \{1, 2, \ldots, n\} \tag{1}$$

$$H(\underline{\sigma}) = \sum_{i=1}^{n} d(\sigma_i, \sigma_{i+1}) \tag{2}$$

where $d(\sigma_\alpha, \sigma_\beta)$ is the distance between cities σ_α and σ_β and $\sigma_{n+1} \equiv \sigma_1$. The constraint that every city is visited only once can be realized by constraining the vector $\underline{\sigma}$ to be a permutation of the sequence $[1, 2, \ldots, n]$.

The optimum order of the cities for a TSP depends on their exact positions, i. e. on the random values of the distance matrix d. It is a feature of all problems we will encounter here that they are characterized by various random parameters. Each random realization of the parameters is called an *instance* of the problem. In general, if we have a collection of optimization problems of the same (general) type, we will call each single problem an instance of the general problem.

Because the values of the random parameters are fixed for each instance of the TSP, one speaks of *frozen* or *quenched* disorder. To obtain information about the general structure of a problem one has to average measurable quantities, like the length of the shortest tour for the TSP, over the disorder.

In this article we give an overview of *methods* how to *solve* these problems, i. e. how to find the optimum. Interestingly, there is no single way to achieve this. For some problems it is very easy while for others it is rather hard, this refers to the time you or a computer will need at least to solve the problem, it does not say anything about the elaborateness of the algorithms which are applied. Additionally, within the class of hard or within the class of easy problems, there is no universal method. Usually, even for each kind of problem there are many different ways to obtain an optimum. Once a problem becomes large, i. e. when the number of variables n is large, it is impossible to find a minimum by hand. Then computers are used to obtain a solution. Only the rapid development in the field of computer science during the last two decades has pushed forward the application of optimization methods to many problems from science and real life.

We will review some of the most fruitful applications of polynomial algorithms from the realm of combinatorial optimization to various problems in the statistical physics of disordered systems. The next section presents the application of Dijkstra's algorithm for finding shortest paths in weighted networks to the model of a non-directed polymer in a disordered environment with isotropical correlations. Then, in the 4th and 5th section, we discuss minimum cost flow problems on weighted graphs and its solution via the successive shortest path algorithm and apply it to the entanglement transition of elastic lines in a disordered environment and to the loop percolation transition

in a vortex glass model. In the 6th section we focus on the minimum cut-maximum flow problem and discuss among its many applications the roughening transition of elastic media in a disordered environment. The 7th section is devoted to the random field Ising model and how its ground states can be computed with maximum-flow-minimum-cut techniques. The spin glass problem is presented in the 8th section with a mapping onto minimum weighted matching in two dimensions and a brief outline of branch and cut methods for the higher dimensional case. The 9th section is devoted to finite temperature properties of the random bond Potts model and how its free energy can be computed in the limit of infinite Potts states. An outlook in the 10th section closes this chapter.

Polymers in a Disordered Environment

A well studied model of a single elastic line [10], like an individual polymer or a single magnetic flux line in a type-II superconductor, in a disordered environment is the following: If one excludes overhangs (and by this also self-overlaps) of the elastic lines one can parametrize its configuration by the longitudinal coordinate z. The line configuration can then be described by the transverse coordinate $\mathbf{r}(z)$ as a function of z. The presence of disorder is usually modeled by a random potential energy $V(\mathbf{r}, z)$ and the ground state configuration of the line is highly nontrivial due to the competition between the elastic energy, that tends to straighten the line, and the random energy, that tries to bend the line into positions of favorable energy:

$$\mathcal{H}_{\text{single-line}} = \mathcal{H}_{\text{elastic}} + \mathcal{H}_{\text{random}} = \int_0^H \mathrm{d}z \left\{ \frac{\gamma}{2} \left[\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}z} \right]^2 + V[\mathbf{r}(z), z] \right\} , \quad (3)$$

where H is the longitudinal length (not the proper length) of the line. The random potential energy is a Gaussian variable with prescribed mean and correlations $\langle\langle V[\mathbf{r}, z] V[\mathbf{r}', z']\rangle\rangle = g(\mathbf{R} - \mathbf{R}')$, where $\mathbf{R} = (\mathbf{r}, z)$ and $\langle\langle \cdots \rangle\rangle$ denotes the average over the disorder.

A lattice version of this continuum model is the *directed* polymer model: The lines correspond to directed paths on a hyper-cubic lattice that start at a specific lattice site, say $(0, 0, \ldots, 0)$ and proceed only in the $(1, 1, \ldots, 1)$ direction along the bonds. The energy contribution for a path passing bond $\mathbf{i}$ of the lattice is a *positive* random variable $e_{\mathbf{i}}$ and the total energy of a path $\mathcal{P}$ is simply

$$\mathcal{H}^{\text{lattice}}_{\text{single-line}} = \sum_{\mathbf{i} \in \mathcal{P}} e_{\mathbf{i}} = \sum_{\mathbf{i}} e_{\mathbf{i}} n_{\mathbf{i}} , \quad (4)$$

where $n_{\mathbf{i}} = 1$ if the path passes bond $\mathbf{i}$ (i. e. $\mathbf{i} \in \mathcal{P}$) and $n_{\mathbf{i}} = 0$ otherwise.

One is interested in isotropically correlated disorder and consider the problem on a *non-directed* (square) lattice (i. e. paths can pass any bond in both directions) in order not too exclude overhangs right from the beginning. In case of uncorrelated disorder overhangs were shown to be irrelevant [12], but for isotropically correlated disorder this is not clear. The latter is defined to decay algebraically with the spatial distance of the bonds

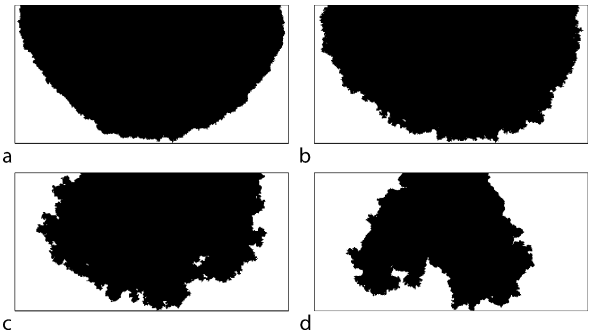
$$\langle\langle e_{\mathbf{i}} - e_{\mathbf{j}}\rangle\rangle = |\mathbf{R}_{\mathbf{i}} - \mathbf{R}_{\mathbf{j}}|^{2\alpha - 1} , \quad (5)$$

where $\mathbf{R}_{\mathbf{i}}$ spatial position of bond $\mathbf{i}$ and α is the correlation exponent: Note that one expects short-range correlations like $\langle\langle e_{\mathbf{i}} - e_{\mathbf{j}}\rangle\rangle \propto \exp(-|\mathbf{R}_{\mathbf{i}} - \mathbf{R}_{\mathbf{j}}|/\lambda)$ with a finite correlation length λ, to be irrelevant and only long-range correlations like (5) to change the universality class of the system. Increasing α imply stronger correlations, uncorrelated disorder corresponds to $\alpha \to -\infty$. The kind of correlated disorder described by (5) can be realized by generating correlated random numbers are generated using a well-established numerical procedure [11].

Exact ground states of the Hamiltonian (4) or optimal paths are calculated using Dijkstra's algorithm (note that all energies $e_{\mathbf{i}}$ are positive). This simple polynomial algorithm works as follows: Let $V = \{1, \ldots, L^d\}$ be the set of lattice sites and $A = \{(i, j) | i, j \in V \text{ nearest neighbors}\}$ the set of bonds. The algorithm increases successively a subset S of sites for which the optimal path starting at the fixed site s are known. Obviously initially $S := \{s\}$. We denote the energy of the optimal path starting at s and terminating at i with $E(i)$ and since all optimal paths can be constructed via a predecessor list, we keep track of this list, too, via an array pred(i), denoting the predecessor site of site i in a **shortest path** from s to i:

algorithm Dijkstra
begin
  $S := \{s\}$; $\overline{S} := V \backslash \{s\}$;
  $E(s) := 0$, pred(s) $:= 0$;
  while $|S| < |V|$ **do**
  begin
    choose (i, j) : $E(j) := \min_{k,m}\{E(k) + e_{(k,m)} | k \in S, m \in \overline{S}, (k, m) \in A\}$;
    $\overline{S} := \overline{S} \backslash \{j\}$; $S := S \cup \{j\}$;
    pred(j) $:= i$;
  end
end

In Fig. 1 we show examples of the set $\{i\}$ of lattice sites that are end-points of optimal paths starting from a fixed



Optimization Problems and Algorithms from Computer Science, Figure 1

Example for the growth front of the non-directed polymer for uncorrelated disorder (a and b) and correlated disorder (c and d; $\alpha = 0.4$). The *black pixels* indicate the lattice sites of the (square) lattice are connected via optimal paths to the offspring (center of the top line) with energy less than a given value (from [13])

initial site and having a total energy $E(i)$ less than a given value $E_{\max}$. For uncorrelated disorder the surface of this set is roughly a semi-circle, whereas for strongly correlated disorder the surface becomes topologically more complicated.

The universal properties of the optimal paths are typically described the scaling of two characteristic quantities: The average transverse fluctuations $\langle\langle \mathbf{r}^2 \rangle\rangle$ and the average energy fluctuations $\langle\langle E^2 \rangle\rangle$. Both are expected to grow algebraically with the the longitudinal distance H between starting point and end point of the paths:

$$\langle\langle \mathbf{r}^2 \rangle\rangle \propto H^{\nu} \quad \text{and} \quad \langle\langle E^2 \rangle\rangle \propto H^{\omega} \,, \tag{6}$$

where ν is called the roughness exponent and ω the energy fluctuation exponent. For uncorrelated disorder ($\alpha \to -\infty$) one knows $\nu = 2/3$ and $\omega = 1/3$. By computing the optimal paths for several thousands of samples for a given disorder correlation exponent α and for a given longitudinal distances H and fitting the resulting data for transverse and energy fluctuations to the expected power laws we can extract the exponents ν and ω (for details see [13]). The resulting estimates in 2d show that the correlations are relevant for $\alpha > 0$ and the roughness exponent increases linearly for $\alpha > 0$ from its value for uncorrelated disorder $\nu = 2/3$. Although the number of overhangs in the optimal paths we computed in the non-directed case increased with α (i. e. increasing correlations) the fraction of bonds contributing to overhangs scaled to zero for all values of α we considered. Hence overhangs appear to be irrelevant also in the presence of correlated disorder.

Many Repulsive Elastic Lines in Random Media

When one puts interacting elastic lines together into a finite system with a given density of lines they will show interesting collective behavior. Examples are the entanglement of magnetic flux lines in high-T_c superconductors in the mixed phase [14] or the entanglement of polymers in materials like rubber [15]. The degree of entanglement of the lines usually manifests itself in various measurable properties like stiffness or shear modulus in the case of polymers and in transport or dynamical properties for magnetic flux lines in superconductors. A theoretical description of these line systems can be based on the single-line Hamiltonian (3) plus an appropriate line interaction term:

$$\mathcal{H}_{\text{many-lines}} = \sum_{i=1}^{N} \mathcal{H}^{(i)}_{\text{single-line}} + \sum_{i<j} \int_0^L \mathrm{d}z \int_0^L \mathrm{d}z' \; V_{\text{int}}[\mathbf{R}_i(z) - \mathbf{R}_j(z')] \,, \tag{7}$$

where $\mathbf{R}_i(z) = (\mathbf{r}_i(z), z)$ is the spatial position of the infinitesimal line segment $\mathrm{d}z$ of the ith line. If the interactions $V_{\text{int}}[\mathbf{R}_i(z) - \mathbf{R}_j(z')]$ are short ranged (i. e. in case of flux lines the screening length small compared to the average line distance) or just hard core repulsive, and the random, δ-correlated disorder potential $V_r[\mathbf{r}_i(z), z]$ in (3) is strong compared to the elastic energy ($\propto \gamma$) this continuum model reduces to a lattice model reminiscent of the single-line lattice model (4):

$$\mathcal{H}^{\text{lattice}}_{\text{many-lines}} = \sum_{\mathbf{i}} e_{\mathbf{i}} n_{\mathbf{i}} \,, \tag{8}$$

where $n_{\mathbf{i}} = 1$ if a line passes bond $\mathbf{i}$ and $n_{\mathbf{i}} = 0$ otherwise and the *positive* random variable $e_{\mathbf{i}}$ is the energy cost for a line segment to occupy bond $\mathbf{i}$. The hard core constraint is thus enforced on the bonds but for the sake of an easier formal description we allow the lines to touch in isolated points, the lattice sites. The lines live on the bonds of a simple cubic lattice with a lateral width L and a longitudinal height H($L \times L \times H$ sites) with free boundary conditions in all directions. Each line starts and ends at an arbitrary position on the bottom respective top planes. The number N of lines threading the sample is fixed by a prescribed density $\rho = N/L^2$. For a single line $N = 1$, one recovers the non-directed polymer model (4). The random bond energies are uniformly distributed over the interval $[0, 1]$.

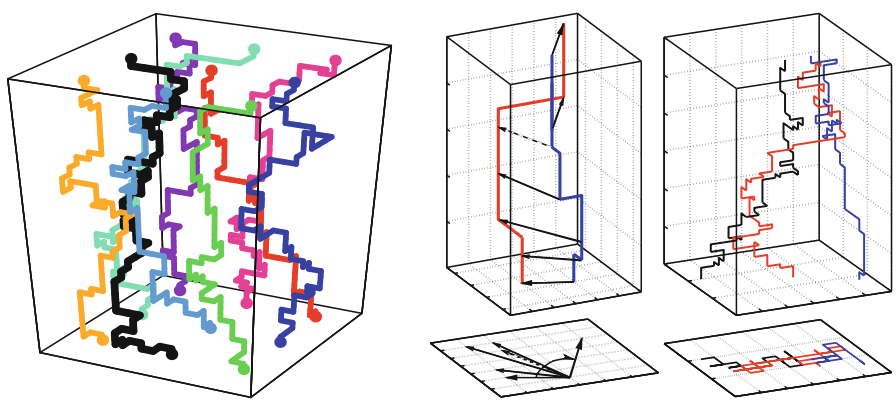
Note that the allowed configurations of the bond variables $n_{\mathbf{i}}$ are only those that can be identified with lines

threading the samples (or loops inside the sample, which, however, cost energy and therefore do not occur in the ground state), which means that the number of occupied bonds connected to a lattice site that lies neither on the top nor on the bottom plane has always to be even. If we connect all sites on the top to an extra site, called the source, an all sites on the bottom to another extra site, called the target, than the latter statement remains true also for the top an bottom plane. We can now say that N lines start at the source node and terminate at the target node, or, in network flow jargon: The feasible configurations of the variables $n_{\mathbf{i}}$ constitute a flow with zero excess on all lattice sites and an excess $+N$ and $-N$ for the source and target node, respectively.

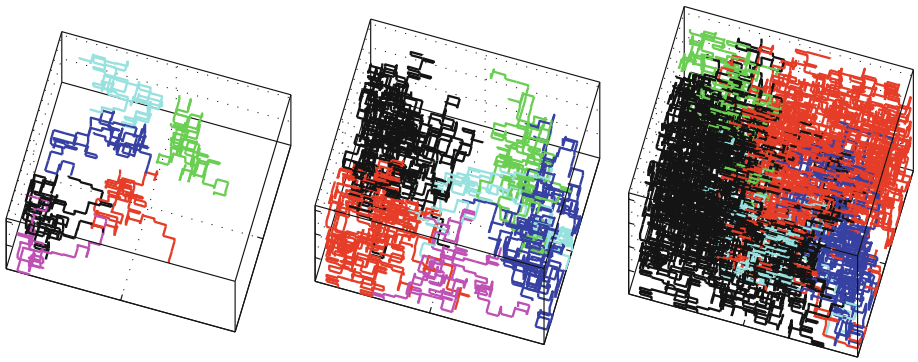
Thus the determination of the ground state configuration of the N-line problem with the Hamiltonian (8) is a **minimum-cost-flow-problem**, which can be solved with a successive shortest path algorithm [1,2,3]. In essence one starts with the zero flow $n_{\mathbf{i}} = 0$, corresponding to zero lines in the system, and sends successively one unit of flow from the source to the target, corresponding to adding one line after the other to the system. This has to happen with the minimal energy, i. e. along optimal paths, which are calculated using Dijkstra's algorithm that we encountered already in the single line problem discussed in the last section. However, when trying to add a line to a system with a number, say M, of lines already present, the existing line configuration sometimes must be changed to minimize the total energy for $M + 1$ line solution. That becomes feasible by allowing flow to be sent *backwards* on already occupied bonds. By this operation one *gains* energy (whereas occupying an empty bond $\mathbf{i}$ always costs energy $e_{\mathbf{i}} \geq 0$), which means one has to operate on a network that has to be adapted to the existing flow configuration and has negative energies on all occupied bonds. Unfortunately Dijkstra's algorithm works only for positive bond energies, and one has either to use a slower (label-correcting) algorithm to find the optimal paths in a graph with negative edge costs [3] or one has to use the concept of node potentials, by which one can make all energies in the adapted network non-negative without changing the actual shortest paths. This procedure is described in full detail in [3].

The resulting line configuration is then analyzed. One computes the winding angle of all line pairs as indicated in Fig. 2 (c.f. [16]). For each z-coordinate the vector connecting the two lines is projected onto that basal plane (*left* part of Fig. 2). $z = 0$ gives the reference line with respect to which the consecutive vectors for increasing z-coordinate have an angle $\phi(z)$. If the two lines intersect one neglects the intersection point and interpolate between the last and the next point in such a way that the global winding angle is minimized. One defines two lines to be *entangled* when $\phi(z) > 2\pi$. This choice is one that measures entanglement from the topological perspective [17], and comes from the requirement that an entangled pair of lines can not be separated by a suitable linear transformation in the basal plane (i. e. the lines almost always would cut each other, if one were shifted). The precise definition of entanglement is not of major relevance, and the one used is useful since it is the computationally easiest.

Sets or *bundles* of pairwise entangled lines are formed so that a line belongs to a bundle if it is entangled at least with one other line in the set. The topological multi-line-entanglement could be characterized by other measures



Optimization Problems and Algorithms from Computer Science, Figure 2
Left: **Ground state configuration of a *N*-line system with $N = 9$ defined by (8). The entry/exit points are fixed in a regular 3×3 array for better visibility. *Right:* Definition of the winding angle of two flux lines. *Right part, top*: A configuration of three lines that are entangled. *Right part, bottom*: The projection of the line configuration on the basal plane, defining a connected cluster**



Optimization Problems and Algorithms from Computer Science, Figure 3
Line configurations for different heights H (*from left to right*: $H = 64, 96, 128$), the lateral size $L = 20$, the line density is $\rho = 0.3$. Only the largest line bundles are shown, indicated by a varying gray scale. *Black* denotes the largest cluster, which eventually percolates

as well; the universal properties of the transition will not depend on these. These line bundles are spaghetti-like – i. e. topologically complicated and knotted sets of one-dimensional objects. To study the size distribution of these objects one projects these bundles on the basal plane, as indicated in Fig. 2, where a bundle projects onto a connected cluster. The probability for two lines to be entangled increases with increasing system height. Consequently one would expect that the bundle size increases with H, and therefore also their projections, the clusters. This scenario is exemplified in Fig. 3, for the largest height the largest cluster spans from one side of the system to the other, i. e. it *percolates*.

Hence, for a given line density ρ one expect that for system heights larger than a critical value H_c an system spanning large entangled bundle occurs, containing an infinite number of lines in the limit $L \to \infty$. One calls this an *entanglement transition* occurring at a finite system height H_c. In the projection plane this appears like a percolation transition and in [18] it was shown that this transition is in the same universality class as conventional bond percolation.

Vortex Glasses and Loop Percolation

Another application of the successive shortest path algorithm for minimum-cost-flow-problems is finding the ground state of the Hamiltonian

$$H = \sum_{\mathbf{i}} (n_{\mathbf{i}} - b_{\mathbf{i}})^2$$
$$\text{with the constraint} \quad \forall k : \sum_{l \text{ n.n. of } k} n_{(kl)} = 0 \,, \tag{9}$$

where the integer variables $n_{\mathbf{i}}$ live on the bonds $\mathbf{i}$ of a d-dimensional hyper-cubic lattice and $b_{\mathbf{i}} \in [-2\sigma, 2\sigma]$ are real valued quenched random variables with $\sigma \geq 0$ setting the strength of the disorder. The constraint $\sum_{l \text{ n.n. of } k} n_{(kl)} = 0$ means that at all lattice sites k the incoming flow has to balance the outgoing flow, i. e. the flow $\{n_{\mathbf{i}}\}$ is divergenceless. The physical motivation of studying models these kind of models is the following:

In 2d the Hamiltonian (9) occurs for instance in the context of the solid-on-solid (SOS) model on a disordered substrate [19]. The SOS representation of a 2d surface is defined by integer height variables u_k for each lattice site k of a square lattice. The disordered substrate is modeled via random offsets $d_k \in [0, 1]$ for each lattice site, such that the total height at lattice site k is $h_k = u_k + d_k$. The the total energy of the surface is

$$\mathcal{H}_{\text{SOS}} = \sum_{(kl)} (h_k - h_l)^2 = \sum_{(\tilde{kl})} \left(n_{(\tilde{kl})} - b_{(\tilde{kl})}\right)^2 \tag{10}$$

where the first sum runs over all nearest neighbor pairs (kl) of the square lattice and the second sum runs over all bonds $(\tilde{kl})$ of the *dual* lattice (being a square lattice, too), which connect the centers of the elementary plaquettes of the original lattice. A dual bond $(\tilde{kl})$ therefore crosses perpendicularly a bond (kl) connecting neighbors k and l on the original lattice. We define $n_{(\tilde{kl})} = n_k - n_l$ and $d_{(\tilde{kl})} = d_l - d_k$ if l is either the right or the upper neighbor of k (i. e. for $k = (x, y)$ either $l = (x + 1, y)$ or $l = (x, y + 1)$ and $n_{(\tilde{kl})} = n_l - n_k$ and $d_{(\tilde{kl})} = d_k - d_l$ if l is either the left or the lower neighbor of k (i. e. for $k = (x, y)$ either $l = (x - 1, y)$ or $l = (x, y - 1)$. In this way the sum over all four dual bond variables attached to one site of the dual lattice corresponds to the sum of original height variables around an elementary plaquettes in the original lattice: $(n_{(x,y)} - n_{(x,y+1)}) + (n_{(x,y+1)} - n_{(x+1,y+1)}) + (n_{(x+1,y+1)}$

$-n_{(x+1,y)})+(n_{(x+1,y)}-n_{(x,y)})=0$, which implies that the flow $\{n_{(\tilde{k}l)}\}$ is divergence free as inferred in (9).

In 3d the Hamiltonian (9) is the strong screening limit of the vortex glass model for disordered superconductors [20,21]

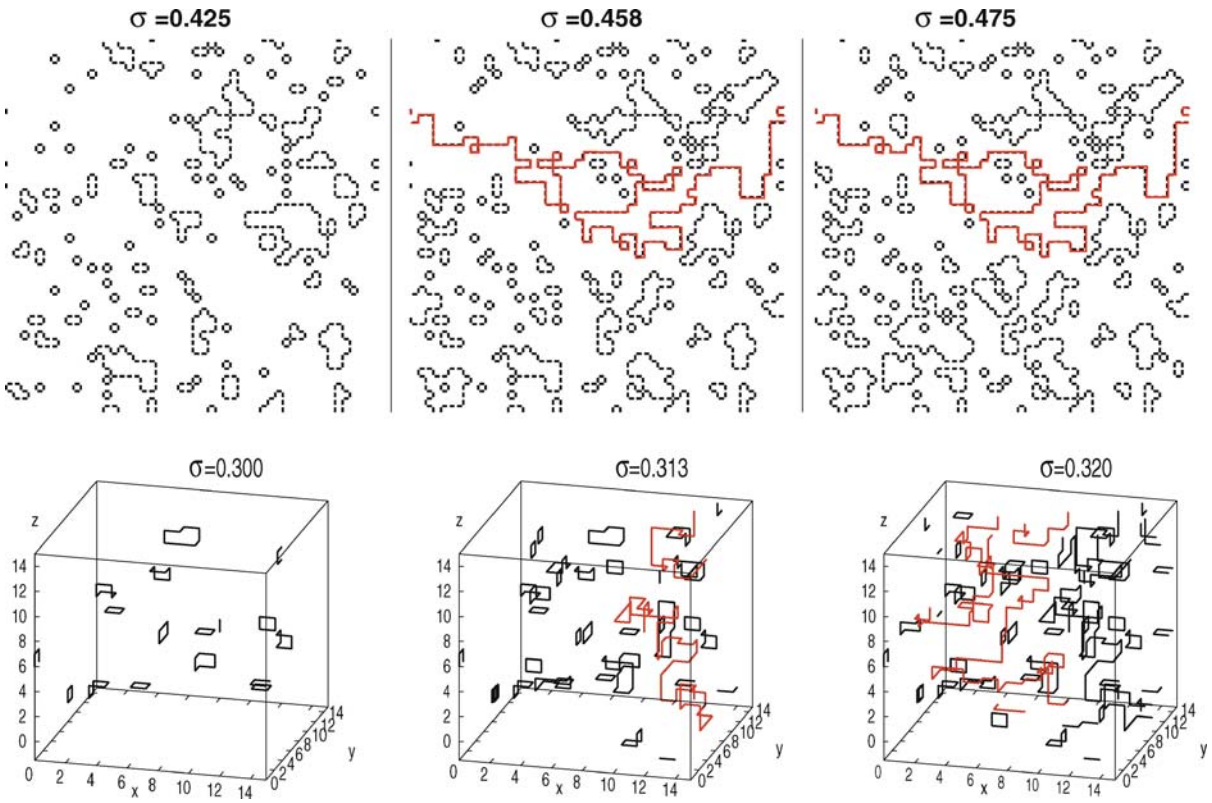
$$\mathcal{H}_{\rm VG}=\sum_{i,j}(n_i-b_i)G_\lambda(\mathbf{r}_i-\mathbf{r}_j)(n_j-b_j)\,, \tag{11}$$

where the integer vortex variables n_i live on the bonds of a simple cubic lattice and have to fulfill the constraint in (9) since they represent magnetic vortex lines that are divergence free. The real valued quenched random variables $b_i\in[-2\sigma,2\sigma]$ are derived from the lattice curl of a random vector potential ($\sigma\geq 0$ being the strength of the disorder). The 3d vector $\mathbf{r}_i$ denotes the spatial positions of bond i in the lattice and the sum runs over all bond pairs of the lattice (not only nearest neighbors). The lattice propagator $G_\lambda(\mathbf{r})$ has the asymptotic form $G_\lambda(\mathbf{r})\propto\exp(-|\mathbf{r}|/\lambda)/|\mathbf{r}|$, where λ is the screening length. In the strong screening limit $\lambda\to 0$ only the on-site repulsion survives [20] and gets

$$\mathcal{H}_{\rm VG}^{\lambda\to 0}=\sum_i(n_i-b_i)^2 \tag{12}$$

which is the Hamiltonian (9) in 3d that we intend to discuss here.

The ground state of (9) can again be computed within polynomial time by a successive shortest path algorithm [3]. As for the N-line problem one starts with a configuration $\{n_{\mathbf{i}}\}$ that optimizes the Hamiltonian in (9) but does not, in general, fulfill the mass balance constraint given in (9). In the N-line problem that was simply the zero-flow $n_{\mathbf{i}}=0$, which does not fulfill the requirement that the source and the target have excess $+N$ and $-N$, respectively. Here we start with $n_{\mathbf{i}}$ the closest integer to the real number $b_{\mathbf{i}}$ for each bond $\mathbf{i}$. Since this solution violates the mass-balance constraint one successively sends flow from nodes that have an excess flow to nodes that have a deficit along optimal paths that are again found using node potentials (to make all costs non-negative) and



Optimization Problems and Algorithms from Computer Science, Figure 4

Examples of ground state configurations of the Hamiltonian (9) for varying disorder strengths σ (for particular disorder realizations). *Top*: 2d, $L=50$, the critical disorder strength is $\sigma_c\approx 0.46$; *Bottom*: 3d, $L=16$, the critical disorder strength is $\sigma_c\approx 0.31$. The occupied bonds ($n_i\neq 0$) are marked *black*, the percolating loop is marked by *light gray (red)*

Dijkstra's algorithm. The details of this algorithm can be found in [1,2,3].

Figure 4 shows three typical ground state configurations for different strength of the disorder σ in 2d and in 3d. For small σ only small isolated loops occur, whereas for larger σ one finds loops that extend through the whole system, they percolate. A finite size scaling study of the underlying percolation transition [22] yields a novel universality class with numerically estimated critical exponents that differ significantly from those for conventional bond- or site-percolation [22].

Interfaces and Elastic Manifolds

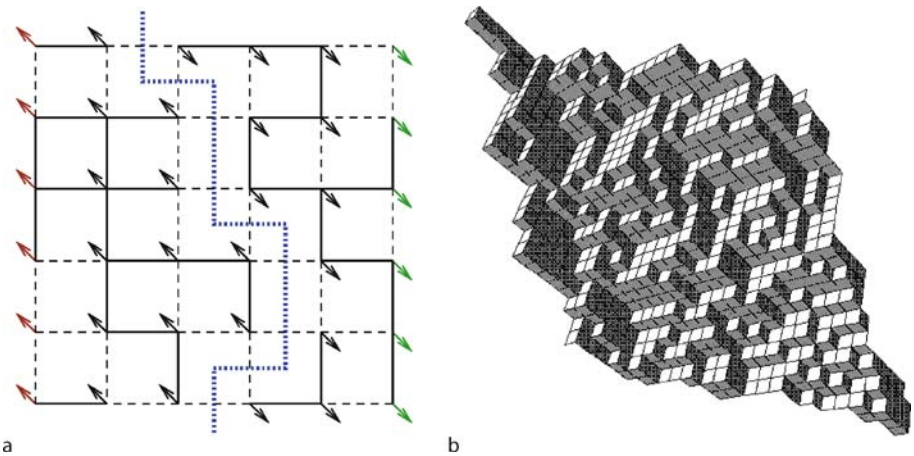
A system of strongly interacting (classical) particles or other objects, like magnetic flux lines in a type-II superconductor (as we discussed in Sect. "Many Repulsive Elastic Lines in Random Media" and for which the starting Hamiltonian would given by (7)), or a charge density wave system in a solid, will order at low temperatures into a regular arrangement a lattice (crystal lattice or flux line lattice). Fluctuations either induced by thermal noise (temperature) or by disorder (impurities, pinning centers) induce deviations of the individual particles from their equilibrium positions. As long as these fluctuations are not too strong an expansion of the potential energy around these equilibrium configuration might be appropriate. An expansion up to 2nd order is called the elastic description or elastic approximation, which in a coarse grained form (where the individual particles that undergo displacements from their equilibrium positions do not occur any more and are replaced by a continuum field $\phi(\mathbf{r})$ reads then

$$\mathcal{H}_{\text{manifold}} = \mathcal{H}_{\text{elastic}} + \mathcal{H}_{\text{random}} = \int \mathrm{d}^d\mathbf{r} \left\{ \frac{\gamma}{2} |\nabla\phi(\mathbf{r})|^2 + V(\phi(\mathbf{r}), \mathbf{r}) \right\} . \quad (13)$$

The random potential energy is a delta-correlated Gaussian variable with mean zero, $\langle\langle V(\phi, \mathbf{r}) V(\phi', \mathbf{r}')\rangle\rangle = D^2\delta(\phi - \phi')\delta(\mathbf{r} - \mathbf{r}')$. The integration extends over the whole space that parameterizes the manifold, for instance $d = 1$ for an elastic line in a random potential, $d = 2$ for an interface or a surface in a disordered environment etc. Note that for $d = 1$ one recovers the single line Hamiltonian (3). The many-line Hamiltonian (7) also allows such an elastic description in the limit, in which the interactions are strong and the the random potential is weak compared to the elastic energy. In this limit the lines will only deviate moderately from a regular, translationally invariant configuration (the Abrikosov flux line lattice). This case is called an elastic periodic medium and one has to modify the φ-part of the disorder correlator such that the Hamiltonian has the correct translational symmetry [26].

Elastic Manifold

The typical example for an elastic manifold in a disordered environment are domain walls in the $d + 1$ dimensional random bond ferromagnet $H = -\sum_{\langle ij\rangle} J_{ij}\sigma_i\sigma_j$ ($J_{ij} \geq 0$, random) in which we fix all spins in the



Optimization Problems and Algorithms from Computer Science, Figure 5
Left: **Sketch of a 2d (RBIM) with antiperiodic boundary conditions.** ***Broken lines*** **represent weak bonds,** ***full lines*** **strong bonds, the spin configuration with the lowest energy defines an interface, as indicated, and corresponds to the minimum cut in the corresponding network flow problem.** ***Right:*** **An optimal interface in the 111-direction of a 3d RBIM corresponding to the ground state configuration of a 2d elastic medium with scalar displacement field (from [23])**

lower (upper) plane, i. e. all σ_i with $i = (x_1 1, \ldots, x_d, y)$ and $y = 1\,(y = H)$, to be $\sigma_i = +1\,(-1)$, c.f. Fig. 5. First one maps it onto a flow problem in a capacitated network. One introduces two extra sites, a source node s, which is connected to all spins of the hyperplane $y = 1$ with bonds $J_{s,(x_1,\ldots,x_d,y=1)} = J_\infty$, and a sink node t, which is connected to all spins of the hyperplane $y = H$ with bonds $J_{s,(x_1,\ldots,x_d,y=H)} = J_\infty$. One chooses $J_\infty = 2\sum_{(ij)} J_{ij}$, i. e. strong enough that the interface cannot pass through a bond involving one of the two extra sites. Now we enforce the aforementioned boundary conditions for the spins in the upper and the lower plane by simply fixing $\sigma_s = +1$ and $\sigma_t = -1$. The graph underlying the capacitated network one has to consider is now defined by the set of vertices (or nodes) $N = \{1, \ldots, H \cdot L^d\} \cup \{s, t\}$ and the set of edges (or arcs) connecting them $A = \{(i, j) | i, j \in N,\ J_{ij} > 0\}$.

The capacities $u_i j$ of the arcs (i, j) is given by the bond strength $J_i j$. For any spin configuration $\mathbf{ff} = (\sigma_1, \ldots, \sigma_N)$ one defines $S = \{i \in N | \sigma_i = +1\}$ and $\overline{S} = \{i \in N | \sigma_i = -1\} = N \backslash S$. Obviously $\sigma_s \in S$ and $\sigma_t \in \overline{S}$. The knowledge of S is sufficient to determine the energy of any spin configuration via $H(S) = -C + 2\sum_{(i,j)\in(S,\overline{S})} J_{ij}$ where $(S, \overline{S}) = \{(i, j) | i \in S, j \in \overline{S}\}$. The constant $C = \sum_{(i,j)\in A} J_{ij}$ is irrelevant, i. e. independent of S. Note that $(S, \overline{S})$ is the set of edges (or arcs) connecting S with $\overline{S}$, this means it cuts N in two disjoint sets. Since $s \in S$ and $t \in \overline{S}$, this is a so called s-t-cut-set, abbreviated $[S, \overline{S}]$. Thus the problem of finding the ground state configuration of an interface in the random bond ferromagnet can be reformulated as a **minimum cut** problem

$$\min_{S\subset N} \{H'(S)\} = \min_{[S,\overline{S}]} \sum_{(i,j)\in(S,\overline{S})} J_{ij}\,. \tag{14}$$

in the above defined capacitated network (with $H' = (H + C)/2$). It does not come as a surprise that this minimum cut is *identical* with the interface between the $(\sigma_i = +1)$-domain and the $(\sigma_i = -1)$-domain that has the lowest energy. Actually any s-t-cut-set defines such an interface, some of them might consist of many components, which is of course energetically unfavorable.

A flow in the network G is a set of nonnegative numbers x_{ij} subject to a capacity constraint and a mass balance constraint for each arc

$$0 \le x_{ij} \le u_{ij}$$

$$\text{and} \sum_{\{j|(i,j)\in A\}} x_{ij} - \sum_{\{j|(j,i)\in A\}} x_{ji} = \begin{cases} -v & \text{for } i = s \\ +v & \text{for } i = t \\ 0 & \text{else}\,. \end{cases} \tag{15}$$

This means that at each node everything that goes in has to go out, too, with the only exception being the source and the sink. What actually flows from s to t is v, the value of the flow. The **maximum flow problem** for the capacitated network G is simply to find the flow $\mathbf{x}$ that has the maximum value v under the constraint (15).

Let $\mathbf{x}$ be a flow, v its value and $[S, \overline{S}]$ an s-t-cut. Then, by adding the mass balances for all nodes in S one has $v = \sum_{(i,j)\in(S,\overline{S})} x_{ij} - \sum_{(i,j)\in(\overline{S},S)} x_{ji}$ and since $x_{ij} \le u_{ij}$ and $x_{ji} \ge 0$ the following inequality holds: $v \le \sum_{(i,j)\in(S,\overline{S})} u_{ij} = u[S, \overline{S}]$. Thus the value of any flow $\mathbf{x}$ is less or equal to the capacity of any cut in the network. If one discovers a flow $\mathbf{x}$ whose value equals to the capacity of some cut $[S, \overline{S}]$, then $\mathbf{x}$ is a maximum flow and the cut is a minimum cut. The following implementation of the augmenting path algorithm constructs a flow whose value is equal to the capacity of a s-t-cut it defines simultaneously. Thus it will solve the maximum flow problem (and, of course, the minimum cut problem).

Given a flow $\mathbf{x}$, the residual capacity $r_i j$ of any arc $(i, j) \in A$ is the maximum additional flow that can be sent from node i to node j using the arcs (i, j) and (j, i). The residual capacity has two components: 1) $u_{ij} - x_{ij}$, the unused capacity of arc (i, j), 2) $x_j i$ the current flow on arc (j, i), which one can cancel to increase the flow from node i to j $r_{ij} = u_{ij} - x_{ij} + x_{ji}$. The residual network $G(\mathbf{x})$ with respect to the flow $\mathbf{x}$ consists of the arcs with *positive* residual capacities. An augmenting path is a directed path from the node s to the node t in the residual network. The *capacity of an augmenting path* is the minimum residual capacity of any arc in this path.

Obviously, whenever there is an augmenting path in the residual network $G(\mathbf{x})$ the flow $\mathbf{x}$ is not optimal. This motivates the following generic augmenting path algorithm:

algorithm Ford–Fulkerson
begin
  Initially set $x_{ij} := 0, x_{ji} := 0$ for all $(i, j) \in A$;
  do
    construct residual network R with capacities r_{ij};
    if there is an augmenting path from s to t in G' **then**
    begin
      Let $r_{\min}$ the minimum capacity of r along this path;
      Increase the flow in N along the path
        by a value of $r_{\min}$;
    end
  until no such path from s to t in G' is found;
end

This algorithm is polynomial in the number of lattice sites if the distribution of capacities is discrete (binary for in-

stance). In the general case it has to be improved and there are indeed more efficient algorithms to solve this problem in polynomial time. One of them is the push/relabel algorithm introduced by Goldberg and Tarjan [24]. It determines the maximal flow by successively improving a "preflow". A preflow is an edge function $f(e)$ that obeys the range constraint $0 \le f(e) \le w(e)$, but the conservation constraint at each node is relaxed: the sum of the $f(e)$ into or out of a node can be nonzero at internal (physical) nodes. The amount of violation of conservation at each node v give "excesses" $e(v)$. The basic operations of the algorithm, push and relabel, are used to rearrange these excesses. When the preflow can no longer be improved, it can, if desired, be converted to a maximal flow, proving the correctness of the algorithm. For details see [24,25]. It can be applied in the way sketched above to compute universal geometrical properties of elastic manifolds in 2 and 3 dimensions [23].

Periodic Medium

The presence of a periodic background potential, like a crystal potential, has a smoothening effect on the elastic manifold and tends to lock it into one of its minima. The competition between the random potential, that roughens the manifold, and such a periodic potential might lead to a roughening transition [27,28]. In 2d this is actually not the case [29], but in 3d there is as we will see. We consider a lattice version of the Hamiltonian

$$\mathcal{H} = \mathcal{H}_{\text{manifold}} + H_{\text{periodic}}$$
$$\text{with} \quad H_{\text{periodic}} = \int \mathrm{d}^d\mathbf{r}\; V_{\text{periodic}}(\phi(\mathbf{r}))\,, \tag{16}$$

where $V_{\text{periodic}}(\phi) = -\cos\phi$ represents the periodic potential.

We introduce a discrete solid-on-solid (SOS) type interface model for the elastic manifold whose continuum Hamiltonian is given in Eq. (16). Locally the EM remains flat in one of periodic potential minima at $\phi = 2\pi h$ with integer h. Due to fluctuations, some regions might shift to a different minimum with another value of h to create a step (or domain wall) separating domains. To minimize the cost of the elastic and periodic potential energy in Eq. (16), the domain-wall width must be finite, say ξ_o. Therefore, if one neglects fluctuations in length scales less than ξ_o, the continuous displacement field $\phi(\mathbf{r})$ can be replaced by the integer height variable $\{h_{\mathbf{x}}\}$ representing a $(3+1)d$ SOS interface on a simple cubic lattice with sites $\mathbf{x} \in \{1,\ldots,L\}^3$. The lattice constant is of order ξ_o and set to unity. The energy of the interface is given by the Hamiltonian

$$\mathcal{H} = \sum_{\langle \mathbf{x},\mathbf{y}\rangle} J_{(h_{\mathbf{x}},\mathbf{x});(h_{\mathbf{y}},\mathbf{y})} |h_{\mathbf{x}} - h_{\mathbf{y}}| - \sum_{\mathbf{x}} V_R(h_{\mathbf{x}},\mathbf{x})\,, \tag{17}$$

where the first sum is over nearest neighbor site pairs. After the coarse graining, the step energy $J > 0$ as well as the random pinning potential energy V_R becomes a quenched random variable distributed independently and randomly. Note a periodic elastic medium has the same Hamiltonian as in Eq. (17) with random but periodic J and V_R in h with periodicity p [30]. In this sense, the elastic manifold emerges as in the limit $p \to \infty$ of the periodic elastic medium.

To find the ground state, one maps the 3D SOS model onto a ferromagnetic random bond Ising model in $(3+1)$d hyper-cubic lattice with anti-periodic boundary conditions in the extra dimension [23] (for the 3 space direction one uses periodic boundary conditions instead). The anti-periodic boundary conditions force a domain wall into the ground state configuration of the $(3+1)$d ferromagnet. Note that bubbles are *not* present in the ground state. A domain wall may contain an overhang which is unphysical in the interface interpretation. Fortunately, one can forbid overhangs in the Ising model representation using a technique described in [23]. If the longitudinal and transversal bond strengths are assigned with $J/2$ and $V_R/2$ occurring in Eq. (17), respectively, this domain wall of the ferromagnet becomes equivalent to the ground state configuration of (17) for the interface with the same energy. The domain wall with the lowest energy is then determined exactly by using again the max-flow/min-cost algorithm.

In elastic media described by (17) the tendency of the periodic potential to lock the displacements competes with the roughening effect of the disorder. Analytically a roughening transition was predicted in [28] and the critical exponents could be numerically estimated in three dimensions [30] with the mapping and algorithm described above.

Random Field Ising Model

The random field Ising model (RFIM, for a review see [31,32]) is defined

$$H = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \tag{18}$$

with $\sigma_i = \pm 1$ Ising spins, ferromagnetic bonds $J_{ij} \ge 0$ (random or uniform), (ij) nearest neighbor pairs on a d-dimensional lattice and at each site i a random field $h_i \in R$

that can be positive and negative. The first term prefers a ferromagnetic order, which means it tries to align all spins. The random field, however, tends to align the spins with the field which points in random directions depending on whether it is positive or negative. This is the source of competition in this model.

Let us suppose for the moment uniform interactions $J_{ij} = J$ and a symmetric distribution of the random fields with mean zero and variance h_r. It is established by now that in 3d (and higher dimensions) the RFIM shows ferromagnetic long range order at low temperatures, provided h_r is small enough. In 1d and 2d there is no ordered phase at any finite temperature. Thus in 3d one has a paramagnetic/ferromagnetic phase transition along a line $h_c(T)$ in the h_r-T-diagram.

The renormalization group picture says that the nature of the transition is the same all along the line $h_c(T)$, with the exception being the pure fixed point at $h_r = 0$ and $T_c \sim 4.51\,J$. The RG flow is dominated by a zero temperature fixed point at $h_c(T = 0)$. As a consequence, the critical exponents determining the critical behavior of the RFIM should be all identical along the phase transition line, in particular identical to those one obtains *at zero temperature* by varying h_r alone. Thus to study the universal properties of the phase transition in the RFIM one needs to calculate its ground state.

This optimization task is again equivalent to a maximum flow problem [33,34], as in the interface model discussed in the last section. Historically the RFIM was the first physical model that has been investigated with a maximum flow algorithm [36]. However, here the minimum-cut is not a geometric object within the original system.

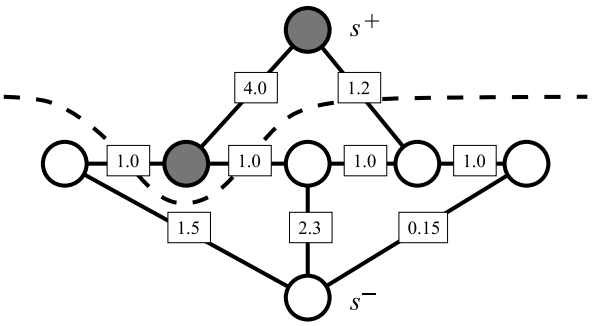
To map the ground state problem for the RFIM onto a max-flow-min-cut problem one proceeds in the same way as in the interface problem: One adds to extra nodes s and t and attaches spins with fixed values there (see Fig. 6):

$$\sigma_s = +1 \qquad \text{and} \qquad \sigma_t = -1 \tag{19}$$

One connects all sites with positive random field to the node s and all sites with negative random field to t:

$$J_{si} = \begin{cases} h_i & \text{if } h_i \geq 0 \\ 0 & \text{if } h_i < 0 \end{cases}$$
$$J_{it} = \begin{cases} |h_i| & \text{if } h_i < 0 \\ 0 & \text{if } h_i \geq 0 \end{cases} \tag{20}$$

The a network is constructed with the set of nodes $N = \{1, \cdots, L^d\} \cup \{s, t\}$ and the set of (forward and backward) arcs $A = \{(i, j) | i, j \in N,\ J_{ij} > 0\}$. Each of them has a capacity $u_{ij} = J_{ij}$. The energy or cost function can



Optimization Problems and Algorithms from Computer Science, Figure 6

Representation of the ground state problem for the RFIM as an RBIM domain wall or minimum-cut problem. The physical spins are the five nodes in the single row in the figure, while the fixed external spins are s^+ and s^-. The physical RFIM coupling $J = 1.0$. A spin with $h_i > 0$ ($h_i < 0$) is connected by an auxiliary coupling of strength h_i ($-h_i$) to s^+ (s^-). The weights of each bond are indicated: the random fields are, *from left to right*, $h = -1.5, +4.0, -2.3, +1.2$, and 0.15. In the ground state, the interfacial energy between up-spin and down-spin domains is minimized, i. e., the spins are partitioned into two sets with minimal total cost for the bonds connecting the two sets. The *dashed* curve indicates the minimal weight cut. The *white (dark)* nodes indicate up (down) spins in the ground state configuration

the be written as

$$E = -\sum_{(i,j)\in A} J_{ij}\sigma_i\sigma_j \tag{21}$$

or, by denoting the set $S = \{i \in N \,|\, S_i = +1\}$ and $\overline{S} = N \backslash S$

$$E(S) = -C + 2 \sum_{(i,j)\in(S,\overline{S})} J_{ij} \tag{22}$$

with $C = \sum_{(i,j)\in A} J_{ij}$. The problem is reduced to the problem of finding a minimum s-t-cut as in (14). The difference to the interface problem is that now the extra bonds connecting the two special nodes s and t with the original lattice do not have infinite capacity: they can lie *in* the cut, namely whenever it is more favorable not to break a ferromagnetic bond but to disalign a spin with its local random field. In the extended graph the s-t-cut again forms connected interface, however, in the original lattice (without the bonds leading to and from the extra nodes) the resulting structure is generally *disconnected*, a multi-component interface. Each single component surrounds a connected region in the original lattice containing spins, which all point in the same direction. In other words, they form ferromagnetically ordered domains separated by domain walls given by the subset of the s-t-cut that lies in the original lattice.

In passing we note that diluted Ising antiferromagnets in a homogeneous external field (DAFF) map straightforwardly onto a RFIM if the underlying lattice is bipartite. The 3d DAFF on a simple cubic lattice is defined by

$$H = +\sum_{(ij)} J_{ij}\varepsilon_i\varepsilon_j\sigma_i\sigma_j - \sum_i h_i\varepsilon_i\sigma_i \tag{23}$$

where $\sigma_i = \pm 1$, $J_{ij} \geq 0$, (ij) are nearest neighbor pairs on a simple cubic lattice, and $\varepsilon_i \in \{0, 1\}$ with $\varepsilon_i = 1$ with probability c, representing the concentration of spins. Because of the plus sign in front of the first term in (23) all interactions are antiferromagnetic, the model represents a diluted antiferromagnet, for which many experimental realizations exist (e. g. $Fe_cZn_{1-c}F_2$). Now that neighboring spins tend to point in opposite directions due to their antiferromagnetic interaction a uniform field competes with this ordering tendency by trying to align them all. On a bipartite lattice in zero external field the ground state would be antiferromagnetic, which means that one can define two bipartite sublattices A and B. One defines new spin and field variables via

$$\sigma_i' = \begin{cases} +\sigma_i & \text{for } i \in A \\ -\sigma_i & \text{for } i \in B \end{cases}$$

$$h_i' = \begin{cases} +\varepsilon_i h_i & \text{for } i \in A \\ -\varepsilon_i h_i & \text{for } i \in B\,. \end{cases}$$

Since $\sigma_i'\sigma_j' = -\sigma_i\sigma_j$ for all nearest neighbor pairs (ij) one can write (23) as

$$H = -\sum_{(ij)} J_{ij}'\sigma_i'\sigma_j' - \sum_i h_i'\sigma_i' \tag{24}$$

with $J_{ij}' = J_{ij}\varepsilon_i\varepsilon_j$. This is again a RFIM and ground states can be computed with the max-flow technique.

The main focus of the application of the max-flow-min-cut algorithm to the RFIM is the phase transition in the three-dimensional model occurring at a critical disorder strength h_c at zero temperature, which separates a paramagnetic phase for large disorder strength from a ferromagnetic phase. The maximum flow algorithm has first been used by Ogielski [36] to calculate the critical exponents of the RFIM via the finite size scaling. More accurate estimates were obtained more recently by Middleton and Fisher [35], where also an detailed discussion of the problems and conflicting results about the RFIM universality class is provided. For Gaussian random fields (with variance h^2) they find for the finite size scaling of magnetization $m = [S_i]\text{av}$ and specific heat $c = N^{-1}\mathrm{d}E/\mathrm{d}T$ and

$$m \sim L^{-\beta/\nu}\,,$$
$$c \sim L^{\alpha/\nu}\,, \tag{25}$$

with the magnetization exponent $x = \beta/\nu = 0.012 \pm 0.004$ the correlation length exponent $\nu = 1.37 \pm 0.09$, and the specific heat exponent $\alpha = -0.07 \pm 0.17$. Note that the magnetization exponent is very close to zero, which means that the transition is hard to discriminate from a first order transition. Also the specific heat exponents is close to zero and slightly negative, implying a lack of divergence of the specific heat at the transition.

The Spin Glass Problem

Spin glasses are the prototypes of (disordered) frustrated systems (see [37]). In the models discussed up to now, the frustration was caused by two separate terms of different physical origin (interactions and external fields or boundary conditions). Spin glasses are magnetic systems in which the magnetic moments interact ferro- or antiferromagnetically in a random way, as in the following Edwards–Anderson Hamiltonian for a short ranged Ising spin glass (SG)

$$H = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j\,, \tag{26}$$

where $\sigma_i = \pm 1$, (ij) are nearest neighbor interactions on a d-dimensional lattice and the interaction strengths $J_{ij} \in R$ are unrestricted in sign. In analogy to Eq. (14) one shows that the problem of finding the ground state is again equivalent to finding a minimal cut $[S, \overline{S}]$ in a network

$$\min_{\underline{\mathbf{ff}}} \{H'(\underline{\mathbf{ff}})\} = \min_{[S,\overline{S}]} \sum_{(i,j)\in(S,\overline{S})} J_{ij}\,, \tag{27}$$

again $H' = (H + C)/2$ with $C = \sum_{(ij)} J_{ij}$. However, now the capacities $u_{ij} = J_{ij}$ of the underlying network are *not* non-negative any more, therefore it is *not* a minimum-cut problem and thus it is also not equivalent to a maximum flow problem, which we know how to handle efficiently.

It turns out that the spin glass problem is *much* harder than the questions we have discussed so far. In general (i. e. in any dimension larger than two and also for 2d in the presence of an external field) the problem of finding the SG ground state is $\mathcal{NP}$-complete [42], which means in essence that no polynomial algorithm for it is known (and also that chances to find one in the future are marginal). Nevertheless, some extremely efficient algorithms for it have been developed [38,39], which have a non-polynomial bound for their worst case running-time but which terminate (i. e. find the optimal solution) after a reasonable computing time for pretty respectable system sizes.

Two Dimensions, Planar Graph

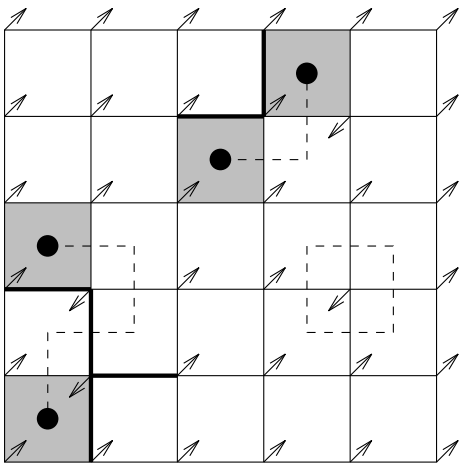
First we discuss the only non-trivial case that can be solved with a polynomial algorithm: the two-dimensional Ising SG on a planar graph. This problem can be shown to be equivalent to finding a minimum weight perfect matching, which can be solved in polynomial time. We do not treat matching problems and the algorithms to solve them in this lecture (see [4,40,41]), however, we would like to present the idea [42]. To be concrete let us consider a square lattice with free boundary conditions. Given a spin configuration $\underline{\mathbf{ff}}$ (which is equivalent to $-\underline{\mathbf{ff}}$) we say that an edge (or arc) (i,j) is satisfied if $J_{ij}\sigma_i\sigma_j > 0$ and it is *unsatisfied* if $J_{ij}\sigma_i\sigma_j < 0$. Furthermore we say a plaquette (i. e. a unit cell of the square lattice) is *frustrated* if it is surrounded by an odd number of negative bonds (i. e. $J_{ij}\cdot J_{jk}\cdot J_{kl}\cdot J_{li} < 0$ with i, j, k and l the four corners of the plaquette)). There is a one-to-one correspondence between equivalent spin configurations ($\underline{\mathbf{ff}}$ and $-\underline{\mathbf{ff}}$) and sets of unsatisfied edges with the property that on each frustrated (unfrustrated) plaquette there is an odd (even) number of unsatisfied edges. See Fig. 7 for illustration.

Note that

$$H(\underline{\mathbf{ff}}) = -C + 2 \sum_{\text{unsatisfied edges}} |J_{ij}| \,. \tag{28}$$

which means that one has to minimize the sum over the weights of unsatisfied edges. A set of unsatisfied edges will be constituted by a set of paths (in the dual lattice) from one frustrated plaquette to another and a set of closed circles (see Fig. 7). Obviously the latter always increase the energy so that we can neglect them. The problem of finding the ground state is therefore equivalent to finding the minimum possible sum of the weights of these paths between the frustrated plaquettes. This means that we have to *match* the black dots in the Fig. 7 with one another in an optimal way. One can map this problem on a minimum weight **perfect matching** problem (a perfect matching of a graph $G = (N, A)$ is a set $M \subseteq A$ such that each node has only has only one edge of M adjacent to it). This can be solved in polynomial time (see [42] for further details).

Note that for binary couplings, i. e. $J_{ij} = \pm J$, where $J_{ij} = +J$ with probability p the weight of a matching is simply proportional to the sum of the lengths of the various paths connecting the centers of the frustrated plaquettes, which simplifies the actual implementation of the algorithm. In [43] the 2d $\pm J$ spin glass and the site disordered SG has been studied extensively with this algorithm. The site disordered spin glass is defined as follows: occupy the sites of a square lattice randomly with A (with concentration c) and B (with concentration $1-c$) atoms. Now define the interactions J_{ij} between neighboring atoms: $J_{ij} = -J$ if on both sites are A-atoms and J_{ij} otherwise.



Optimization Problems and Algorithms from Computer Science, Figure 7
Two-dimensional Ising spin glass with ±-J couplings: *Thin* lines, are positive interactions, *thick* lines are negative interactions, ↗ means $\sigma_i = +1$, ↙ means $\sigma_i = -1$, *shaded* faces are frustrated plaquettes, *broken* lines cross unsatisfied edges

The main application of this algorithm is directed towards studying domain walls in spin glasses since they provide informations on the low temperature behavior and the stability of the ground state with respect to thermal fluctuations. Domain walls can be induced by applying two different boundary conditions to the system (usually periodic and anti-periodic), their energy is simply the difference between the energies of the ground states with the two different boundary conditions. The domain wall energy of the two-dimensional spin glass model with Gaussian couplings scales like

$$\Delta E \sim L^{\theta} \,, \tag{29}$$

where the stiffness exponent is $\theta = -0.282$ (see [44] for a survey). The negativity of this exponent indicates the absence of stable spin glass phase at any non-vanishing temperature in the 2d spin glass model. Recently also the fractal properties of the domain walls in 2d spin glasses with Gaussian couplings became important: They have a fractal dimension of $d_f = 1.27(1)$ and it was argued [45] that they might be a realization of a stochastic Loewner evolution (see [46] for a review) realized in disordered systems.

Three Dimensions, Non-planar Graphs

As we mentioned, in any other case except the planar lattice situation discussed above the spin glass problem is

$\mathcal{NP}$-hard. In what follows we would like to sketch the idea of an efficient but non-polynomial algorithm [39]. To avoid confusion with the minimum cut problem we discussed in connection with maximum flows one calls the problem (27) a max-cut problem (since finding the minimum of H is equivalent to finding the maximum of $-H$).

Let us consider the vector space R^A. For each cut $[S, \overline{S}]$ define $\chi^{(S,\overline{S})} \in R^A$, the incidence vector of the cut, by $\chi_e^{(S,\overline{S})} = 1$ for each edge $e = (i, j) \in (S, \overline{S})$ and $\chi_e^{(S,\overline{S})} = 0$ otherwise. Thus there is a one-to-one correspondence between cuts in G and their $\{0, 1\}$-incidence vectors in R^A. The *cut-polytope* $P_C(G)$ of G is the convex hull of all incidence vectors of cuts in G: $P_C(G) = \text{conv}\{\chi^{(S,\overline{S})} \in R^A \mid S \subseteq A\}$. Then the max-cut problem can be written as a *linear program*

$$\max \{\underline{u}^T \underline{x} \mid \underline{x} \in P_C(G)\} \tag{30}$$

since the vertices of $P_C(G)$ are cuts of G and vice versa. Linear programs usually consist of a linear cost function $\underline{u}^T \underline{x}$ that has to be maximized under the constraint of various inequalities defining a polytope in R^n (i. e. the convex hull of finite subsets of R^n) and can be solved for example by the simplex method, which proceeds from corner to corner of that polytope to find the maximum (see e. g. [40,41,48]). The crucial problem in the present case is that it is $\mathcal{NP}$-hard to write down all inequalities that represent the cut polytope $P_C(G)$.

It turns out that also *partial* systems are useful, and this is the essential idea for an efficient algorithm to solve the general spin glass problem as well as the traveling salesman problem or other so called mixed integer problems (i. e. linear programs where some of the variables x are only allowed to take on some integer values, like 0 and 1 in our case) [7,47]. One chooses a system of linear inequalities L whose solution set $P(L)$ contains $P_C(G)$ and for which $P_C(G) = \text{convex hull } \{\mathbf{x} \in P(L) | x \text{ integer}\}$. In the present case these are $0 \le x \le 1$, which is trivial, and the so called cycle inequalities, which are based on the observation that all cycles in G have to intersect a cut an even number of times. The most remarkable feature of this set L of inequalities is the following:

The separation problem for a set of inequalities L consists in either proving that a vector x satisfies all inequalities of this class or to find an inequality that is violated by $\mathbf{x}$. A linear program can be solved in polynomial time if and only if the separation problem is solvable in polynomial time [49]. The separation problem for the cycle inequalities can be solved in polynomial time by the *cutting plane algorithm* which, starting from some small initial set of inequalities, generates iteratively new inequalities until the optimal solution for the actual subset of inequalities is feasible. Note that one does not solve this linear program by the simplex method since the cycle inequalities are still too numerous for this to work efficiently.

Due to the insufficient knowledge of the inequalities that are necessary to describe $P_C(G)$ completely, one may end up with a non-integral solution $\mathbf{x}^*$. In this case one *branches* on some fractional variable x_e (i. e. a variable with $x_e^* \notin \{0, 1\}$), creating two subproblems in one of which x_e is set to 0 and in the other x_e is set to 1. Then one applies the cutting plane algorithm recursively for both subproblems, which is the origin of the name *branch-and-cut*. Note that in principle this algorithm is not restricted to any dimension, boundary conditions, or to the fieldless case. However, there are realizations of it that run fast (e. g. in 2d) and others that run slow (e. g. in 3d) and it is ongoing research to improve on the latter, for an overview over the current status see [47].

Potts Free Energy and Submodular Functions

The problem addressed in this chapter is not a low temperature problem but concerns the computation of the free energy of a Potts model (see [50] for a review) at *any temperature*, including some phase transition temperatures. To transform the problem of computing the free energy into an optimization problem (i. e. find a minimum in a finite set), one needs to take some limit. Usually this is a zero temperature limit as it was for all applications discussed so far in this article. Here this will be the limit of an *infinite number of states*.

Consider the q-state Potts model on a d-dimensional hyper-cubic lattice with periodic boundary conditions defined by the Hamiltonian:

$$H = -\sum_{\langle ij \rangle} J_{ij} \delta(\sigma_i, \sigma_j) \,, \tag{31}$$

where σ_i are q-state Potts variables ($\sigma_i \in \{1, \ldots, q\}$ located at lattice sites i, the sum goes over all nearest neighbor pairs $\langle ij \rangle$ of the lattice, and $J_{ij} > 0$ are ferromagnetic couplings (not that $\delta(\sigma, \sigma')$ is the Kronecker-delta, which means $\delta(\sigma, \sigma') = 1$ for $\sigma = \sigma'$ and $\delta(\sigma, \sigma') = 0$ for $\sigma \neq \sigma'$). The case $q = 2$ corresponds to the Ising model. In the random bond Potts model, which is of interest here, the couplings J_{ij} are random variables. In $d \le 2$ dimensions the Potts model has phase transition at some critical temperature T from a paramagnetic to a ferromagnetic phase. Thermodynamic properties of the q-state Potts model are computed via its partition function

$$\mathcal{Z} = \sum_{\{\underline{\sigma}\}} \exp\left(\sum_{ij} -\beta J_{ij} \delta(\sigma_i, \sigma_j)\right) \,. \tag{32}$$

The first sum runs over all possible spin configuration, i. e. it involves q^N terms, where N is the number of spins in the system and $\beta = 1/T$ is the inverse temperature.

In the so-called random cluster representation [51] the partition sum can be written as a sum over all subsets $U \subseteq E$ of the set of edges (or bonds)

$$\begin{aligned} \mathcal{Z} &= \sum_{\{\sigma\}} \prod_{ij} \exp\left(-\beta J_{ij} \delta(\sigma_i, \sigma_j)\right) \\ &= \sum_{\{\sigma\}} \prod_{ij} \left(1 + v_{ij} \delta(\sigma_i, \sigma_j)\right) \end{aligned}$$

where $v_{ij} = \exp(\beta K_{ij}) - 1$. Note that the Kronecker-delta can only take on the values zero and one by which it is possible to identify $\exp(J\delta) = 1 + \delta(\exp(J) - 1) = 1 + v\delta$. Again one can regard the lattice as a graph $G = (V, E)$, where the sites and the bonds of the lattice are the vertices V and the edges E of the graph. Then a careful bookkeeping of the terms in the development of the above expression leads to:

$$Z = \sum_{G' \subseteq G} q^{c(G')} \prod_{e \in G'} v_e \,, \tag{33}$$

where G' denotes any subgraph of G, i. e. a graph, possibly not connected (but all vertices are kept), where some edges of G have been deleted (there are 2^m subgraphs where m is the number of edges of G). $c(G')$ is the number of connected components of the subgraph G'. For example for the empty subgraph $G' = \emptyset$ the number of connected components is the number of sites, while for $G' = G$ it is one. The product in (33) is over all the edges in G' with the convention that the product over an empty set is one. If the parameter β is small (i. e. high temperature) then the parameters v_{ij} are small and, summing in (33), only the subgraphs with few edges provides an approximation to the partition function: this is a high temperature development. Note also the way the parameter q appears in (33): it can be extended to non integer values, relating the Potts model to other problems (percolation, etc ...) [58].

Following [52] one can map the computation of the partition function Z of any ferromagnetic Potts model in the limit $q \to \infty$ onto an optimization problem by introducing another parametrization of the couplings with new variables w_e defined by

$$v_e = q^{w_e} \,.$$

Inserting this expression in (33) one gets $Z = \sum_{G' \subseteq G} q^{c(G') + \sum_{e \in G'} w_e}$, and defining $f(G) = c(G) + \sum_{e \in G} w_e$:

$$Z = \sum_{G' \subseteq G} q^{f(G')} \,.$$

In the limit $q \to \infty$ only the subgraphs $G^\star$ maximizing $f(G)$ will contribute, and computing the partition function of the Potts model in the infinite number of states limit amounts to finding the subgraphs G' of the graph G maximizing the function f, i. e. minimizing the function [52]:

$$f_P(G') = -\left(c(G') + \sum_{e \in G'} w_e\right) \,. \tag{34}$$

It turns out that this function has a property which allows to minimize it very efficiently: it is a *submodular function*.

Submodular Functions

The concept of a submodular function in discrete optimization appears to be in several respects analogous to that of a convex function in continuous optimization. In many combinatorial theorems and problems, submodularity is involved, in one form or another, and submodularity often plays an essential role in a proof or an algorithm. Moreover, analogous to the fast methods for convex function minimization, it turns out that submodular functions can also be minimized fast, i. e. in polynomial time.

Submodularity is a special property of *set functions*, which are defined as follows: Let V be a finite set and $2^V = \{X \mid X \subseteq V\}$ be the set of all the subsets of V. A function $f \colon 2^V \to \mathbb{R}$ is called a set function.

Now a set function f is **submodular** if for all subsets $A \subseteq V$ and $B \subseteq V$:

$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B) \,. \tag{35}$$

It is simple to show that a function f is submodular if and only if for any subsets $S \subseteq R \subseteq V$ and for any $x \in V$:

$$f(S \cup \{x\}) - f(S) \geq f(R \cup \{x\}) - f(R) \,. \tag{36}$$

This means intuitively that adding an element to a "small" ensemble S (since $S \subseteq R$) has more effect than adding to a "large" ensemble R.

The function (34) $f_P(A) = -(c(A) + w(A))$ is submodular, because the function $-c(A)$ is submodular (and the function $w(A)$ is modular: Take two sets of edges $A \subseteq B$ and an edge e. Inspecting the three possible cases: $e \in A$, $e \notin A$ and $e \in B$, $e \notin A$ and $e \notin B$ one sees that $c(A \cup \{e\}) - c(A) \leq c(B \cup \{e\}) - c(B)$, which is the reverse of (36), so that the function $-c$ is a submodular function. Note that $c(E')$ with $E' \subseteq E$ counts the number of connected components of the graph G' that contains *all* vertices V of the complete graph but only the edges in E'. Thus adding an edge will never increase the number of components.

On the other hand it is straightforward to see that the function $w(G) = \sum_{e \in G} w_e$ verifies $w(A \cup C) + w(A \cap C) = w(A) + w(C)$. It is a so-called *modular* function. Consequently the function (34) f_P is a submodular function. In summary we are looking for the sets of edges minimizing the submodular function f_P for which a *strongly polynomial* algorithm has been recently discovered.

In passing we note that we encountered other examples of submodular functions already in the preceding sections, namely the function that defines the costs of cuts in a graph with positive edge weights, which occurs the interface problem and the random field Ising model in the last sections: Take a graph $G = (V, E)$ and define C to be a function of the subsets of the V and $C(U \subseteq V)$ is the number of edges having exactly one end in U. This function can be generalized to the case where the edges are directed and weighted, i. e. each edge carries an arrow and a positive number. The function $C(U \subseteq V)$ is then the sum of the weight of the edges having the beginning vertex in U and the ending vertex not in U. This kind of function is generally called a "cut" and is submodular.

Minimization of Submodular Function

The minimization of any submodular function can be done in polynomial time. This was first published in reference [54] in 1981. In this paper the authors utilize the so-called ellipsoid method. However this method is not a combinatorial one and is far from being efficient. In that respect this result was not quite satisfactory at least for the practical applications. Eighteen years later, Iwata–Fleischer–Fujishige [55], and independently Schrijver [56] discovered a combinatorial method which is fully satisfactory from the theoretical, as well as from the practical, point of view.

The general method uses a mathematical programming formulation. The problem is algebraically expressed as a linear program, i. e. a set of variables y_S associated to each subset $S \subset V$ is introduced, these variables are subjected to constraints and a linear function F of these variables is to be minimized. The constraints include a set of linear equations and the condition that each of the y_S is zero or one. This last condition is in general extremely difficult to realize. However, it turns out that a theorem due to Edmonds [57] indicates this condition can be simply dropped, and that automatically the set of values y_S which minimize F will all be zero or one! Actually only one variable $y_{S^\star} = 1$ will be non zero and it is precisely associated to the optimal set. Combined with the dual version of this linear program, it provides a characterization of the optimal set.

The general algorithm mentioned above can be applied to minimize (34), however, due to the specific form of the function to minimize, a more suitable method does exist. For this a property that is true for any submodular function is useful. To emphasize that the function f to minimize is defined on all the subsets of a set E we will label f with the index E as f_E. Let us now consider a subset $F \subseteq E$; one can define a set function on F by $f_F(A) = f_E(A)$ for any $A \subseteq F$. If the function f_E is submodular then its restriction f_F is also submodular. We have the following property:

Let $F \subseteq E$ and $e \in E$, if A_F is an optimal set of the set function f_F defined on F, then there will be an optimal set $A_{F \cup \{e\}}$ of the function $f_{F \cup \{e\}}$ defined on $F \cup \{e\}$ such that $A_F \subseteq A_{F \cup \{e\}}$.

To make the notation simpler we denote the function $f_{F \cup \{e\}}$ on $F \cup \{e\}$ by f_1. Let A be an optimal set of f_F on F and B an optimal set of f_1 on $F \cup \{e\}$. One has

$$f_1(A \cup B) \leq f_1(A) + f_1(B) - f_1(A \cap B) \tag{37}$$

since f_1 is submodular. But $f_1(A) = f_F(A)$ and $f_1(A \cap B) = f_F(A \cap B)$ since both A and $A \cap B$ are in A. Since A is an optimal set one has $f_F(A) \leq f_F(A \cap B)$ and consequently $f_1(A) - f_1(A \cap B) \leq 0$. Inserting this last inequality into (37) one finds that $f_1(A \cup B) \leq f_1(B)$ which proves that $A \cup B$ is one of the optimal sets (Q.E.D.).

This property has an important consequence. Indeed let us suppose that the optimal set has been found for a subset F of E. Then all the elements of E which have been selected as belonging to the optimal set of F will still belong to one optimal set of all the sets $G \supseteq F$. In other words, let us find the optimal set for $\{e_0, e_1\}$ where e_0 and e_1 are *arbitrary* elements of E; then if we find that any of these two elements belongs to the optimal set, it will belong to one optimal set for $F \subseteq E$! Such an algorithm which makes a definitive choice at each step is called a *greedy* algorithm.

Based on this observation an efficient algorithm for the minimization of (34) was developed in [59], see also [60].

Results

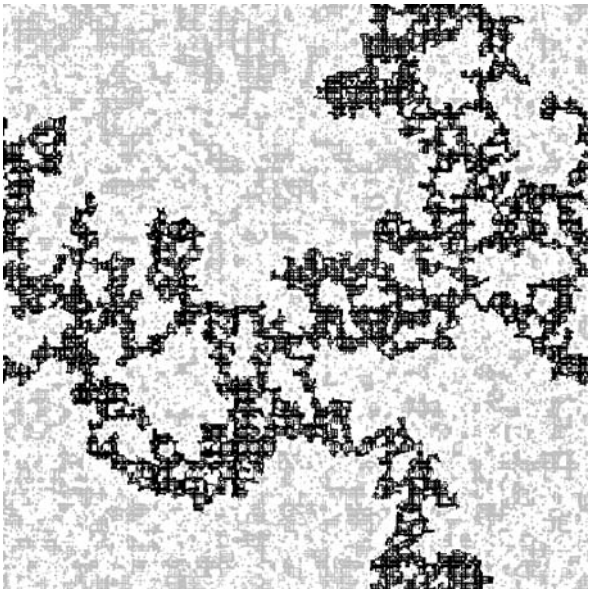
The algorithm based on the ideas mentioned before and presented in detail in [59,60], was applied to various two dimensional and three dimensional lattices. A realization of the disorder is chosen accordingly to a probability distribution. In practice all the weights $w(e)$ on the edge e are rational numbers with a common integer denominator q. In other words, we choose an integer $p(e)$ for each

edge and set $w(e) = p(e)/q$. To work only with integers one maximizes the product qf:

$$qf(A) = qC(A) + \sum_{e \in A} p(e) \, .$$

It is clear that if q is small compare to all the $p(e)$, then all the weights $w(e)$ will be large and the optimal set will be the set of all edges. On the contrary if q is large all the weights will be small and the optimal set will be empty. These two situations are easy to handle. Between this two limits the optimal set grows, and for a precise value q_c of q, which depends on the lattice, the optimal set percolates. This value corresponds to a phase transition. Depending on the lattice under consideration and on the distribution of the random variables $p(e)$ this transition can be first or second order.

In Fig. 8, one optimal set is shown for a lattice where each edge carries a weight 1/6 or 5/6 with probability one half (i. e. it is a critical point). The edges from the optimal set belonging to the percolation cluster are shown in black, while the others are shown in gray. The percolation cluster, which is the largest connected component in the optimal subgraph $G' \subseteq G$ is fractal with a fractal dimension $d_f = 1.809$ that is related to the critical exponent $x = \beta/\nu$ for the magnetization of the random bond $q \to \infty$ Potts model (31) in two dimensions via $x = 2 - d_f = 0.191$. Surprisingly this agrees within the error bars with the magnetization exponent $x = (3 - \sqrt{5})/4$ of the random transverse Ising chain [62], which is a one-dimensional quantum spin model. A discussion of this observation and details of the computations can be found in [61].



Optimization Problems and Algorithms from Computer Science, Figure 8

A 512 × 512 lattice. The edges of the optimal set belonging to the percolating cluster are shown in *black*, and the edges of the optimal set not belonging to the optimal set are in *gray* (from [60])

Future Directions

We have reviewed several applications of polynomial optimization algorithms from computer science to disordered systems in statistical physics. They were used extensively in the recent years to compute numerically universal properties like critical exponents, domain wall exponents and geometrical features like roughness and stiffness with much higher precision than with Monte-Carlo methods, which suffer notoriously from equilibration problems. A number of important issues, which were controversially debated within different analytical could be clarified, numerically, in this way – as for instance the nature of the low temperature phase of the superrough phase in the two-dimensional Bragg glass [19,63], the absence of a stable glass phase in the strongly screened vortex glass model [21] and the issue of many states in various two-dimensional glassy models [64]. Other questions still remain to be answered, as for example the phenomenon of an apparent non-universality in the three-dimensional random field Ising model [65].

NP-hard problems occurring in the statistical physics of disordered systems, still remain a challenge: Examples are the computation of ground states of spin glass models on non-planar graphs, like the three-dimensional spin glass or the random field Potts model for three or more Potts states [66]. Stochastic optimization techniques like hysteretic optimization [67] or extremal optimization [68] have reached a high level of sophistication but naturally suffer from the lack of a proof of optimality of the resulting solution. Progress in the development of exact and efficient algorithm that can handle sufficiently large system sizes to perform a reliable finite size scaling analysis is being made [47] and highly rewarding.

The cross-fertilization between computer science and statistical physics is also fruitful in the other direction: Phase transitions that occur in some combinatorial optimization problems like the satisfiability problem (SAT) were studied intensively in recent years by physicists and remarkable progress has been achieved in understanding it and inventing efficient algorithms. These developments were not covered in this article, excellent introductions can be found in [69].

Bibliography

Primary Literature

1. Rieger H (1998) Frustrated systems: Ground state properties via combinatorial optimization. In: Kertesz J, Kondor I (eds) Lect Note Phys 501:122–158
2. Alava M, Duxbury P, Moukarzel C, Rieger H (2000) Exact combinatorial algorithms: Ground states of disordered systems. In: Domb C and Lebowitz JL (eds) Phase Transit Crit Phenom 18:141–317
3. Hartmann A, Rieger H (2002) Optimization in physics. Wiley VCH, Darmstadt
4. Papadimitriou CH Steiglitz K (1998) Combinatorial Optimization. Dover Publications, Mineola (NY)
5. Cook WJ, Cunningham WH, Pulleyblank WR, Schrijver A (1998) Combinatorial Optimization. Wiley, New York
6. Korte B, Vygen J (2000) Combinatorial Optimization. Springer, Berlin
7. Lawler EL, Lenstra JK, Rinnooy Kan AHG, Shmoys DB (1990) The Travelling Salesman Problem. Wiley, Chichester
8. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1995) Numerical Recipes in C. Cambridge University Press, Cambridge
9. Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671
10. Halpin-Healy T, Zhang Y-C (1995) Kinetic roughening phenomena, stochastic growth directed polymers and all that – aspects of multidisciplinary statistical-mechanics. Phys Rep 254:215
11. Peng C-K, Havlin S, Schwartz M, Stanley HE (1991) Directed-polymer and ballistic-deposition growth with correlated noise. Phys Rev A 44:2239; Pang N-N, Yu Y-K, Halpin-Healy T (1995) Interfacial kinetic roughening with correlated noise. Phys Rev E 52:3224
12. Marsili M, Zhang Y-C (1998) Overhangs in interface growth and ground-state paths. Phys Rev E 57:4814; Schwartz N, Nazaryev AL, Havlin S (1998) Optimal path in two and three dimensions. Phys Rev E 58:7642
13. Schorr R, Rieger H (2003) Universal properties of shortest paths in isotropically correlated random potentials. Europ Phys J 33:347
14. For a review see Blatter G et al (1994) Vortices in high-temperature superconductors. Rev Mod Phys 66:1125
15. Doi M, Edwards SF (1986) The Theory of Polymer Dynamics. Oxford University Press, Oxford
16. Drossel B, Kardar M (1996) Winding angle distributions for random walks and flux lines. Phys Rev E 53:5861
17. Bikbov R, Nechaev S (2001) Topological Relaxation of Entangled Flux Lattices: Single versus Collective Line Dynamics. Phys Rev Lett 87:150602
18. Petäjä V, Alava M, Rieger H (2004) Entanglement transition of elastic lines in a strongly disordered environment. Europhys Lett 66:778
19. Rieger H, Blasum U (1997) Ground state properties of solid-on-solid models with disordered substrates. Phys Rev B 55:7394R; Pfeiffer F, Rieger H (2000) Dislocations in the ground state of the solid-on-solid model on a disordered substrate. J Phys A 33:2489
20. Bokil HS, Young AP (1995) Absence of a phase transition in a three-dimensional vortex glass model with screening. Phys Rev Lett 74:3021
21. Kisker J, Rieger H (1998) Application of a minimum cost flow algorithm to the three-dimensional gauge glass model with screening. Phys Rev B 58:R8873; Pfeiffer F, Rieger H (1999) Numerical study of the strongly screened vortex glass model in an external field. Phys Rev B 60:6304
22. Pfeiffer FO, Rieger H (2002) Superconductor-to-normal phase transition in a vortex glass model: a new percolation universality glass. J Phys C 14:2361; Pfeiffer FO, Rieger H (2003) Critical properties of loop percolation models with optimization constraints. Phys Rev E 67:056113
23. Middleton AA (1995) Numerical results for the ground-state interface in a random medium. Phys Rev E 52:R3337; McNamara D, Middleton AA, Zeng C (1999) Simulation of the zero-temperature behavior of a three-dimensional elastic medium. Phys Rev B 60:10062
24. Goldberg AV, Tarjan RE (1988) A new approach to the maximum-flow problem. J Assoc Comput Mach 35:921
25. Ahuja RK, Magnati TL, Orlin JB (1993) Network Flows. Prentice Hall, London
26. Nattermann T (1990) Scaling approach to pinning: Charge density waves and giant flux creep in superconductors. Phys Rev Lett 64:2454; Giarmachi T, Le Doussal P (1994) Elastic theory of pinned flux lattices. Phys Rev Lett 72:1530; (1995) Phys Rev B 52:1242
27. Bouchaud J-P, Georges A (1992) Competition between lattice pinning and impurity pinning: Variational theory and physical realizations. Phys Rev Lett 68:3908
28. Emig T, Nattermann T (1997) A new disorder-driven roughening transition of charge-density waves and flux-line lattices. Phys Rev Lett 79:5090; (1999) Disorder driven roughening transitions of elastic manifolds and periodic elastic media. Eur J Phys B 8:525
29. Seppälä ET, Alava MJ, Duxbury PM (2001) Intermittence and roughening of periodic elastic media. Phys Rev E 63:036126
30. Noh JD, Rieger H (2001) Disorder driven critical behavior of periodic elastic media in a crystal potential. Phys Rev Lett 87:176102; (2002) Numerical study of the disorder-driven roughening transition in an elastic manifold in a periodic potential. Phys Rev E 66:036117
31. Rieger H (1995) Monte Carlo simulations of Ising spin glasses and random field systems. In: Annual Reviews of Computational Physics II. World Scientific, Singapore, pp 295–341
32. Nattermann T (1998) In: Young AP (ed) Spin Glasses and Random Fields. World Scientific, Singapore
33. Anglés d'Auriac JC, Preissman M, Rammal R (1985) The random field Ising-model - algorithmic complexity and phase-transition. J Phys (France) Lett 46:L173
34. Barahona F (1985) Finding ground-states in random-field Ising-ferromagnets. J Phys A 18:L673
35. Middleton AA, Fisher DS (2002) Three-dimensional random-field Ising magnet: Interfaces, scaling, and the nature of states. Phys Rev B 65:13411
36. Ogielski AT (1986) Integer optimization and zero-temperature fixed point in Ising random-field systems. Phys Rev Lett 57:1251
37. Kawashima N, Rieger H (2004) In: Diep HT (ed) Frustrated Spin Systems. World Scientific, Singapore
38. Grötschel M, Jünger M, Reinelt G (1985) In: van Hemmen L, Morgenstern I (eds) Heidelberg Colloqium on Glassy dynamics and Optimization. Springer, Heidelberg
39. de Simone C, Diehl M, Jünger M, Mutzel P, Reinelt G, Rinaldi G (1995) Exact ground-states of Ising spin-glasses - new exper-

imental results with a branch-and-cut algorithm. J Stat Phys 80:487
40. Lawler EL (1976) Combinatorial optimization: Networks and matroids. Holt, Rinehart and Winston, New York
41. Derigs U (1988) Programming in networks and graphs. In: Springer Series: Lecture Notes in Economics and Mathematical Systems, vol 300. Springer, Berlin
42. Barahona F (1982) On the computational-complexity of Ising spin-glass models. J Phys A 15:3241; Barahona F, Maynard R, Rammal R, Uhry JP (1982) Morphology of ground-states of two-dimensional frustration model. J Phys A 15:673
43. Kawashima N, Rieger H (1997) Finite size scaling analysis of exact ground states for $\pm J$ spin glass models. Europhys Lett 39:85
44. Hartmann AK, Young AP (2002) Large-scale low-energy excitations in the two-dimensional Ising spin glass. Phys Rev B 66:094419; Hartmann AK, Bray AJ, Carter AC, Moore MA, Young AP (2002) Stiffness exponent of two-dimensional Ising spin glasses for nonperiodic boundary conditions using aspect-ratio scaling. Phys Rev B 66:224401
45. Amoruso C, Hartmann AK, Hastings MB, Moore MA (2006) Conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. Phys Rev Lett 97:267202; Bernard D, LeDoussal P, Middleton AA (2007) Possible description of domain walls in two-dimensional spin glasses by stochastic Loewner evolutions. Phys Rev B 76:020403(R)
46. Cardy J (2005) SLE for theoretical physicists. Ann Phys 318:81; Bauer M, Bernard D (2006) 2D growth processes: SLE and Loewner chains. Phys Rep 432:115
47. Liers F, Jünger M, Reinelt G, Rinaldi G (2004) Computing exact ground states of hard Ising spin glass problems by branch-and-cut. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
48. Chvátal V (1983) Linear programming. Freeman, San Francisco
49. Grötschel M, Lovász L, Schrijver A (1988) Geometric algorithms and combinatorial optimization. Springer, Berlin
50. Wu FY (1982) The Potts Model. Rev Mod Phys 54:235
51. Kasteleyn PW, Fortuin CM (1969) Phase transitions in lattice systems with random local properties. J Phys Soc Jpn 46:11
52. Juhász R, Rieger H, Iglói F (2001) The random-bond Potts model in the large-q limit. Phys Rev E 64:056122
53. Schrijver A (2003) Combinatorial Optimization – Polyhedra and Efficiency, vol B. Springer, Berlin
54. Grötschel M, Lovász L, Schrijver A (1981) The ellipsoid method and its consequences in combinatorial optimization. Comb 1:169
55. Iwata S, Fleischer L, Fujishige S (2001) A combinatorial strongly polynomial algorithm for minimizing submodular functions. J ACM 48(4):761
56. Schrijver A (2000) A combinatorial algorithm minimizing submodular functions in strongly polynomial time. J Comb Theory Ser B 80:346
57. Edmonds J (1977) In: Guy R, Hannani H, Sauer N, Schóonheim J (eds) Combinatorial Structures and Their Applications. Gordon and Breach, New York
58. Kasteleyn PW, Fortuin CM (1969) Phase transitions in lattice systems with random local properties. J Phys Soc Jpn 26:11
59. Anglés d'Auriac JC, Iglói F, Preissmann M, Sebö A (2002) Optimal cooperation and submodularity for computing Potts' partition functions with a large number of states. J Phys A 85:6973
60. Anglés d'Auriac JC (2004) Computing the Potts free energy and submodular functions. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
61. Anglés d'Auriac JC, Iglói F (2003) Phase transition in the 2D random Potts model in the large-q limit. Phys Rev Lett 90:190601; Mercaldo MT, Anglés d'Auriac J-C, Iglói F (2004) Disorder-induced rounding of the phase transition in the large- q - state Potts model. Phys Rev E 69:056112; Mercaldo MT, Anglés d'Auriac J-C, Iglói F (2005) Disorder-driven phase transitions of the large q-state Potts model in three dimensions. Europhys Lett 70:733
62. Fisher DS (1992) Random transverse field Ising spin chains. Phys Rev Lett 69:534; (1995) Critical behavior of random transverse-field Ising spin chains. Phys Rev B 51:6411
63. Zeng C, Middleton AA, Shapir Y (1996) Ground-state roughness of the disordered substrate and flux lines in d = 2. Phys Rev Lett 77:3204
64. Middleton AA (1999) Numerical investigation of the thermodynamic limit for ground states in models with quenched disorder. Phys Rev Lett 83:1672
65. Anglés d'Auriac J-C, Sourlas N (1997) The 3d random field Ising model at zero temperature. Europhys Lett 39:473
66. Anglés d'Auriac J-C, Preissmann M, Sebö A (1997) Optimal cuts in graphs and statistical mechanics. Math Comput Model 26:1
67. Pal KF (2004) Hysteretic opimization. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
68. Boettcher S (2004) Extremal optimization. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
69. Weigt M (2004) The random 3-satisfiability problem: From the phase transition to the efficient generation of hard, but satisfiable instances. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin; Cocco S, Ein-Dor L, Monasson R (ibid) Analysis of backtracking procedures for random decision problems; Zecchina R (ibid) New iterative algorithms for hard combinatorial problems

Books and Reviews

Alava M, Duxbury P, Moukarzel C, Rieger H (2000) Combinatorial optimization and disordered systems. In: Domb C, Lebowitz JL (eds) Phase Transition and Critical Phenomena, vol 18. Academic Press, Cambridge

Hartmann A, Rieger H (2002) Optimization Algorithms in Physics. Wiley VCH, Berlin

Hartmann A, Rieger H (2004) New Optimization Algorithms in Physics. Wiley VCH, Berlin

Hartmann AK, Weigt M (2005) Phase Transitions in Combinatorial Optimization Problems. Wiley-VCH, Berlin

Orbital Dynamics, Chaos in

JOHN D. HADJIDEMETRIOU
Department of Physics, University of Thessaloniki, Thessaloniki, Greece

Article Outline

Glossary

Reference frame A *reference frame* can be determined by a set of solid bodies, through which we can define a three dimensional geometric figure, for example a triedron (three non planar axes starting from a point). The surface of the Earth can be used to define a reference system. A moving car can be also used to define a reference frame, different from the first one.

Inertial frame An *inertial frame* is a special class of reference frames, in which the basic laws of motion (*Newton's laws*) are valid. According to *Galileos' Principle of Relativity*, any frame of reference moving uniformly (with constant velocity without rotation) with respect to an inertial frame is also an inertial frame. A frame of reference which is rotating with respect to an inertial frame is not inertial. The criterion for a frame to be inertial is Newton's first law to be valid. This means that in an inertial frame a free body is either at rest or moves in a straight line with constant velocity. The best approximation in nature of an inertial frame is that frame which is defined by a triedron whose origin is at the center of mass of our Solar System and its three axes are in three fixed directions in space, defined by three distant stars.

Degrees of freedom The number of independent variables that are needed to determine the position of a dynamical system is called the number of *degrees of freedom*. For example, a particle moving freely in space has three degrees of freedom, since its position is determined by its three Cartesian coordinates (x_1, x_2, x_3), which are independent.

Phase space Consider a space whose coordinates determine exactly the state of the system. This space is called the *state space* or the *phase space* of the system. Each point of the phase space determines uniquely the initial conditions of the motion. The evolution of the system in the phase space is represented by a smooth curve, which is called the *phase curve*. The phase curves do not intersect, otherwise the point of intersection would correspond to two different solutions. The set of all phase curves is called the *phase diagram* and gives important information of the stable and unstable regions of the phase space. For gravitational systems, the phase space is the space of coordinates and velocities of all the bodies. Usually, instead of the velocities, the moments are used in the definition of the phase space. In a gravitational system with n degrees of freedom, the phase space has $2n$ dimensions. For example, a body moving in the plane under the action of a force, has two degrees of freedom (coordinates x, y) and its phase space is the four-dimensional space $x, y, p_x = m\dot{x}, p_y = m\dot{y}$.

Orbit An *orbit* of a body, or a set of bodies, considered as point masses, is the path that the bodies describe in a reference frame. The orbit of the *same* body or set of bodies is different in different frames of reference.

Periodic orbit A *periodic orbit* is the orbit of one or more bodies that repeats itself after a certain time, which is called the *period* of the periodic orbit. The periodicity property is closely related to the frame of reference to which the motion is referred to. For example, an orbit may be periodic in a rotating frame, but not in the inertial frame. In this latter case, for two or more bodies, it is the relative configuration that is repeated in the inertial frame.

Poincaré map The Poincaré map is a method by which we transform the continuous flow of a dynamical system in its n-dimensional phase space, into a discrete map in a reduced phase space. The map is obtained by taking the intersections of the continuous flow in the original phase space with a *surface of section*, defined properly. This surface of section is $(n-1)$-dimensional, in general, or $(n-2)$-dimensional if an integral of motion exists, which is the energy integral in gravitational systems. These will be explained in detail in Sects. "The Poincaré Map," "Poincaré Map in Hamiltonian Systems." A periodic orbit appears as a fixed point on the Poincaré surface of section. The Poincaré map is very useful in the study of ordered and chaotic motion in a dynamical system, especially in systems with few degrees of freedom.

Stability The notion of *stability* refers to the behavior of the orbits in the vicinity of a periodic orbit. If a slight change in the initial conditions results to a new orbit, called the *perturbed orbit*, which deviates much from the periodic orbit, then the periodic orbit is called *unstable*. In the gravitational systems that we will study,

this deviation is exponential. If on the other hand, the perturbed orbit stays close to the periodic orbit, the periodic orbit is called *stable*. But there are different aspects of stability. For example, if the perturbed orbit, considered as a geometrical figure, is close to the periodic orbit, then the periodic orbit is called *orbitally stable*. However, in this latter case it may happen that two bodies, one on the periodic orbit and one on the perturbed orbit, which start very close to each other, may deviate much as each one moves on its own orbit, although the geometric figures of the two orbits are close to each other. In this aspect, the orbit is considered as unstable. A Keplerian elliptic orbit, in the two-body problem, belongs to this latter category. A different type of stability is the *asymptotic stability*. In this case any perturbed orbit, not only stays in the vicinity of the periodic orbit, but tends asymptotically to the periodic orbit. In gravitational systems asymptotic stability does not appear, unless there exists a dissipation to the system.

Ordered and chaotic motion The notion of *chaoticity* is related to the deviation of a perturbed orbit from a given orbit. It may happen that the perturbed orbit does not deviate much as time goes on. In this case we say that we are in an *ordered region*. The prediction of the evolution of the system in this case is possible. In some cases however, the perturbed orbit deviates exponentially from the original orbit. Prediction is not possible for a long time. In this latter case we are in a *chaotic region*. In general, both ordered and chaotic regions exist in the same dynamical system.

Definition of the Subject

By the term *orbital dynamics* we mean the study of the motion of one or more bodies. Motion is one of the first things that a human being noticed, since the very early stages of human life. Apart from the motion of himself, walking around, he also realized that everything around him is not still, but changes position, being it a wild animal, a dry leaf drifting in the wind, the motion of clouds in the sky, or the change of the position of the celestial bodies, most notably of the Sun and the Moon.

Evidently, motion is one of the most important aspects in everyday life. By the term *motion* we mean the change of the position of one or more bodies in space, with respect to the other bodies in that region. If only one body existed in the universe, motion could not be defined. This makes necessary the introduction of an important notion in physics, the *frame of reference*. The surface of the Earth, for example, defines a frame of reference, with respect to which we determine the position of a body and its motion, as the body changes position. But a bus moving on the road is also a frame of reference, different from that defined by the surface of the Earth, i. e., the road. And it is a different thing if the bus moves on a straight line with constant velocity, or makes rapid turns following a difficult mountain road. Among all possible frames of reference, the *inertial frames of reference* have a special status in the study of motion. It is in these frames that the basic laws of motion (Newton's laws) are valid.

If the dimensions of the body can be considered as negligible, with respect to its surroundings, we can consider it as a point mass. However, in many cases, the finite dimensions of a body cannot be ignored. In this case its motion cannot be described by the motion that a point describes, but we have to consider also the rotation of the body. Whether we consider a body as a point mass or as a body with finite dimensions, depends on the particular study. For example the Earth is considered as a point mass in the study of its motion around the Sun, but as a finite body when we study the motion of an artificial satellite. In the present study we restrict ourselves in the motion of point masses. The path that such a body describes, is called the *orbit* of the body.

The motion of the bodies takes place under the action of forces which follow definite laws. In everyday life, the dominant force is the gravitational interaction between the bodies, according to *Newton's law of gravity*. Although it is, by far, the weakest force in nature, it is the main force that we feel in everyday life, in addition to the electromagnetic force, which also manifests itself in macroscopic phenomena. These forces affect the motion of the bodies through definite laws, expressed by differential equations, which are *deterministic* equations, i. e. to a certain set of initial conditions there exists one and only one final result, which in our case is a definite orbit. In classical physics, these laws are *Newton's laws of motion*. They are expressed by differential equations of the *second order*, which implies that the initial conditions that define the motion are the *initial position* and the *initial velocity*. This is the essential property of Newton's laws. In classical physics we assume that it is possible to know exactly, *at the same time*, the position and the velocity of a body. In some other world, where the laws of motion were expressed by differential equations of the third order, the initial acceleration would be also necessary to define the motion. Alternatively, if the laws of motion were expressed as differential equations of the first order, only the initial position would be enough to determine the motion.

As we mentioned, the equations of motion are deterministic. This implies that motion would be exactly de-

fined and that we could predict the motion of one or more bodies, for example the motion of an asteroid in our Solar System, if we knew its initial conditions. This idea prevailed classical physics until the sixties of the 20th century. But what will happen if we make a small error in the initial conditions? Does it have a great effect on the final position, after a certain time (for example a few million years for the asteroid), or the final error will be of the same order as the initial error? In the latter case the small error, due for example to a not very accurate measurement of the position and the velocity of the asteroid, is not important. In many cases however, including an asteroid in certain regions in the asteroid belt close to some mean motion resonance with Jupiter, the orbit is very sensitive to a change in the initial conditions. In this latter case, after a certain time, the orbit which corresponds to the slightly changed initial conditions, differs very much from the original orbit, because it deviates exponentially. These orbits, which are very sensitive to the initial conditions, are called *chaotic orbits*. In such a case prediction of the final position of the body, after a certain time, is not possible, because a very small error in the initial conditions, beyond the accuracy of the observations for the initial conditions, will give a completely different final position, due to the exponential deviation between the two orbits. As we will see, all the physical systems are non integrable and consequently they present chaotic behavior, at least for some initial conditions, and for this reason prediction of the evolution of such a system is not possible, after a certain time interval. This time interval is different in different systems and may be two weeks for meteorological systems or some million years for the motion of an asteroid.

Among all possible orbits in a dynamical system, the *periodic orbits* play a dominant role in the study of the evolution of the system, although it is known that they form a set of measure zero. This is so because, as it will become clear in the following, the periodic orbits are the "backbone" of the topology of the phase space, because their position and their stability character (stable or unstable) determine the structure of the phase space. It is close to the unstable periodic orbits that chaotic motion appears. A special class of periodic orbits in dynamical systems that describe the motion in the Solar System are the *resonant* periodic orbits, because around the stable resonant orbits islands of stable motion exist and the system can be trapped in these regions. In addition, since in a system there exist more that one resonances, the overlap of these resonances, as a perturbation increases, will generate chaotic motion.

In the following we restrict ourselves to the study of motion under gravitational forces, focusing on our Solar System and on extrasolar planetary systems, but the theory is applicable in all cases of motion, under any force field.

Introduction

The Newtonian gravitational force is the dominant force in the N-Body systems in the universe, as for example in a planetary system, a planet with its satellites, a multiple stellar system, or a galaxy.

In many cases, there is only one massive body, whose gravitational attraction provides the dominant force, as is the case with a planetary system, where the Sun is the main attracting body, or a planet surrounded by satellites. In this case the motion of the small bodies (planets or satellites) follow Keplerian orbits, perturbed by the gravitational interaction between the small bodies. This is a *nearly integrable* dynamical system. In these systems resonances exist between the small bodies in their motion around the massive body, as will be explained in the following. These correspond to periodic motion, and this makes clear the importance of the resonances in the dynamical properties of a nearly integrable system.

The simplest model of a gravitational system is a system of two bodies moving in Keplerian orbits around their common center of mass. This is an integrable system. In such systems all motion is ordered and chaos never appears. We consider now a hierarchy of models, starting from the above mentioned integrable system and adding more bodies to the system. We have different models, which are used to study particular systems. All these systems are not integrable, although they are nearly integrable. In these latter systems both ordered and chaotic regions appear, as we will see in the following. We consider two basic models:

The restricted three-body problem: Two bodies of finite masses, called *primaries*, revolve around their common center of mass in *circular* or *elliptic* orbits and a third body with *negligible* mass moves under their gravitational attraction, but does not affect the orbits of the two primaries. In most astronomical applications the second primary has a small mass compared to the first primary (the Sun), and consequently the motion of the third, massless, body is a perturbed Keplerian orbit. This is a model for the study of an asteroid (Jupiter being the second primary) or a trans-Neptunian object (Neptune being the second primary).

The general planetary three-body problem: Three bodies with finite masses moving under their gravitational attraction. This is a model for a triple stellar system. In many astronomical applications one of the three bodies has a large mass and the other two bodies have small,

but not negligible masses. This is a model for an extrasolar planetary system, or a system of two satellites moving around a major planet. In the latter two cases the two small bodies move in perturbed Keplerian orbits.

The long term evolution of the system depends on the topology of its phase space and the existence of ordered or chaotic regions. The topology of the phase space is determined by the position and the stability character of the periodic orbits of the system (fixed points on a Poincaré map on a surface of section). Islands of stable motion exist around the stable periodic orbits. Chaotic motion appears at the unstable periodic orbits. This makes clear the importance of the periodic orbits in the study of the dynamics of such systems.

We will start with the basic elements of gravitational systems in general. Then we will focus our attention to the study of systems of two degrees of freedom, and then extend the results to three degrees of freedom. The study will be for a general dynamical system, with particular emphasis on Hamiltonian systems, because the gravitational systems are Hamiltonian.

Basic Equations and Integrals of Motion

The gravitational force between two bodies, N_i, N_j, is given by Newton's law of gravitation

$$F_{ij} = -\frac{Gm_i m_j}{r_{ij}^2} ,$$

where G is the gravitational constant, m_i, m_j are the masses of the bodies N_i and N_j and r_{ij} is their distance. The minus sign indicates attraction. We have $3N$ degrees of freedom and the evolution in space is given by the system of differential equations

$$m_i \ddot{\vec{r}}_i = \vec{F}_i ,$$

where

$$\vec{F}_i = -\sum_{j=1}^{N} \frac{Gm_i m_j(\vec{r}_i - \vec{r}_j)}{r_{ij}^3} = -\frac{\partial V}{\partial \vec{r}_i} ,$$

and

$$\vec{r}_i(x_i, y_i, z_i) \quad (i = 1, 2, \ldots N)$$

is the position vector of the body N_i. The system is conservative, and the potential function is

$$V(\vec{r}_m - \vec{r}_n) = -\sum_{ij} \frac{Gm_i m_j}{r_{ij}} . \quad 1 \le i < j \le N .$$

The gravitational system of N bodies can be formulated in Hamiltonian dynamics, and the Hamiltonian function is

$$H = \sum \frac{\vec{p}_i^2}{2m} + V , \quad \vec{p}_i = m_i \dot{\vec{r}}_i .$$

We have the following integrals of motion:

$$\begin{aligned} \vec{r}_{\rm cm} &= \left(\sum m_i \vec{r}_i\right) && \text{Center of mass} \\ \vec{p} &= \sum m_i \vec{v}_i = \text{constant} && \text{Linear momentum} \\ \vec{L} &= \sum \vec{r}_i \times m_i \vec{v}_i = \text{constant} && \text{Angular momentum} \\ E &= T + V = \text{constant} && \text{Energy integral.} \end{aligned}$$

The total momentum of the system is equal to $\vec{p} = m\vec{v}_{\rm cm}$, $\vec{v}_{\rm cm}$ being the velocity of the center of mass. We can assume that the total momentum is equal to zero, $\vec{p} = 0$, which implies that the center of mass is at rest, $\vec{v}_{\rm cm} = 0$. Consequently, in the system where the center of mass is at rest, we have $3N - 3$ degrees of freedom. In this latter case we can take $\vec{r}_{\rm cm} = 0$.

Periodic Orbits in Systems with Two Degrees of Freedom

Periodic Orbits

The periodic orbits play an important role in understanding the dynamics of a system, because they determine critically the topology of the phase space. This will become clear in the following. For this reason it is important to know the basic families of periodic orbits in a dynamical system, because they are the "backbone" of the phase space.

Let us consider a dynamical system with two degrees of freedom, defined by the set of two second order differential equations

$$\begin{aligned} \ddot{x}_1 &= F_1(x_1, x_2, \dot{x}_1, \dot{x}_2) , \\ \ddot{x}_2 &= F_2(x_1, x_2, \dot{x}_1, \dot{x}_2) . \end{aligned} \tag{1}$$

The initial conditions that determine a solution are $(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20})$ and the corresponding solution has the form

$$\begin{aligned} &x_1(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t) , \\ &x_2(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t) . \end{aligned}$$

The solution is periodic, with period T, if

$$\begin{aligned} &x_i(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t + T) \\ &\qquad = x_i(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t) , \end{aligned}$$

for every t.

Existence of Symmetric Periodic Orbits

We assume that the differential equations (1) are invariant under the transformation

$$x_1 \to x_1\,, \quad x_2 \to -x_2\,, \quad t \to -t.$$

This property appears in several models that are of astronomical interest. This means that if $x_1(t), x_2(t)$ is a solution, then $x_1(t), -x_2(-t)$ is also a solution. Note that this second solution is the symmetric of the first solution with respect to the x_1-axis. Consequently, if an orbit starts from the x_1-axis perpendicularly, $\dot{x}_{10} = 0$, and crosses again the x_1-axis perpendicularly, $\dot{x}_1 = 0$, the orbit is closed and is a *symmetric* periodic orbit with respect to the x_1-axis.

The initial conditions of a symmetric periodic orbit are $(x_{10}, x_{20} = 0, \dot{x}_{10} = 0, \dot{x}_{20})$, which means that a symmetric periodic orbit is determined only by the two nonzero initial conditions $x_{10}, \dot{x}_{20}$. From the above we see that the *periodicity conditions* are

$$x_2(x_{10}, 0, 0, \dot{x}_{20};\, T/2) = 0\,,$$
$$\dot{x}_1(x_{10}, 0, 0, \dot{x}_{20};\, T/2) = 0\,,$$

which imply that the orbit starts perpendicularly from the x_1-axis ($x_{20} = 0, \dot{x}_{10} = 0$) and crosses again perpendicularly the x_1-axis after a time interval equal to half the period T. We remark that the second *perpendicular* crossing may take place after several (non perpendicular) crossings from the x_1-axis.

The periodic orbits are not isolated, in general. They belong to families, along which the period varies. A family of symmetric periodic orbits is represented by a continuous curve in the space of initial conditions $x_{10}, \dot{x}_{20}$. This curve is called a *characteristic curve*.

Variational Equations

We study now the behavior of the system in the vicinity of a specific orbit, by considering perturbed initial conditions, i. e. initial conditions in the vicinity of the initial conditions of this orbit.

We express the system of differential equations (1) as a system of four differential equations of the first order,

$$\dot{x}_i = f_i(x_1, x_2, x_3, x_4)\,, \quad (i = 1, \ldots 4) \tag{2}$$

where $x_3 = \dot{x}_1, x_4 = \dot{x}_2$. Let $x_i = x_i(x_{10}, x_{20}, x_{30}, x_{40};\, t)$, $(i = 1, \ldots 4)$ be a solution of the system (2), nonperiodic in general, corresponding to the initial conditions $x_1(0), x_2(0), x_3(0), x_4(0)$. We consider new initial conditions, in the vicinity of these initial conditions, of the form $x_i(0) + \xi_i(0)$, where $\xi_i(0)$ are small. The new solution can be expressed in the form

$$x_i'(t) = x_i(t) + \xi_i(t)\,, \quad (i = 1, \ldots 4)$$

where $\xi(t)$ is the deviation vector between the initial solution $x_i(t)$ and the perturbed solution $x_i'(t)$, at the *same time* t, $\xi(t) = x_i'(t) - x_i(t)$. The behavior of the system in the vicinity of the solution $x_i(t)$ depends on the deviation vector $\xi(t)$.

We assume that the initial perturbation $\xi(0)$ is small, and consequently, for continuity reasons, the deviation $\xi(t)$ should be also small, at least for a finite time interval. For this reason we linearize the system of differential equations (2), to first order terms in the $\xi_i(t)$, by substituting the perturbed solution $x_i'(t)$ into the system (2) and keeping only the first order terms in ξ_i. We obtain the system of *variational equations*,

$$\dot{\xi}_i = \sum_{k=1}^{4} p_{ik}\xi_k\,, \quad p_{ik} = \left(\frac{\partial f_i}{\partial x_k}\right)_{x_i(t)}\,, \quad (i = 1, \ldots 4) \tag{3}$$

which describes the evolution of the system (2) in the neighborhood of the orbit $x_i(t)$, to first order terms in the deviations. The partial derivatives are computed for the solution $x_i(t)$. This is a linear system with time dependent coefficients.

The general solution of the linear system (3) is expressed as a linear combination of four linearly independent solutions. In particular, let us consider a 4×4 matrix $\Delta(t)$ whose columns are four linearly independent solutions corresponding to the initial conditions $\Delta(0) = I_4$, where I_4 is the 4×4 unit matrix. This matrix is called *fundamental matrix of solutions* and the general solution of the variational equations is expressed in the form

$$\xi(t) = \Delta(t)\xi(0)\,. \tag{4}$$

A basic property of the matrix $\Delta(t)$ is the Liouville–Jacobi formula [24,58].

$$\det \Delta(t) = \det \Delta(0) \exp \int_0^t \mathrm{trace}(P)\mathrm{d}t\,, \tag{5}$$

where P is the matrix of the coefficients of the variational equations (3), with elements p_{ij}. This relation gives the change in time of the determinant of the matrix $\Delta(t)$, which describes important properties of the evolution of the system in phase space, as we shall see in the following. Of special importance is the case where $\mathrm{trace}(P) = 0$, because in this case the determinant of the matrix $\Delta(t)$ is constant.

Another important property is that the columns of the matrix $\Delta(t)$ are the partial derivatives of the solution $x_i(x_{10}, x_{20}, x_{30}, x_{40}, t)$ with respect to the initial conditions: The solution $x_i(x_0;\ t)$ satisfies the system (2),

$$\frac{\partial x_i(x_0,t)}{\partial t} = f_i(x_1(x_0,t), x_2(x_0,t), x_3(x_0,t), x_4(x_0,t))\,, \quad (i = 1,\ldots 4)\,.$$

If we apply to the above equations the operator $\partial/\partial x_{j0}$, $j = 1,\ldots 4$, we obtain

$$\frac{\partial}{\partial t}\left(\frac{\partial x_i}{\partial x_{j0}}\right) = \sum_k \left(\frac{\partial f_i}{\partial x_k}\right)\frac{\partial x_k}{\partial x_{j0}}\,, \quad (i = 1,\ldots 4) \tag{6}$$

for each x_{j0}. We note that the system (6) is the system of variational equations (3) satisfied by the vector $(\partial x_1/\partial x_{j0}, \partial x_2/\partial x_{j0}, \partial x_3/\partial x_{j0}, \partial x_4/\partial x_{j0})$. In addition, we note that $\partial x_i/\partial x_{j0} = \delta_{ij}$ for $t = 0$, which implies that these vectors, for $j = 1,\ldots 4$, are the four columns of the fundamental matrix of solutions $\Delta(t)$. This means that the fundamental matrix of solutions $\Delta(t)$ is the Jacobian of the solution $x(t)$ with respect to the initial conditions,

$$\Delta(t) = \frac{\partial(x_1, x_2, x_3, x_4)}{\partial(x_{10}, x_{20}, x_{30}, x_{40})}\,. \tag{7}$$

Linear Stability of a Periodic Orbit

The variational equations (3) are a system of four linear differential equations with time dependent coefficients. If the solution $x(t)$ is T-periodic, then the partial derivatives are also T-periodic. In this latter case the system of variational equations is a *linear system with periodic coefficients*. The theory related to the study of such systems is the *Floquet theory* [24] and some elements of it will be presented in the following sections.

Existence of a Periodic Solution

We shall prove that the derivative $\dot{x}_i(t)$ of the periodic solution $x_i(t)$ is a solution of the variational equations. Indeed, the solution $x_i(t)$ satisfies the system (2)

$$\dot{x}_i(t) = f_i(x_1(t), x_2(t), x_3(t), x_4(t))\,, \quad (i = 1,\ldots 4)$$

and if we apply the operator $\mathrm{d}/\mathrm{d}t$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\dot{x}_i(t)) = \sum_{j=1}^{4}\left(\frac{\partial f_i}{\partial x_j}\right)_{x_i(t)} \dot{x}_j(t)\,.$$

This is the system of variational equations (3), for the solution $\xi_i = \dot{x}_i(t)$. So we come to the conclusion that *the variational equations that correspond to a T-periodic orbit have always a T-periodic solution, which is the derivative $\dot{x}_i(t)$ of the periodic solution.*

Mapping at Integral Multiples of the Period. The Monodromy Matrix

Let $x_i(t)$ be a periodic orbit and $x'(t)$ a perturbed orbit, which, to a linear approximation, can be expressed in the form

$$x_i'(t) = x_i(t) + \xi_i(t)\,,$$

where $\xi_i(t)$ is the solution of the variational equations. This latter solution is expressed in the form

$$\xi(t) = \Delta(t)\xi(0)\,, \tag{8}$$

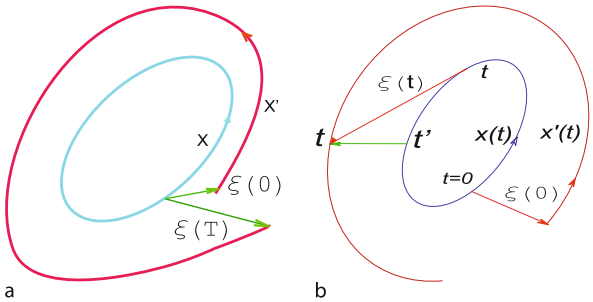
and for $t = T$,

$$\xi(T) = \Delta(T)\xi(0)\,. \tag{9}$$

From this expression we obtain, by induction,

$$\xi(nT) = [\Delta(T)]^n\xi(0)\,. \tag{10}$$

Equations (9) and (10) give the deviation, to a linear approximation, of the perturbed orbit $x'(t)$ from the periodic orbit $x(t)$ after a time interval equal to n times the period T, due to an initial deviation $\xi(0) = x'(0) - x(0)$ at $t = 0$. In fact Eq. (10) is a mapping of the initial deviation $\xi(0)$ at integral multiples of the period T (see Fig. 1a). This is a linear mapping defined by the matrix $\Delta(T)$. It is clear that the stability of the periodic orbit $x(t)$ depends on the properties of the mapping (10), i. e. on the eigenvalues of the matrix $\Delta(T)$. The matrix $\Delta(T)$ is called the *monodromy matrix*.



Orbital Dynamics, Chaos in, Figure 1
a Mapping at integral multiples of the period. b Orbital stability: The distance between two points *at the same time t*, $\xi(t) = x'(t) - x(t)$, is not small, but the distance between the points at two *different times*, t and t', $x'(t) - x(t')$, remains bounded

Unit Eigenvalue of the Monodromy Matrix

Existence of an Integral of Motion We shall prove that if the system of differential equations (2) has an integral of

motion,

$$G(x_1, x_2, x_3, x_4) = \text{constant}\,,$$

the system of variational equations (3) has a unit eigenvalue: Let $x_i(x_0, t)$ be a T-periodic solution. Since $G(x_1, x_2, x_3, x_4)$ is an integral, we have the relation

$$G(x_1(x_0,t), x_2(x_0,t), x_3(x_0,t), x_4(x_0,t)) = G(x_{10}, x_{20}, x_{30}, x_{40})\,.$$

We apply to the above relation the operator $\partial/\partial x_{j0}$, and we obtain

$$\sum_{k=1}^{4} \left(\frac{\partial G}{\partial x_k}\right)_t \left(\frac{\partial x_k}{\partial x_{j0}}\right)_t = \left(\frac{\partial G}{\partial x_{j0}}\right).$$

We set now $t = T$ and taking into account that

$$\left(\frac{\partial G}{\partial x_k}\right)_{t=T} = \left(\frac{\partial G}{\partial x_k}\right)_{t=0},$$

due to the fact that $x(t)$ is periodic, we obtain

$$\left(\Delta^{\tau}(T) - I\right) \nabla G = 0\,, \tag{11}$$

where Δ^{τ} is the transpose of Δ. From this relation we obtain that if $\nabla G \neq 0$ then $\Delta(T)^{\tau}$ has a unit eigenvalue. Thus finally, we come to the conclusion that *if the dynamical system has an integral of motion, which is not stationary along the periodic orbit, the monodromy matrix $\Delta(T)$ has a unit eigenvalue.*

Existence of a Periodic Orbit of the Variational Equations We shall also prove that if the system of variational equations has a periodic solution $\xi(t)$, the monodromy matrix has a unit eigenvalue: We have $\xi(t + T) = \xi(t)$, for any t and consequently, for $t = 0$, $\xi(T) = \xi(0)$. Due to this latter relation, Eq. (8) takes the form, for $t = T$, $\xi(0) = \Delta(T)\xi(0)$, and finally

$$(\Delta(T) - I)\,\xi(0) = 0\,. \tag{12}$$

Thus we come to the conclusion that *if the system of variational equations has a periodic solution, the monodromy matrix has a unit eigenvalue.*

We have proved above that the system of variational equations has the T-periodic solution $\xi(t) = \dot{x}_i(t)$, where $\dot{x}_i(t)$ is the periodic solution corresponding to the variational equations. Thus, the monodromy matrix $\Delta(T)$ has always a unit eigenvalue. The corresponding eigenvector is the vector $\xi(0) = \dot{x}_i(0)$, which is the tangent vector to the periodic orbit, in the phase space.

Vertical Stability of Planar Periodic Orbits

In the previous sections we studied the stability of a planar periodic orbit with respect to perturbations of the initial conditions *in the plane*. We study now the stability with respect to perturbations of the initial conditions *perpendicular* to the plane of motion. This type of stability we call *vertical stability* and completes the study of the stability of a planar periodic orbit.

Consider a dynamical system of three degrees of freedom,

$$\begin{aligned} \ddot{x}_1 &= f_1(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)\,,\\ \ddot{x}_2 &= f_2(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)\,,\\ \ddot{x}_3 &= x_3\, f_3(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)\,. \end{aligned} \tag{13}$$

This is the form of the differential equations of many gravitational systems, for example of the 3-dimensional restricted three body problem, described in Sect. "Application to the Solar System" (see p. 67 in [41]). In this model, a small body with negligible mass moves under the gravitational attraction of two main bodies, which describe Keplerian orbits around their center of mass, under their mutual gravitational attraction. The plane $x_1\,x_2$ is the plane of motion of the two main bodies (in the inertial frame) and the small body moves in the three dimensional space $x_1\,x_2\,x_3$. It is intuitively clear that if the small body starts from a position in the $x_1\,x_2$ plane and its velocity is *in* this plane, then its motion is restricted in the $x_1\,x_2$ plane, since the gravitational attraction from the two main bodies is in this plane. This physical property of the motion is described by the special mathematical form of the equations of motion (13) (third equation).

It is easy to verify that the Eqs. (13) admit a planar solution, which we will assume to be periodic: $x_1(t)$, $x_2(t), x_3(t) = 0$, corresponding to the initial conditions $x_{10}, x_{20}, x_{30} = 0, \dot{x}_{10}, \dot{x}_{20}, \dot{x}_{30} = 0$. We consider now a small perturbation ϵ_3, ϵ_6 along the x_3 axis, $x_{10} + \epsilon_1$, $x_{20} + \epsilon_2$, $x_{30} = 0 + \epsilon_3$, $\dot{x}_{10} + \epsilon_4$, $\dot{x}_{20} + \epsilon_5$, $\dot{x}_{30} = 0 + \epsilon_6$, where ϵ_i are small, and we want to study the behavior of the perturbed solution. We define new variables $x_4 = \dot{x}_1, x_5 = \dot{x}_2, x_6 = \dot{x}_3$, and a simple calculation shows that the system of variational equations of the system (13), for the periodic solution $x_i(t)$, breaks into two uncoupled systems: a system in the planar displacements $\xi_1, \xi_2, \xi_4, \xi_5$, corresponding to the variational equations of the planar motion, and a system in the vertical displacements (along the x_3 axis) ξ_3, ξ_6. This latter system is

$$\begin{aligned} \dot{\xi}_3 &= \xi_6\,,\\ \dot{\xi}_6 &= f_{30}(t)\xi_3\,, \end{aligned} \tag{14}$$

where the function $f_{30}(t)$ is the T-periodic function $f_3(x_1(t), x_2(t), x_3 = 0, x_4(t), x_5(t), x_6 = 0)$, computed for the planar T-periodic solution $x_i(t)$. The system (14) is the system of variational equations for the displacements along the x_3 axis. The vertical stability depends on the eigenvalues λ_5, λ_6 of the monodromy matrix $\Delta_2(T)$ of this system.

Hamiltonian Systems

The gravitational systems are Hamiltonian. For this reason, we study in this section the special properties that a Hamiltonian system has, in addition to the general properties obtained in the previous sections. We start with systems with two degrees of freedom.

A Hamiltonian system is defined by the Hamiltonian function

$$H(x_1, x_2, x_3, x_4)\,,$$

where x_1, x_2 are the coordinates and x_3, x_4 the momenta.

The Hamiltonian equations are

$$\dot{x}_1 = \partial H/\partial x_3\,, \quad \dot{x}_2 = \partial H/\partial x_4\,,$$
$$\dot{x}_3 = -\partial H/\partial x_1\,, \quad \dot{x}_4 = -\partial H/\partial x_2\,,$$

or

$$\dot{x} = -J\nabla H\,, \tag{15}$$

where ∇H is a column vector with elements $\partial H/\partial x_i$ and J the 4×4 symplectic matrix

$$J = \begin{pmatrix} 0 & -I_2 \\ +I_2 & 0 \end{pmatrix}\,,$$

where I_2 is the 2×2 unit matrix. Note that $J^{-1} = -J$.

Variational Equations of Hamiltonian Systems

The variational equations of a Hamiltonian system (15) have the special form given by

$$\dot{\xi} = -JA\xi\,, \tag{16}$$

where the elements a_{ij} of the 4×4 matrix A are

$$a_{ij} = \frac{\partial^2 H}{\partial x_i \partial x_j}\,. \quad (i, j = 1, \ldots 4) \tag{17}$$

Note that the matrix A is symmetric. The system (16) is called a *linear Hamiltonian system*. A complete study of such systems can be found in [58]. It is easy to see that it can be expressed in the Hamiltonian form (15) with Hamiltonian

$$H = \frac{1}{2}\xi^\tau A\xi = \frac{1}{2}\sum_{i,j=1}^{4} a_{ij}\xi_i\xi_j\,.$$

From the relations (16), (17) we can verify that the trace of the matrix of the coefficients of the linear Hamiltonian system (16) is equal to zero. Consequently, due to the general property (5), the determinant of the fundamental matrix of solutions $\Delta(t)$ is equal to unity (see also [34]), $\det \Delta(t) = \det \Delta(0) = 1$. For $t = T$ we obtain

$$\det \Delta(T) = 1\,,$$

from which we see that the determinant of the monodromy matrix is equal to one.

Using now the results of Sect. "Variational Equations," we find that

$$\det \Delta(t) = \det \left| \frac{\partial(x_1, x_2, x_3, x_4)}{\partial(x_{10}, x_{20}, x_{30}, x_{40})} \right| = 1\,. \tag{18}$$

This means that the determinant of the Jacobian of the flow in phase space is equal to one. Consequently, *the volume in phase space is conserved* (Liouville theorem).

The monodromy matrix of a Hamiltonian system is symplectic (see for example, [20])

$$\Delta^\tau(T) J \Delta(T) = J\,, \tag{19}$$

where the superscript τ means transpose. This is an important property of the monodromy matrix of a Hamiltonian system, which is called the *symplectic* property. Thus we come to the conclusion that *the monodromy matrix of a Hamiltonian system is symplectic*.

The eigenvalues of a symplectic matrix have some special properties. We express the property (19) as

$$\Delta^\tau(T) = J\Delta^{-1}(T)J^{-1}\,,$$

from which we see that the matrix $\Delta^\tau(T)$ is related to the matrix $\Delta^{-1}(T)$ by a similarity transformation. Consequently, they have the same set of eigenvalues. Thus finally, we come to the conclusion that the eigenvalues of $\Delta(T)$ are in reciprocal pairs. In addition, due to the fact that the matrix $\Delta(T)$ is real, they are also in complex conjugate pairs.

From the above we see that the four eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ of the monodromy matrix have the property

$$\lambda_1\lambda_2 = 1\,, \quad \lambda_3\lambda_4 = 1\,.$$

We note that the variational equations correspond to a periodic orbit $x(t)$. So, $\xi(t) = \dot{x}(t)$ is a periodic solution of

the variational equations and according to Sect. "Variational Equations," one eigenvalue is equal to one, $\lambda_1 = 1$. Using now relation $\lambda_1\lambda_2 = 1$, we come to the conclusion that *the monodromy matrix of a Hamiltonian system corresponding to a periodic orbit has a double unit eigenvalue*,

$$\lambda_1 = 1\,, \quad \lambda_2 = 1\,.$$

Stability of Hamiltonian Systems

The stability of the periodic orbit depends on the eigenvalues of the monodromy matrix, as we showed in Sect. "Linear Stability of a Periodic Orbit." Instability appears if at least one eigenvalue is outside the unit circle in the complex plane. Since two of the eigenvalues are always equal to unity, it is the other two eigenvalues, λ_3, λ_4, that determine the stability. As we proved, these eigenvalues are reciprocal and also complex conjugate, so they are either *on the unit circle*, or *on the real axis*, one inside the unit circle and the other outside. If they are real, the orbit is unstable, because one of them will be larger than + 1 or smaller than − 1. A special case is $\lambda_3 = \lambda_4 = +1$ or $\lambda_3 = \lambda_4 = -1$.

A remark is necessary at this point for the double unit eigenvalue. In Hamiltonian systems, in general, to the double unit eigenvalue there exists only one eigenvector. This introduces a secular term in the general solution of the variational equations. The two linearly independent solutions corresponding to the double unit eigenvalue are

$$\begin{aligned} \xi^1 &= f_1(t)\,, \\ \xi^2 &= f_2(t) + t f_1(t)\,, \end{aligned} \tag{20}$$

where $f_1(t)$, $f_2(t)$ are T-periodic. This implies that the orbit is always unstable, due to the secular term $t f_1(t)$. We will show however that this secular term introduces a time shift only along the perturbed orbit, and thus we have *orbital stability*, provided that the other two eigenvalues are on the unit circle: Taking into account that $\xi^1(t) = \dot{x}(t)$, where $x(t)$ is the periodic solution corresponding to the unit eigenvalue, we note that the perturbed orbit has a term $\epsilon\xi^2$ and the corresponding part of the the solution is expressed as $x'(t) = x(t) + \epsilon t\dot{x}_i(t) + \epsilon f_2(t)$ and, to a linear approximation in ϵ,

$$x'(t) = x(t + \epsilon) + \epsilon f_2(t + \epsilon t)\,.$$

Thus, if we define a new time $t' = t + \epsilon t$, we obtain (see Fig. 1b)

$$x'(t) - x(t') = \text{bounded}\,.$$

Thus we come to the conclusion that the secular term introduces a phase shift only along the orbit. This means that the two orbits, $x(t)$ and $x'(t)$, considered as geometrical curves, are close to each other. In this case we say that we have *orbital stability*, provided that the eigenvalues λ_3, λ_4 are on the unit circle and consequently the corresponding solution is bounded.

For the other two eigenvalues λ_3, λ_4 we have the solutions

- Eigenvalues real and positive: $\xi^{3,4} = f_{3,4}(t)\,\mathrm{e}^{\pm\alpha t}$,
- Eigenvalues real and negative: $\xi^{3,4} = f_{3,4}(t)\,\mathrm{e}^{\pm\alpha t}\,\mathrm{e}^{\pm i\pi t/T}$,
- Eigenvalues complex conjugate on the unit circle $\xi^{3,4} = f_{3,4}(t)\,\mathrm{e}^{\pm i\beta t}$,

where α, β are real and the functions $f_3(t)$, $f_4(t)$ are T-periodic. The exponent α is called the *characteristic exponent*. The general solution in the vicinity of the periodic solution is a linear combination of the above four solutions $\xi^1, \xi^2, \xi^3, \xi^4$.

The stability criteria can be obtained from the elements of the monodromy matrix as follows: The eigenvalues are the roots of the characteristic equation of $\Delta(T)$ and consequently

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = \text{trace}\,\Delta(T)\,,$$

$$\lambda_1\lambda_2\lambda_3\lambda_4 = \det\Delta(T) = 1\,.$$

Taking into account that $\lambda_1 = \lambda_2 = 1$ we find that the two nonzero eigenvalues λ_3, λ_4 are the roots of the quadratic equation

$$\lambda^2 - K\lambda + 1 = 0\,,$$

where

$$K = \text{trace}\,\Delta(T) - 2\,.$$

The stability depends on the value of K, which is called the *stability index*. Note that the stability index depends only on the trace of the monodromy matrix.

Asymptotic stability never appears, because it is not possible for the eigenvalues λ_3, λ_4 to be *both* inside the unit circle. This is also a consequence of the fact that the volume in phase space is conserved.

Let us assume that a periodic orbit is stable, which implies that the eigenvalues λ_3, λ_4 are on the unit circle and we assume that they are not equal to + 1 or − 1. If a parameter varies, then the eigenvalues λ_3, λ_4 are restricted to move *on the unit circle*, because they must be both inverse, $\lambda_3 = 1/\lambda_4$ and complex conjugate. Consequently, the stability is conserved. However, if λ_3, λ_4 meet at the points + 1 or − 1, then it is possible for them to go outside the unit

circle and thus generate instability. For this reason the orbits with $\lambda_3 = \lambda_4 = \pm 1$ are called *critical* as far as the stability is concerned. This is the mechanism by which instability is generated at the 3:1 resonance in the asteroid belt, where the eigenvalues λ_3, λ_4 meet at the point -1 [13].

Extension to Three or More Degrees of Freedom

All the above results concerning the eigenvalues and the stability of a periodic orbit, obtained for Hamiltonian systems with two degrees of freedom, can be easily extended to three or more degrees of freedom.

In a Hamiltonian system with n degrees of freedom the monodromy matrix is a $2n \times 2n$ symplectic matrix, and the eigenvalues are in reciprocal pairs (because of the symplectic property), and in complex conjugate pairs (because the elements of the matrix are real).

There is always a unit pair of eigenvalues, due to the existence of the energy integral $H = h =$ constant (see Sect. "Linear Stability of a Periodic Orbit"). For the other eigenvalues we have the following possibilities:

- Complex conjugate, on the unit circle, $e^{\pm i\phi}$: STABILITY.
- Real, on the real axis, in reciprocal pairs (positive or negative), $\lambda, 1/\lambda$: INSTABILITY.
- Complex, inside and outside the unit circle, in reciprocal and in complex conjugate pairs, $Re^{i\phi}$, $Re^{-i\phi}$, $R^{-1}e^{i\phi}$, $R^{-1}e^{-i\phi}$: COMPLEX INSTABILITY.

Note that in three, or more, degrees of freedom we have a new type of instability, the *complex instability*, which cannot appear in systems with two degrees of freedom.

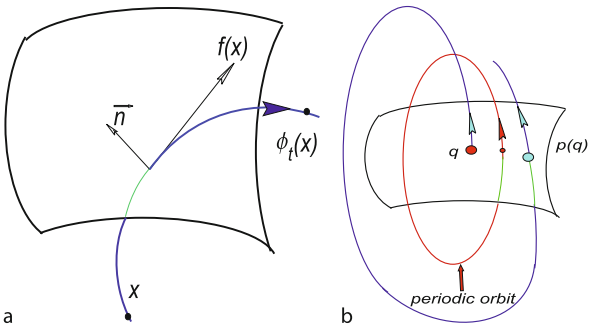
The Poincaré Map

This a very useful method in the study of the evolution of a dynamical system. By the Poincaré map we transform the continuous flow in the n-dimensional phase space of a dynamical system to an equivalent discrete flow (map) in a phase space of $(n-1)$-dimensions (or $(n-2)$-dimensions for Hamiltonian flows).

Consider the dynamical system in $\mathcal{R}^n$: $\dot{x} = f(x)$, where $x, f(x)$: vectors in $\mathcal{R}^n$ and $\phi_t(x)$ is the flow. We consider the surface of section

$$\Sigma \subset \mathcal{R}^n : (n-1) - \text{dim}$$

and we assume that the flow is transverse: The velocity vector of the flow is not tangent to this surface (Fig. 2a): $f(x) \cdot n(x) \neq 0$, where $n(x)$ is the normal unit vector to the surface.



Orbital Dynamics, Chaos in, Figure 2
a The surface of section, b The Poincaré map on a surface of section

The Poincaré map is defined as:

$$q \to p(q)\,,$$
$$p(q) = \phi_\tau(q)\,,$$

where q is the position on the surface of section at a $t = 0$ and $p(q)$ is the position on this surface at the next intersection at $t = \tau$, (Fig. 2b).

The following properties apply:

- The vector $p(q)$ defines accurately the state.
- The vector $p(q)$ is a continuous function of q.
- If $\tilde{x}(t)$ is a T-periodic orbit, the corresponding Poincaré map is a *fixed point*, maybe multiple (it repeats itself after several intersections) as seen in Fig. 2b (for the simple case).

Poincaré Map in Hamiltonian Systems

In this case the differential equations of motion are the canonical equations

$$\dot{q} = \partial H/\partial p\,, \quad \dot{p} = -\partial H/\partial q\,, \quad q, p \subset \mathcal{R}^n\,.$$

Let us consider the $(2n-2)$-dimensional surface of section Σ, defined as

$$H = h\,, \quad f(q,p) = 0 \quad (\text{for example} \quad q_2 = 0)\,.$$

The continuous Hamiltonian flow in the $2n$-dimensional phase space is transformed to an equivalent discrete flow (map), on a $(2n-2)$-dimensional surface of section. In addition to the general properties of the Poincaré map mentioned above, we also have the properties:

- The Poinaré map of a Hamiltonian flow is symplectic.
- The stability of the fixed points of the Poincaré map is the same as the stability of the corresponding periodic orbit. We have the same set of eigenvalues, except the double unit eigenvalue which corresponds to

the periodic orbit (the phase space now has two dimensions less). Note that this double unit eigenvalue is responsible for the phase shift along the perturbed orbit, which implies that this shift is eliminated by the Poincaré map. Thus, in the Poincaré map, the stability of the fixed point (periodic orbit) means *orbital stability*.

Hamiltonian Systems with Two Degrees of Freedom

The Poincaré map is particularly useful in systems with two degrees of freedom, where the phase space is four dimensional and the Poincaré map is in a two dimensional phase space. This makes the study very easy because we present the evolution of the system in a two dimensional space, where we can have a direct view.

We define the variables x_i as $x_1 = q_1$, $x_2 = q_2$, $x_3 = p_1$, $x_4 = p_2$. The energy integral is $H(x_1, x_2, x_3, x_4) =$ constant. We consider the surface of section

$$H(x_1, x_2, x_3, x_4) = h\,, \qquad x_2 = 0\,, \quad \text{with } x_4 > 0\,.$$

The map is in the space $x_1\,x_3$. The consecutive points of the map may lie on a smooth curve, called *invariant curve* (ordered motion), or be scattered (chaotic motion).

Let us assume that another first integral of motion exists, in addition to the energy integral $H = h =$ constant:

$$G(x_1, x_2, x_3, x_4) = c\,.$$

Then all the consecutive points of the map lie on smooth invariant curves: Let (x_1, x_3) be a point of the map on the two-dimensional surface of section. We have $x_2 = 0$ and x_4 is expressed in terms of x_1, x_3, through the energy integral $H = h$, as $x_4 = x_4(x_1, x_2 = 0, x_3)$. The points x_1, x_3 satisfy also the integral $G(x_1, x_2 = 0, x_3, x_4\,(x_1, x_2 = 0, x_3)) = c$, or

$$F(x_1, x_3) = 0\,,$$

which implies that the consecutive points (x_1, x_3) of the map lie on a smooth curve.

The Gravitational Two-Body Problem

The differential equations of the *relative motion* of two point masses m_1, m_2 are given by

$$\ddot{\vec{r}} = -\frac{GM}{r^2}\,\vec{e}_r\,,$$

where $M = m_1 + m_2$. The orbit is a conic section and in particular, for bounded motion, it is a Keplerian, elliptic orbit. The two bodies describe in the inertial frame two similar orbits around their common center of mass, whose dimensions are inversely proportional to their masses. This is one of the few integrable problems in nature. Its importance is that many real systems, as for example the asteroid problem, or the planetary systems, can be considered as perturbed two-body problems. For this reason it is important to know the basic properties of this simple two body problem and then study the evolution as a perturbation is applied.

The Two-Body Problem in a Rotating Frame

Consider a body, S, with mass m_1 and a second body, J, with mass m_2, which describe circular orbits around their common center of mass. We define a rotating frame of reference xOy, whose x-axis is the line SJ, the origin is at their center of mass and the xy plane is the orbital plane of the circular motion of the these two bodies (Fig. 3b).

Our aim is to study the motion of a massless body A in the rotating frame xOy, under the gravitational attraction of S and J. We start with a zero mass of the body J, $m_2 = 0$. In this approximation, the second body J is used only to define the rotating frame xOy, which rotates with constant angular velocity n'. Evidently, the motion of the body A is a Keplerian orbit, presented in the rotating frame. We shall give later mass to the body J, thus perturbing the Keplerian orbit of A.

The Hamiltonian function H that describes the unperturbed motion of A, in polar coordinates, r, ϕ (in the rotating frame), is

$$H_0 = \frac{p_r^2}{2} + \frac{p_\phi^2}{2r^2} - n' p_\phi - \frac{GM}{r}\,. \tag{21}$$

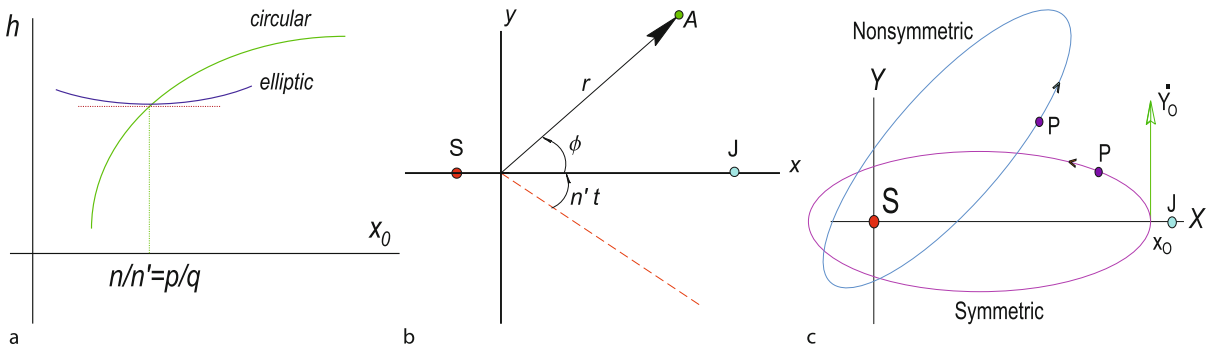
The moments are $p_r = \dot{r}$ and $p_\phi = r^2(\dot{\phi} + n')$. Note that the angle ϕ is an ignorable coordinate and consequently, in addition to the energy integral $H_0 = h =$ constant, we also have the angular momentum integral $p_\phi =$ constant.

The orbit of the body A (in the inertial frame) is a Keplerian orbit, which we assume to be elliptic. In terms of the elements of the orbit, the Hamiltonian (in the rotating frame) and the angular momentum are expressed as

$$H_0 = -\frac{GM}{2a} - n' p_\phi\,, \quad p_\phi = \sqrt{GMa(1-e^2)}\,.$$

Circular Orbits In the rotating frame there exist circular orbits of the body A with an arbitrary radius r_0, which correspond to the periodic solution

$$r = r_0\,, \quad p_r = 0\,, \quad \dot{\phi} = n - n'\,, \quad p_\phi = n r_0^2\,,$$



Orbital Dynamics, Chaos in, Figure 3

a The families of circular and of resonant elliptic periodic orbits in the unperturbed problem. The tangent to the elliptic family at the bifurcation point is parallel to the *x*-axis. b The rotating frame of the restricted problem. The mass, μ, of the second body, J, is equal to zero in the unperturbed problem, and is used only to define the rotating frame. In this case, the first body, S, is at the origin. c Two elliptic orbits of the small body, in the inertial frame, for $\mu = 0$. One is symmetric, corresponding to $\omega = 0^0$, and the other is asymmetric, corresponding to an arbitrary value of ω

where

$$n = p_\phi / r_0^2$$

is the angular velocity of the circular orbit (in the inertial frame). The following relations also hold:

$$\frac{p_{\phi 0}^2}{r_0^3} = \frac{GM}{r_0^2} \rightarrow \frac{GM}{r_0^3} = n^2 .$$

The period of the circular orbit in the rotating frame is

$$T = \frac{2\pi}{(n - n')} .$$

A circular orbit in xOy is a Keplerian orbit in the inertial frame, with semi major axis $a = r_0$ for any r_0. Consequently, a *family of circular periodic orbits* exists, which evidently is symmetric with respect to the x-axis. The parameter along the family is the semi major axis a (the radius), or the angular velocity (in the inertial frame) n. This family is represented by a smooth curve, in the space $h - r_0$, given by

$$-\frac{GM}{2a} - n' \sqrt{GMa} = h ,$$

obtained from the energy integral for $e = 0$ (Fig. 3a). Note that from the energy integral $H_0 = h$ we can obtain the value of $\dot{y}_0$, which together with x_0 define exactly the initial state, because $y_0 = 0$ and $\dot{x}_0 = 0$, due to the symmetry of the orbit with respect to the x-axis.

Elliptic Orbits An elliptic orbit in the inertial frame is periodic in the rotating frame only if it is resonant:

$$\frac{n}{n'} = \frac{p}{q} = \text{rational} ,$$

which means that the semi major axis must be given by

$$\frac{(GM)^{1/2} a^{-3/2}}{n'} = \frac{p}{q} .$$

Let us consider now a particular resonance p/q, which means that we keep fixed the semi major axis $a_{p/q}$. The orbit is resonant periodic for *any* eccentricity e, so a *family of elliptic periodic orbits* exists, with the eccentricity as a parameter along the family. There is however another parameter, defining the *orientation* of the elliptic orbit, which is the angle ω of the line of apsides with a fixed direction. In general, an elliptic orbit is not symmetric with respect to the rotating x-axis, contrary to the circular orbits, which are symmetric.

In the space $h - r_0$, where $r_0 = a_{p/q}(1 - e)$ is the pericenter distance ($r_0 = x_0$), an elliptic family is represented by a smooth curve (Fig. 3a), given by the energy integral

$$-\frac{GM}{2a} - n' \sqrt{GMa(1 - e^2)} = h .$$

The value of a is fixed, equal to the corresponding resonance and the eccentricity is a parameter along the family. Note that this presentation is not unique: a point on the elliptic family represents *all* the elliptic resonant orbits with the same eccentricity, but arbitrary orientation ω. An elliptic periodic orbit in the rotating frame is also periodic in the inertial frame. The resonant families of periodic orbits bifurcate from the family of circular orbits, at those points corresponding to the resonant values of the radius $a = a_{p/q}$.

Note that along the circular family the value of the semimajor axis varies, and consequently the ratio n/n'

varies and passes through resonant (rational) values. It is at these points that we have a bifurcation to an elliptic family. Evidently, all the circular and the elliptic orbits are stable, as they are Keplerian orbits.

In the following we study how the above mentioned nice picture of the families of periodic orbits change, when a perturbation is applied, and how instabilities and chaotic regions appear. We consider two cases: the restricted 3-body problem, both circular and elliptic, and the planetary problem, including our Solar System and the extrasolar planetary systems. In all these cases the Hamiltonian is expressed in the form

$$H = H_0 + \epsilon H_1 \,, \tag{22}$$

where H_0 is the integrable Hamiltonian of the two body problem.

Application to the Solar System

A Global View of the Families of Periodic Orbits

We consider the Sun and Jupiter revolving around their common center of mass in *circular orbits* or in *elliptic orbits* and a third body, with negligible mass, moving under the gravitational attraction of these two bodies. We make the approximation that the small body does not affect the motion of the two main bodies, Sun and Jupiter, which we will call *primaries*. This model is the *restricted three-body problem* and an extended study is in the books of Szebehely [46] and Roy [44]. This is a non integrable system, which is a good model to study the motion of a small body in our Solar System, for example an asteroid, a comet, or a small body in the Kuiper belt, at the edge of our Solar System (Jupiter is replaced by Neptune in this latter case).

Let us start with the study of the motion of an asteroid in the asteroid belt, a zone of small bodies between the orbits of Mars and Jupiter. For this reason we define a *rotating* frame xOy, with O the center of mass of the Sun and Jupiter and the x-axis along the line Sun–Jupiter (Fig. 3b). We start our study with the simplest case, considering that the orbits of the Sun and of Jupiter are circular (circular restricted three body problem). In this case the system xOy rotates with *constant* angular velocity n'. We start with planar motion of the asteroid and then extend the study to motion in space. Based on this model, we extend our study by assuming that the orbits of the Sun and Jupiter are elliptic.

In our study we normalize the units of length, mass and time by the relations

$$G = 1\,, \quad (m_{\rm sun} = 1-\mu\,,\ m_{\rm J} = \mu)\,, \quad r_0 = 1\,, \quad \text{which implies} \quad n' = 1\,,$$

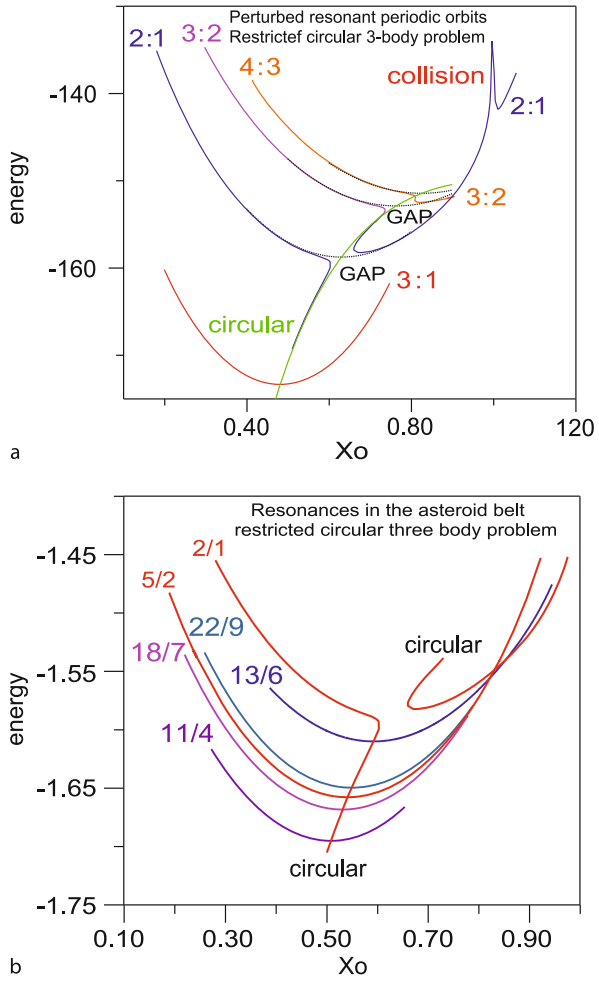
where G is the gravitational constant, r_0 is the radius of the circular orbit of Jupiter around the Sun, and μ, the mass of Jupiter, is considered a small parameter, of the order of 10^{-3} in our case. The Hamiltonian for the motion of the small body is of the form (22), with $\epsilon = \mu$, where H_0 is the Hamiltonian (21) of the two body problem in the rotating frame.

Planar Orbits Let us start with the unperturbed problem, $\mu = 0$, for planar motion, which is the two-body problem in the rotating frame. As we mentioned in the previous section, there exists a family of circular orbits, along which the resonance n/n' varies (n is the mean motion (angular velocity) of the orbit of the asteroid) and families of resonant *elliptic* periodic orbits, which bifurcate from the circular family at all the resonant circular orbits $n/n' = p/q$, as shown schematically in Fig. 3a. Evidently, all the orbits of these families are stable, as they are Keplerian, elliptic, orbits. We study now how all these families evolve and where instabilities appear, when $\mu > 0$, i.e, when the gravitational effect of Jupiter in taken into account. A complete analysis is given in [20].

As we mentioned before, there is an infinite set of resonant periodic orbits along the circular unperturbed family. The continuation of the above mentioned circular family from $\mu = 0$ to $\mu > 0$ and the generation of instabilities depends on the resonances that appear on this family. These resonances belong to three topologically different cases, as far as the continuation to $\mu > 0$ is concerned. These are the cases (i) $n/n' = (\nu+1)/\nu$, (ii) $n/n' = (2\nu+1)/(2\nu-1)$, ($\nu = 1, 2, 3, \dots$) and (iii) all other resonances.

(i) All the circular orbits that are not at the resonance $n/n' = 2/1, 3/2, \dots$, are continued as nearly circular orbits in the rotating frame. The resonant circular orbits $n/n' = 2/1, 3/2, \dots$, are not continued as periodic orbits in the rotating frame. At these resonances, a gap appears and the single unperturbed family of circular orbits breaks into a set of disconnected families of periodic orbits. From these gaps we have a bifurcation of two families of resonant elliptic periodic orbits (Fig. 4a). The stability of the circular orbits at $\mu = 0$ is preserved, except at the resonances 3/1, 5/3, ..., as we explain below. The resonant elliptic families at the 2/1, 3/2, ... resonances may be stable or unstable, depending on the phase (perihelion or aphelion at $t = 0$) and other factors (for example, close approaches).

(ii) At the circular orbits at the resonances $n/n' = 3/1, 5/3, \dots$ the continuation to nearly circular orbits is



Orbital Dynamics, Chaos in, Figure 4

a The circular family and the bifurcation to resonant elliptic families at the resonances 2/1, 3/2, 4/3, . . . for $\mu = 0.000954786$. The orbits on the elliptic families are symmetric, corresponding to $\omega = 0^0$ or $\omega = 180^0$. These are the only orbits that were continued to $\mu \neq 0$. All the other unperturbed orbits, corresponding to any other value of ω (see Fig. 3c), did not survive the perturbation, as a consequence of the Poincaré–Birkhoff fixed point theorem. b A closer look at some resonant families, for different higher order resonances

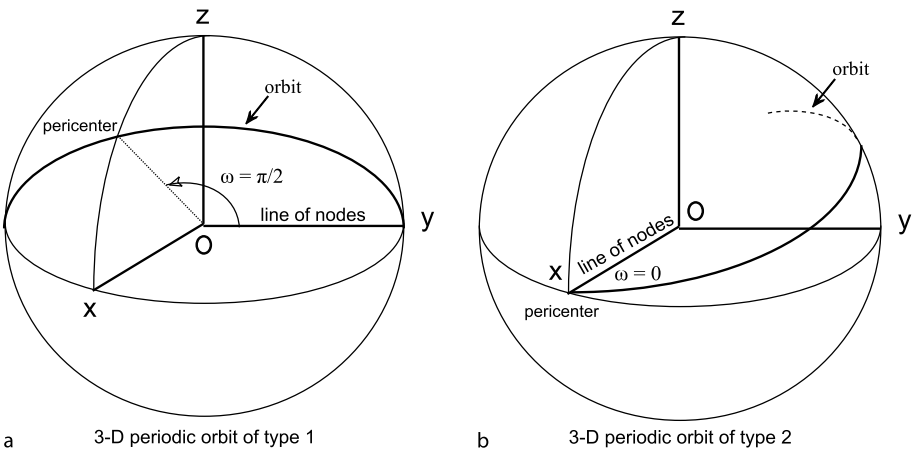
possible, but the stability is destroyed. A small unstable region appears at these resonances, on the family of circular orbits. At the critical points, at the two ends of this unstable region, we have a bifurcation of two families of symmetric resonant elliptic periodic orbits, which differ in phase. One is stable and the other is unstable (Fig. 13a).

(iii) In all other resonances on the circular family, for example 5/2, 4/1, 7/3, … the circular orbits are continued as nearly circular orbits and in addition the stability is preserved. At these points we have a bifurcation of two families of symmetric resonant elliptic periodic orbits which differ in phase (see Fig. 4b). In this case also, one family is stable and the other unstable (but the stability may change along a family).

A remark is necessary at this point for the families of elliptic periodic orbits for $\mu > 0$. The elliptic unperturbed resonant families, shown in Fig. 3a, are two parametric, with the eccentricity e and the angle of apsides ω as the two parameters. The eccentricity increases along the elliptic family, starting from zero values, but to a fixed eccentricity there corresponds an infinity of values of ω (Fig. 3c). What happens to this two-parametric family of unperturbed periodic orbits as $\mu > 0$? It is proved [20] that, for a fixed eccentricity, out of the infinite set of periodic orbits, for different omegas, only a finite, even, number survive (usually just two), half stable and half unstable. This is a consequence of the Poincaré–Birkhoff fixed point theorem (see [1]). This theorem refers to perturbed twist mappings: In the unperturbed case there exist resonant invariant curves where *all* points are fixed points, so that on this unperturbed invariant curve there exists an infinite number of fixed points. As soon a a perturbation is applied, only a finite number of fixed points survive, half of them stable and half unstable.

Thus, all the elliptic resonant families are monoparametric families along which the eccentricity increases, starting from zero values. In most cases the orbits are symmetric with respect to the rotating x-axis ($\omega = 0$ or $\omega = \pi$) and the eccentricity can be considered as a parameter. Some of these families are stable and others are unstable. The stability depends on the phase, i. e. on whether the asteroid is at perihelion or aphelion when it crosses the x-axis, but also on other factors as, for example, to a close encounter with Jupiter. Along a family of resonant elliptic periodic orbits the resonance is almost constant. A global picture of the circular and the elliptic families is shown in Fig. 4a,b.

Three-Dimensional Orbits in the Circular Model We study now three-dimensional periodic orbits in the model of the circular restricted problem. These families bifurcate from the planar families at those points which are critical with respect to the vertical stability. It is only at these points that the vertical deviations $\xi_3(t)$ of a perturbed orbit, given by the variational equations (14), have a period equal to the period of the planar periodic orbit. We remark at this point that along a resonant family of elliptic periodic orbits, the vertical stability index is very close to crit-



Orbital Dynamics, Chaos in, Figure 5
The two different symmetries in three-dimensional orbits: a Type 1. b Type 2

ical. (It is exactly critical on the unperturbed elliptic family). Depending on the particular resonance, such a critical point may or may not exist.

The three-dimensional periodic orbits are, in general, symmetric and there exist two types of symmetry. Their initial conditions are given below and are shown in Fig. 5:

$$\underline{\textit{Type 1}}: \quad x_{20} = \dot{x}_{10} = \dot{x}_{30}, \quad x_{10}, x_{30}, \dot{x}_{20} \neq 0 .$$

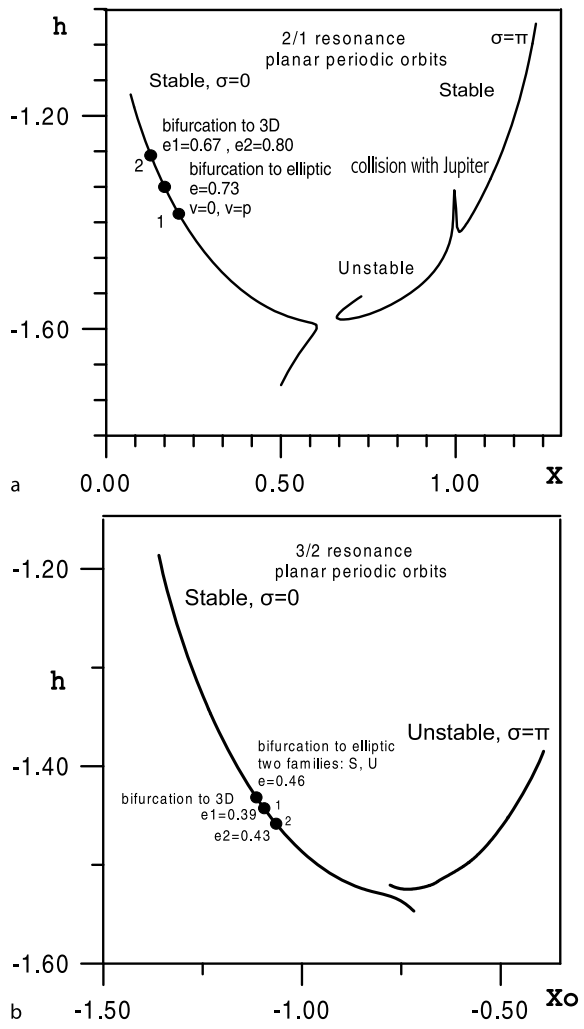
$$\underline{\textit{Type 2}}: \quad x_{20} = x_{30} = \dot{x}_{10} = 0, \quad x_{10}, \dot{x}_{20}, \dot{x}_{30} \neq 0 .$$

In Fig. 6a we present the two families of 2/1 resonant planar periodic orbits, corresponding to $\mu = 0.000954786$. The stable family, for $\omega = 0$, corresponds to the case where the asteroid is at perihelion at $t = 0$. The other family, for $\omega = \pi$, corresponds to position of asteroid at aphelion at $t = 0$ and starts as unstable up to the point where we have a collision with Jupiter. After that point the family continues, for larger eccentricities, as stable. On the stable family, corresponding to perihelion at $t = 0$, there exist two critical points, at high eccentricities, $e = 0.67$ and $e = 0.80$, as far as the vertical stability is concerned, as shown in Fig. 6a. From each one of these two points we have a bifurcation of a family of three-dimensional periodic orbits. One family, starting from $e = 0.67$, belongs to type 1 and is stable (Fig. 7a), while the other family, starting from $e = 0.80$, belongs to type 2 and is unstable (Fig. 8a). Typical three-dimensional periodic orbits on these two families are shown in Fig. 7b and Fig. 8b. Also, on the stable family in Fig. 6a, there exists a bifurcation point, at $e = 0.73$, to two families of periodic orbits of the elliptic problem (see Fig. 9), as explained in the next paragraph. Note that the bifurcation points to three dimensional periodic orbits and to the elliptic problem exist only on the stable family. On the unstable families such bifurcation points do not exist. A remark is necessary at this point: In the unperturbed case ($\mu = 0$) all points on the families of elliptic periodic orbits are critical as far as the vertical stability is concerned and also critical as far as the bifurcation to the elliptic problem is concerned (period equal to 2π). The existence or not of such critical points when $\mu \neq 0$ depends on the particular resonance. In the case we studied here, only the above critical points appeared. In other resonances the situation may be quite different.

A similar situation exists for the 3/2 resonance, as shown in Fig. 6b.

Families in the Elliptic Restricted Problem Families of resonant periodic orbits in the case where the orbits of the Sun and Jupiter are *elliptic*, with eccentricity e_J (elliptic restricted three-body problem) exist, which bifurcate from the families of the circular model, either the circular family or the elliptic families. The bifurcation can take place only at those points where the period of the periodic orbit on the families of the circular model is equal to the period of Jupiter (or a multiple of it). For a fixed value $e_J > 0$ the periodic orbits are isolated. We obtain a family by varying e_J.

Continuation from the Family of Circular Orbits Let us start from the unperturbed circular family. The period of a circular orbit ($\mu = 0$) is $T = 2\pi/(n - n')$ in the rotating frame and the period of Jupiter is $T_J = 2\pi/n'$, where n' is its mean angular velocity. We have $T = T_J/(n/n' - 1)$. At the resonance $n/n' = p/q$ we have $T = T_J q/(p - q)$ and if this orbit is described $p - q$ times, the period T^* of this orbit is an integral multiple of T_J, $T^* = qT_J$. If



Orbital Dynamics, Chaos in, Figure 6

a The two elliptic families of resonant periodic orbits at the 2/1 resonance and the bifurcation points to three-dimensional orbits, at $e = 0.67$ and $e = 0.80$ and also to the elliptic model, at $e = 0.72$ on the stable family. b The resonant elliptic periodic family at the 3/2 resonance and the bifurcation points to three-dimensional orbits, at $e = 0.39$ and $e = 0.43$ and also to the elliptic model, at $e = 0.46$

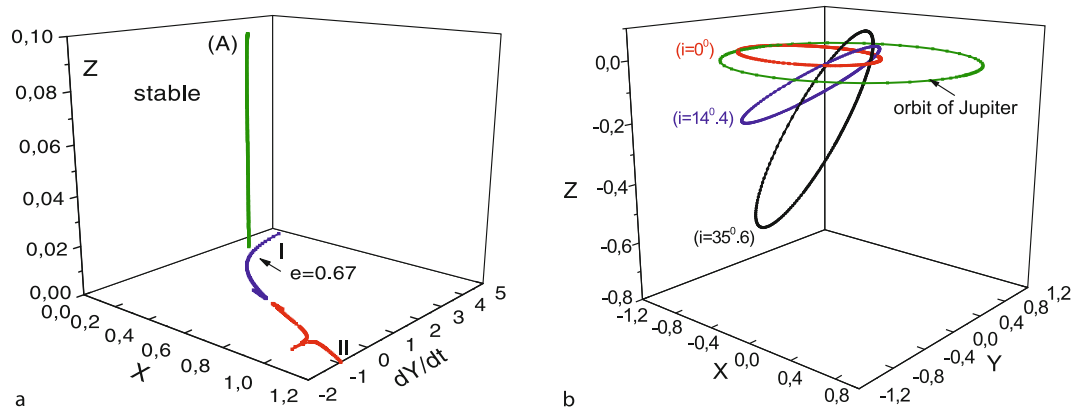
$n/n' \neq 2/1, 3/2, \ldots$ the region around this resonant orbit is continued to $\mu > 0$, as mentioned above. On the continued circular family there exists an orbit which, if described $p - q$ times, has a period exactly equal to qT_J. This means that a bifurcation from the circular family $\mu > 0$ to a family of the elliptic problem can take place close to a resonance. This is the case with the 3/1 resonance (Fig. 13a).

Continuation from the Family of Nearly Elliptic Orbits Consider a family of $n/n' = p/q$ resonant elliptic periodic orbits of the circular planar problem, for $\mu = 0$. The period all along the family is constant, and according to the above, if the orbits of the family are described $(p - q)$ times, the period is $T^* = qT_J$. This family is continued, when $\mu > 0$, to two families of elliptic periodic orbits, differing in phase. Along each family the eccentricity increases, starting from zero values. For continuity reasons, the (multiple) period along the continued family is close to qT_J. If at a certain point, corresponding to a value e of the eccentricity, it happens to be *exactly* equal to qT_J, then a bifurcation to the elliptic problem can take place. Two families of periodic orbits exist, along which the eccentricity of Jupiter increases. For a fixed eccentricity of Jupiter, for example $e_J = 0.048$, only two *isolated* periodic orbits exist. The above mentioned two families differ in the initial phase of Jupiter on its elliptic orbit at $t = 0$, i. e. if it is at perihelion or aphelion. In general, one family is stable and the other is unstable.

The numerical computations have shown that in certain resonances, for example 2/1, 3/2, 3/1, such bifurcation points do exist, at quite large values of the eccentricity (see Fig. 6a,b for the 2/1 and 3/2 resonances and Fig. 13a for the 3/1 resonance). But in other resonances, for example 7/3, such bifurcation points do not exist. This plays an important role on the topology of the phase space close to a particular resonance, because the existence of a resonant periodic orbit/fixed point of the Poincaré map, determines the topology of the phase space. The non existence of periodic orbits in a region implies that the phase space is smooth and ordered regions exist. A systematic study along these lines has been made by Tsiganis et al. [47,48,49]. We present in Fig. 9, as an example, two resonant families of the elliptic problem, at the 2/1 resonance, one stable and one unstable. These two orbits bifurcate from the point on the stable branch of the 2/1 resonant family of the circular problem, at $e = 0.72$, as shown in Fig. 6a.

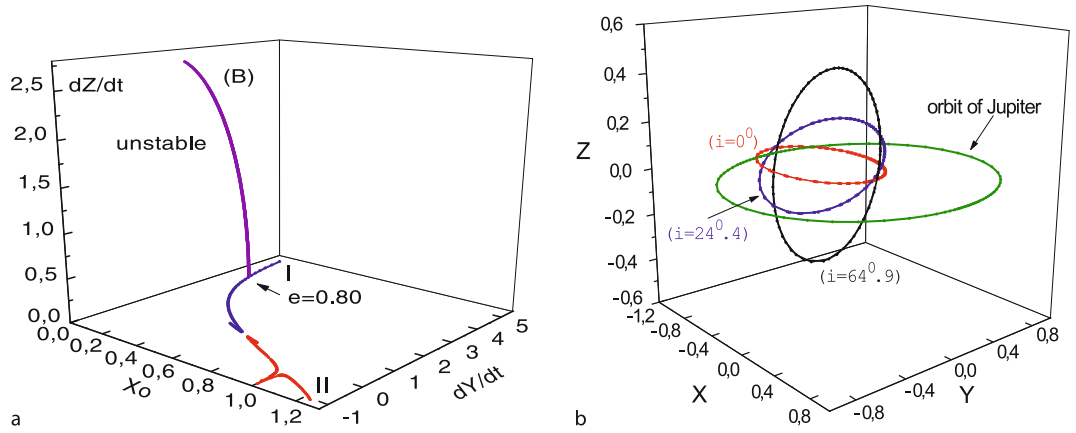
Generation of Chaos at the Unstable Periodic Orbits

Let us consider the simplest model, the circular restricted three-body problem, and study the topology of the phase space at the 2/1 and the 3/2 resonances, using as a guide the families of periodic orbits as presented in Fig. 4a and Fig. 6a,b. We compute the Poincaré map on the surface of section $y = 0$, $H = h$, for different values of the energy h. These energy levels can be visualized by considering lines parallel to the x_0 axis in Figs. 4a (or 6a,b) at different values of h. Note that these lines intersect the circular family and the resonant families, and these intersections correspond to the fixed points of the Poincaré map. For a better understanding of the physics, we mark on each map the



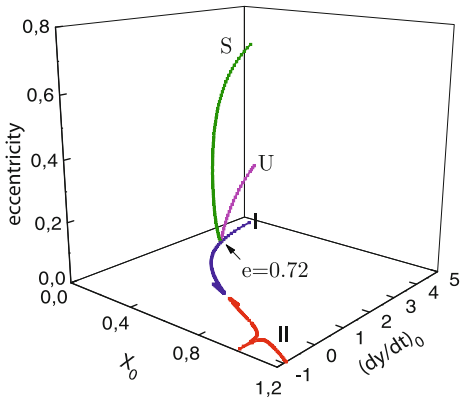
Orbital Dynamics, Chaos in, Figure 7

a The stable family of three-dimensional periodic orbits bifurcating at $e = 0.67$ b Three-dimensional periodic orbits on the stable family (type 1)



Orbital Dynamics, Chaos in, Figure 8

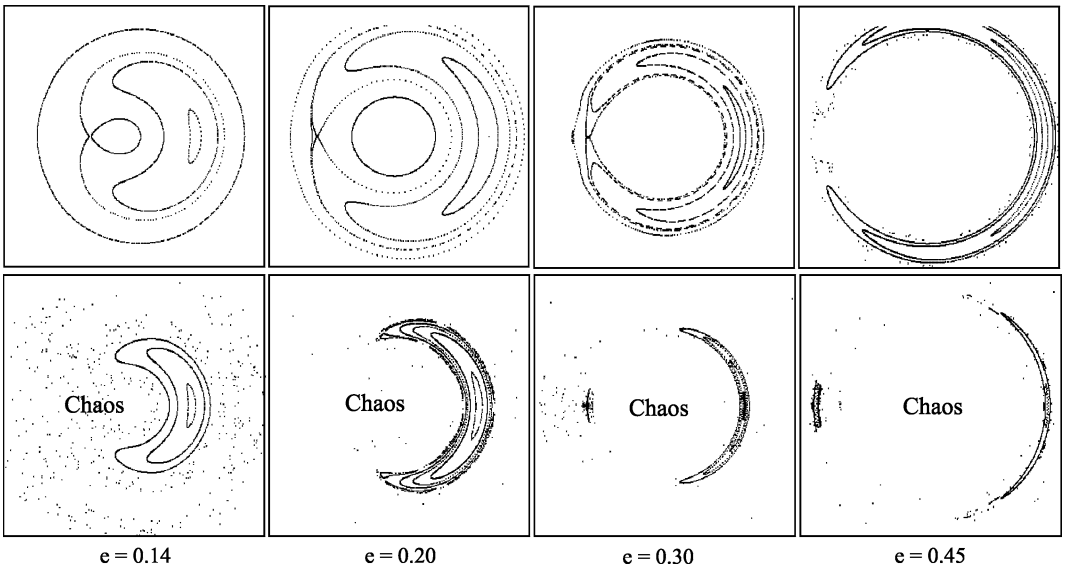
a The unstable family of three-dimensional periodic orbits bifurcating at $e = 0.80$. b Three-dimensional periodic orbits on the unstable family (type 2)



Orbital Dynamics, Chaos in, Figure 9

The two families of periodic orbit of the elliptic restricted three-body problem, bifurcating from the 2/1 resonant family at $e = 0.72$. One is stable and the other unstable

value of the eccentricity of the stable resonant fixed point, instead of the energy h, since along the family the eccentricity increases. In Fig. 10 we present several surfaces of section, at different energy levels, corresponding to different eccentricities, at the resonances 2/1 and 3/2. The fixed points corresponding to the circular periodic orbit (in the middle of the diagram) and to the stable and unstable resonant periodic orbits are clearly seen. The stable fixed points are surrounded by islands (closed invariant curves) of ordered motion, while the mapping close to the unstable fixed points is hyperbolic. Chaotic motion starts at these unstable points as the eccentricity increases as we move along the family. The chaotic orbits appear as scattered points, in contrast to the regular orbits, which are represented by smooth invariant curves. This phenomenon is stronger at the 3/2 resonance.



Orbital Dynamics, Chaos in, Figure 10
The Poincaré map at the 2/1 resonance (*upper row*) and the 3/2 resonance (*lower row*) for different energy levels, presented here by the corresponding eccentricity *e* of the resonant stable fixed point on the elliptic family. The stable and the unstable fixed points are clearly seen. The generation of chaotic motion at the unstable fixed points is evident. Note that the chaotic motion starts from the unstable fixed points

Note that the topology of the phase space, on the Poincaré map, is critically determined by the position of the fixed points and their stability character. This is the reason that the knowledge of the basic families of periodic orbits is so important for the study of the dynamics of the system.

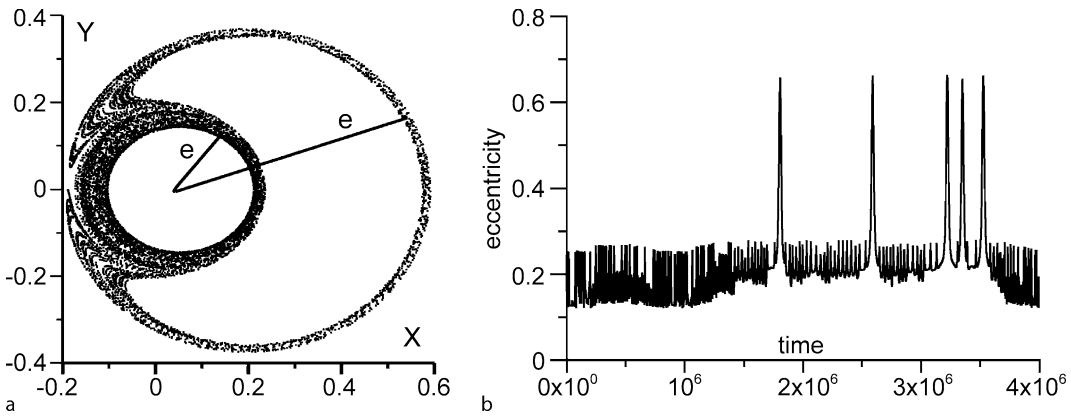
Asteroid Motion Close to a Resonance

It is known that in the region between the orbits of Mars and Jupiter there exists a zone of small bodies, revolving around the Sun, called *asteroid belt*. It has been observed that the distribution of these bodies is not smooth, but gaps exist at several resonances between the mean motion of the asteroid and Jupiter, the famous Kirkwood gaps (see Fig. 12). The explanation of these gaps was an open question for many decades, and their existence was explained by realizing that the motion at the 3/1 resonance (and in many other resonances) is chaotic and consequently an asteroid could not stay in this region for a long time. The first study was made by Wisdom [53,54,55]. The study was based on the construction of a symplectic mapping model, by making use of the averaged Hamiltonian of the elliptic restricted three-body problem at the 3/1 resonance. It was shown that, due to the existence of chaos at this region, the eccentricity of an asteroid that starts its motion in a nearly circular orbit undergoes sudden jumps, after a period which may be several million years (the semi major axis remaining almost constant), and thus the orbit of the asteroid may by Mars or even Earth crossing and thus undergo additional perturbations that will eventually drive it outside the 3/1 resonance region. Several papers followed this study, for many resonances in the asteroid belt, which used mapping models based on an averaged Hamiltonian at the corresponding resonance [1,15,17,21]. For the different methods used to transform the continuous flow to a mapping model see [16]. Much work on the asteroid belt has been also made by making use of the averaging method or a combination of this method and numerical integrations [22,35,37,38,39].

The variables used in the averaged models are the Delaunay variables, transformed to resonant action angle-variables (see [41]). For example, for the 3/1 resonance, for planar motion these variables are

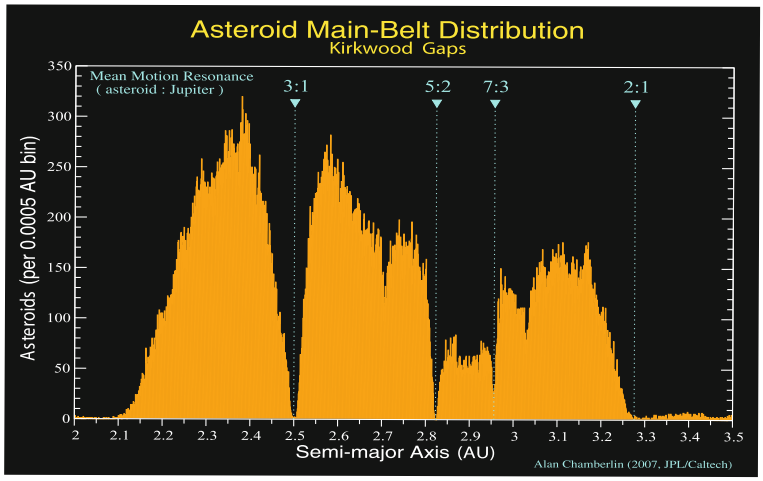
$$S = \sqrt{\mu_1 a}\left(1 - \sqrt{1-e^2}\right), \quad \sigma = \frac{1}{3}(3\lambda' - \lambda) - \omega,$$
$$N = \sqrt{\mu_1 a}\left(3 - \sqrt{1-e^2}\right), \quad \nu = -\frac{1}{3}(3\lambda' - \lambda) + \omega',$$

where $\mu_1 = 1 - \mu$, $e' = 0.048$, and λ, ω, a are the mean longitude, the longitude of perihelion and the semimajor axis of the asteroid and the corresponding primed quantities refer to Jupiter. The variables used to present



Orbital Dynamics, Chaos in, Figure 11

a The mapping, in the variables $X = e\cos(\sigma)$ — $Y = e\sin(\sigma)$ for the motion of an asteroid at the 3/1 resonance. b The evolution of the eccentricity. Chaotic jumps of the eccentricity appear, at unpredictable times



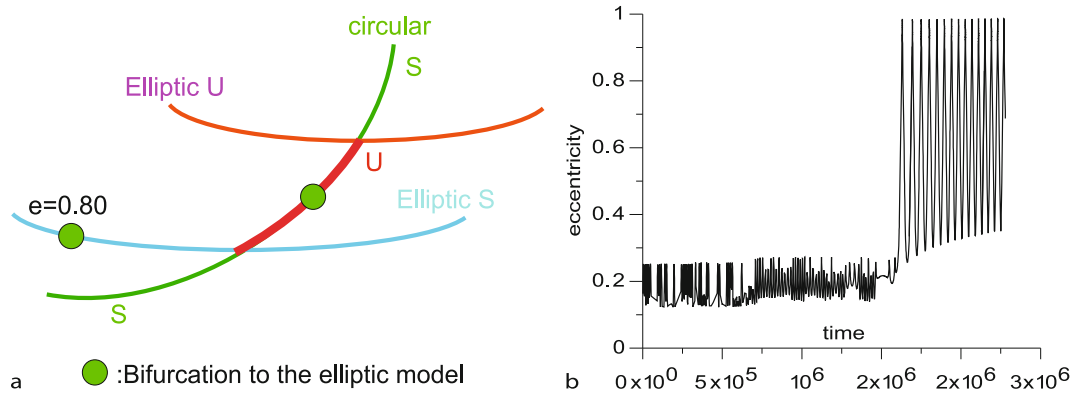
Orbital Dynamics, Chaos in, Figure 12

The distribution of the asteroids, obtained from 156 929 asteroids, as given by JPL/Caltech in 2007. The Kirkwood gaps at the 3/1, 5/2, 7/3 and 2/1 mean motion resonances are clearly seen

the mapping are the Poincaré variables $X = \sqrt{2S}\cos(\sigma)$, $Y = \sqrt{2S}\sin(\sigma)$, which are also canonical variables.

In Fig. 11a we present the mapping for an asteroid at the 3/1 resonance, from a mapping model used by Hadjidemetriou [15], equivalent to the map used by Wisdom [53]. The variables are similar to the Poincaré variables, but instead of $\sqrt{2S}$ we used the eccentricity e (note that $\sqrt{2S}$ is proportional to e, for small values of e). At the beginning the asteroid moves along the inner "diffused" circle, with small radius, corresponding to low values of the eccentricity, but eventually comes close to the chaotic region which connects the inner circle with an outer circle, with larger radius, corresponding to larger values of the eccentricity. So, through this chaotic window we have a connection between the low eccentricity regions and the high eccentricity regions. This results to a chaotic jump between small and large eccentricities, in a chaotic, unpredictable, way. This behavior is called *intermittency*. This is clearly shown in Fig. 11b.

At this point we draw the attention to an important point when we use the averaged Hamiltonian in the dynamical study. Since the averaging method is based on series expansions in a small parameter (in our case it is the eccentricity), it is not valid for high values of the parameter. In the present case, in the study of the asteroid at the 3/1 resonance, the averaged Hamiltonian used to construct the mapping which gives the evolution of the asteroid eccentricity does not contain the high eccentricity res-



Orbital Dynamics, Chaos in, Figure 13

a The families of circular and elliptic periodic orbits of the circular model at the 3/1 resonance, and the bifurcation points to the elliptic model (schematically). S stands for stable and U for unstable. b The evolution of the eccentricity when the high eccentricity resonances are included in the model. The chaotic jumps are now up to eccentricities equal to 1

onances. This is the case with the evolution of Fig. 11. For this reason it is important, in constructing the averaged model, to know the topology of the *whole* phase space, and this can be done only if we know *all* the resonant families of periodic orbits at the 3/1 resonance (and of course in all other similar studies in other resonances). *A necessary criterion for the validity of the averaged Hamiltonian is its fixed points to coincide with the periodic orbits (fixed points of the Poincaré map) of the original model* (the elliptic restricted three-body problem in this case). This shows the importance of the periodic orbits in orbital dynamics. In Fig. 13a we show, schematically, for the model of the circular restricted problem, the family of circular orbits and the unstable region which appears at the 3/1 resonance on the circular family, and also the two families of elliptic periodic orbits that bifurcate from the critical points at the two ends of this unstable region. One family is stable and the other is unstable. It is found that on the unstable part of the circular family there exists a bifurcation point to two families of 3/1 resonant periodic orbits of the elliptic model, which start with zero eccentricities. Both of them are unstable. These are the *low eccentricity resonances* that are included in the model of Fig. 11. However, there exists one more bifurcation point, at the eccentricity $e = 0.80$ on the stable family of elliptic periodic orbits (of the circular model), from which two 3/1 resonant periodic orbits of the elliptic model appear, one stable and the other unstable, starting with high eccentricities, equal to $e = 0.80$. For a full description of the resonant structure of the restricted three-body problem at the 3/1 resonance see [14,15]. It is these *high eccentricity resonances* that are missing from the model of Fig. 11. If these high eccentricity resonances are also included in the model, the jumps in the eccentricity are higher, up to $e = 1$ and thus the asteroid not only approaches the inner planets, but may also fall on the Sun. This evolution is shown in Fig. 13b.

The study of the ordered and chaotic regions in the asteroid belt is not the only such study in our Solar System. A zone of small bodies, similar to the asteroid belt, exists at the edge of our Solar System, after the orbit of Neptune. This is the Kuiper belt, whose existence was conjectured to explain the source of low period comets. Since the last decade of the 20th century many small bodies were observed in the Kuiper belt and it was realized that ordered and chaotic regions exist in this region also, similar to those in the asteroid belt, at several resonances with Neptune. *Pluto* is one such body in the Kuiper belt, trapped at the 3/2 resonance with Neptune, together with many other smaller bodies at the same resonance, called *plutinos*. A good view of the dynamical structure in the Kuiper belt is given in [6].

All major planets, Jupiter, Saturn, Uranus, Neptune, have planetary rings, the most well known being the rings of Saturn. Although many of the properties of the ring systems can be understood by a fluid dynamics approach, several of their features are explained by resonant dynamics, as in the case of the asteroid belt or the Kuiper belt. The fine structure of the rings can be explained by resonances between the ring particles and small satellites of the planet. A description of the dynamics of the planetary rings can be found in Chap. 10 in [41].

The chaotic behavior of the Solar System, as a whole, is yet another interesting subject, and there are several numerical works on this problem, notably by Laskar [25,26,27] and by Wisdom [56]. There are not large scale chaotic orbits of the planets, although the system is non

integrable and some chaos is expected. Especially the large planets do not show any significant change and their orbits stay, for some billion years, close to their present orbits. The inner planets, especially Mercury, have shown large deviations, but due to the chaotic nature of the Solar System and the fact that the numerical integrations are not with infinite accuracy, these results may not represent the actual evolution of the Solar System. It seems that the Solar System is stable and any chaotic motion is in a small scale and is bounded. Studies on the stability of extrasolar planetary systems have started recently, with many interesting results, as we explain in the next section.

Another interesting case of chaotic motion in the Solar System refers to the rotational motion of celestial bodies. Although the rotation of the planets is regular, there are small bodies, with irregular shape, that show chaotic rotation. Such a case is the satellite of Saturn, Hyperion, with approximate dimensions 180 km × 140 km × 112.5 km. Although its orbit is stable, due to the fact that it is at the 4:3 resonance with the more massive satellite of Saturn, Titan, its rotation is chaotic [57].

Extrasolar Planetary Systems

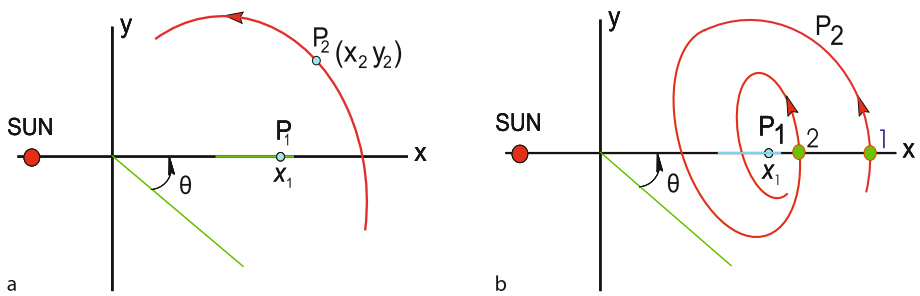
Some General Remarks

In the last decade of the 20th century it was discovered that our Solar System is not the only planetary system in the universe. Up to the present (May 2008) there are 281 observed extrasolar planetary systems, with 25 of them having two or more planets. In many planetary systems with two planets close to each other, the two planets are in mean motion resonance. Examples are: HD 82943 [23,33], GLIESE 876 [31,43], at the 2:1 resonance and 55Cnc at the 3:1 resonance [32]. Some of these systems have large eccentricities and are evidently stable.

There are different approaches to the study of the dynamical evolution of a planetary system and on the mechanisms that stabilize the system, or generate chaotic motion and instability: Beaugé and Michtchenko [2], Beaugé et al. [3,4,5], Ferraz-Mello et al. [8], Gozdjiewski et al. [10], Malhotra [30], Lee and Peale [29], Lee [28]. In these papers different methods have been applied, as the averaging method, direct numerical integrations of orbits, or various numerical methods which provide indicators for the exponential growth of nearby orbits. In this way the regions where stable motion exists have been detected, in the orbital elements space.

We present briefly a global view of the structure of the phase space of a planetary system with two planets, moving in the plane, as obtained from the set of the families of periodic orbits. As we have already mentioned before, the periodic orbits play a dominant role in understanding the dynamics of a system, because they determine critically the structure of the phase space. In this way, we can detect the regions where stable librations could exist. These will be the regions where a real planetary system could exist in nature. As we will see, stable regions corresponding to elliptic orbits of the two planets with relatively large eccentricities are associated with mean motion resonances. An early work on periodic orbits of the planetary type is by Hadjidemetriou [12], well before the first extrasolar planetary systems were observed. Many papers followed on these lines, after the first extrasolar planetary systems were observed [18,42,51,52]. We remark that stable motion could also exist far from resonances, if the eccentricities are small. This latter motion is close to a stable periodic orbit of the *circular family* of periodic orbits. We also remark that it is possible to have stable motion far from a periodic orbit, but in this latter case the two planets are not close to each other, so that their gravitational interaction is not very significant.

It can be proved [11] that families of periodic orbits in the planar general three body problem exist, in a *rotating* frame xOy, whose x-axis is the line $S - P_1$, with origin



Orbital Dynamics, Chaos in, Figure 14
a The rotating frame. The planet P_1 moves on the x-axis and the planet P_2 in the xOy plane. The angle θ is an ignorable coordinate. b The Poincaré map at $y_2 = 0$

at the center of mass of these two bodies, where S is the Sun and P_1 the inner planet. We assume that the center of mass of the whole system is at rest with respect to an inertial frame. We have four degrees of freedom, for planar motion, with generalized variables x_1, x_2, y_2, θ (Fig. 14a). This is a non uniformly rotating frame, and the second planet P_2 moves in the plane xy. It turns out [11] that the angle θ is ignorable, so we have three degrees of freedom in the rotating frame, with variables x_1, x_2, y_2. In the planetary three body problem (one big body, the star and two small bodies, the planets) the periodic orbits are similar to the families of the restricted problem as shown in Fig. 4. There are two types of periodic orbits:

- Non resonant periodic orbits with nearly circular orbits of the two planets.
- Resonant periodic orbits with nearly elliptic orbits of the two planets.

The circular orbits are all symmetric but the elliptic orbits may be symmetric or asymmetric. There exist families of elliptic periodic orbits for every mean motion resonance. Close to a stable periodic orbit there exists a region of stable librations, and it is at these regions that a planetary system could be trapped.

Concerning the continuation of the unperturbed family of periodic orbits ($m_1 = m_2 = 0$), the situation is similar to that explained in the restricted three body problem. There are three topologically different resonant cases:

- The resonances of the form $(n + 1)/n$, $(2/1, 3/2, \ldots)$ (Gaps on the circular family).
- The resonances $(2n + 1)/(2n - 1)$, $(3/1, 5/3, \ldots)$ (Instability on the circular family).
- All other resonances, $(5/2, 7/3, 8/3, \ldots)$ (Preservation of the stability on the circular family).

A global view of the resonant families of elliptic periodic orbits, for each one of the above resonance types can be found in [19]. There exist both *symmetric* and *asymmetric* families. The ratio of the planetary masses plays an important role on the stability and the existence of asymmetric families of periodic orbits. The sum of the masses of the planets also plays an important role on the stability and the existence of families of resonant periodic orbits. The stability of a symmetric periodic orbit depends, all other things being the same (semimajor axes, eccentricities), on the phase of the two planets, that is on whether the line of apsides are aligned or antialigned and on the position of the two planets at perihelion of aphelion at some epoch. The proper phase generates a *phase protection mechanism* so that stable planetary systems exist even for large eccentricities.

The properties of motion close to a periodic orbit are studied by considering a Poincaré map on the surface of section $y_2 = 0$, $(\dot{y}_2 > 0)$, $H = h =$ constant (Fig. 14b). The phase space of the Poincaré map is the four dimensional space $x_1, \dot{x}_1, x_2, \dot{x}_2$ ($y_2 = 0$ and $\dot{y}_2$ is obtained from $H = h, \dot{y}_2 > 0$). Close to a stable periodic orbit we have stable librations and the motion in phase space takes place on a torus. On the contrary, close to an unstable periodic orbit we have irregular, chaotic, motion and in many cases the system disrupts into a binary system (the star and one planet) and an escaping planet.

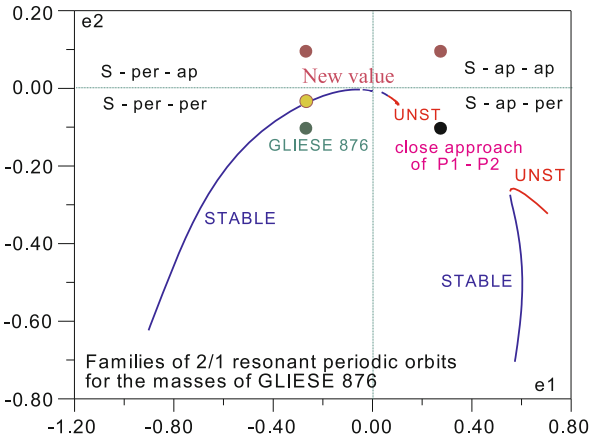
The position of some real extrasolar planetary systems is compared with the above mentioned regions of stable librations. A detailed analysis of the dynamics of extrasolar planetary systems based on the families of periodic orbits is presented in [20].

In the following, we present, as an example, the dynamics of a real extrasolar planetary system, Gliese 876, on the whole phase space, and study the stable configurations and the regions where chaotic motion appears.

A Real Extrasolar Planetary System: Gliese 876

Studies on the dynamical evolution of a planetary system, both theoretical and for real extrasolar planetary systems, have been made by different methods. One way to study the problem is to compute many orbits, for a set of initial conditions and study their behavior for a long time. A different method is to use the averaging method in order to obtain an averaged Hamiltonian, thus reducing the number of degrees of freedom. Analytic and numerical studies can then be made to find the stable regions in phase space [2,3,4,5] Ferraz-Mello et al., Gozdjiewski et al. [2], [28,29,30,45]. A systematic study of the orbital dynamics in planetary systems can be made by finding all the basic families of periodic orbits. As we mentioned before, the position and the stability character of the periodic orbits define the topology of the phase space, and in this way we find all the stable regions, close to the stable periodic orbits, where a planetary system can be trapped, and we also find the chaotic regions, close to the unstable periodic orbits, where planetary system could not exist [19].

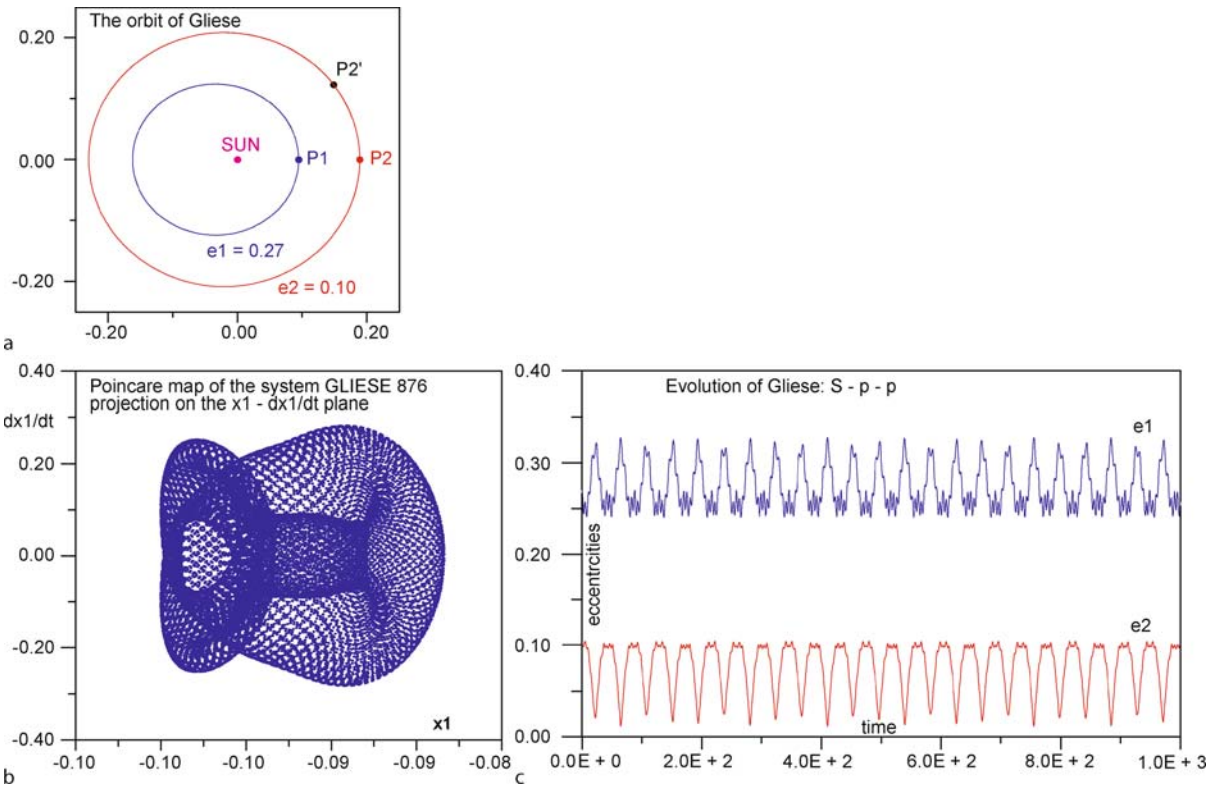
The ordered and chaotic regions in an extrasolar planetary system, the factors that affect the stability and the mechanism of generation of chaos, will be presented here by an example from a real extrasolar planetary system, Gliese 876 [31]. This is a planetary system 15.4 light years far from our solar system. The mass of the sun in this system is equal to $m_0 = 0.32$ solar masses and the masses of the planets P_1, P_2 are $m_1 \sin i = 1.89\,\mathrm{MJ}$ and $m_2 \sin i = 0.56\,\mathrm{MJ}$, where MJ stands for the mass of



Orbital Dynamics, Chaos in, Figure 15
The families of periodic orbits at the 2/1 resonance in the space of the eccentricities. $e_i > 0$ means position at aphelion and $e_i < 0$ position at perihelion

Jupiter (i is the inclination of the orbital plane of this system with respect to the line of sight from us, and it is not known). The semimajor axes, the eccentricities and the periods of the planetary orbits are: $a_1 = 0.13$ AU, $a_2 = 0.21$ AU, $e_1 = 0.27$, $e_2 = 0.10$, $T_1 = 30.1$ days and $T_2 = 61.02$ days. The perihelia of the two planetary orbits are in the same direction. This is a system very close to the 2/1 resonance, $T_2/T_1 = 2.03$, and for this reason we study all the families of resonant periodic orbits at the 2/1 resonance, for the masses of this system (assuming $\sin i = 1$).

In Fig. 15 we present the families of resonant 2/1 periodic orbits for the masses of Gliese 876, in the space of the planetary eccentricities e_1 e_2. We used the convention that $e_i > 0$ means position of the planet at aphelion and $e_i < 0$ position at perihelion. In this way the space of the eccentricities is divided into four sections, according to the sign of the eccentricities, as shown in Fig. 15. For $e_1 < 0$, $e_2 < 0$ and $e_1 > 0$, $e_2 > 0$ the perihelia of both planets are in the same direction, while for $e_1 > 0$, $e_2 < 0$ and $e_1 < 0$, $e_2 > 0$ the perihelia are in opposite directions. We may also note that due to the 2/1 resonance, the phases where, for the

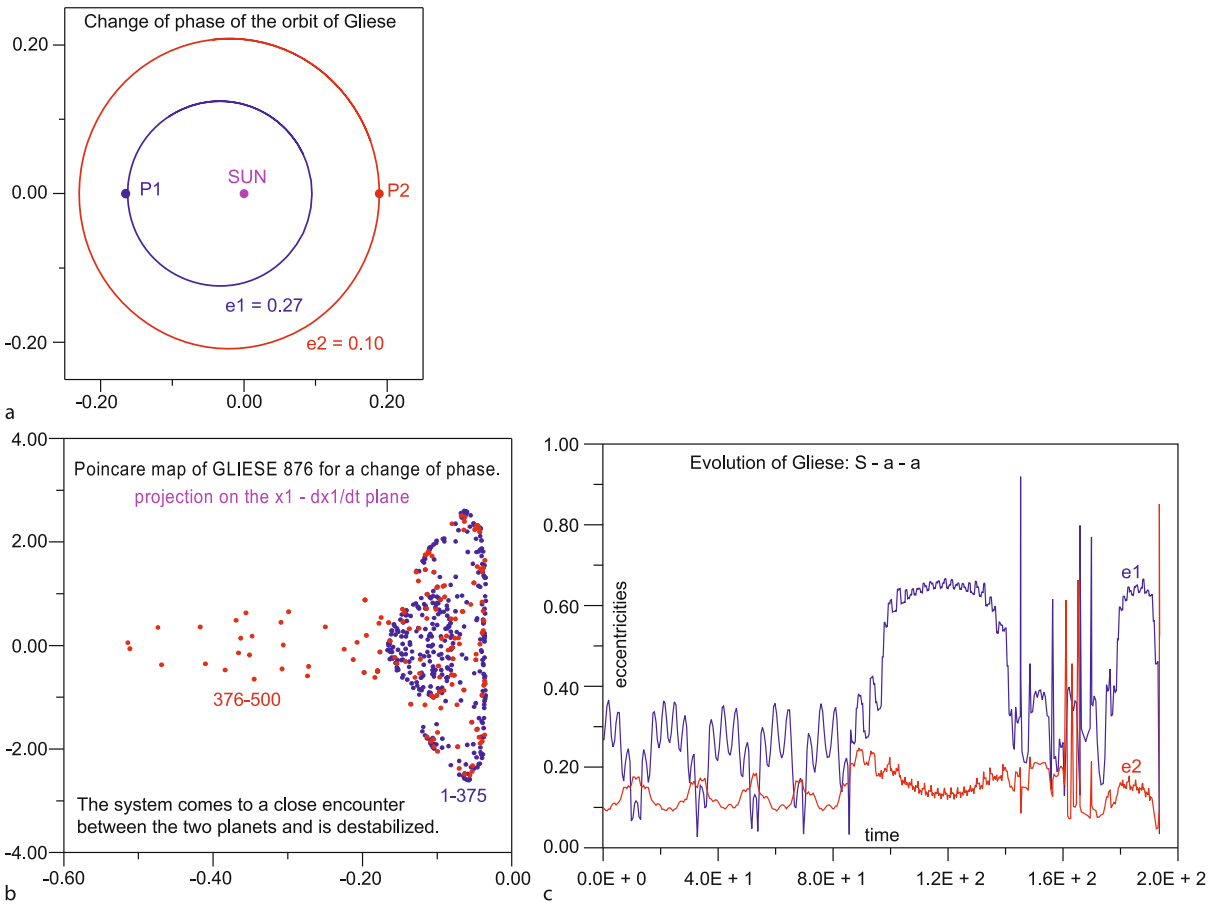


Orbital Dynamics, Chaos in, Figure 16
a The orbit, corresponding to $e_1 < 0$, $e_2 < 0$. b The Poincaré map: projection on the line $x_1\dot{x}_1$. The motion is ordered. b The evolution of the eccentricities.

same position of P_1, the position of P_2 is at perihelion or in aphelion are equivalent, corresponding to $t = 0$ and to $t = T/2$, respectively, where T is the period.

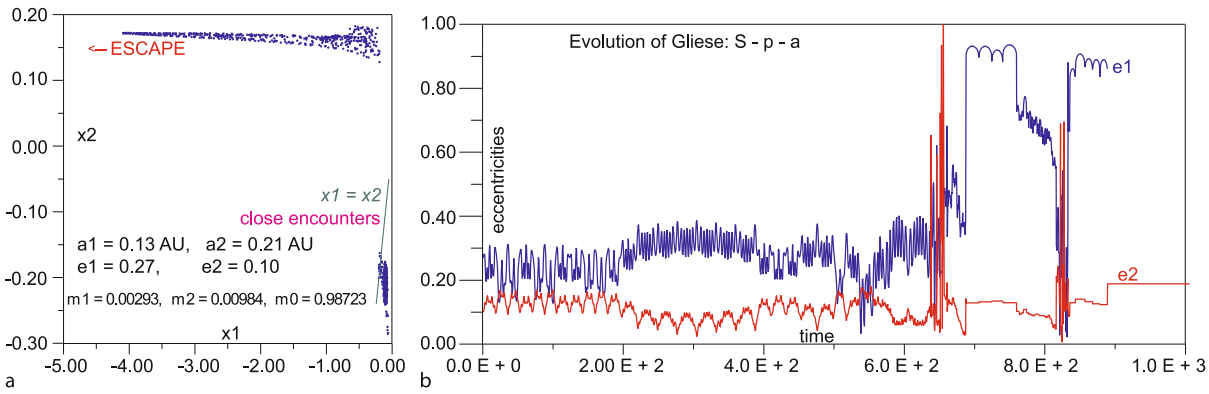
There are two families that start from the region $e_1 \approx 0$, $e_2 \approx 0$. At $e_1 = e_2 = 0$ there is a gap, similar to the gap on the family of circular orbits of the restricted three body problem, as shown in Fig. 6. The first family corresponds to $e_1 < 0$, $e_2 < 0$, along which the eccentricities of the two planets increase. In all orbits of this family the perihelia are in the same direction and at $t = 0$ both planets are at perihelia. This family is stable, even for large values of the orbital eccentricities. Another family exists, for $e_1 > 0$ and $e_2 < 0$. In this family the perihelia of the two planets are in opposite directions and at $t = 0$ the planet P_1 is at aphelion and the planet P_2 is at perihelion. This family presents a gap at the region $e_1 = -0.2$, $e_2 = 0.4$, because the two planets are close to each other and the gravitational attraction between them is so strong (for the given masses) that a resonant 2/1 orbit cannot survive. This part of the family, from zero eccentricities up to the gap, which corresponds to small eccentricities, is unstable. But after this close approach region, the family continues with large eccentricities, and this part is now stable.

In the space of the eccentricities of Fig. 15 we placed a planetary system with the same semimajor axes and eccentricities as Gliese 876, but with different phases. One of these positions, for $e_1 < 0$, $e_2 < 0$, is very close to the stable family. In Figs. 16, 17, and 18 we present the evolution of each of these systems (with the same elements a_i, e_i as Gliese 876), by making use of the Poinaré map on the surface of section defined in Fig. 14b. We note that the real system (green circle in Fig. 15) is in an ordered region (Fig. 16), with the eccentricities undergoing quasi periodic variations, and the projection of the Poincaré map on the



Orbital Dynamics, Chaos in, Figure 17

a The orbit, corresponding to $e_1 > 0$, $e_2 > 0$. b The Poincaré map: projection on the plane $x_1\dot{x}_1$. The motion is chaotic. c The evolution of the eccentricities



Orbital Dynamics, Chaos in, Figure 18

The orbit, corresponding to $e_1 < 0, e_2 > 0$. a The Poincaré map: projection on the plane $x_1\, x_2$. There exist points close to the $x_2 = x_1$ line, corresponding to close encounters between the two planets. The motion is chaotic. b The evolution of the eccentricities

$x_1\dot{x}_1$ plane is a nice surface (the same holds for the projection in all other planes of the four dimensional phase space of the Poincaré map). All other configurations however are unstable and present chaotic behavior, although the orbital elements are the same and only the phase differs. In Fig. 17, corresponding to $e_1 > 0$, $e_2 > 0$, chaotic behavior develops after a rather ordered motion, and the system disrupts. In Fig. 18, corresponding to $e_1 < 0$, $e_2 > 0$, also chaotic behavior develops after a long time of rather ordered motion. From the Poincaré map, which is given in its projection in the $x_1\, x_2$ space, we see that the mechanism of generation of chaos is the close encounters between the two planets, shown by the several points of intersection close to the line $x_2 = x_1$ (this is a real encounter and not just due to the projection from the four dimensional space $x_1\dot{x}_1\, x_2\dot{x}_2$ to the two dimensional plane $x_1\, x_2$, because P_1 is always on the x-axis and P_2 is also on the x-axis, which implies $y_2 = 0$, due to the definition of the map (see Fig. 14b)).

From the above we see that the phase of the two planets (perihelia in the same or in opposite directions, position of the planet at perihelion or aphelion at $t = 0$) plays a crucial role on the stability of the system. As we have seen, the stable regions are close to the stable periodic orbits, and this makes clear the importance of knowing all the families of periodic orbits. In this way we are in a position to know in what regions of the orbital elements a planetary system could exist in nature and what are the regions where a planetary system cannot exist. We note also that the orbital elements for Gliese 876 that we used in the above study were revised, as more accurate observations were taken into account. The new values correspond to a position almost *on* the stable family, as we show in Fig. 15 (yellow circle).

Future Directions

The model of the restricted three-body problem has been studied for almost a century and most of its dynamical aspects are now known. This is not so for the general, planetary, three-body problem, where several aspects of the dynamics are not yet well studied. One reason is that the phase space has more dimensions than the restricted problem. Though the motion in the plane is quite well understood, because all the basic resonant and non resonant periodic orbits (symmetric and asymmetric) are well known, the three dimensional motion is not completely studied. The main reason for this is that the observational data for the extrasolar systems are not yet accurate enough to give information on three dimensional planetary motion. The knowledge of the basic three dimensional families will give a clear picture of the topology of the phase space and of the regions where a three dimensional planetary system could be trapped.

Another problem in the study of the extrasolar planetary systems is the explanation of large planetary eccentricities. Evidently, such systems are stable, since they are observed in nature, and we know from the studies up to now that such high eccentricity planetary systems can be stable, provided we have the right phase. But how did these systems reach their present configuration? It has been proposed that they were generated as low eccentricity systems and reached the present configuration following a *migration* process. A dissipation is needed for such an evolution and several mechanisms have been proposed. It is possible that a planetary system can trapped in a stable configuration, possibly with high eccentricities, due to the migration process. It is the stable periodic orbits that correspond to

these stable configurations. Some work has been done on this problem [36], but more work is needed.

An open problem in orbital dynamics is the study of the early history of our Solar System. This study involves calculations of the N-body problem. It is believed that the orbits of the giant planets of our Solar System, from Jupiter and beyond, migrated due to the planetesimals which were left after the dispersal of the gas disk, in which the Solar System was formed. The idea is that the giant planets ejected the planetesimals and this resulted to a change of their orbits. Recent studies by Tsiganis et al. [50], Gomes et al. [9] and Morbidelli et al. [40] suggest that all outer planets started in a different configuration than the present one, with Jupiter slightly further from the Sun than its present distance, while the rest giant planets were in a distance less than 15 AU from the Sun. This is the so called *Nice model*, from the observatory of Nice where this group works. It is assumed that the planets were surrounded by a disk of planetesimals, which were ejected by the planets, and this resulted to a migration of their orbits. The assumption was made that Saturn was initially inside the 2/1 resonance with Jupiter, and as Saturn crossed this resonance, the eccentricity of the planets increased very much and the planets entered the outer planetesimal disk. This resulted to a heavy scattering of the planetesimals, which reached the inner Solar System and are responsible for the *Late Heavy Bombardment* on the surface of the Moon, which created its craters. More work is still to be done on this problem, including the effect of the giant planets on the orbits of the inner Solar System.

It has been realized recently that very small nonconservative forces, as the effect from mass loss of the sun, or the effects from theory of general relativity, must be included in the study of the past history or the long term evolution of the solar system, for billions of years. This is important for the study of the evolution of the inner planets and most notably of Mercury. Work has now started on this matter, and it is expected to give interesting results.

Bibliography

Primary Literature

1. Arnold VA, Avez A (1968) Ergodic Problems of Classical Mechanics. WA Benjamin, New York
2. Beaugé C, Michtchenko T (2003) Modelling the high-eccentricity planetary three-body problem. Application to the GJ876 planetary system. MNRAS 341:760
3. Beaugé C, Ferraz-Mello S, Michtchenko T (2003) Extrasolar Planets in Mean-Motion Resonance, Apses Alignment and Asymmetric Stationary Solutions. Ap J 593:1124
4. Beaugé C, Callegari N, Ferraz-Mello S, Michtchenko T (2005) Resonances and stability of extra-solar planetary systems. In: Knezevic Z, Milani A (eds) Dynamics of Populations of Planetary Systems. Cambridge Univ Press, p 3
5. Beaugé C, Ferraz-Mello S, Michtchenko TA (2006) Planetary Migration and Extrasolar Planets in the 2/1 Mean-motion Resonance. MNRAS 365:1160–1170
6. Celletti A, Kotoulas T, Voyatzis G, Hadjidemetriou JD (2007) The dynamical stability of a Kuiper belt-like region. MNRAS 378:1153–1164
7. Froeschlé C (1991) Modelling: An aim and a tool for the study of the chaotic behaviour of asteroidal and cometary orbits. In: Roy AE (ed) Predictability, Stability and Chaos in N-Body Dynamical Systems. Plenum Press, pp 125–155
8. Ferraz-Mello S, Beaugé C, Michtchenko T (2003) Evolution of migrating planet pairs in resonance. Cel Mech Dyn Astr 87:99–112
9. Gomes R, Levison HF, Tsiganis K, Morbidelli A (2005) Origin of the Cataclysmic Late Heavy Bombardment period of the terrestial planets. Nature 435:466
10. Gozdziewski K, Bois E, Maciejewski A (2002) Global dynamics of the Gliese 876 planetary system. MNRAS 332:839
11. Hadjidemetriou JD (1975) The continuation of periodic orbits from the restricted to the general three-body problem. Cel Mech 12:155–174
12. Hadjidemetriou JD (1976) Families of Periodic Planetary Type Orbits in the Three-Body Problem and their Stability. Astrophys Sp Sci 40:201–224
13. Hadjidemetriou JD (1982) On the Relation between Resonance and Instability in Planetary Systems. Cel Mech Dyn Astr 27:305–322
14. Hadjidemetriou JD (1992) The Elliptic Restricted Problem at the 3:1 Resonance. Celest Mech 53:151–183
15. Hadjidemetriou JD (1993) Asteroid Motion near the 3:1 Resonance. Cel Mec Dyn Astr 56:563–599
16. Hadjidemetriou JD (1998) Symplectic Maps and their use in Celestial Mechanics. In: Benest D, Froeschlé C (eds) Analysis and Modeling of Discrete Dynamical Systems, ch 9. Gordon and Breach Publ, pp 249–282
17. Hadjidemetriou JD (1999) A symplectic mapping model as a tool to understand the dynamics of the 2/1 resonant asteroid motion. Cel Mec Dyn Astr 73:65–76
18. Hadjidemetriou JD (2002) Resonant periodic motion and the stability of extrasolar planetary systems. Cel Mech Dyn Astr 83:141–154
19. Hadjidemetriou JD (2006) Symmetric and Asymmetric Librations in Extrasolar Planetary Systems: A global view. Cel Mech Dyn Astron 95:225–244
20. Hadjidemetriou JD (2006) Periodic orbits in gravitational systems. In: Steves BA et al (eds) Chaotic Worlds: From Order to Disorder in Gravitational N-Body Dynamical Systems. Springer, pp 43–79
21. Hadjidemetriou JD, Voyatzis G (2000) The 2/1 and 3/2 resonant asteroid motion: A symplectic mapping approach. Cel Mech Dyn Astron 78:137–150
22. Henrard J, Watanabe N, Moons M (1995) A bridge between Secondary and Secular Resonances inside the Hecuba Gap. Icarus 115:336–346
23. Israelian G, Santos N, Mayor M, Rebolo R (2001) Evidence for planet engulfment by the star HD82943. Nature 411:163
24. Jordan DW, Smith P (1988) Nonlinear Ordinary Differential Equations. Clarendon Press, Oxford

25. Laskar J (1988) Secular evolution of the Solar System over 10 million years. Astron Astrophys 198:341–362
26. Laskar J (1989) A numerical experiment on the chaotic behaviour of the Solar System. Nature 338:237–238
27. Laskar J (1994) Large scale chaos in the Solar System. Aston Astrophys 287:l9–12
28. Lee MH (2004) Diversity and Origin of 2:1 Orbital Resonance in Extrasolar Planetary Systems. Ap J 611:517
29. Lee MH, Peale S (2002) Dynamic and origin of the 2:1 orbital resonances of the GJ 876 planets. Ap J 567:596–609
30. Malhotra R (2002) A dynamical mechanism for establishing apsidal resonance. Ap J 575:L33–36
31. Marcy G, Butler P, Fischer D, Vogt S, Lissauer J, Rivera E (2001) Pair A of Resonant Planets Orbiting GJ 876. Ap J 556:296
32. Marcy GW, Butler RP, Fischer DA, Laughlin G, Vogt SS, Henry GW, Pourbaix D (2002) A planet at 5AU around 55Cnc. Ap J 581:1375–1388
33. Mayor M, Udry S, Naef D, Pepe F, Queloz D, Santos NC, Burnet M (2004) The CORALIE survey for southern extra-solar planets. XII Orbital solutions for 16 extra-solar planets discovered with CORALIE. A& A 415:291
34. Meyer KR, Hall GR (1992) Introduction to Hamiltonian Dynamical Systems and the N-Body Problem. Springer
35. Michtchenko TA, Ferraz-Mello S (1996) Comparative study of the asteroidal motion in the 3:2 and 2:1 resonances with Jupiter. I planar model. Astron Astrophys 310:1021–1035
36. Michtchenko TA, Beaugé C, Ferraz-Mello S (2006) Stationary Orbits in Resonant Extrasolar Planetary Systems. Cel Mech Dyn Astron 94:381–397
37. Morbidelli A (1996) On the Kirkwood Gap at the 2/1 Commensurability with Jupiter: numerical results. Astron J 111:2453–2461
38. Morbidelli A, Giorgilli A (1990) On the Dynamics in the Asteroids belt. Part I: General Theory. Celest Mech 47:145–172
39. Morbidelli A, Giorgilli A (1990) On the Dynamics in the Asteroids belt. Part II: Detailed Study of the Main Resnances. Celest Mech 47:173–1204
40. Morbidelli A, Levison HF, Tsiganis K, Gomes R (2005) Chaotic capture of Jupiter's Trojan Asteroids in the early Solar System. Nature 435:462
41. Murray CD, Dermott SF (1999) Solar System Dynamics. Cambridge University Press
42. Psychoyos D, Hadjidemetriou JD (2005) Dynamics of 2/1 resonant extrasolar systems. Application to HD82943 and Gliese 876. Cel Mech Dyn Astr 92:135–156
43. Rivera EJ, Lissauer JJ (2001) Dynamical models of the resonant pair of planets orbiting the star GJ 876. Ap J 558:392–402
44. Roy AE (1982) Orbital Motion, 2nd edn. Adam Hilder
45. Sandor ZS, Suli A, Erdi B, Pilat-Lohinger E, Dvorak R (2007) A stability catalogue of the habitable zones in extrasolar planetary systems. MNRAS 375:1495–1502
46. Szebehely V (1967) Theory of Orbits. Academic Press
47. Tsiganis K, Varvoglis H, Hadjidemetriou JD (2000) Stable chaos in the 12:7 mean motion resonance and its relation to the stickiness effect. Icarus 146:240–252
48. Tsiganis K, Varvoglis H, Hadjidemetriou JD (2002) Stable chaos in Higher-order Jovian Resonances. Icarus 155:454–474
49. Tsiganis K, Varvoglis H, Hadjidemetriou JD (2002) Stable chaos versus Kirkwood Gaps in the Asteroid Belt: A comparative Study of mean Motion Resonances. Icarus 159:284–299
50. Tsiganis K, Gomes R, Morbidelli K, Levison HF (2005) Origin of the orbital architecture of the Giant Planets of the Solar System. Nature 435:459
51. Voyatzis G, Hadjidemetriou HD (2005) Symmetric and asymmetric librations in planetary and satellite systems at the 2/1 resonance. Cel Mech Dyn Astr 93:263–294
52. Voyatzis G, Hadjidemetriou HD (2006) Symmetric and asymmetric 3:1 resonant periodic orbits. An application to the 55Cnc extra-solar system. Cel Mech Dyn Astr 95:259–271
53. Wisdom J (1982) The Origin of the Kirkwood gaps. Astron J 87:577–593
54. Wisdom J (1983) Chaotic behaviour and the Origin of the Kirkwood gaps. Icarus 56:51–74
55. Wisdom J (1985) A Perturbative Treatment of Motion Near the 3/1 Commensurability. Icarus 63:272–289
56. Wisdom J (1987) Urey Prize lecture. Chaotic Dynamics in the Solar System. Icarus 72:241–275
57. Wisdom J, Peale SJ, Mignard F (1984) The chaotic rotation of Hyperion. Icarus 58:137–152
58. Yakubovich VA, Starzhinskii VM (1975) Linear Differential Equations with Periodic Coefficients, vol 1,2. Halsted Press

Books and Reviews

Contopoulos G (2002) Order and Chaos in Dynamical Systems. Springer
Dvorak R, Freistetter F, Kurths J (eds) (2005) Chaos and Stability in Planetary Systems. Lecture Notes in Physics. Springer
Ferraz-Mello S (2007) Canonical Perturbation Theories. Degenerate Systems and resonance. Springer
Hagihara Y (1970–1972) Celestial mechanics vol 1, 2(I), 2(II). MIT Press, Cambridge
Hagihara Y (1974–1976) Celestial mechanics, vol 3(I), 3(II), 4(I), 4(II), 5(I), 5(II). Japan Society For the Promotion of Science, Tokyo