

H

Hamiltonian Perturbation Theory (and Transition to Chaos)

HENK W. BROER¹, HEINZ HANSSMANN²

¹ Instituut voor Wiskunde en Informatica,
Rijksuniversiteit Groningen, Groningen,
The Netherlands

² Mathematisch Instituut, Universiteit Utrecht,
Utrecht, The Netherlands

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[One Degree of Freedom](#)

[Perturbations of Periodic Orbits](#)

[Invariant Curves of Planar Diffeomorphisms](#)

[KAM Theory: An Overview](#)

[Splitting of Separatrices](#)

[Transition to Chaos and Turbulence](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Bifurcation In parametrized dynamical systems a bifurcation occurs when a qualitative change is invoked by a change of parameters. In models such a qualitative change corresponds to transition between dynamical regimes. In the generic theory a finite list of cases is obtained, containing elements like ‘saddle-node’, ‘period doubling’, ‘Hopf bifurcation’ and many others.

Cantor set, Cantor dust, Cantor family, Cantor stratification Cantor dust is a separable locally compact space that is perfect, i.e. every point is in the closure of its complement, and totally disconnected. This determines Cantor dust up to homeomorphisms. The term Cantor set (originally reserved for the specific form of Cantor dust obtained by repeatedly deleting the mid-

dle third from a closed interval) designates topological spaces that locally have the structure $\mathbb{R}^n \times \text{Cantor dust}$ for some $n \in \mathbb{N}$. Cantor families are parametrized by such Cantor sets. On the real line \mathbb{R} one can define Cantor dust of positive measure by excluding around each rational number p/q an interval of size

$$\frac{2\gamma}{q^\tau}, \quad \gamma > 0, \quad \tau > 2.$$

Similar Diophantine conditions define Cantor sets in \mathbb{R}^n . Since these Cantor sets have positive measure their Hausdorff dimension is n . Where the unperturbed system is stratified according to the co-dimension of occurring (bifurcating) tori, this leads to a Cantor stratification.

Chaos An evolution of a dynamical system is chaotic if its future is badly predictable from its past. Examples of non-chaotic evolutions are periodic or multi-periodic. A system is called chaotic when many of its evolutions are. One criterion for chaoticity is the fact that one of the Lyapunov exponents is positive.

Diophantine condition, Diophantine frequency vector

A frequency vector $\omega \in \mathbb{R}^n$ is called Diophantine if there are constants $\gamma > 0$ and $\tau > n - 1$ with

$$|\langle k, \omega \rangle| \geq \frac{\gamma}{|k|^\tau} \quad \text{for all } k \in \mathbb{Z}^n \setminus \{0\}.$$

The Diophantine frequency vectors satisfying this condition for fixed γ and τ form a Cantor set of half lines. As the Diophantine parameter γ tends to zero (while τ remains fixed), these half lines extend to the origin. The complement in any compact set of frequency vectors satisfying a Diophantine condition with fixed τ has a measure of order $O(\gamma)$ as $\gamma \downarrow 0$.

Integrable system A Hamiltonian system with n degrees of freedom is (Liouville)-integrable if it has n functionally independent commuting integrals of motion. Locally this implies the existence of a torus action, a feature that can be generalized to dissipative sys-

tems. In particular a mapping is integrable if it can be interpolated to become the stroboscopic mapping of a flow.

KAM theory Kolmogorov–Arnold–Moser theory is the perturbation theory of (Diophantine) quasi-periodic tori for nearly integrable Hamiltonian systems. In the format of quasi-periodic stability, the unperturbed and perturbed system, restricted to a Diophantine Cantor set, are smoothly conjugated in the sense of Whitney. This theory extends to the world of reversible, volume-preserving or general dissipative systems. In the latter KAM theory gives rise to families of quasi-periodic attractors. KAM theory also applies to torus bundles, in which case a global Whitney smooth conjugation can be proven to exist, that keeps track of the geometry. In an appropriate sense invariants like monodromy and Chern classes thus also can be defined in the nearly integrable case. Also compare with ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#).

Nearly integrable system In the setting of perturbation theory, a nearly integrable system is a perturbation of an integrable one. The latter then is an integrable approximation of the former. See an above item.

Normal form truncation Consider a dynamical system in the neighborhood of an equilibrium point, a fixed or periodic point, or a quasi-periodic torus, reducible to Floquet form. Then Taylor expansions (and their analogues) can be changed gradually into normal forms, that usually reflect the dynamics better. Often these display a (formal) torus symmetry, such that the normal form truncation becomes an integrable approximation, thus yielding a perturbation theory setting. See above items. Also compare with ► [Normal Forms in Perturbation Theory](#).

Persistent property In the setting of perturbation theory, a property is persistent whenever it is inherited from the unperturbed to the perturbed system. Often the perturbation is taken in an appropriate topology on the space of systems, like the Whitney C^k -topology [72].

Perturbation problem In perturbation theory the unperturbed systems usually are transparent regarding their dynamics. Examples are integrable systems or normal form truncations. In a perturbation problem things are arranged in such a way that the original system is well-approximated by such an unperturbed one. This arrangement usually involves both changes of variables and scalings.

Resonance If the frequencies of an invariant torus with multi- or conditionally periodic flow are rationally dependent, this torus divides into invariant sub-tori. Such resonances $\langle h, \omega \rangle = 0$, $h \in \mathbb{Z}^k$, define hyper-

planes in ω -space and, by means of the frequency mapping, also in phase space. The smallest number $|h| = |h_1| + \dots + |h_k|$ is the order of the resonance. Diophantine conditions describe a measure-theoretically large complement of a neighborhood of the (dense!) set of all resonances.

Separatrices Consider a hyperbolic equilibrium, fixed or periodic point or invariant torus. If the stable and unstable manifolds of such hyperbolic elements are codimension one immersed manifolds, then they are called separatrices, since they separate domains of phase space, for instance, basins of attraction.

Singularity theory A function $H: \mathbb{R}^n \rightarrow \mathbb{R}$ has a critical point $z \in \mathbb{R}^n$ where $DH(z)$ vanishes. In local coordinates we may arrange $z = 0$ (and similarly that it is mapped to zero as well). Two germs $K: (\mathbb{R}^n, 0) \rightarrow (\mathbb{R}, 0)$ and $N: (\mathbb{R}^n, 0) \rightarrow (\mathbb{R}, 0)$ represent the same function H locally around z if and only if there is a diffeomorphism η on \mathbb{R}^n satisfying

$$N = K \circ \eta.$$

The corresponding equivalence class is called a singularity.

Structurally stable A system is structurally stable if it is topologically equivalent to all nearby systems, where ‘nearby’ is measured in an appropriate topology on the space of systems, like the Whitney C^k -topology [72]. A family is structurally stable if for every nearby family there is a re-parametrization such that all corresponding systems are topologically equivalent.

Definition of the Subject

The fundamental problem of mechanics is to study Hamiltonian systems that are small perturbations of integrable systems. Also, perturbations that destroy the Hamiltonian character are important, be it to study the effect of a small amount of friction, or to further the theory of dissipative systems themselves which surprisingly often revolves around certain well-chosen Hamiltonian systems. Furthermore there are approaches like KAM theory that historically were first applied to Hamiltonian systems. Typically perturbation theory explains only part of the dynamics, and in the resulting gaps the orderly unperturbed motion is replaced by random or chaotic motion.

Introduction

We outline perturbation theory from a general point of view, illustrated by a few examples.

The Perturbation Problem

The aim of perturbation theory is to approximate a given dynamical system by a more familiar one, regarding the former as a perturbation of the latter. The problem then is to deduce certain dynamical properties from the unperturbed to the perturbed case.

What is familiar may or may not be a matter of taste, at least it depends a lot on the dynamical properties of one's interest. Still the most frequently used unperturbed systems are:

- Linear systems
- Integrable Hamiltonian systems, compare with ► [Dynamics of Hamiltonian Systems](#) and references therein
- Normal form truncations, compare with ► [Normal Forms in Perturbation Theory](#) and references therein
- Etc.

To some extent the second category can be seen as a special case of the third. To avoid technicalities in this section we assume all systems to be sufficiently smooth, say of class C^∞ or real analytic. Moreover in our considerations ε will be a real parameter. The unperturbed case always corresponds to $\varepsilon = 0$ and the perturbed one to $\varepsilon \neq 0$ or $\varepsilon > 0$.

Examples of Perturbation Problems To begin with consider the autonomous differential equation

$$\ddot{x} + \varepsilon \dot{x} + \frac{dV}{dx}(x) = 0,$$

modeling an oscillator with small damping. Rewriting this equation of motion as a planar vector field

$$\begin{aligned}\dot{x} &= y \\ \dot{y} &= -\varepsilon y - \frac{dV}{dx}(x),\end{aligned}$$

we consider the energy $H(x, y) = \frac{1}{2}y^2 + V(x)$. For $\varepsilon = 0$ the system is Hamiltonian with Hamiltonian function H . Indeed, generally we have $\dot{H}(x, y) = -\varepsilon y^2$, implying that for $\varepsilon > 0$ there is dissipation of energy. Evidently for $\varepsilon \neq 0$ the system is no longer Hamiltonian.

The reader is invited to compare the phase portraits of the cases $\varepsilon = 0$ and $\varepsilon > 0$ for $V(x) = -\cos x$ (the pendulum) or $V(x) = \frac{1}{2}\lambda x^2 + \frac{1}{24}bx^4$ (Duffing).

Another type of example is provided by the non-autonomous equation

$$\ddot{x} + \frac{dV}{dx}(x) = \varepsilon f(x, \dot{x}, t),$$

which can be regarded as the equation of motion of an oscillator with small external forcing. Again rewriting as

a vector field, we obtain

$$\begin{aligned}\dot{t} &= 1 \\ \dot{x} &= y \\ \dot{y} &= -\frac{dV}{dx}(x) + \varepsilon f(x, y, t),\end{aligned}$$

now on the generalized phase space $\mathbb{R}^3 = \{t, x, y\}$. In the case where the t -dependence is periodic, we can take $\mathbb{S}^1 \times \mathbb{R}^2$ for (generalized) phase space.

Remark

- A small variation of the above driven system concerns a parametrically forced oscillator like

$$\ddot{x} + (\omega^2 + \varepsilon \cos t) \sin x = 0,$$

which happens to be entirely in the world of Hamiltonian systems.

- It may be useful to study the Poincaré or period mapping of such time periodic systems, which happens to be a mapping of the plane. We recall that in the Hamiltonian cases this mapping preserves area. For general reference in this direction see, e. g., [6,7,27,66].

There are lots of variations and generalizations. One example is the solar system, where the unperturbed case consists of a number of uncoupled two-body problems concerning the Sun and each of the planets, and where the interaction between the planets is considered as small [6,9,107,108].

Remark

- One variation is a restriction to fewer bodies, for example only three. Examples of this are systems like Sun–Jupiter–Saturn, Earth–Moon–Sun or Earth–Moon–Satellite.
- Often Sun, Moon and planets are considered as point masses, in which case the dynamics usually are modeled as a Hamiltonian system. It is also possible to extend this approach taking tidal effects into account, which have a non-conservative nature.
- The Solar System is close to resonance, which makes application of KAM theory problematic. There exist, however, other integrable approximations that take resonance into account [3,63].

Quite another perturbation setting is local, e. g., near an equilibrium point. To fix thoughts consider

$$\dot{x} = Ax + f(x), \quad x \in \mathbb{R}^n$$

with $A \in \text{gl}(n, \mathbb{R})$, $f(0) = 0$ and $D_x f(0) = 0$. By the scaling $x = \varepsilon \bar{x}$ we rewrite the system to

$$\dot{\bar{x}} = A\bar{x} + \varepsilon g(\bar{x}).$$

So, here we take the linear part as an unperturbed system. Observe that for small ε the perturbation is small on a compact neighborhood of $\bar{x} = 0$.

This setting also has many variations. In fact, any normal form approximation may be treated in this way ► [Normal Forms in Perturbation Theory](#). Then the normalized truncation forms the unperturbed part and the higher order terms the perturbation.

Remark In the above we took the classical viewpoint which involves a perturbation parameter controlling the size of the perturbation. Often one can generalize this by considering a suitable topology (like the Whitney topologies) on the corresponding class of systems [72]. Also compare with ► [Normal Forms in Perturbation Theory](#), ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#) and ► [Dynamics of Hamiltonian Systems](#).

Questions of Persistence

What are the kind of questions perturbation theory asks? A large class of questions concerns the *persistence* of certain dynamical properties as known for the unperturbed case. To fix thoughts we give a few examples.

To begin with consider equilibria and periodic orbits. So we put

$$\dot{x} = f(x, \varepsilon), \quad x \in \mathbb{R}^n, \quad \varepsilon \in \mathbb{R}, \quad (1)$$

for a map $f: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$. Recall that equilibria are given by the equation $f(x, \varepsilon) = 0$. The following theorem that continues equilibria in the unperturbed system for $\varepsilon \neq 0$, is a direct consequence of the implicit function theorem.

Theorem 1 (Persistence of equilibria) *Suppose that $f(x_0, 0) = 0$ and that*

$$D_x f(x_0, 0) \text{ has maximal rank.}$$

Then there exists a local arc $\varepsilon \mapsto x(\varepsilon)$ with $x(0) = x_0$ such that

$$f(x(\varepsilon), \varepsilon) \equiv 0.$$

Periodic orbits can be approximated in a similar way. Indeed, let the system (1) for $\varepsilon = 0$ have a periodic orbit γ_0 . Let Σ be a local transversal section of γ_0 and $P_0: \Sigma \rightarrow \Sigma$ the corresponding Poincaré map. Then P_0 has a fixed

point $x_0 \in \Sigma \cap \gamma_0$. By transversality, for $|\varepsilon|$ small, a local Poincaré map $P_\varepsilon: \Sigma \rightarrow \Sigma$ is well-defined for (1). Observe that fixed points x_ε of P_ε correspond to periodic orbits γ_ε of (1). We now have, again as another direct consequence of the implicit function theorem.

Theorem 2 (Persistence of periodic orbits) *In the above assume that*

$$P_0(x_0) = x_0 \text{ and } D_x P_0(x_0) \text{ has no eigenvalue } 1.$$

Then there exists a local arc $\varepsilon \mapsto x(\varepsilon)$ with $x(0) = x_0$ such that

$$P_\varepsilon(x(\varepsilon)) \equiv x_\varepsilon.$$

Remark

- Often the conditions of Theorem 2 are not easy to verify. Sometimes it is useful here to use Floquet Theory, see [97]. In fact, if T_0 is the period of γ_0 and Ω_0 its Floquet matrix, then $D_x P_0(x_0) = \exp(T_0 \Omega_0)$.
- The format of the Theorems 1 and 2 with the perturbation parameter ε directly allows for algorithmic approaches. One way to proceed is by perturbation series, leading to asymptotic formulae that in the real analytic setting have positive radius of convergence. In the latter case the names of Poincaré and Lindstedt are associated with the method, cf. [10]. Also numerical continuation programmes exist based on the Newton method.
- The Theorems 1 and 2 can be seen as special cases of a general theorem for *normally hyperbolic invariant manifolds* [73], Theorem 4.1. In all cases a contraction principle on a suitable Banach space of graphs leads to persistence of the invariant dynamical object.

This method in particular yields existence and persistence of stable and unstable manifolds [53,54].

Another type of dynamics subject to perturbation theory is quasi-periodic. We emphasize that persistence of (Diofantine) quasi-periodic invariant tori occurs both in the conservative setting and in many others, like in the reversible and the general (dissipative) setting. In the latter case this leads to persistent occurrence of families of quasi-periodic attractors [125]. These results are in the domain of Kolmogorov–Arnold–Moser (KAM) theory. For details we refer to Sect. “[KAM Theory: An Overview](#)” below or to [24], ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#), the former reference containing more than 400 references in this area.

Remark

- Concerning the Solar System, KAM theory always has aimed at proving that it contains many quasi-periodic motions, in the sense of positive Liouville measure. This would imply that there is positive probability that a given initial condition lies on such a stable quasi-periodic motion [3,63], however, also see [85].
- Another type of result in this direction compares the distance of certain individual solutions of the perturbed and the unperturbed system, with coinciding initial conditions over time scales that are long in terms of ε . Compare with [24].

Apart from persistence properties related to invariant manifolds or individual solutions, the aim can also be to obtain a more global persistence result. As an example of this we mention the Hartman–Grobman Theorem, e. g., [7,116,123]. Here the setting once more is

$$\dot{x} = Ax + f(x), \quad x \in \mathbb{R}^n,$$

with $A \in \text{gl}(n, \mathbb{R})$, $f(0) = 0$ and $D_x f(0) = 0$. Now we assume A to be hyperbolic (i. e., with no purely imaginary eigenvalues). In that case the full system, near the origin, is topologically conjugated to the linear system $\dot{x} = Ax$. Therefore all global, *qualitative* properties of the unperturbed (linear) system are persistent under perturbation to the full system. For details on these notions see the above references, also compare with, e. g., [30].

It is said that the hyperbolic linear system $\dot{x} = Ax$ is (locally) *structurally stable*. This kind of thinking was introduced to the dynamical systems area by Thom [133], with a first, successful application to catastrophe theory. For further details, see [7,30,69,116].

General Dynamics

We give a few remarks on the general dynamics in a neighborhood of Hamiltonian KAM tori. In particular this concerns so-called superexponential stickiness of the KAM tori and adiabatic stability of the action variables, involving the so-called Nekhoroshev estimate.

To begin with, emphasize the following difference between the cases $n = 2$ and $n \geq 3$ in the classical KAM theorem of Subject. “[Classical KAM Theory](#)”. For $n = 2$ the level surfaces of the Hamiltonian are three-dimensional, while the Lagrangian tori have dimension two and hence codimension one in the energy hypersurfaces. This means that for open sets of initial conditions, the evolution curves are forever trapped in between KAM tori, as these tori foliate over nowhere dense sets of positive measure. This im-

plies perpetual adiabatic stability of the action variables. In contrast, for $n \geq 3$ the Lagrangian tori have codimension $n - 1 > 1$ in the energy hypersurfaces and evolution curves may escape.

This actually occurs in the case of so-called *Arnold diffusion*. The literature on this subject is immense, and we here just quote [5,9,93,109], for many more references see [24].

Next we consider the motion in a neighborhood of the KAM tori, in the case where the systems are real analytic or at least Gevrey smooth. For a definition of Gevrey regularity see [136]. First we mention that, measured in terms of the distance to the KAM torus, nearby evolution curves generically stay nearby over a *superexponentially* long time [102,103]. This property often is referred to as superexponential stickiness of the KAM tori, see [24] for more references.

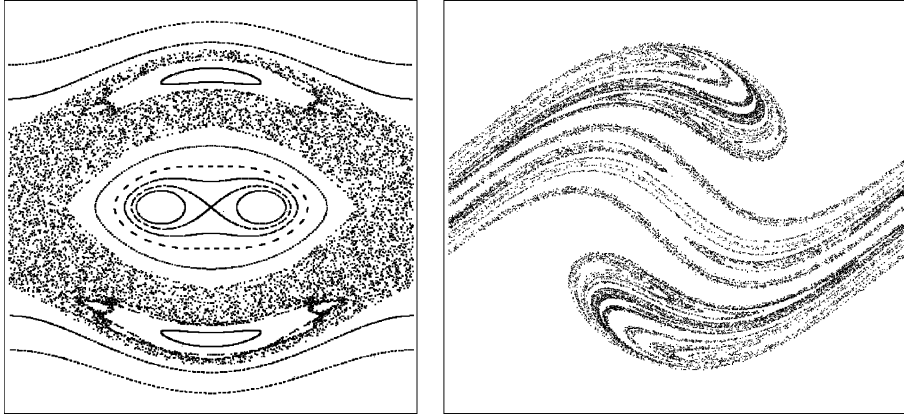
Second, nearly integrable Hamiltonian systems, in terms of the perturbation size, generically exhibit *exponentially* long adiabatic stability of the action variables, see e. g. [15,88,89,90,93,103,109,110,113,120], ► [Nekhoroshev Theory](#) and many others, for more references see [24]. This property is referred to as the *Nekhoroshev estimate* or the *Nekhoroshev theorem*. For related work on perturbations of so-called superintegrable systems, also see [24] and references therein.

Chaos

In the previous subsection we discussed persistent and some non-persistent features of dynamical systems under small perturbations. Here we discuss properties related to splitting of separatrices, caused by generic perturbations.

A first example was met earlier, when comparing the pendulum with and without (small) damping. The unperturbed system is the undamped one and this is a Hamiltonian system. The perturbation however no longer is Hamiltonian. We see that the equilibria are persistent, as they should be according to Theorem 1, but that none of the periodic orbits survives the perturbation. Such qualitative changes go with perturbing away from the Hamiltonian setting.

Similar examples concern the breaking of a certain symmetry by the perturbation. The latter often occurs in the case of normal form approximations. Then the normalized truncation is viewed as the unperturbed system, which is perturbed by the higher order terms. The truncation often displays a reasonable amount of symmetry (e. g., toroidal symmetry), which *generically* is forbidden for the class of systems under consideration, e. g. see [25].



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 1

Chaos in the parametrically forced pendulum. *Left:* Poincaré map $P_{\omega,\varepsilon}$ near the 1 : 2 resonance $\omega = \frac{1}{2}$ and for $\varepsilon > 0$ not too small. *Right:* A dissipative analogue

To fix thoughts we reconsider the conservative example

$$\ddot{x} + (\omega^2 + \varepsilon \cos t) \sin x = 0$$

of the previous section. The corresponding (time dependent, Hamiltonian [6]) vector field reads

$$\begin{aligned} \dot{t} &= 1 \\ \dot{x} &= y \\ \dot{y} &= -(\omega^2 + \varepsilon \cos t) \sin x. \end{aligned}$$

Let $P_{\omega,\varepsilon}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be the corresponding (area-preserving) Poincaré map. Let us consider the unperturbed map $P_{\omega,0}$ which is just the flow over time 2π of the free pendulum $\ddot{x} + \omega^2 \sin x = 0$. Such a map is called *integrable*, since it is the stroboscopic map of a two-dimensional vector field, hence displaying the \mathbb{R} -symmetry of a flow. When perturbed to the *nearly integrable* case $\varepsilon \neq 0$, this symmetry generically is broken. We list a few of the generic properties for such maps [123]:

- The homoclinic and heteroclinic points occur at transversal intersections of the corresponding stable and unstable manifolds.
- The periodic points of period less than a given bound are isolated.

This means generically that the separatrices split and that the resonant invariant circles filled with periodic points with the same (rational) rotation number fall apart. In any concrete example the issue remains whether or not it satisfies appropriate genericity conditions. One method to check this is due to Melnikov, compare [66,137], for

more sophisticated tools see [65]. Often this leads to elliptic (Abelian) integrals.

In nearly integrable systems chaos can occur. This fact is at the heart of the celebrated non-integrability of the three-body problem as addressed by Poincaré [12,59,107,108,118]. A long standing open conjecture is that the clouds of points as visible in Fig. 1, left, densely fill sets of positive area, thereby leading to ergodicity [9].

In the case of dissipation, see Fig. 1, right, we conjecture the occurrence of a Hénon-like strange attractor [14,22,126].

Remark

- The persistent occurrence of periodic points of a given rotation number follows from the Poincaré–Birkhoff fixed point theorem [74,96,107], i. e., on topological grounds.
- The above arguments are not restricted to the conservative setting, although quite a number of unperturbed systems come from this world. Again see Fig. 1.

One Degree of Freedom

Planar Hamiltonian systems are always integrable and the orbits are given by the level sets of the Hamiltonian function. This still leaves room for a perturbation theory. The recurrent dynamics consists of periodic orbits, equilibria and asymptotic trajectories forming the (un)stable manifolds of unstable equilibria. The equilibria organize the phase portrait, and generically all equilibria are elliptic (purely imaginary eigenvalues) or hyperbolic (real eigenvalues), i. e. there is no equilibrium with a vanishing eigenvalue. If the system depends on a parameter such van-

ishing eigenvalues may be unavoidable and it becomes possible that the corresponding dynamics persist under perturbations.

Perturbations may also destroy the Hamiltonian character of the flow. This happens especially where the starting point is a dissipative planar system and e. g. a scaling leads for $\varepsilon = 0$ to a limiting Hamiltonian flow. The perturbation problem then becomes twofold. Equilibria still persist by Theorem 1 and hyperbolic equilibria moreover persist as such, with the sum of eigenvalues of order $\mathcal{O}(\varepsilon)$. Also for elliptic eigenvalues the sum of eigenvalues is of order $\mathcal{O}(\varepsilon)$ after the perturbation, but here this number measures the dissipation whence the equilibrium becomes (weakly) attractive for negative values and (weakly) unstable for positive values. The one-parameter families of periodic orbits of a Hamiltonian system do not persist under dissipative perturbations, the very fact that they form families imposes the corresponding fixed point of the Poincaré mapping to have an eigenvalue one and Theorem 2 does not apply. Typically only finitely many periodic orbits survive a dissipative perturbation and it is already a difficult task to determine their number.

Hamiltonian Perturbations

The Duffing oscillator has the Hamiltonian function

$$H(x, y) = \frac{1}{2}y^2 + \frac{1}{24}bx^4 + \frac{1}{2}\lambda x^2 \quad (2)$$

where b is a constant distinguishing the two cases $b = \pm 1$ and λ is a parameter. Under variation of the parameter the equations of motion

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -\frac{1}{6}bx^3 - \lambda x \end{aligned}$$

display a Hamiltonian pitchfork bifurcation, supercritical for positive b and subcritical in case b is negative. Correspondingly, the linearization at the equilibrium $x = 0$ of the anharmonic oscillator $\lambda = 0$ is given by the matrix

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

whence this equilibrium is parabolic.

The typical way in which a parabolic equilibrium bifurcates is the center-saddle bifurcation. Here the Hamiltonian reads

$$H(x, y) = \frac{1}{2}ay^2 + \frac{1}{6}bx^3 + c\lambda x \quad (3)$$

where $a, b, c \in \mathbb{R}$ are nonzero constants, for instance $a = b = c = 1$. Note that this is a completely different unfolding of the parabolic equilibrium at the origin. A closer look at the phase portraits and in particular at the Hamiltonian function of the Hamiltonian pitchfork bifurcation reveals the symmetry $x \mapsto -x$ of the Duffing oscillator. This suggests the addition of the non-symmetric term μx . The resulting two-parameter family

$$H_{\lambda, \mu}(x, y) = \frac{1}{2}y^2 + \frac{1}{24}bx^4 + \frac{1}{2}\lambda x^2 + \mu x$$

of Hamiltonian systems is indeed structurally stable. This implies not only that all equilibria of a Hamiltonian perturbation of the Duffing oscillator have a local flow equivalent to the local flow near a suitable equilibrium in this two-parameter family, but that every one-parameter family of \mathbb{Z}_2 -symmetric Hamiltonian systems that is a perturbation of (2) has equivalent dynamics. For more details see [36] and references therein.

This approach applies mutatis mutandis to every non-degenerate planar singularity, cf. [69, 130]. At an equilibrium all partial derivatives of the Hamiltonian vanish and the resulting singularity is called non-degenerate if it has finite multiplicity, which implies that it admits a versal unfolding H_λ with finitely many parameters. The family of Hamiltonian systems defined by this versal unfolding contains all possible (local) dynamics that the initial equilibrium may be perturbed to. Imposing additional discrete symmetries is immediate, the necessary symmetric versal unfolding is obtained by averaging

$$H_\lambda^G = \frac{1}{|G|} \sum_{g \in G} H_\lambda \circ g$$

along the orbits of the symmetry group G .

Dissipative Perturbations

In a generic dissipative system all equilibria are hyperbolic. Qualitatively, i. e. up to topological equivalence, the local dynamics is completely determined by the number of eigenvalues with positive real part. Those hyperbolic equilibria that can appear in Hamiltonian systems (the eigenvalues forming pairs $\pm \nu$) do not play an important role. Rather, planar Hamiltonian systems become important as a tool to understand certain bifurcations triggered off by non-hyperbolic equilibria. Again this requires the system to depend on external parameters.

The simplest example is the Hopf bifurcation, a co-dimension one bifurcation where an equilibrium loses stability as the pair of eigenvalues crosses the imaginary axis,

say at $\pm i$. At the bifurcation the linearization is a Hamiltonian system with an elliptic equilibrium (the co-dimension one bifurcations where a single eigenvalue crosses the imaginary axis through 0 do not have a Hamiltonian linearization). This limiting Hamiltonian system has a one-parameter family of periodic orbits around the equilibrium, and the non-linear terms determine the fate of these periodic orbits. The normal form of order three reads

$$\begin{aligned}\dot{x} &= -y(1 + b(x^2 + y^2)) + x(\lambda + a(x^2 + y^2)) \\ \dot{y} &= x(1 + b(x^2 + y^2)) + y(\lambda + a(x^2 + y^2))\end{aligned}$$

and is Hamiltonian if and only if $(\lambda, a) = (0, 0)$. The sign of the coefficient distinguishes between the supercritical case $a > 0$, in which there are no periodic orbits coexisting with the attractive equilibria (i. e. when $\lambda < 0$) and one attracting periodic orbit for each $\lambda > 0$ (coexisting with the unstable equilibrium), and the subcritical case $a < 0$, in which the family of periodic orbits is unstable and coexists with the attractive equilibria (with no periodic orbits for parameters $\lambda > 0$). As $\lambda \rightarrow 0$ the family of periodic orbits shrinks down to the origin, so also this Hamiltonian feature is preserved.

Equilibria with a double eigenvalue 0 need two parameters to persistently occur in families of dissipative systems. The generic case is the Takens–Bogdanov bifurcation. Here the linear part is too degenerate to be helpful, but the nonlinear Hamiltonian system defined by (3) with $a = 1 = c\lambda$ and $b = -3$ provides the periodic and heteroclinic orbit(s) that constitute the nontrivial part of the bifurcation diagram. Where discrete symmetries are present, e. g. for equilibria in dissipative systems originating from other generic bifurcations, the limiting Hamiltonian system exhibits that same discrete symmetry. For more details see [54,66,82] and references therein.

The continuation of certain periodic orbits from an unperturbed Hamiltonian system under dissipative perturbation can be based on Melnikov-like methods, again see [66,137]. As above, this often leads to Abelian integrals, for instance to count the number of periodic orbits that branch off.

Reversible Perturbations

A dynamical system that admits a reflection symmetry R mapping trajectories $\varphi(t, z_0)$ to trajectories $\varphi(-t, R(z_0))$ is called reversible. In the planar case we may restrict to the reversing reflection

$$R: \begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ (x, y) & \mapsto & (x, -y). \end{array} \quad (4)$$

All Hamiltonian functions $H = \frac{1}{2}y^2 + V(x)$ which have an interpretation “kinetic + potential energy” are reversible, and in general the class of reversible systems is positioned between the class of Hamiltonian systems and the class of dissipative systems. A guiding example is the perturbed Duffing oscillator (with the roles of x and y exchanged so that (4) remains the reversing symmetry)

$$\begin{aligned}\dot{x} &= -\frac{1}{6}y^3 - y + \varepsilon xy \\ \dot{y} &= x\end{aligned}$$

that combines the Hamiltonian character of the equilibrium at the origin with the dissipative character of the two other equilibria. Note that all orbits outside the homoclinic loop are periodic.

There are two ways in which the reversing symmetry (4) imposes a Hamiltonian character on the dynamics. An equilibrium that lies on the symmetry line $\{y = 0\}$ has a linearization that is itself a reversible system and consequently the eigenvalues are subject to the same constraints as in the Hamiltonian case. (For equilibria z_0 that do not lie on the symmetry line the reflection $R(z_0)$ is also an equilibrium, and it is to the union of their eigenvalues that these constraints still apply.) Furthermore, every orbit that crosses $\{y = 0\}$ more than once is automatically periodic, and these periodic orbits form one-parameter families. In particular, elliptic equilibria are still surrounded by periodic orbits.

The dissipative character of a reversible system is most obvious for orbits that do not cross the symmetry line. Here R merely maps the orbit to a reflected counterpart. The above perturbed Duffing oscillator exemplifies that the character of an orbit crossing $\{y = 0\}$ exactly once is undetermined. While the homoclinic orbit of the saddle at the origin has a Hamiltonian character, the heteroclinic orbits between the other two equilibria behave like in a dissipative system.

Perturbations of Periodic Orbits

The perturbation of a one-degree-of-freedom system by a periodic forcing is a perturbation that changes the phase space. Treating the time variable t as a phase space variable leads to the extended phase space $\mathbb{S}^1 \times \mathbb{R}^2$ and equilibria of the unperturbed system become periodic orbits, inheriting the normal behavior. Furthermore introducing an action conjugate to the “angle” t yields a Hamiltonian system in two degrees of freedom.

While the one-parameter families of periodic orbits merely provide the typical recurrent motion in one degree of freedom, they form special solutions in two or more de-

degrees of freedom. Arcs of elliptic periodic orbits are particularly instructive. Note that these occur generically in both the Hamiltonian and the reversible context.

Conservative Perturbations

Along the family of elliptic periodic orbits a pair $e^{\pm i\Omega}$ of Floquet multipliers passes regularly through roots of unity. Generically this happens on a dense set of parameter values, but for fixed denominator q in $e^{\pm i\Omega} = e^{\pm 2\pi i p/q}$ the corresponding energy values are isolated. The most important of such resonances are those with small denominators q .

For $q = 1$ generically a periodic center-saddle bifurcation takes place where an elliptic and a hyperbolic periodic orbit meet at a parabolic periodic orbit. No periodic orbit remains under further variation of a suitable parameter.

The generic bifurcation for $q = 2$ is the period-doubling bifurcation where an elliptic periodic orbit turns hyperbolic (or vice versa) when passing through a parabolic periodic orbit with Floquet multipliers -1 . Furthermore, a family of periodic orbits with twice the period emerges from the parabolic periodic orbit, inheriting the normal linear behavior from the initial periodic orbit.

In case $q = 3$, and possibly also for $q = 4$, generically two arcs of hyperbolic periodic orbits emerge, both with three (resp. four) times the period. One of these extends for lower and the other for higher parameter values. The initial elliptic periodic orbit momentarily loses its stability due to these approaching unstable orbits.

Denominators $q \geq 5$ (and also the second possibility for $q = 4$) lead to a pair of subharmonic periodic orbits of q times the period emerging either for lower or for higher parameter values. This is (especially for large q) comparable to the behavior at Diophantine $e^{\pm i\Omega}$ where a family of invariant tori emerges, cf. Sect. “Invariant Curves of Planar Diffeomorphisms” below.

For a single pair $e^{\pm i\Omega}$ of Floquet multipliers this behavior is traditionally studied for the (iso-energetic) Poincaré-mapping, cf. [92] and references therein. However, the above description remains true in higher dimensions, where additionally multiple pairs of Floquet multipliers may interact. An instructive example is the Lagrange top, the sleeping motion of which is gyroscopically stabilized after a periodic Hamiltonian Hopf bifurcation; see [56] for more details.

Dissipative Perturbations

There exists a large class of local bifurcations in the dissipative setting, that can be arranged in a perturbation theory setting, where the unperturbed system is Hamil-

tonian. The arrangement consists of changes of variables and rescaling. An early example of this is the Bogdanov–Takens bifurcation [131,132]. For other examples regarding nilpotent singularities, see [23,40] and references therein.

To fix thoughts, consider families of planar maps and let the unperturbed Hamiltonian part contain a center (possibly surrounded by a homoclinic loop). The question then is which of these persist when adding the dissipative perturbation.

Usually only a definite finite number persists. As in Subsect. “Chaos”, a Melnikov function can be invoked here, possibly again leading to elliptic (Abelian) integrals, Picard Fuchs equations, etc. For details see [61,124] and references therein.

Invariant Curves of Planar Diffeomorphisms

This section starts with general considerations on circle diffeomorphisms, in particular focusing on persistence properties of quasi-periodic dynamics. Our main references are [2,24,29,31,70,71,139,140]. For a definition of rotation number, see [58]. After this we turn to area preserving maps of an annulus where we discuss Moser’s twist map theorem [104], also see [24,29,31]. The section is concluded by a description of the holomorphic linearization of a fixed point in a planar map [7,101,141,142].

Our main perspective will be perturbative, where we consider circle maps near a rigid rotation. It turns out that generally parameters are needed for persistence of quasi-periodicity under perturbations. In the area preserving setting we consider perturbations of a pure twist map.

Circle Maps

We start with the following general problem. Given a two-parameter family

$$P_{\alpha,\varepsilon}: \mathbb{T}^1 \rightarrow \mathbb{T}^1, \quad x \mapsto x + 2\pi\alpha + \varepsilon a(x, \alpha, \varepsilon)$$

of circle maps of class C^∞ . It turns out to be convenient to view this two-parameter family as a one-parameter family of maps

$$P_\varepsilon: \mathbb{T}^1 \times [0, 1] \rightarrow \mathbb{T}^1 \times [0, 1], \\ (x, \alpha) \mapsto (x + 2\pi\alpha + \varepsilon a(x, \alpha, \varepsilon), \alpha)$$

of the cylinder. Note that the unperturbed system P_0 is a family of rigid circle rotations, viewed as a cylinder map, where the individual map $P_{\alpha,0}$ has rotation number α . The question now is what will be the fate of this rigid dynamics for $0 \neq |\varepsilon| \ll 1$.

The classical way to address this question is to look for a conjugation Φ_ε , that makes the following diagram commute

$$\begin{array}{ccc} \mathbb{T}^1 \times [0, 1] & \xrightarrow{P_\varepsilon} & \mathbb{T}^1 \times [0, 1] \\ \uparrow \Phi_\varepsilon & & \uparrow \Phi_\varepsilon \\ \mathbb{T}^1 \times [0, 1] & \xrightarrow{P_0} & \mathbb{T}^1 \times [0, 1], \end{array}$$

i. e., such that

$$P_\varepsilon \circ \Phi_\varepsilon = \Phi_\varepsilon \circ P_0.$$

Due to the format of P_ε we take Φ_ε as a skew map

$$\Phi_\varepsilon(x, \alpha) = (x + \varepsilon U(x, \alpha, \varepsilon), \alpha + \varepsilon \sigma(\alpha, \varepsilon)),$$

which leads to the *nonlinear* equation

$$\begin{aligned} U(x + 2\pi\alpha, \alpha, \varepsilon) - U(x, \alpha, \varepsilon) \\ = 2\pi\sigma(\alpha, \varepsilon) + a(x + \varepsilon U(x, \alpha, \varepsilon), \alpha + \varepsilon\sigma(\alpha, \varepsilon), \varepsilon) \end{aligned}$$

in the unknown maps U and σ . Expanding in powers of ε and comparing at lowest order yields the linear equation

$$U_0(x + 2\pi\alpha, \alpha) - U_0(x, \alpha) = 2\pi\sigma_0(\alpha) + a_0(x, \alpha)$$

which can be directly solved by Fourier series. Indeed, writing

$$\begin{aligned} a_0(x, \alpha) &= \sum_{k \in \mathbb{Z}} a_{0k}(\alpha) e^{ikx}, \\ U_0(x, \alpha) &= \sum_{k \in \mathbb{Z}} U_{0k}(\alpha) e^{ikx} \end{aligned}$$

we find $\sigma_0 = -1/(2\pi)a_{00}$ and

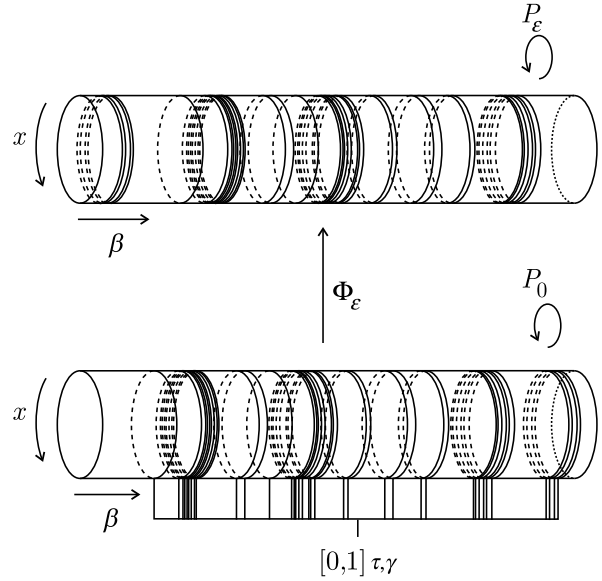
$$U_{0k}(\alpha) = \frac{a_{0k}(\alpha)}{e^{2\pi i k \alpha} - 1}.$$

It follows that in general a formal solution exists if and only if $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Still, the accumulation of $e^{2\pi i k \alpha} - 1$ on 0 leads to the celebrated *small divisors* [9,108], also see [24,29,31,55].

The classical solution considers the following Diophantine non-resonance conditions. Fixing $\tau > 2$ and $\gamma > 0$ consider $\alpha \in [0, 1]$ such that for all rationals p/q

$$\left| \alpha - \frac{p}{q} \right| \geq \gamma q^{-\tau}. \quad (5)$$

This subset of such α s is denoted by $[0, 1]_{\tau, \gamma}$ and is well-known to be nowhere dense but of large measure as $\gamma > 0$ gets small [115]. Note that Diophantine numbers are irrational.



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 2

Skew cylinder map, conjugating (Diophantine) quasi-periodic invariant circles of P_0 and P_ε

Theorem 3 (Circle Map Theorem) For γ sufficiently small and for the perturbation εa sufficiently small in the C^∞ -topology, there exists a C^∞ transformation $\Phi_\varepsilon: \mathbb{T}^1 \times [0, 1] \rightarrow \mathbb{T}^1 \times [0, 1]$, conjugating the restriction $P_0|_{[0,1]_{\tau, \gamma}}$ to a subsystem of P_ε .

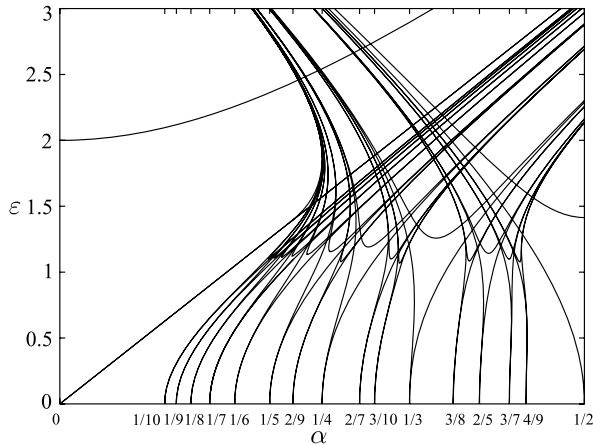
Theorem 3 in the present structural stability formulation (compare with Fig. 2) is a special case of the results in [29,31]. We here speak of *quasi-periodic stability*. For earlier versions see [2,9].

Remark

- Rotation numbers are preserved by the map Φ_ε and irrational rotation numbers correspond to quasi-periodicity. Theorem 3 thus ensures that *typically* quasi-periodicity occurs with *positive* measure in the parameter space. Note that since Cantor sets are perfect, quasi-periodicity typically has a non-isolated occurrence.
- The map Φ_ε has no dynamical meaning inside the gaps. The gap dynamics in the case of circle maps can be illustrated by the Arnold family of circle maps [2,7,58], given by

$$P_{\alpha, \varepsilon}(x) = x + 2\pi\alpha + \varepsilon \sin x$$

which exhibits a countable union of open resonance tongues where the dynamics is periodic, see Fig. 3. Note that this map only is a diffeomorphism for $|\varepsilon| < 1$.



Hamiltonian Perturbation Theory (and Transition to Chaos),
Figure 3

Arnold resonance tongues; for $\varepsilon \geq 1$ the maps are endomorphic

- We like to mention that non-perturbative versions of Theorem 3 have been proven in [70,71,139].
- For simplicity we formulated Theorem 3 under C^∞ -regularity, noting that there exist many ways to generalize this. On the one hand there exist C^k -versions for finite k and on the other hand there exist fine tunings in terms of real-analytic and Gevrey regularity. For details we refer to [24,31] and references therein. This same remark applies to other results in this section and in Sect. “KAM Theory: An Overview” on KAM theory.

A possible application of Theorem 3 runs as follows. Consider a system of weakly coupled Van der Pol oscillators

$$\begin{aligned}\ddot{y}_1 + c_1 \dot{y}_1 + a_1 y_1 + f_1(y_1, \dot{y}_1) &= \varepsilon g_1(y_1, y_2, \dot{y}_1, \dot{y}_2) \\ \ddot{y}_2 + c_2 \dot{y}_2 + a_2 y_2 + f_2(y_2, \dot{y}_2) &= \varepsilon g_2(y_1, y_2, \dot{y}_1, \dot{y}_2).\end{aligned}$$

Writing $\dot{y}_j = z_j$, $j = 1, 2$, one obtains a vector field in the four-dimensional phase space $\mathbb{R}^2 \times \mathbb{R}^2 = \{(y_1, z_1), (y_2, z_2)\}$. For $\varepsilon = 0$ this vector field has an invariant two-torus, which is the product of the periodic motions of the individual Van der Pol oscillations. This two-torus is normally hyperbolic and therefore persistent for $|\varepsilon| \ll 1$ [73]. In fact the torus is an attractor and we can define a Poincaré return map within this torus attractor. If we include some of the coefficients of the equations as parameters, Theorem 3 is directly applicable. The above statements on quasi-periodic circle maps then directly translate to the case of quasi-periodic invariant two-tori. Concerning the resonant cases, generically a tongue structure like in Fig. 3 occurs; for the dynamics corresponding to parameter values inside such a tongue one speaks of *phase lock*.

Remark

- The celebrated synchronization of Huygens’ clocks [77] is related to a $1 : 1$ resonance, meaning that the corresponding Poincaré map would have its parameters in the main tongue with rotation number 0. Compare with Fig. 3.
- There exist direct generalizations to cases with n -oscillators ($n \in \mathbb{N}$), leading to families of invariant n -tori carrying quasi-periodic flow, forming a nowhere dense set of positive measure. An alteration with resonance occurs as roughly sketched in Fig. 3. In higher dimension the gap dynamics, apart from periodicity, also can contain strange attractors [112,126]. We shall come back to this subject in a later section.

Area-Preserving Maps

The above setting historically was preceded by an area preserving analogue [104] that has its origin in the Hamiltonian dynamics of frictionless mechanics.

Let $\Delta \subseteq \mathbb{R}^2 \setminus \{(0, 0)\}$ be an annulus, with symplectic polar coordinates $(\varphi, I) \in \mathbb{T}^1 \times \mathbf{K}$, where \mathbf{K} is an interval. Moreover, let $\sigma = d\varphi \wedge dI$ be the area form on Δ .

We consider a σ -preserving smooth map $P_\varepsilon : \Delta \rightarrow \Delta$ of the form

$$P_\varepsilon(\varphi, I) = (\varphi + 2\pi\alpha(I), I) + O(\varepsilon),$$

where we assume that the map $I \mapsto \alpha(I)$ is a (local) diffeomorphism. This assumption is known as the *twist condition* and P_ε is called a *twist map*. For the unperturbed case $\varepsilon = 0$ we are dealing with a pure twist map and its dynamics are comparable to the unperturbed family of cylinder maps as met in Subsect. “Circle Maps”. Indeed it is again a family of rigid rotations, parametrized by I and where $P_0(\cdot, I)$ has rotation number $\alpha(I)$. In this case the question is what will be the fate of this family of invariant circles, as well as with the corresponding rigidly rotational dynamics.

Regarding the rotation number we again introduce Diophantine conditions. Indeed, for $\tau > 2$ and $\gamma > 0$ the subset $[0, 1]_{\tau, \gamma}$ is defined as in (5), i.e., it contains all $\alpha \in [0, 1]$, such that for all rationals p/q

$$\left| \alpha - \frac{p}{q} \right| \geq \gamma q^{-\tau}.$$

Pulling back $[0, 1]_{\tau, \gamma}$ along the map α we obtain a subset $\Delta_{\tau, \gamma} \subseteq \Delta$.

Theorem 4 (Twist Map Theorem [104]) For γ sufficiently small, and for the perturbation $O(\varepsilon)$ sufficiently

small in C^∞ -topology, there exists a C^∞ transformation $\Phi_\varepsilon: \Delta \rightarrow \Delta$, conjugating the restriction $P_0|_{\Delta_{\tau,\gamma}}$ to a subsystem of P_ε .

As in the case of Theorem 3 again we chose the formulation of [29,31]. Largely the remarks following Theorem 3 also apply here.

Remark

- Compare the format of the Theorems 3 and 4 and observe that in the latter case the role of the parameter α has been taken by the action variable I . Theorem 4 implies that *typically* quasi-periodicity occurs with positive measure in phase space.
- In the gaps typically we have coexistence of periodicity, quasi-periodicity and chaos [6,9,35,107,108,123,137]. The latter follows from transversality of homo- and heteroclinic connections that give rise to positive topological entropy. Open problems are whether the corresponding Lyapunov exponents also are positive, compare with the discussion at the end of the introduction.

Similar to the applications of Theorem 3 given at the end of Subsect. “Circle Maps”, here direct applications are possible in the conservative setting. Indeed, consider a system of weakly coupled pendula

$$\begin{aligned}\ddot{y}_1 + \alpha_1^2 \sin y_1 &= \varepsilon \frac{\partial U}{\partial y_1}(y_1, y_2) \\ \ddot{y}_2 + \alpha_2^2 \sin y_2 &= \varepsilon \frac{\partial U}{\partial y_2}(y_1, y_2).\end{aligned}$$

Writing $\dot{y}_j = z_j$, $j = 1, 2$ as before, we again get a vector field in the four-dimensional phase space $\mathbb{R}^2 \times \mathbb{R}^2 = \{(y_1, y_2), (z_1, z_2)\}$. In this case the energy

$$\begin{aligned}H_\varepsilon(y_1, y_2, z_1, z_2) \\ = \frac{1}{2}z_1^2 + \frac{1}{2}z_2^2 - \alpha_1^2 \cos y_1 - \alpha_2^2 \cos y_2 + \varepsilon U(y_1, y_2)\end{aligned}$$

is a constant of motion. Restricting to a three-dimensional energy surface $H_\varepsilon^{-1} = \text{const.}$, the iso-energetic Poincaré map P_ε is a twist map and application of Theorem 4 yields the conclusion of quasi-periodicity (on invariant two-tori) occurring with positive measure in the energy surfaces of H_ε .

Remark As in the dissipative case this example directly generalizes to cases with n oscillators ($n \in \mathbb{N}$), again leading to invariant n -tori with quasi-periodic flow. We shall return to this subject in a later section.

Linearization of Complex Maps

The Subjects. “Circle Maps” and “Area-Preserving Maps” both deal with smooth circle maps that are conjugated to rigid rotations. Presently the concern is with planar holomorphic maps that are conjugated to a rigid rotation on an open subset of the plane. Historically this is the first time that a small divisor problem was solved [7,101,141,142] and ► [Perturbative Expansions, Convergence of](#).

Complex Linearization Given is a holomorphic germ $F: (\mathbb{C}, 0) \rightarrow (\mathbb{C}, 0)$ of the form $F(z) = \lambda z + f(z)$, with $f(0) = f'(0) = 0$. The problem is to find a biholomorphic germ $\Phi: (\mathbb{C}, 0) \rightarrow (\mathbb{C}, 0)$ such that

$$\Phi \circ F = \lambda \cdot \Phi.$$

Such a diffeomorphism Φ is called a *linearization* of F near 0.

We begin with the formal approach. Given the series $f(z) = \sum_{j \geq 2} f_j z^j$, we look for $\Phi(z) = z + \sum_{j \geq 2} \phi_j z^j$. It turns out that a solution always exists whenever $\lambda \neq 0$ is not a root of unity. Indeed, direct computation reveals the following set of equations that can be solved recursively:

For $j = 2$ get the equation $\lambda(1 - \lambda)\phi_2 = f_2$

For $j = 3$ get the equation $\lambda(1 - \lambda^2)\phi_3 = f_3 + 2\lambda f_2 \phi_2$

For $j = n$ get the equation $\lambda(1 - \lambda^{n-1})\phi_n = f_n + \text{known}$.

The question now reduces to whether this formal solution has a positive radius of convergence.

The hyperbolic case $0 < |\lambda| \neq 1$ was already solved by Poincaré, for a description see [7]. The elliptic case $|\lambda| = 1$ again has small divisors and was solved by Siegel when for some $\gamma > 0$ and $\tau > 2$ we have the Diophantine non-resonance condition

$$|\lambda - e^{2\pi i \frac{p}{q}}| \geq \gamma |q|^{-\tau}.$$

The corresponding set of λ constitutes a set of full measure in $\mathbb{T}^1 = \{\lambda\}$.

Yoccoz [141] completely solved the elliptic case using the Bruno-condition. If

$$\lambda = e^{2\pi i \alpha} \quad \text{and when} \quad \frac{p_n}{q_n}$$

is the n th convergent in the continued fraction expansion of α then the Bruno-condition reads

$$\sum_n \frac{\log(q_{n+1})}{q_n} < \infty.$$

This condition turns out to be necessary and sufficient for Φ having positive radius of convergence [141,142].

Cremer's Example in Herman's Version As an example consider the map

$$F(z) = \lambda z + z^2,$$

where $\lambda \in \mathbb{T}^1$ is not a root of unity.

Observe that a point $z \in \mathbb{C}$ is a periodic point of F with period q if and only if $F^q(z) = z$, where obviously

$$F^q(z) = \lambda^q z + \dots + z^{2^q}.$$

Writing

$$F^q(z) - z = z \left(\lambda^q - 1 + \dots + z^{2^q-1} \right),$$

the period q periodic points exactly are the roots of the right hand side polynomial. Abbreviating $N = 2^q - 1$, it directly follows that, if z_1, z_2, \dots, z_N are the nontrivial roots, then for their product we have

$$z_1 \cdot z_2 \cdot \dots \cdot z_N = \lambda^q - 1.$$

It follows that there exists a nontrivial root within radius

$$|\lambda^q - 1|^{1/N}$$

of $z = 0$.

Now consider the set of $\Lambda \subset \mathbb{T}^1$ defined as follows: $\lambda \in \Lambda$ whenever

$$\liminf_{q \rightarrow \infty} |\lambda^q - 1|^{1/N} = 0.$$

It can be directly shown that Λ is *residual*, again compare with [115]. It also follows that for $\lambda \in \Lambda$ linearization is impossible. Indeed, since the rotation is irrational, the existence of periodic points in any neighborhood of $z = 0$ implies zero radius of convergence.

Remark

- Notice that the residual set Λ is in the complement of the full measure set of all Diophantine numbers, again see [115].
- Considering $\lambda \in \mathbb{T}^1$ as a parameter, we see a certain analogy of these results on complex linearization with the Theorems 3 and 4. Indeed, in this case for a full measure set of λ s on a neighborhood of $z = 0$ the map $F = F_\lambda$ is conjugated to a rigid irrational rotation. Such a domain in the z -plane often is referred to as a Siegel disc. For a more general discussion of these and of Herman rings, see [101].

KAM Theory: An Overview

In Sect. “Invariant Curves of Planar Diffeomorphisms” we described the persistent occurrence of quasi-periodicity in the setting of diffeomorphisms of the circle or the plane. The general perturbation theory of quasi-periodic motions is known under the name Kolmogorov–Arnold–Moser (or KAM) theory and discussed extensively elsewhere in this encyclopedia ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#). Presently we briefly summarize parts of this KAM theory in broad terms, as this fits in our considerations, thereby largely referring to [4,80,81,119,121,143,144], also see [20,24,55].

In general quasi-periodicity is defined by a smooth conjugation. First on the n -torus $\mathbb{T}^n = \mathbb{R}^n / (2\pi\mathbb{Z})^n$ consider the vector field

$$\mathbb{X}_\omega = \sum_{j=1}^n \omega_j \frac{\partial}{\partial \varphi_j},$$

where $\omega_1, \omega_2, \dots, \omega_n$ are called frequencies [43,106]. Now, given a smooth (say, of class C^∞) vector field X on a manifold M , with $T \subseteq M$ an invariant n -torus, we say that the restriction $X|_T$ is *parallel* if there exists $\omega \in \mathbb{R}^n$ and a smooth diffeomorphism $\Phi: T \rightarrow \mathbb{T}^n$, such that $\Phi_*(X|_T) = \mathbb{X}_\omega$. We say that $X|_T$ is *quasi-periodic* if the frequencies $\omega_1, \omega_2, \dots, \omega_n$ are independent over \mathbb{Q} .

A quasi-periodic vector field $X|_T$ leads to an integer affine structure on the torus T . In fact, since each orbit is dense, it follows that the self conjugations of \mathbb{X}_ω exactly are the translations of \mathbb{T}^n , which completely determine the affine structure of \mathbb{T}^n . Then, given $\Phi: T \rightarrow \mathbb{T}^n$ with $\Phi_*(X|_T) = \mathbb{X}_\omega$, it follows that the self conjugations of $X|_T$ determines a natural affine structure on the torus T . Note that the conjugation Φ is unique modulo translations in T and \mathbb{T}^n .

Note that the composition of Φ by a translation of \mathbb{T}^n does not change the frequency vector ω . However, the composition by a linear invertible map $S \in \text{GL}(n, \mathbb{Z})$ yields $S_*\mathbb{X}_\omega = \mathbb{X}_{S\omega}$. We here speak of an *integer affine structure* [43].

Remark

- The transition maps of an integer affine structure are translations and elements of $\text{GL}(n, \mathbb{Z})$.
- The current construction is compatible with the integrable affine structure on the Liouville tori of an integrable Hamiltonian system [6]. Note that in that case the structure extends to all parallel tori.

Classical KAM Theory

The classical KAM theory deals with smooth, nearly integrable Hamiltonian systems of the form

$$\begin{aligned}\dot{\varphi} &= \omega(I) + \varepsilon f(I, \varphi, \varepsilon) \\ \dot{I} &= \varepsilon g(I, \varphi, \varepsilon),\end{aligned}\quad (6)$$

where I varies over an open subset of \mathbb{R}^n and φ over the standard torus \mathbb{T}^n . Note that for $\varepsilon = 0$ the phase space as an open subset of $\mathbb{R}^n \times \mathbb{T}^n$ is foliated by invariant tori, parametrized by I . Each of the tori is parametrized by φ and the corresponding motion is parallel (or multi-periodic or conditionally periodic) with frequency vector $\omega(I)$.

Perturbation theory asks for persistence of the invariant n -tori and the parallelity of their motion for small values of $|\varepsilon|$. The answer that KAM theory gives needs two essential ingredients. The first ingredient is that of *Kolmogorov non-degeneracy* which states that the map $I \in \mathbb{R}^n \mapsto \omega(I) \in \mathbb{R}^n$ is a (local) diffeomorphism. Compare with the twist condition of Sect. “Invariant Curves of Planar Diffeomorphisms”. The second ingredient generalizes the Diophantine conditions (5) of that section as follows: for $\tau > n - 1$ and $\gamma > 0$ consider the set

$$\mathbb{R}_{\tau, \gamma}^n = \{\omega \in \mathbb{R}^n \mid |\langle \omega, k \rangle| \geq \gamma |k|^{-\tau}, k \in \mathbb{Z}^n \setminus \{0\}\}. \quad (7)$$

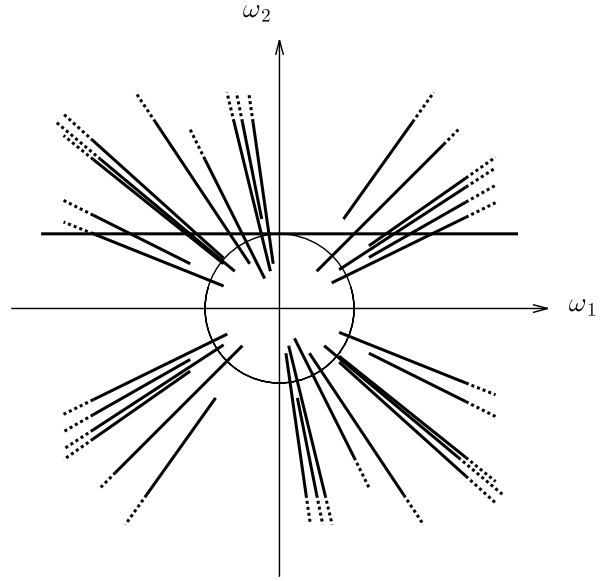
The following properties are more or less direct. First $\mathbb{R}_{\tau, \gamma}^n$ has a closed half line geometry in the sense that if $\omega \in \mathbb{R}_{\tau, \gamma}^n$ and $s \geq 1$ then also $s\omega \in \mathbb{R}_{\tau, \gamma}^n$. Moreover, the intersection $\mathbb{S}^{n-1} \cap \mathbb{R}_{\tau, \gamma}^n$ is a Cantor set of measure $\mathbb{S}^{n-1} \setminus \mathbb{R}_{\tau, \gamma}^n = O(\gamma)$ as $\gamma \downarrow 0$, see Fig. 4.

Completely in the spirit of Theorem 4, the classical KAM theorem roughly states that a Kolmogorov non-degenerate nearly integrable system (6) _{ε} , for $|\varepsilon| \ll 1$ is smoothly conjugated to the unperturbed version (6)₀, provided that the frequency map ω is co-restricted to the Diophantine set $\mathbb{R}_{\tau, \gamma}^n$. In this formulation smoothness has to be taken in the sense of Whitney [119, 136], also compare with [20, 24, 29, 31, 55, 121].

As a consequence we may say that in Hamiltonian systems of n degrees of freedom typically quasi-periodic invariant (Lagrangian) n -tori occur with positive measure in phase space. It should be said that also an iso-energetic version of this classical result exists, implying a similar conclusion restricted to energy hypersurfaces [6, 9, 21, 24]. The Twist Map Theorem 4 is closely related to the iso-energetic KAM Theorem.

Remark

- We chose the quasi-periodic stability format as in Sect. “Invariant Curves of Planar Diffeomorphisms”.



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 4

The Diophantine set $\mathbb{R}_{\tau, \gamma}^n$ has the closed half line geometry and the intersection $\mathbb{S}^{n-1} \cap \mathbb{R}_{\tau, \gamma}^n$ is a Cantor set of measure $\mathbb{S}^{n-1} \setminus \mathbb{R}_{\tau, \gamma}^n = O(\gamma)$ as $\gamma \downarrow 0$

For regularity issues compare with a remark following Theorem 3.

- For applications we largely refer to the introduction and to [24, 31] and references therein.
- Continuing the discussion on affine structures at the beginning of this section, we mention that by means of the symplectic form, the domain of the I -variables in \mathbb{R}^n inherits an affine structure [60], also see [91] and references therein.

Statistical Mechanics deals with particle systems that are large, often infinitely large. The *Ergodic Hypothesis* roughly says that in a bounded energy hypersurface, the dynamics are ergodic, meaning that any evolution in the energy level set comes near every point of this set.

The taking of limits as the number of particles tends to infinity is a notoriously difficult subject. Here we discuss a few direct consequences of classical KAM theory for many degrees of freedom. This discussion starts with Kolmogorov's papers [80, 81], which we now present in a slightly rephrased form. First, we recall that for Hamiltonian systems (say, with n degrees of freedom), typically the union of Diophantine quasi-periodic Lagrangian invariant n -tori fills up positive measure in the phase space and also in the energy hypersurfaces. Second, such a collection of KAM tori immediately gives rise to non-ergodic-

ity, since it clearly implies the existence of distinct invariant sets of positive measure. For background on Ergodic Theory, see e. g. [9,27] and [24] for more references. Apparently the KAM tori form an obstruction to ergodicity, and a question is how bad this obstruction is as $n \rightarrow \infty$. Results in [5,78] indicate that this KAM theory obstruction is not too bad as the size of the system tends to infinity. In general the role of the Ergodic Hypothesis in Statistical Mechanics has turned out to be much more subtle than was expected, see e. g. [18,64].

Dissipative KAM Theory

As already noted by Moser [105,106], KAM theory extends outside the world of Hamiltonian systems, like to volume preserving systems, or to equivariant or reversible systems. This also holds for the class of general smooth systems, often called dissipative. In fact, the KAM theorem allows for a Lie algebra proof, that can be used to cover all these special cases [24,29,31,45]. It turns out that in many cases parameters are needed for persistent occurrence of (Diophantine) quasi-periodic tori.

As an example we now consider the dissipative setting, where we discuss a parametrized system with normally hyperbolic invariant n -tori carrying quasi-periodic motion. From [73] it follows that this is a persistent situation and that, up to a smooth (in this case of class C^k for large k) diffeomorphism, we can restrict to the case where \mathbb{T}^n is the phase space. To fix thoughts we consider the smooth system

$$\begin{aligned}\dot{\varphi} &= \omega(\mu) + \varepsilon f(\varphi, \mu, \varepsilon) \\ \dot{\mu} &= 0,\end{aligned}\tag{8}$$

where $\mu \in \mathbb{R}^n$ is a multi-parameter. The results of the classical KAM theorem regarding $(6)_\varepsilon$ largely carry over to $(8)_{\mu,\varepsilon}$.

Now, for $\varepsilon = 0$ the product of phase space and parameter space as an open subset of $\mathbb{T}^n \times \mathbb{R}^n$ is completely foliated by invariant n -tori and since the perturbation does not concern the $\dot{\mu}$ -equation, this foliation is persistent. The interest is with the dynamics on the resulting invariant tori that remains parallel after the perturbation; compare with the setting of Theorem 3. As just stated, KAM theory here gives a solution similar to the Hamiltonian case. The analogue of the Kolmogorov non-degeneracy condition here is that the frequency map $\mu \mapsto \omega(\mu)$ is a (local) diffeomorphism. Then, in the spirit of Theorem 3, we state that the system $(8)_{\mu,\varepsilon}$ is smoothly conjugated to $(8)_{\mu,0}$, as before, provided that the map ω is co-restricted to the Diophantine set $\mathbb{R}_{\tau,\gamma}^n$. Again the smoothness has to be

taken in the sense of Whitney [29,119,136,143,144], also see [20,24,31,55].

It follows that the occurrence of normally hyperbolic invariant tori carrying (Diophantine) quasi-periodic flow is typical for families of systems with sufficiently many parameters, where this occurrence has positive measure in parameter space. In fact, if the number of parameters equals the dimension of the tori, the geometry as sketched in Fig. 4 carries over in a diffeomorphic way.

Remark

- Many remarks following Subsect. “Classical KAM Theory” and Theorem 3 also hold here.
- In cases where the system is degenerate, for instance because there is a lack of parameters, a path formalism can be invoked, where the parameter path is required to be a generic subfamily of the Diophantine set $\mathbb{R}_{\tau,\gamma}^n$, see Fig. 4. This amounts to the Rüssmann non-degeneracy, that still gives positive measure of quasi-periodicity in the parameter space, compare with [24,31] and references therein.
- In the dissipative case the KAM theorem gives rise to families of quasi-periodic attractors in a typical way. This is of importance in center manifold reductions of infinite dimensional dynamics as, e. g., in fluid mechanics [125,126]. In Sect. “Transition to Chaos and Turbulence” we shall return to this subject.

Lower Dimensional Tori

We extend the above approach to the case of lower dimensional tori, i. e., where the dynamics transversal to the tori is also taken into account. We largely follow the set-up of [29,45] that follows Moser [106]. Also see [24,31] and references therein. Changing notation a little, we now consider the phase space $\mathbb{T}^n \times \mathbb{R}^m = \{x(\bmod 2\pi), y\}$, as well a parameter space $\{\mu\} = P \subset \mathbb{R}^s$. We consider a C^∞ -family of vector fields $X(x, y, \mu)$ as before, having $\mathbb{T}^n \times \{0\} \subset \mathbb{T}^n \times \mathbb{R}^m$ as an invariant n -torus for $\mu = \mu_0 \in P$.

$$\begin{aligned}\dot{x} &= \omega(\mu) + f(y, \mu) \\ \dot{y} &= \Omega(\mu) y + g(y, \mu) \\ \dot{\mu} &= 0,\end{aligned}\tag{9}$$

with $f(y, \mu_0) = O(|y|)$ and $g(y, \mu_0) = O(|y|^2)$, so we assume the invariant torus to be of Floquet type.

The system $X = X(x, y, \mu)$ is integrable in the sense that it is \mathbb{T}^n -symmetric, i. e., x -independent [29]. The interest is with the fate of the invariant torus $\mathbb{T}^n \times \{0\}$ and

its parallel dynamics under small perturbation to a system $\tilde{X} = \tilde{X}(x, y, \mu)$ that no longer needs to be integrable.

Consider the smooth mappings $\omega: P \rightarrow \mathbb{R}^n$ and $\Omega: P \rightarrow \mathfrak{gl}(m, \mathbb{R})$. To begin with we restrict to the case where all eigenvalues of $\Omega(\mu_0)$ are simple and nonzero. In general for such a matrix $\Omega \in \mathfrak{gl}(m, \mathbb{R})$, let the eigenvalues be given by $\alpha_1 \pm i\beta_1, \dots, \alpha_{N_1} \pm i\beta_{N_1}$ and $\delta_1, \dots, \delta_{N_2}$, where all α_j, β_j and δ_j are real and hence $m = 2N_1 + N_2$. Also consider the map $\text{spec}: \mathfrak{gl}(m, \mathbb{R}) \rightarrow \mathbb{R}^{2N_1+N_2}$, given by $\Omega \mapsto (\alpha, \beta, \delta)$. Next to the internal frequency vector $\omega \in \mathbb{R}^n$, we also have the vector $\beta \in \mathbb{R}^{N_1}$ of normal frequencies.

The present analogue of Kolmogorov non-degeneracy is the Broer–Huiteima–Takens (BHT) non-degeneracy condition [29,127], which requires that the product map $\omega \times (\text{spec}) \circ \Omega: P \rightarrow \mathbb{R}^n \times \mathfrak{gl}(m, \mathbb{R})$ at $\mu = \mu_0$ has a surjective derivative and hence is a local submersion [72].

Furthermore, we need Diophantine conditions on both the internal and the normal frequencies, generalizing (7). Given $\tau > n - 1$ and $\gamma > 0$, it is required for all $k \in \mathbb{Z}^n \setminus \{0\}$ and all $\ell \in \mathbb{Z}^{N_1}$ with $|\ell| \leq 2$ that

$$|\langle k, \omega \rangle + \langle \ell, \beta \rangle| \geq \gamma |k|^{-\tau}. \quad (10)$$

Inside $\mathbb{R}^n \times \mathbb{R}^{N_1} = \{\omega, \beta\}$ this yields a Cantor set as before (compare Fig. 4). This set has to be pulled back along the submersion $\omega \times (\text{spec}) \circ \Omega$, for examples see Subsects. “*(n-1)-Tori*” and “*Quasi-periodic Bifurcations*” below.

The KAM theorem for this setting is quasi-periodic stability of the n -tori under consideration, as in Subsect. “*Dissipative KAM Theory*”, yielding typical examples where quasi-periodicity has positive measure in parameter space. In fact, we get a little more here, since the normal linear behavior of the n -tori is preserved by the Whitney smooth conjugations. This is expressed as normal linear stability, which is of importance for quasi-periodic bifurcations, see Subsect. “*Quasi-periodic Bifurcations*” below.

Remark

- A more general set-up of the normal stability theory [45] adapts the above to the case of non-simple (multiple) eigenvalues. Here the BHT non-degeneracy condition is formulated in terms of versal unfolding of the matrix $\Omega(\mu_0)$ [7]. For possible conditions under which vanishing eigenvalues are admissible see [29,42,69] and references therein.
- This general set-up allows for a structure preserving formulation as mentioned earlier, thereby including the Hamiltonian and volume preserving case, as well as equivariant and reversible cases. This allows

us, for example, to deal with quasi-periodic versions of the Hamiltonian and the reversible Hopf bifurcation [38,41,42,44].

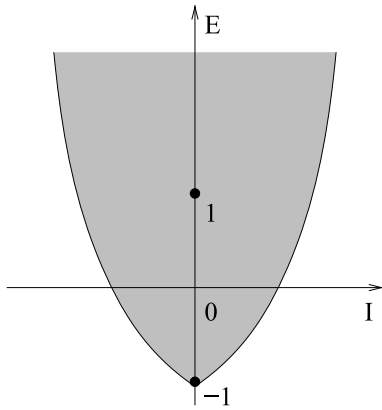
- The Parameterized KAM Theory discussed here a priori needs many parameters. In many cases the parameters are distinguished in the sense that they are given by action variables, etc. For an example see Subsect. “*(n-1)-Tori*” on Hamiltonian $(n-1)$ -tori. Also see [127] and [24,31] where the case of Rüssmann non-degeneracy is included. This generalizes a remark at the end of Subsect. “*Dissipative KAM Theory*”.

Global KAM Theory

We stay in the Hamiltonian setting, considering Lagrangian invariant n -tori as these occur in a Liouville integrable system with n degrees of freedom. The union of these tori forms a smooth \mathbb{T}^n -bundle $f: M \rightarrow B$ (where we leave out all singular fibers). It is known that this bundle can be non-trivial [56,60] as can be measured by monodromy and Chern class. In this case global action angle variables are not defined. This non-triviality, among other things, is of importance for semi-classical versions of the classical system at hand, in particular for certain spectrum defects [57,62,134,135], for more references also see [24].

Restricting to the classical case, the problem is what happens to the (non-trivial) \mathbb{T}^n -bundle f under small, non-integrable perturbation. From the classical KAM theory, see Subsect. “*Classical KAM Theory*” we already know that on trivializing charts of f Diophantine quasi-periodic n -tori persist. In fact, at this level, a Whitney smooth conjugation exists between the integrable system and its perturbation, which is even Gevrey regular [136]. It turns out that these local KAM conjugations can be glued together so to obtain a global conjugation at the level of quasi-periodic tori, thereby implying global quasi-periodic stability [43]. Here we need unicity of KAM tori, i. e., independence of the action-angle chart used in the classical KAM theorem [26]. The proof uses the integer affine structure on the quasi-periodic tori, which enables taking convex combinations of the local conjugations subjected to a suitable partition of unity [72,129]. In this way the geometry of the integrable bundle can be carried over to the nearly-integrable one.

The classical example of a Liouville integrable system with non-trivial monodromy [56,60] is the spherical pendulum, which we now briefly revisit. The configuration space is $\mathbb{S}^2 = \{q \in \mathbb{R}^3 \mid \langle q, q \rangle = 1\}$ and the phase space $T^*\mathbb{S}^2 \cong \{(q, p) \in \mathbb{R}^6 \mid \langle q, q \rangle = 1 \text{ and } \langle q, p \rangle = 0\}$. The two integrals $I = q_1 p_2 - q_2 p_1$ (angular momentum) and $E = \frac{1}{2} \langle p, p \rangle + q_3$ (energy) lead to



Hamiltonian Perturbation Theory (and Transition to Chaos),
Figure 5

Range of the energy-momentum map of the spherical pendulum

an energy momentum map $\mathcal{EM}: T^*\mathbb{S}^2 \rightarrow \mathbb{R}^2$, given by $(q, p) \mapsto (I, E) = (q_1 p_2 - q_2 p_1, \frac{1}{2}(p, p) + q_3)$. In Fig. 5 we show the image of the map \mathcal{EM} . The shaded area B consists of regular values, the fiber above which is a Lagrangian two-torus; the union of these gives rise to a bundle $f: M \rightarrow B$ as described before, where $f = \mathcal{EM}|_M$. The motion in the two-tori is a superposition of Huygens' rotations and pendulum-like swinging, and the non-existence of global action angle variables reflects that the three interpretations of 'rotating oscillation', 'oscillating rotation' and 'rotating rotation' cannot be reconciled in a consistent way. The singularities of the fibration include the equilibria $(q, p) = ((0, 0, \pm 1), (0, 0, 0)) \mapsto (I, E) = (0, \pm 1)$. The boundary of this image also consists of singular points, where the fiber is a circle that corresponds to Huygens' horizontal rotations of the pendulum. The fiber above the upper equilibrium point $(I, E) = (0, 1)$ is a pinched torus [56], leading to non-trivial monodromy, in a suitable bases of the period lattices, given by

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \in \text{GL}(2, \mathbb{R}).$$

The question here is what remains of the bundle f when the system is perturbed. Here we observe that locally Kolmogorov non-degeneracy is implied by the non-trivial monodromy [114,122]. From [43,122] it follows that the non-trivial monodromy can be extended in the perturbed case.

Remark

- The case where this perturbation remains integrable is covered in [95], but presently the interest is with the

nearly integrable case, so where the axial symmetry is broken. Also compare [24] and many of its references.

- The global conjugations of [43] are Whitney smooth (even Gevrey regular [136]) and near the identity map in the C^∞ -topology [72]. Geometrically speaking these diffeomorphisms also are \mathbb{T}^n -bundle isomorphisms between the unperturbed and the perturbed bundle, the basis of which is a Cantor set of positive measure.

Splitting of Separatrices

KAM theory does not predict the fate of close-to-resonant tori under perturbations. For fully resonant tori the phenomenon of frequency locking leads to the destruction of the torus under (sufficiently rich) perturbations, and other resonant tori disintegrate as well. In the case of a single resonance between otherwise Diophantine frequencies the perturbation leads to quasi-periodic bifurcations, cf. Sect. "Transition to Chaos and Turbulence".

While KAM theory concerns the fate of most trajectories and for all times, a complementary theorem has been obtained in [93,109,110,113]. It concerns all trajectories and states that they stay close to the unperturbed tori for *long* times that are exponential in the inverse of the perturbation strength. For trajectories starting close to surviving tori the diffusion is even superexponentially slow, cf. [102,103]. Here a form of smoothness exceeding the mere existence of infinitely many derivatives of the Hamiltonian is a necessary ingredient, for finitely differentiable Hamiltonians one only obtains polynomial times.

Solenoids, which cannot be present in integrable systems, are constructed for generic Hamiltonian systems in [16,94,98], yielding the simultaneous existence of representatives of all homeomorphism-classes of solenoids. Hyperbolic tori form the core of a construction proposed in [5] of trajectories that venture off to distant points of the phase space. In the unperturbed system the union of a family of hyperbolic tori, parametrized by the actions conjugate to the toral angles, form a normally hyperbolic manifold. The latter is persistent under perturbations, cf. [73,100], and carries a Hamiltonian flow with fewer degrees of freedom. The main difference between integrable and non-integrable systems already occurs for periodic orbits.

Periodic Orbits

A sharp difference to dissipative systems is that it is generic for hyperbolic periodic orbits on compact energy shells in Hamiltonian systems to have homoclinic orbits, cf. [1] and references therein. For integrable systems these form to-

gether a pinched torus, but under generic perturbations the stable and unstable manifold of a hyperbolic periodic orbit intersect transversely. It is a nontrivial task to actually check this genericity condition for a given non-integrable perturbation, a first-order condition going back to Poincaré requires the computation of the so-called Mel'nikov integral, see [66,137] for more details. In two degrees of freedom normalization leads to approximations that are integrable to all orders, which implies that the Melnikov integral is a flat function. In the real analytic case the Melnikov criterion is still decisive in many examples [65].

Genericity conditions are traditionally formulated in the universe of smooth vector fields, and this makes the whole class of analytic vector fields appear to be non-generic. This is an overly pessimistic view as the conditions defining a certain class of generic vector fields may certainly be satisfied by a given analytic system. In this respect it is interesting that the generic properties may also be formulated in the universe of analytic vector fields, see [28] for more details.

$(n-1)$ -Tori

The $(n-1)$ -parameter families of invariant $(n-1)$ -tori organize the dynamics of an integrable Hamiltonian system in n degrees of freedom, and under small perturbations the parameter space of persisting analytic tori is Cantorized. This still allows for a global understanding of a substantial part of the dynamics, but also leads to additional questions.

A hyperbolic invariant torus \mathbb{T}^{n-1} has its Floquet exponents off the imaginary axis. Note that \mathbb{T}^{n-1} is not a normally hyperbolic manifold. Indeed, the normal linear behavior involves the $n-1$ zero eigenvalues in the direction of the parametrizing actions as well; similar to (9) the format

$$\begin{aligned}\dot{x} &= \omega(y) + O(y) + O(z^2) \\ \dot{y} &= O(y) + O(z^3) \\ \dot{z} &= \Omega(y)z + O(z^2)\end{aligned}$$

in Floquet coordinates yields an x -independent matrix Ω that describes the symplectic normal linear behavior, cf. [29]. The union $\{z=0\}$ over the family of $(n-1)$ -tori is a normally hyperbolic manifold and constitutes the center manifold of \mathbb{T}^{n-1} . Separatrices splitting yields the dividing surfaces in the sense of Wiggins et al. [138].

The persistence of elliptic tori under perturbation from an integrable system involves not only the internal frequencies of \mathbb{T}^{n-1} , but also the normal frequencies. Next

to the internal resonances the necessary Diophantine conditions (10) exclude the normal-internal resonances

$$\langle k, \omega \rangle = \alpha_j \quad (11)$$

$$\langle k, \omega \rangle = 2\alpha_j \quad (12)$$

$$\langle k, \omega \rangle = \alpha_i + \alpha_j \quad (13)$$

$$\langle k, \omega \rangle = \alpha_i - \alpha_j. \quad (14)$$

The first three resonances lead to the quasi-periodic center-saddle bifurcation studied in Sect. “[Transition to Chaos and Turbulence](#)”, the frequency-halving (or quasi-periodic period doubling) bifurcation and the quasi-periodic Hamiltonian Hopf bifurcation, respectively. The resonance (14) generalizes an equilibrium in $1:1$ resonance whence \mathbb{T}^{n-1} persists and remains elliptic, cf. [78]. When passing through resonances (12) and (13) the lower-dimensional tori lose ellipticity and acquire hyperbolic Floquet exponents. Elliptic $(n-1)$ -tori have a single normal frequency whence (11) and (12) are the only normal-internal resonances. See [35] for a thorough treatment of the ensuing possibilities.

The restriction to a single normal-internal resonance is dictated by our present possibilities. Indeed, already the bifurcation of equilibria with a fourfold zero eigenvalue leads to unfoldings that simultaneously contain all possible normal resonances. Thus, a satisfactory study of such tori which already may form one-parameter families in integrable Hamiltonian systems with five degrees of freedom has to await further progress in local bifurcation theory.

Transition to Chaos and Turbulence

One of the main interests over the second half of the twentieth century has been the transition between orderly and complicated forms of dynamics upon variation of either initial states or of system parameters. By ‘orderly’ we here mean equilibrium and periodic dynamics and by complicated quasi-periodic and chaotic dynamics, although we note that only chaotic dynamics is associated to unpredictability, e.g. see [27]. As already discussed in the introduction systems like a forced nonlinear oscillator or the planar three-body problem exhibit coexistence of periodic, quasi-periodic and chaotic dynamics, also compare with Fig. 1.

Similar remarks go for the onset of turbulence in fluid dynamics. Around 1950 this led to the scenario of Hopf-Landau-Lifschitz [75,76,83,84], which roughly amounts to the following. Stationary fluid motion corresponds to an equilibrium point in an ∞ -dimensional state space

of velocity fields. The first transition is a Hopf bifurcation [66,75,82], where a periodic solution branches off. In a second transition of similar nature a quasi-periodic two-torus branches off, then a quasi-periodic three-torus, etc. The idea is that the motion picks up more and more frequencies and thus obtains an increasingly complicated power spectrum. In the early 1970s this idea was modified in the Ruelle–Takens route to turbulence, based on the observation that, for flows, a three-torus can carry chaotic (or strange) attractors [112,126], giving rise to a broad band power spectrum. By the quasi-periodic bifurcation theory [24,29,31] as sketched below these two approaches are unified in a generic way, keeping track of measure theoretic aspects. For general background in dynamical systems theory we refer to [27,79].

Another transition to chaos was detected in the quadratic family of interval maps

$$f_{\mu}(x) = \mu x(1 - x),$$

see [58,99,101], also for a holomorphic version. This transition consists of an infinite sequence of period doubling bifurcations ending up in chaos; it has several universal aspects and occurs persistently in families of dynamical systems. In many of these cases also homoclinic bifurcations show up, where sometimes the transition to chaos is immediate when parameters cross a certain boundary, for general theory see [13,14,30,117]. There exist quite a number of case studies where all three of the above scenarios play a role, e.g., see [32,33,46] and many of their references.

Quasi-periodic Bifurcations

For the classical bifurcations of equilibria and periodic orbits, the bifurcation sets and diagrams are generally determined by a classical geometry in the product of phase space and parameter space as already established by, e.g., [8,133], often using singularity theory. Quasi-periodic bifurcation theory concerns the extension of these bifurcations to invariant tori in nearly-integrable systems, e.g., when the tori lose their normal hyperbolicity or when certain (strong) resonances occur. In that case the dense set of resonances, also responsible for the small divisors, leads to a Cantorization of the classical geometries obtained from Singularity Theory [29,35,37,38,39,41,44,45,48,49,67,68,69], also see [24,31,52,55]. Broadly speaking, one could say that in these cases the Preparation Theorem [133] is partly replaced by KAM theory. Since the KAM theory has been developed in several settings with or without preservation of structure, see Sect. “KAM Theory: An

Overview”, for the ensuing quasi-periodic bifurcation theory the same holds.

Hamiltonian Cases To fix thoughts we start with an example in the Hamiltonian setting, where a robust model for the quasi-periodic center-saddle bifurcation is given by

$$\begin{aligned} H_{\omega_1, \omega_2, \mu, \varepsilon}(I, \varphi, p, q) \\ = \omega_1 I_1 + \omega_2 I_2 + \frac{1}{2} p^2 + V_{\mu}(q) + \varepsilon f(I, \varphi, p, q) \end{aligned} \quad (15)$$

with $V_{\mu}(q) = \frac{1}{3} q^3 - \mu q$, compare with [67,69]. The unperturbed (or integrable) case $\varepsilon = 0$, by factoring out the \mathbb{T}^2 -symmetry, boils down to a standard center-saddle bifurcation, involving the fold catastrophe [133] in the potential function $V = V_{\mu}(q)$. This results in the existence of two invariant two-tori, one elliptic and the other hyperbolic. For $0 \neq |\varepsilon| \ll 1$ the dense set of resonances complicates this scenario, as sketched in Fig. 6, determined by the Diophantine conditions

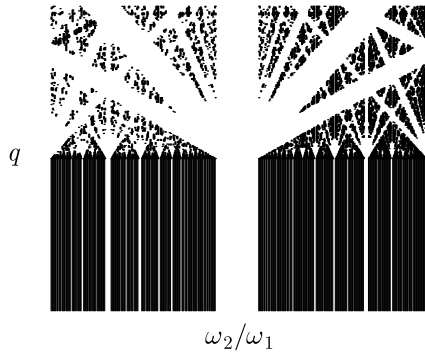
$$\begin{aligned} |\langle k, \omega \rangle| &\geq \gamma |k|^{-\tau}, \quad \text{for } q < 0, \\ |\langle k, \omega \rangle + \ell \beta(q)| &\geq \gamma |k|^{-\tau}, \quad \text{for } q > 0 \end{aligned} \quad (16)$$

for all $k \in \mathbb{Z}^n \setminus \{0\}$ and for all $\ell \in \mathbb{Z}$ with $|\ell| \leq 2$. Here $\beta(q) = \sqrt{2q}$ is the normal frequency of the elliptic torus given by $q = \sqrt{\mu}$ for $\mu > 0$. As before, (cf. Sects. “Invariant Curves of Planar Diffeomorphisms”, “KAM Theory: An Overview”), this gives a Cantor set of positive measure [24,29,31,45,69,105,106].

For $0 < |\varepsilon| \ll 1$ Fig. 6 will be distorted by a near-identity diffeomorphism; compare with the formulations of the Theorems 3 and 4. On the Diophantine Cantor set the dynamics is quasi-periodic, while in the gaps generically there is coexistence of periodicity and chaos, roughly comparable with Fig. 1, at left. The gaps at the border furthermore lead to the phenomenon of parabolic resonance, cf. [86].

Similar programs exist for all cuspid and umbilic catastrophes [37,39,68] as well as for the Hamiltonian Hopf bifurcation [38,44]. For applications of this approach see [35]. For a reversible analogue see [41]. As so often within the gaps generically there is an infinite regress of smaller gaps [11,35]. For theoretical background we refer to [29,45,106], for more references also see [24].

Dissipative Cases In the general dissipative case we basically follow the same strategy. Given the standard bifurcations of equilibria and periodic orbits, we get more complex situations when invariant tori are involved as well.



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 6

Sketch of the Cantorized Fold, as the bifurcation set of the quasi-periodic center-saddle bifurcation for $n = 2$ [67], where the horizontal axis indicates the frequency ratio $\omega_2 : \omega_1$, cf. (15). The lower part of the figure corresponds to hyperbolic tori and the upper part to elliptic ones. See the text for further interpretations

The simplest examples are the quasi-periodic saddle-node and quasi-periodic period doubling [29] also see [24,31].

To illustrate the whole approach let us start from the Hopf bifurcation of an equilibrium point of a vector field [66,75,82,116] where a hyperbolic point attractor loses stability and branches off a periodic solution, cf. Subsect. “Dissipative Perturbations”. A topological normal form is given by

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - (y_1^2 + y_2^2) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (17)$$

where $y = (y_1, y_2) \in \mathbb{R}^2$, ranging near $(0, 0)$. In this representation usually one fixes $\beta = 1$ and lets $\alpha = \mu$ (near 0) serve as a (bifurcation) parameter, classifying modulo topological equivalence. In polar coordinates (17) so gets the form

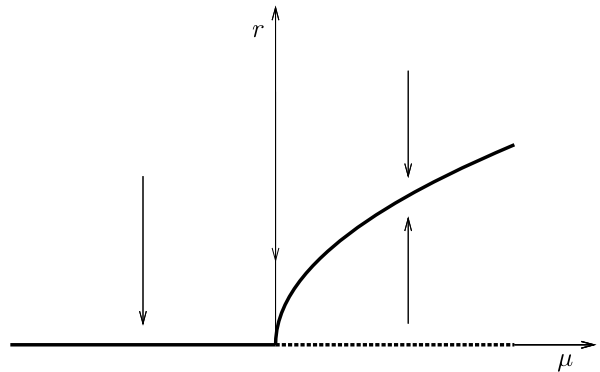
$$\begin{aligned} \dot{\varphi} &= 1, \\ \dot{r} &= \mu r - r^3. \end{aligned}$$

Figure 7 shows an amplitude response diagram (often called the bifurcation diagram). Observe the occurrence of the attracting periodic solution for $\mu > 0$ of amplitude $\sqrt{\mu}$.

Let us briefly consider the Hopf bifurcation for fixed points of diffeomorphisms. A simple example has the form

$$P(y) = e^{2\pi(\alpha + i\beta)y} + O(|y|^2), \quad (18)$$

$y \in \mathbb{C} \cong \mathbb{R}^2$, near 0. To start with β is considered a constant, such that β is not rational with denominator less



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 7

Bifurcation diagram of the Hopf bifurcation

than five, see [7,132], and where $O(|y|^2)$ should contain generic third order terms. As before, we let $\alpha = \mu$ serve as a bifurcation parameter, varying near 0. On one side of the bifurcation value $\mu = 0$, this system has by normal hyperbolicity and [73], an invariant circle. Here, due to the invariance of the rotation numbers of the invariant circles, no topological stability can be obtained [111]. Still this bifurcation can be characterized by many persistent properties. Indeed, in a generic two-parameter family (18), say with both α and β as parameters, the periodicity in the parameter plane is organized in resonance tongues [7,34,82]. (The tongue structure is hardly visible when only one parameter, like α , is used.) If the diffeomorphism is the return map of a periodic orbit for flows, this bifurcation produces an invariant two-torus. Usually this counterpart for flows is called Neimark–Sacker bifurcation. The periodicity as it occurs in the resonance tongues, for the vector field is related to phase lock. The tongues are contained in gaps of a Cantor set of quasi-periodic tori with Diophantine frequencies. Compare the discussion in Subsect. “Circle Maps”, in particular also regarding the Arnold family and Fig. 3. Also see Sect. “KAM Theory: An Overview” and again compare with [115].

Quasi-periodic versions exist for the saddle-node, the period doubling and the Hopf bifurcation. Returning to the setting with $\mathbb{T}^n \times \mathbb{R}^m$ as the phase space, we remark that the quasi-periodic saddle-node and period doubling already occur for $m = 1$, or in an analogous center manifold. The quasi-periodic Hopf bifurcation needs $m \geq 2$. We shall illustrate our results on the latter of these cases, compare with [19,31]. For earlier results in this direction see [52]. Our phase space is $\mathbb{T}^n \times \mathbb{R}^2 = \{x(\bmod 2\pi), y\}$, where we are dealing with the parallel invariant torus $\mathbb{T}^n \times \{0\}$. In the integrable case, by \mathbb{T}^n -symmetry we can

reduce to $\mathbb{R}^2 = \{y\}$ and consider the bifurcations of relative equilibria. The present interest is with small non-integrable perturbations of such integrable models.

We now discuss the *quasi-periodic Hopf bifurcation* [17,29], largely following [55]. The unperturbed, integrable family $X = X_\mu(x, y)$ on $\mathbb{T}^n \times \mathbb{R}^2$ has the form

$$\begin{aligned} X_\mu(x, y) \\ = [\omega(\mu) + f(y, \mu)]\partial_x + [\Omega(\mu)y + g(y, \mu)]\partial_y, \end{aligned} \quad (19)$$

where $f = O(|y|)$ and $g = O(|y|^2)$ as before. Moreover $\mu \in P$ is a multi-parameter and $\omega: P \rightarrow \mathbb{R}^n$ and $\Omega: P \rightarrow \text{gl}(2, \mathbb{R})$ are smooth maps. Here we take

$$\Omega(\mu) = \begin{pmatrix} \alpha(\mu) & -\beta(\mu) \\ \beta(\mu) & \alpha(\mu) \end{pmatrix},$$

which makes the ∂_y component of (19) compatible with the planar Hopf family (17). The present form of Kolmogorov non-degeneracy is Broer–Huiteima–Takens stability [29,42,45], requiring that there is a subset $\Gamma \subseteq P$ on which the map

$$\mu \in P \mapsto (\omega(\mu), \Omega(\mu)) \in \mathbb{R}^n \times \text{gl}(2, \mathbb{R})$$

is a submersion. For simplicity we even assume that μ is replaced by

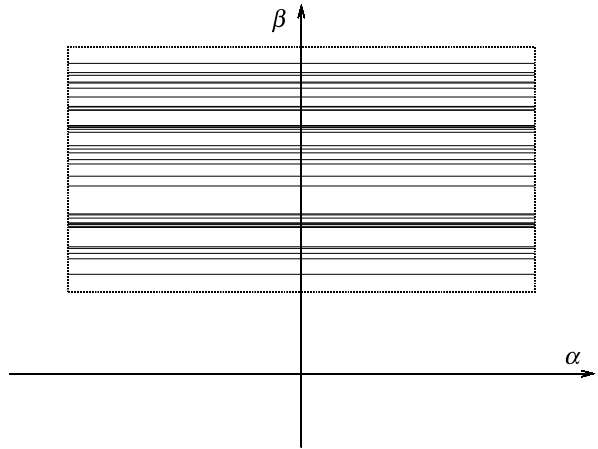
$$(\omega, (\alpha, \beta)) \in \mathbb{R}^n \times \mathbb{R}^2.$$

Observe that if the non-linearity g satisfies the well-known Hopf non-degeneracy conditions, e.g., compare [66,82], then the relative equilibrium $y = 0$ undergoes a standard planar Hopf bifurcation as described before. Here α again plays the role of bifurcation parameter and a closed orbit branches off at $\alpha = 0$. To fix thoughts we assume that $y = 0$ is attracting for $\alpha < 0$, and that the closed orbit occurs for $\alpha > 0$, and is attracting as well. For the integrable family X , qualitatively we have to multiply this planar scenario with \mathbb{T}^n , by which all equilibria turn into invariant attracting or repelling n -tori and the periodic attractor into an attracting invariant $(n+1)$ -torus. Presently the question is what happens to both the n - and the $(n+1)$ -tori, when we apply a small near-integrable perturbation.

The story runs much like before. Apart from the BHT non-degeneracy condition we require Diophantine conditions (10), defining the Cantor set

$$\begin{aligned} \Gamma_{\tau, \gamma}^{(2)} = \{(\omega, (\alpha, \beta)) \in \Gamma \mid | \langle k, \omega \rangle + \ell \beta | \geq \gamma |k|^{-\tau}, \\ \forall k \in \mathbb{Z}^n \setminus \{0\}, \forall \ell \in \mathbb{Z} \text{ with } |\ell| \leq 2\}, \end{aligned} \quad (20)$$

In Fig. 8 we sketch the intersection of $\Gamma_{\tau, \gamma}^{(2)} \subset \mathbb{R}^n \times \mathbb{R}^2$



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 8

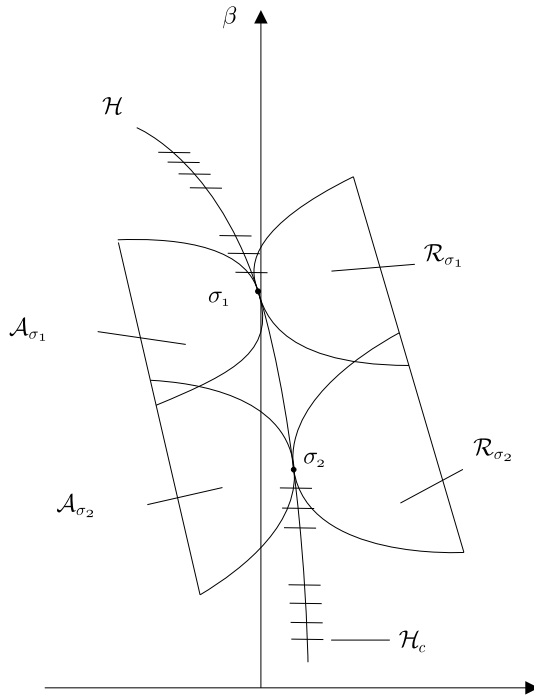
Planar section of the Cantor set $\Gamma_{\tau, \gamma}^{(2)}$

with a plane $\{\omega\} \times \mathbb{R}^2$ for a Diophantine (internal) frequency vector ω , cf. (7).

From [17,29] it now follows that for any family \tilde{X} on $\mathbb{T}^n \times \mathbb{R}^2 \times P$, sufficiently near X in the C^∞ -topology a near-identity C^∞ -diffeomorphism $\Phi: \mathbb{T}^n \times \mathbb{R}^2 \times \Gamma \rightarrow \mathbb{T}^n \times \mathbb{R}^2 \times \Gamma$ exists, defined near $\mathbb{T}^n \times \{0\} \times \Gamma$, that conjugates X to \tilde{X} when further restricting to $\mathbb{T}^n \times \{0\} \times \Gamma_{\tau, \gamma}^{(2)}$. So this means that the Diophantine quasi-periodic invariant n -tori are persistent on a diffeomorphic image of the Cantor set $\Gamma_{\tau, \gamma}^{(2)}$, compare with the formulations of the Theorems 3 and 4.

Similarly we can find invariant $(n+1)$ -tori. We first have to develop a \mathbb{T}^{n+1} symmetric normal form approximation [17,29] and ▶ [Normal Forms in Perturbation Theory](#). For this purpose we extend the Diophantine conditions (20) by requiring that the inequality holds for all $|\ell| \leq N$ for $N = 7$. We thus find another large Cantor set, again see Fig. 8, where Diophantine quasi-periodic invariant $(n+1)$ -tori are persistent. Here we have to restrict to $\alpha > 0$ for our choice of the sign of the normal form coefficient, compare with Fig. 7.

In both the cases of n -tori and of $(n+1)$ -tori, the nowhere dense subset of the parameter space containing the tori can be fattened by normal hyperbolicity to open subsets. Indeed, the quasi-periodic n - and $(n+1)$ -tori are infinitely normally hyperbolic [73]. Exploiting the normal form theory [17,29] and ▶ [Normal Forms in Perturbation Theory](#) to the utmost and using a more or less standard contraction argument [17,53], a fattening of the parameter domain with invariant tori can be obtained that leaves out only small ‘bubbles’ around the resonances, as sketched



Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 9

Fattening by normal hyperbolicity of a nowhere dense parameter set with invariant n -tori in the perturbed system. The curve \mathcal{H} is the Whitney smooth (even Gevrey regular [136]) image of the β -axis in Fig. 8. \mathcal{H} interpolates the Cantor set \mathcal{H}_c that contains the non-hyperbolic Diophantine quasi-periodic invariant n -tori, corresponding to $\Gamma_{\tau, \gamma}^{(2)}$, see (20). To the points $\sigma_{1,2} \in \mathcal{H}_c$ discs $\mathcal{A}_{\sigma_{1,2}}$ are attached where we find attracting normally hyperbolic n -tori and similarly in the discs $\mathcal{R}_{\sigma_{1,2}}$ repelling ones. The contact between the disc boundaries and \mathcal{H} is infinitely flat [17,29]

and explained in Fig. 9 for the n -tori. For earlier results in the same spirit in a case study of the quasi-periodic saddle-node bifurcation see [49,50,51], also compare with [11].

A Scenario for the Onset of Turbulence

Generally speaking, in many settings quasi-periodicity constitutes the order in between chaos [31]. In the Hopf–Landau–Lifschitz–Ruelle–Takens scenario [76,83,84,126] we may consider a sequence of typical transitions as given by quasi-periodic Hopf bifurcations, starting with the standard Hopf or Hopf–Neimark–Sacker bifurcation as described before. In the gaps of the Diophantine Cantor sets generically there will be coexistence of periodicity, quasi-periodicity and chaos in infinite regress. As said earlier, period doubling sequences and homoclinic bifurcations may accompany this.

As an example consider a family of maps that undergoes a generic quasi-periodic Hopf bifurcation from circle to two-torus. It turns out that here the Cantorized fold of Fig. 6 is relevant, where now the vertical coordinate is a bifurcation parameter. Moreover compare with Fig. 3, where also variation of ε is taken into account. The Cantor set contains the quasi-periodic dynamics, while in the gaps we can have chaos, e. g., in the form of Hénon like strange attractors [46,112]. A fattening process as explained above, also can be carried out here.

Future Directions

One important general issue is the mathematical characterization of chaos and ergodicity in dynamical systems, in conservative, dissipative and in other settings. This is a tough problem as can already be seen when considering two-dimensional diffeomorphisms. In particular we refer to the still unproven ergodicity conjecture of [9] and to the conjectures around Hénon like attractors and the principle ‘Hénon everywhere’, compare with [22,32]. For a discussion see Subsect. “A Scenario for the Onset of Turbulence”. In higher dimension this problem is even harder to handle, e. g., compare with [46,47] and references therein. In the conservative case a related problem concerns a better understanding of Arnold diffusion.

Somewhat related to this is the analysis of dynamical systems without an explicit perturbation setting. Here numerical and symbolic tools are expected to become useful to develop computer assisted proofs in extended perturbation settings, diagrams of Lyapunov exponents, symbolic dynamics, etc. Compare with [128]. Also see [46,47] for applications and further reference. This part of the theory is important for understanding concrete models, that often are not given in perturbation format.

Regarding nearly-integrable Hamiltonian systems, several problems have to be considered. Continuing the above line of thought, one interest is the development of Hamiltonian bifurcation theory without integrable normal form and, likewise, of KAM theory without action angle coordinates [87]. One big related issue also is to develop KAM theory outside the perturbation format.

The previous section addressed persistence of Diophantine tori involved in a bifurcation. Similar to Cremer’s example in Subsect. “Cremer’s Example in Herman’s Version” the dynamics in the gaps between persistent tori displays new phenomena. A first step has been made in [86] where internally resonant parabolic tori involved in a quasi-periodic Hamiltonian pitchfork bifurcation are considered. The resulting large dynamical instabilities may be further amplified for tangent (or flat)

parabolic resonances, which fail to satisfy the iso-energetic non-degeneracy condition.

The construction of solenoids in [16,94] uses elliptic periodic orbits as starting points, the simplest example being the result of a period-doubling sequence. This construction should carry over to elliptic tori, where normal-internal resonances lead to encircling tori of the same dimension, while internal resonances lead to elliptic tori of smaller dimension and excitation of normal modes increases the torus dimension. In this way one might be able to construct solenoid-type invariant sets that are limits of tori with varying dimension.

Concerning the global theory of nearly-integrable torus bundles [43], it is of interest to understand the effects of quasi-periodic bifurcations on the geometry and its invariants. Also it is of interest to extend the results of [134] when passing to semi-classical approximations. In that case two small parameters play a role, namely Planck's constant as well as the distance away from integrability.

Bibliography

1. Abraham R, Marsden JE (1978) Foundations of Mechanics, 2nd edn. Benjamin
2. Arnold VI (1961) Small divisors I: On mappings of the circle onto itself. *Izv Akad Nauk SSSR Ser Mat* 25:21–86 (in Russian); English translation: *Am Math Soc Transl Ser* 2(46):213–284 (1965); Erratum: *Izv Akad Nauk SSSR Ser Mat* 28:479–480 (1964, in Russian)
3. Arnold VI (1962) On the classical perturbation theory and the stability problem of the planetary system. *Dokl Akad Nauk SSSR* 145:487–490
4. Arnold VI (1963) Proof of a theorem by A.N. Kolmogorov on the persistence of conditionally periodic motions under a small change of the Hamilton function. *Russ Math Surv* 18(5):9–36 (English; Russian original)
5. Arnold VI (1964) Instability of dynamical systems with several degrees of freedom. *Sov Math Dokl* 5:581–585
6. Arnold VI (1978) Mathematical Methods of Classical Mechanics, GTM 60. Springer, New York
7. Arnold VI (1983) Geometrical Methods in the Theory of Ordinary Differential Equations. Springer
8. Arnold VI (ed) (1994) Dynamical Systems V: Bifurcation Theory and Catastrophe Theory. *Encyclopedia of Mathematical Sciences*, vol 5. Springer
9. Arnold VI, Avez A (1967) Problèmes Ergodiques de la Mécanique classique, Gauthier-Villars; English edition: Arnold VI, Avez A (1968) Ergodic problems of classical mechanics. Benjamin
10. Arnold VI, Kozlov VV, Neishtadt AI (1988) Mathematical Aspects of Classical and Celestial Mechanics. In: Arnold VI (ed) *Dynamical Systems*, vol III. Springer
11. Baesens C, Guckenheimer J, Kim S, MacKay RS (1991) Three coupled oscillators: Mode-locking, global bifurcation and toroidal chaos. *Phys D* 49(3):387–475
12. Barrow-Green J (1997) Poincaré and the Three Body Problem. In: *History of Mathematics*, vol 11. Am Math Soc, Providence; London Math Soc, London
13. Benedicks M, Carleson L (1985) On iterations of $1 - ax^2$ on $(-1, 1)$. *Ann Math* 122:1–25
14. Benedicks M, Carleson L (1991) The dynamics of the Hénon map. *Ann Math* 133:73–169
15. Benettin G (2005) Physical applications of Nekhoroshev theorem and exponential estimates. In: Giorgilli A (ed) *Hamiltonian dynamics theory and applications*, Cetraro 1999, *Lecture Notes in Mathematics*, vol 1861. Springer, pp 1–76
16. Birkhoff BD (1935) Nouvelles recherches sur les systèmes dynamiques. *Mem Pont Acad Sci Novi Lyncae* 1(3):85–216
17. Braaksma BLJ, Broer HW (1987) On a quasi-periodic Hopf bifurcation. *Ann Inst Henri Poincaré, Anal non linéaire* 4(2): 115–168
18. Bricmont J (1996) Science of chaos or chaos in science? In: Gross PR, Levitt N, Lewis MW (eds) *The Flight from Science and Reason* (New York, 1995), *Ann New York Academy of Sciences*, vol 775. New York Academy of Sciences, New York, pp 131–175; Also appeared in: *Phys Mag* 17:159–208 (1995)
19. Broer HW (2003) Coupled Hopf-bifurcations: Persistent examples of n -quasiperiodicity determined by families of 3-jets. *Astérisque* 286:223–229
20. Broer HW (2004) KAM theory: the legacy of Kolmogorov's 1954 paper. *Bull Am Math Soc (New Series)* 41(4):507–521
21. Broer HW, Huitema GB (1991) A proof of the isoenergetic KAM-theorem from the "ordinary" one. *J Differ Equ* 90:52–60
22. Broer HW, Krauskopf B (2000) Chaos in periodically driven systems. In: Krauskopf B, Lenstra D (eds) *Fundamental Issues of Nonlinear Laser Dynamics*. American Institute of Physics Conference Proceedings 548:31–53
23. Broer HW, Roussarie R (2001) Exponential confinement of chaos in the bifurcation set of real analytic diffeomorphisms. In: Broer HW, Krauskopf B, Vegter G (eds) *Global Analysis of Dynamical Systems*, Festschrift dedicated to Floris Takens for his 60th birthday. Bristol and Philadelphia IOP, pp 167–210
24. Broer HW, Sevryuk MB (2007) KAM Theory: quasi-periodicity in dynamical systems. In: Broer HW, Hasselblatt B, Takens F (eds) *Handbook of Dynamical Systems*, vol 3. North-Holland (to appear)
25. Broer HW, Takens F (1989) Formally symmetric normal forms and genericity. *Dyn Rep* 2:36–60
26. Broer HW, Takens F (2007) Unicity of KAM tori. *Ergod Theory Dyn Syst* 27:713–724
27. Broer HW, Takens F (2008) *Dynamical Systems and Chaos*. To be published by Epsilon Uitgaven
28. Broer HW, Tangerman FM (1986) From a differentiable to a real analytic perturbation theory, applications to the Kupka Smale theorems. *Ergod Theory Dyn Syst* 6:345–362
29. Broer HW, Huitema GB, Takens F, Braaksma BLJ (1990) Unfoldings and bifurcations of quasi-periodic tori. In: *Memoir AMS*, vol 421. Amer Math Soc, Providence
30. Broer HW, Dumortier F, van Strien SJ, Takens F (1991) Structures in dynamics, finite dimensional deterministic studies. In: de Jager EM, van Groesen EWC (eds) *Studies in Mathematical Physics*, vol II. North-Holland
31. Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-Periodic Motions in Families of Dynamical Systems: Order amidst Chaos. In: *Lecture Notes in Mathematics*, vol 1645. Springer
32. Broer HW, Simó C, Tatjer JC (1998) Towards global models

- near homoclinic tangencies of dissipative diffeomorphisms. *Nonlinearity* 11(3):667–770
33. Broer HW, Simó C, Vitolo R (2002) Bifurcations and strange attractors in the Lorenz-84 climate model with seasonal forcing. *Nonlinearity* 15(4):1205–1267
 34. Broer HW, Golubitsky M, Vegter G (2003) The geometry of resonance tongues: a singularity theory approach. *Nonlinearity* 16:1511–1538
 35. Broer HW, Hanßmann H, Jorba À, Villanueva J, Wagener FOO (2003) Normal-internal resonances in quasi-periodically forced oscillators: a conservative approach. *Nonlinearity* 16:1751–1791
 36. Broer HW, Hoveijn I, Lunter G, Vegter G (2003) Bifurcations in Hamiltonian systems: Computing Singularities by Gröbner Bases. In: *Lecture Notes in Mathematics*, vol 1806. Springer
 37. Broer HW, Hanßmann H, You J (2005) Bifurcations of normally parabolic tori in Hamiltonian systems. *Nonlinearity* 18:1735–1769
 38. Broer HW, Hanßmann H, Hoo J, Naudot V (2006) Nearly-integrable perturbations of the Lagrange top: applications of KAM theory. In: Denteneer D, den Hollander F, Verbitskiy E (eds) *Dynamics & Stochastics: Festschrift in Honor of MS Keane* Lecture Notes, vol 48. Inst. of Math. Statistics, pp 286–303
 39. Broer HW, Hanßmann H, You J (2006) Umbilical torus bifurcations in Hamiltonian systems. *J Differ Equ* 222:233–262
 40. Broer HW, Naudot V, Roussarie R (2006) Catastrophe theory in Dulac unfoldings. *Ergod Theory Dyn Syst* 26:1–35
 41. Broer HW, Ciocci MC, Hanßmann H (2007) The quasi-periodic reversible Hopf bifurcation. In: Doedel E, Krauskopf B, Sanders J (eds) *Recent Advances in Nonlinear Dynamics: Theme section dedicated to André Vanderbauwhede*. Intern J Bifurc Chaos 17:2605–2623
 42. Broer HW, Ciocci MC, Hanßmann H, Vanderbauwhede A (2009) Quasi-periodic stability of normally resonant tori. *Phys D* 238:309–318
 43. Broer HW, Cushman RH, Fassò F, Takens F (2007) Geometry of KAM tori for nearly integrable Hamiltonian systems. *Ergod Theory Dyn Syst* 27(3):725–741
 44. Broer HW, Hanßmann H, Hoo J (2007) The quasi-periodic Hamiltonian Hopf bifurcation. *Nonlinearity* 20:417–460
 45. Broer HW, Hoo J, Naudot V (2007) Normal linear stability of quasi-periodic tori. *J Differ Equ* 232:355–418
 46. Broer HW, Simó C, Vitolo R (2008) The Hopf–Saddle-Node bifurcation for fixed points of 3D-diffeomorphisms, the Arnol'd resonance web. *Bull Belg Math Soc Simon Stevin* 15:769–787
 47. Broer HW, Simó C, Vitolo R (2008) The Hopf–Saddle-Node bifurcation for fixed points of 3D-diffeomorphisms, analysis of a resonance ‘bubble’. *Phys D Nonlinear Phenom* (to appear)
 48. Broer HW, Hanßmann H, You J (in preparation) On the destruction of resonant Lagrangean tori in Hamiltonian systems
 49. Chenciner A (1985) Bifurcations de points fixes elliptiques I, courbes invariantes. *Publ Math IHÉS* 61:67–127
 50. Chenciner A (1985) Bifurcations de points fixes elliptiques II, orbites périodiques et ensembles de Cantor invariants. *Invent Math* 80:81–106
 51. Chenciner A (1988) Bifurcations de points fixes elliptiques III, orbites périodiques de “petites” périodes et élimination résonnante des couples de courbes invariantes. *Publ Math IHÉS* 66:5–91
 52. Chenciner A, Iooss G (1979) Bifurcations de tores invariants. *Arch Ration Mech Anal* 69(2):109–198; 71(4):301–306
 53. Chow S-N, Hale JK (1982) *Methods of Bifurcation Theory*. Springer
 54. Chow S-N, Li C, Wang D (1994) *Normal Forms and Bifurcation of Planar Vector Fields*. Cambridge University Press, Cambridge
 55. Ciocci MC, Litvak-Hinenzon A, Broer HW (2005) Survey on dissipative KAM theory including quasi-periodic bifurcation theory based on lectures by Henk Broer. In: Montaldi J, Ratiu T (eds) *Geometric Mechanics and Symmetry: the Peyresq Lectures*, LMS Lecture Notes Series, vol 306. Cambridge University Press, Cambridge, pp 303–355
 56. Cushman RH, Bates LM (1997) *Global Aspects of Classical Integrable Systems*. Birkhäuser, Basel
 57. Cushman RH, Dullin HR, Giacobbe A, Holm DD, Joyeux M, Lynch P, Sadovskii DA and Zhilinskii BI (2004) CO₂ molecule as a quantum realization of the 1 : 1 : 2 resonant swing-spring with monodromy. *Phys Rev Lett* 93:024302
 58. Devaney RL (1989) *An Introduction to Chaotic Dynamical Systems*, 2nd edn. Addison-Wesley, Redwood City
 59. Diacu F, Holmes P (1996) *Celestial Encounters. The Origins of Chaos and Stability*. Princeton University Press, Princeton
 60. Duistermaat JJ (1980) On global action-angle coordinates. *Commun Pure Appl Math* 33:687–706
 61. Dumortier F, Roussarie R, Sotomayor J (1991) Generic 3-parameter families of vector fields, unfoldings of saddle, focus and elliptic singularities with nilpotent linear parts. In: Dumortier F, Roussarie R, Sotomayor J, Zoladek H (eds) *Bifurcations of Planar Vector Fields: Nilpotent Singularities and Abelian Integrals*. LNM 1480, pp 1–164
 62. Efsthafiou K (2005) Metamorphoses of Hamiltonian systems with symmetries. *LNM*, vol 1864. Springer, Heidelberg
 63. Féjóz J (2004) Démonstration du “théorème d’Arnold” sur la stabilité du système planétaire (d’après Herman). *Ergod Theory Dyn Syst* 24:1–62
 64. Gallavotti G, Bonetto F, Gentile G (2004) *Aspects of Ergodic, Qualitative and Statistical Theory of Motion*. Springer
 65. Gelfreich VG, Lazutkin VF (2001) Splitting of Separatrices: perturbation theory and exponential smallness. *Russ Math Surv* 56:499–558
 66. Guckenheimer J, Holmes P (1983) *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer
 67. Hanßmann H (1988) The quasi-periodic centre-saddle bifurcation. *J Differ Equ* 142:305–370
 68. Hanßmann H (2004) Hamiltonian Torus Bifurcations Related to Simple Singularities. In: Ladde GS, Medhin NG, Sambandham M (eds) *Dynamic Systems and Applications*, Atlanta 2003. Dynamic Publishers, pp 679–685
 69. Hanßmann H (2007) Local and Semi-Local Bifurcations in Hamiltonian Dynamical Systems – Results and Examples. In: *Lecture Notes in Mathematics*, vol 1893. Springer, Berlin
 70. Herman M (1977) Mesure de Lebesgue et nombre de rotation. In: Palis J, do Carmo M (eds) *Geometry and Topology*. In: *Lecture Notes in Mathematics*, vol 597. Springer, pp 271–293
 71. Herman MR (1979) Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. *Publ Math IHÉS* 49:5–233
 72. Hirsch MW (1976) *Differential Topology*. Springer
 73. Hirsch MW, Pugh CC, Shub M (1977) Invariant Manifolds. In: *Lecture Notes in Mathematics*, vol 583. Springer
 74. Hofer H, Zehnder E (1994) *Symplectic invariants and Hamiltonian dynamics*. Birkhäuser

75. Hopf E (1942) Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems. *Ber Math-Phys Kl Sächs Akad Wiss Leipzig* 94:1–22
76. Hopf E (1948) A mathematical example displaying features of turbulence. *Commun Appl Math* 1:303–322
77. Huygens C *Œuvres complètes de Christiaan Huygens*, (1888–1950), vol 5, pp 241–263 and vol 17, pp 156–189. Martinus Nijhoff, The Hague
78. de Jong HH (1999) Quasiperiodic breathers in systems of weakly coupled pendulums: Applications of KAM theory to classical and statistical mechanics. Ph.D. Thesis, Univ. Groningen
79. Katok A, Hasselblatt B (1995) *Introduction to the Modern Theory of Dynamical Systems*. Cambridge University Press, Cambridge
80. Kolmogorov AN (1954) On the persistence of conditionally periodic motions under a small change of the Hamilton function. *Dokl Akad Nauk SSSR* 98:527–530 (in Russian); English translation: *Stochastic Behavior in Classical and Quantum Hamiltonian Systems*, Volta Memorial Conference (Como, 1977). In: Casati G, Ford J (eds) *Lecture Notes in Physics*, vol 93. Springer, Berlin pp 51–56 (1979); Reprinted in: Bai Lin Hao (ed) *Chaos*. World Scientific, Singapore, pp 81–86 (1984)
81. Kolmogorov AN (1957) The general theory of dynamical systems and classical mechanics. In: Gerretsen JCH, de Groot J (eds) *Proceedings of the International Congress of Mathematicians*, vol 1 (Amsterdam, 1954), North-Holland, Amsterdam, pp 315–333 (in Russian); Reprinted in: *International Mathematical Congress in Amsterdam*, (1954) (Plenary Lectures). Fizmatgiz, Moscow, pp 187–208 (1961); English translation as Appendix D in: Abraham RH (1967) *Foundations of Mechanics*. Benjamin, New York, pp 263–279; Reprinted as Appendix in [1], pp 741–757
82. Kuznetsov YA (2004) *Elements of Applied Bifurcation Theory*, 3rd edn. In: *Applied Mathematical Sciences*, vol 112. Springer, New York
83. Landau LD (1944) On the problem of turbulence. *Akad Nauk* 44:339
84. Landau LD, Lifschitz EM (1959) *Fluid Mechanics*. Pergamon, Oxford
85. Laskar J (1995) Large scale chaos and marginal stability in the Solar System, XIth International Congress of Mathematical Physics (Paris, 1994). In: Iagolnitzer D (ed) *Internat Press*, Cambridge, pp 75–120
86. Litvak-Hinzenon A, Rom-Kedar V (2002) Parabolic resonances in 3 degree of freedom near-integrable Hamiltonian systems. *Phys D* 164:213–250
87. de la Llave R, González A, Jorba À, Villanueva J (2005) KAM theory without action-angle variables. *Nonlinearity* 18:855–895
88. Lochak P (1999) Arnold diffusion; a compendium of remarks and questions. In: Simó C (ed) *Hamiltonian systems with three or more degrees of freedom* (S'Agaró, 1995), NATO ASI Series C: *Math Phys Sci*, vol 533. Kluwer, Dordrecht, pp 168–183
89. Lochak P, Marco J-P (2005) Diffusion times and stability exponents for nearly integrable analytic systems, *Central Eur J Math* 3:342–397
90. Lochak P, Neishtadt AI (1992) Estimates of stability time for nearly integrable systems with a quasiconvex Hamiltonian. *Chaos* 2:495–499
91. Lukina O (2008) *Geometry of torus bundles in Hamiltonian systems*, Ph.D. Thesis, Univ. Groningen
92. MacKay RS (1993) *Renormalisation in area-preserving maps*. World Scientific
93. Marco J-P, Sauzin D (2003) Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian systems. *Publ Math Inst Hautes Etud Sci* 96:199–275
94. Markus L, Meyer KR (1980) Periodic orbits and solenoids in generic Hamiltonian dynamical systems. *Am J Math* 102: 25–92
95. Matveev VS (1996) Integrable Hamiltonian systems with two degrees of freedom. Topological structure of saturated neighborhoods of points of focus-focus and saddle-saddle types. *Sb Math* 187:495–524
96. McDuff D, Salamon D (1995) *Introduction to Symplectic Geometry*. Clarendon/Oxford University Press
97. Meyer KR, Hall GR (1992) *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*. In: *Applied Mathematical Sciences*, vol 90. Springer
98. Meiss JD (1986) Class renormalization: Islands around islands. *Phys Rev A* 34:2375–2383
99. de Melo W, van Strien SJ (1991) *One-Dimensional Dynamics*. Springer
100. Mielke A (1991) *Hamiltonian and Lagrangian Flows on Center Manifolds – with Applications to Elliptic Variational Problems*. In: *Lecture Notes in Mathematics*, vol 1489. Springer
101. Milnor JW (2006) *Dynamics in One Complex Variable*, 3rd edn. In: *Ann. Math. Studies*, vol 160. Princeton University Press, Princeton
102. Morbidelli A, Giorgilli A (1995) Superexponential Stability of KAM Tori. *J Stat Phys* 78:1607–1617
103. Morbidelli A, Giorgilli A (1995) On a connection between KAM and Nekhoroshev's theorems. *Physica D* 86:514–516
104. Moser JK (1962) On invariant curves of area-preserving mappings of an annulus. *Nachr Akad Wiss Göttingen II, Math-Phys Kl* 1:1–20
105. Moser JK (1966) On the theory of quasiperiodic motions. *SIAM Rev* 8(2):145–172
106. Moser JK (1967) Convergent series expansions for quasi-periodic motions. *Math Ann* 169:136–176
107. Moser JK (1968) Lectures on Hamiltonian systems. *Mem Am Math Soc* 81:1–60
108. Moser JK (1973) Stable and random motions in dynamical systems, with special emphasis to celestial mechanics. In: *Ann. Math. Studies*, vol 77. Princeton University Press, Princeton
109. Nekhoroshev NN (1977) An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems. *Russ Math Surv* 32:1–65
110. Nekhoroshev NN (1985) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems II. In: Oleinik OA (ed) *Topics in Modern Mathematics*, Petrovskii Seminar No.5. Consultants Bureau, pp 1–58
111. Newhouse SE, Palis J, Takens F (1983) Bifurcations and stability of families of diffeomorphisms. *Publ Math IHÉS* 57:5–71
112. Newhouse SE, Ruelle D, Takens F (1978) Occurrence of strange Axiom A attractors near quasi-periodic flows on \mathbb{T}^m , $m \geq 3$. *Commun Math Phys* 64:35–40
113. Niederman L (2004) Prevalence of exponential stability among nearly-integrable Hamiltonian systems. *Ergod Theory Dyn Syst* 24(2):593–608
114. Nguyen Tien Zung (1996) Kolmogorov condition for integrable systems with focus-focus singularities. *Phys Lett A* 215(1/2):40–44

115. Oxtoby J (1971) *Measure and Category*. Springer
116. Palis J, de Melo M (1982) *Geometric Theory of Dynamical Systems*. Springer
117. Palis J, Takens F (1993) *Hyperbolicity & Sensitive Chaotic Dynamics at Homoclinic Bifurcations*. Cambridge University Press, Cambridge
118. Poincaré H (1980) Sur le problème des trois corps et les équations de la dynamique. *Acta Math* 13:1–270
119. Pöschel J (1982) Integrability of Hamiltonian systems on Cantor sets. *Commun Pure Appl Math* 35(5):653–696
120. Pöschel J (1993) Nekhoroshev estimates for quasi-convex Hamiltonian systems. *Math Z* 213:187–216
121. Pöschel J (2001) A lecture on the classical KAM Theorem. In: *Proc Symp Pure Math* 69:707–732
122. Rink BW (2004) A Cantor set of tori with monodromy near a focus-focus singularity. *Nonlinearity* 17:347–356
123. Robinson C (1995) *Dynamical Systems*. CRC Press
124. Roussarie R (1997) Smoothness properties of bifurcation diagrams. *Publ Mat* 41:243–268
125. Ruelle D (1989) *Elements of Differentiable Dynamics and Bifurcation Theory*. Academic Press
126. Ruelle D, Takens F (1971) On the nature of turbulence. *Commun Math Phys* 20:167–192; 23:343–344
127. Sevryuk MB (2007) Invariant tori in quasi-periodic non-autonomous dynamical systems via Herman’s method. *DCDS-A* 18(2/3):569–595
128. Simó C (2001) Global dynamics and fast indicators. In: Broer HW, Krauskopf B, Vegter G (eds) *Global Analysis of Dynamical Systems*, *Festschrift dedicated to Floris Takens for his 60th birthday*. IOP, Bristol and Philadelphia, pp 373–390
129. Spivak M (1970) *Differential Geometry*, vol I. Publish or Perish
130. Takens F (1973) Introduction to Global Analysis. *Comm. 2 of the Math. Inst. Rijksuniversiteit Utrecht*
131. Takens F (1974) Singularities of vector fields. *Publ Math IHÉS* 43:47–100
132. Takens F (1974) Forced oscillations and bifurcations. In: *Applications of Global Analysis I*, *Comm 3 of the Math Inst Rijksuniversiteit Utrecht* (1974); In: Broer HW, Krauskopf B, Vegter G (eds) *Global Analysis of Dynamical Systems*, *Festschrift dedicated to Floris Takens for his 60th birthday*. IOP, Bristol and Philadelphia, pp 1–62
133. Thom R (1989) *Structural Stability and Morphogenesis*. An Outline of a General Theory of Models, 2nd edn. Addison-Wesley, Redwood City (English; French original)
134. Vũ Ngọc San (1999) Quantum monodromy in integrable systems. *Commun Math Phys* 203:465–479
135. Waalkens H, Junge A, Dullin HR (2003) Quantum monodromy in the two-centre problem. *J Phys A Math Gen* 36:L307–L314
136. Wagener FOO (2003) A note on Gevrey regular KAM theory and the inverse approximation lemma. *Dyn Syst* 18:159–163
137. Wiggins S (1990) *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer
138. Wiggins S, Wiesenfeld L, Jaffe C, Uzer T (2001) Impenetrable barriers in phase-space. *Phys Rev Lett* 86(24):5478–5481
139. Yoccoz J-C (1983) C^1 -conjugaisons des difféomorphismes du cercle. In: Palis J (ed) *Geometric Dynamics*, *Proceedings, Rio de Janeiro* (1981) *Lecture Notes in Mathematics*, vol 1007, pp 814–827
140. Yoccoz J-C (1992) Travaux de Herman sur les tores invariants. In: *Séminaire Bourbaki*, vol 754, 1991–1992. *Astérisque* 206:311–344
141. Yoccoz J-C (1995) Théorème de Siegel, nombres de Bruno et polynômes quadratiques. *Astérisque* 231:3–88
142. Yoccoz J-C (2002) Analytic linearization of circle diffeomorphisms. In: Marmi S, Yoccoz J-C (eds) *Dynamical Systems and Small Divisors*, *Lecture Notes in Mathematics*, vol 1784. Springer, pp 125–174
143. Zehnder E (1974) An implicit function theorem for small divisor problems. *Bull Am Math Soc* 80(1):174–179
144. Zehnder E (1975) Generalized implicit function theorems with applications to some small divisor problems, I and II. *Commun Pure Appl Math* 28(1):91–140; (1976) 29(1):49–111

Hamilton–Jacobi Equations and Weak KAM Theory

ANTONIO SICONOLFI

Dip. di Matematica, “La Sapienza” Università di Roma, Roma, Italy

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Subsolutions](#)

[Solutions](#)

[First Regularity Results for Subsolutions](#)

[Critical Equation and Aubry Set](#)

[An Intrinsic Metric](#)

[Dynamical Properties of the Aubry Set](#)

[Long-Time Behavior of Solutions to the Time-Dependent Equation](#)

[Main Regularity Result](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Hamilton–Jacobi equations This class of first-order partial differential equations has a central relevance in several branches of mathematics, both from a theoretical and an application point of view. It is of primary importance in classical mechanics, Hamiltonian dynamics, Riemannian and Finsler geometry, and optimal control theory, as well. It furthermore appears in the classical limit of the Schrödinger equation. A connection with Hamilton’s equations, in the case where the Hamiltonian has sufficient regularity, is provided by the classical Hamilton–Jacobi method which shows that the graph of the differential of any regular, say C^1 ,

global solution to the equation is an invariant subset for the corresponding Hamiltonian flow. The drawback of this approach is that such regular solutions do not exist in general, even for very regular Hamiltonians. See the next paragraph for more comments on this issue.

Viscosity solutions As already pointed out, Hamilton–Jacobi equations do not have in general *global* classical solutions, i. e. everywhere differentiable functions satisfying the equation pointwise. The method of characteristics just yields *local* classical solutions. This explains the need of introducing weak solutions. The idea for defining those of viscosity type is to consider C^1 functions whose graph, up to an additive constant, touches that of the candidate solution at a point and then stay locally above (resp. below) it. These are the viscosity test functions, and it is required that the Hamiltonian satisfies suitable inequalities when its first-order argument is set equal to the differential of them at the first coordinate of the point of contact. Similarly it is defined the notion of viscosity sub, supersolution. Clearly a viscosity solution satisfies pointwise the equation at any differentiability points. A peculiarity of the definition is that a viscosity solution can admit no test function at some point, while the nonemptiness of both classes of test functions is equivalent to the solution being differentiable at the point. Nevertheless powerful existence, uniqueness and stability results hold in the framework of viscosity solution theory. The notion of viscosity solutions was introduced by Crandall and Lions at the beginning of the 1980s. We refer to Bardi and Capuzzo Dolcetta [2], Barles [3], Koike [24] for a comprehensive treatment of this topic.

Semiconcave and semiconvex functions These are the appropriate regularity notions when working with viscosity solution techniques. The definition is given by requiring some inequalities, involving convex combinations of points, to hold. These functions possess viscosity test functions of one of the two types at any point. When the Hamiltonian enjoys coercivity properties ensuring that any viscosity solution is locally Lipschitz-continuous then a semiconcave or semiconcave function is the solution if and only if it is classical solution almost everywhere, i. e. up to a set of zero Lebesgue measure.

Metric approach This method applies to stationary Hamilton–Jacobi equations with the Hamiltonian only depending on the state and momentum variable. This consists of defining a length functional, on the set of Lipschitz-continuous curves, related to the corre-

sponding sublevels of the Hamiltonian. The associated length distance, obtained by performing the infimum of the intrinsic length of curves joining two given points, plays a crucial role in the analysis of the equation and, in particular, enters in representation formulae for any viscosity solution. One important consequence is that only the sublevels of the Hamiltonian matter for determining such solutions. Accordingly the convexity condition on the Hamiltonian can be relaxed, just requiring quasiconvexity, i. e. convexity of sublevels. Note that in this case the metric is of Finsler type and the sublevels are the unit cotangent balls of it.

Critical equations To any Hamiltonian is associated a one-parameter family of Hamilton–Jacobi equations obtained by fixing a constant level of Hamiltonian. When studying such a family, one comes across a threshold value under which no subsolutions may exist. This is called the critical value and the same name is conferred to the corresponding equation. If the ground space is compact then the critical equation is unique among those of the family for which viscosity solutions do exist. When, in particular, the underlying space is a torus or, in other terms, the Hamiltonian is \mathbb{Z}^N -periodic then such functions play the role of correctors in related homogenization problems.

Aubry set The analysis of the critical equation shows that the obstruction for getting subsolutions at subcritical levels is concentrated on a special set of the ground space, in the sense that no critical subsolution can be strict around it. This is precisely the Aubry set. This is somehow compensated by the fact that critical subsolutions enjoy extra regularity properties on the Aubry set.

Definition of the Subject

The article aims to illustrate some applications of weak KAM theory to the analysis of Hamilton–Jacobi equations. The presentation focuses on two specific problems, namely the existence of C^1 classical subsolutions for a class of stationary (i. e. independent of the time) Hamilton–Jacobi equations, and the long-time behavior of viscosity solutions of an evolutive version of it.

The Hamiltonian is assumed to satisfy mild regularity conditions, under which the corresponding Hamilton equations cannot be written. Consequently PDE techniques will be solely employed in the analysis, since the powerful tools of the Hamiltonian dynamics are not available.

Introduction

Given a continuous or more regular Hamiltonian $H(x, p)$ defined on the cotangent bundle of a boundaryless manifold M , where x and p are the state and the momentum variable, respectively, and satisfying suitable convexity and coercivity assumptions, is considered the family of Hamilton–Jacobi equations

$$H(x, Du) = a \quad x \in M, \quad (1)$$

with a a real parameter, as well as the time-dependent version

$$w_t + H(x, Dw) = 0 \quad x \in M, \quad t \in (0, +\infty), \quad (2)$$

where Du and u_t stand for the derivative with respect to state and time variable, respectively. As a matter of fact, it will be set, for the sake of simplicity, to either $M = \mathbb{R}^N$ (noncompact case) or $M = \mathbb{T}^N$ (compact case), where \mathbb{T}^N indicates the flat torus endowed with the Euclidean metric and with the cotangent bundle identified to $\mathbb{T}^N \times \mathbb{R}^N$.

The main scope of this article is to study the existence of the C^1 classical subsolution to (1), and the long-time behavior of viscosity solutions to (2) by essentially employing tools issued from weak KAM theory.

Some of the results that will be outlined are valid with the additional assumption of compactness for the underlying manifold, in particular those concerning the asymptotics of solutions to (2).

For both issues it is crucial to perform a qualitative analysis of (1) for a known value of the parameter a , qualified as *critical*; accordingly Eq. (1) is called critical when a is equal to the critical value. This analysis leads to detection of a special closed subset of the ground space, named after Aubry, where any locally Lipschitz-continuous subsolution to (1) enjoys some additional regularity properties, and behaves in a peculiar way. This will have a central role in the presentation.

The requirements on H will be strengthened to obtain some theorems, but we remain in a setting where the corresponding Hamilton equations cannot be written. Consequently PDE techniques will be solely employed in the analysis, since the powerful tools of Hamiltonian dynamics are not available.

Actually the notion of critical value was independently introduced by Ricardo Mañé at the beginning of the 1980s, in connection with the analysis of integral curves of the Euler–Lagrange flow enjoying some global minimizing properties, and by P.L. Lions, S.R.S. Varadhan and G. Papanicolaou [25] in 1987 in the framework of viscosity so-

lutions theory, for studying the periodic homogenization of Hamilton–Jacobi equations.

The Aubry set was determined and analyzed by Serge Aubry, in a pure dynamical way, as the union of the supports of integral curves of Euler–Lagrange flow possessing suitable minimality properties. John Mather defined (1986), in a more general framework, a set, contained in the Aubry set, starting from special probability measures invariant with respect to the flow. See Contreras and Iturriaga, [9], for an account on this theory.

The first author to point out the link between Aubry–Mather theory and weak solutions to the critical Hamilton–Jacobi equation was Albert Fathi, see [16,17], with the so-called weak KAM theory (1996); he thoroughly investigated the PDE counterpart of the dynamical phenomena occurring at the critical level. However, his investigation is still within the framework of the dynamical systems theory, requires the Hamiltonian to be at least C^2 , and requires the existence of associated Hamiltonian flow as well.

The work of Fathi and Siconolfi (2005) [19,20], completely bypassed such assumptions and provided a geometrical analysis of the critical equation independent of the flow, which made it possible to deal with nonregular Hamiltonians. The new idea being the introduction of a length functional, intrinsically related to H , for any curve, and of the related distance, as well. The notion of the Aubry set was suitably generalized to this broader setting.

Other important contributions in bridging the gap between PDE and the dynamical viewpoint have been made by Evans and Gomes [14,15].

The material herein is organized as follows: the Sects. “Subsolutions”, “Solutions” are of introductory nature and illustrate the notions of viscosity (sub)solution with their basic properties. Some fundamental techniques used in this framework are introduced as well. Section “First Regularity Results for Subsolutions” deals with issues concerning regularity of subsolutions to (1). The key notion of the Aubry set is introduced in Sect. “Critical Equation and Aubry Set” in connection with the investigation of the critical equation, and a qualitative analysis of it, specially devoted to looking into metric and dynamical properties, is performed in Sects. “An Intrinsic Metric”, “Dynamical Properties of the Aubry Set”. Sections “Long-Time Behavior of Solutions to the Time-Dependent Equation”, “Main Regularity Result” present the main results relative to the long-time behavior of solutions to (2) and the existence of C^1 subsolutions to (1). Finally, Sect. “Future Directions” gives some ideas of possible developments in the topic.

Subsolutions

First I will detail the basic conditions postulated throughout the paper for H . Additional properties required for obtaining some particular result, will be introduced when needed. The Hamiltonian is assumed to be

$$\text{continuous in both variables,} \quad (3)$$

to satisfy the coercivity assumption

$$\{(x, p): H(y, p) \leq a\} \text{ is compact for any } a \quad (4)$$

and the following quasiconvexity conditions for any $x \in M$, $a \in \mathbb{R}$

$$\{p: H(x, p) \leq a\} \text{ is convex} \quad (5)$$

$$\partial\{p: H(x, p) \leq a\} = \{p: H(x, p) = a\} \quad (6)$$

where ∂ , in the above formula, indicates the boundary. The a -sublevel of the Hamiltonian appearing in (5) will be denoted by $Z_a(x)$. It is a consequence of the coerciveness and convexity assumptions on H that the set-valued map $x \mapsto Z_a(x)$ possesses convex compact values and, in force of (3), is upper semicontinuous; it is in addition continuous at any point x where $\text{int } Z_a(x) \neq \emptyset$. Here (semi)continuous must be understood with respect to the Hausdorff metric.

Next will be given four different definitions of weak subsolutions to Eq. (1) and their equivalence will be proved. From this it can be seen that the family of functions so detected is intrinsically related to the equation. As a matter of fact, it will be proved, under more stringent assumptions, that this family is the closure of the classical (i. e. C^1) subsolutions in the locally uniform topology.

Some notations and definitions must be preliminarily introduced. Given two continuous functions u and v , it is said that v is a (strict) *supertangent* to u at some point x_0 if such point is a (strict) local maximizer of $u - v$. The notion of *subtangent* is obtained by replacing a maximizer with a minimizer. Since (sub, super)tangents are involved in the definition of viscosity solution, they will be called in the sequel viscosity test functions. It is necessary to check:

Proposition 1 *Let u be a continuous function possessing both C^1 supertangents and subtangents at a point x_0 , then u is differentiable at x_0 .*

Recall that by Rademacher Theorem a locally Lipschitz function is differentiable almost everywhere (for short a.e.), with respect to the Lebesgue measure. For such a function w the (Clarke) generalized gradient at any

point x is defined by

$$\partial w(x) = \text{co}\{p = \lim_i D w(x_i) : x_i \text{ differentiability point of } u, \lim_i x_i = x\},$$

where co denotes the convex hull.

Remark 2 Record for later use that this set of weak derivatives can be retrieved even if the differentiability points are taken not in the whole ground space, but just outside a set of vanishing Lebesgue measure.

The generalized gradient is nonempty at any point; if it reduces to a singleton at some x , i. e. it is differentiable and Du is continuous at x . The set-valued function $x \mapsto \partial w(x)$ possesses convex compact values and is upper semicontinuous. The following variational property holds:

$$0 \in \partial w(x) \quad \text{at any local minimizer or maximizer of } w, \quad (7)$$

furthermore, if ψ is C^1 then

$$\partial(w - \psi)(x) = \partial w(x) - D\psi(x). \quad (8)$$

First definition of weak subsolution A function u is said to be an *a.e. subsolution* to (1) if it is locally Lipschitz-continuous and satisfies

$$H(x, Du(x)) \leq a$$

for x in a subset of M with full measure.

Second definition of weak subsolution A function u is said to be a *viscosity subsolution of first type* to (1) if it is continuous and

$$H(x_0, D\psi(x_0)) \leq a,$$

or equivalently

$$D\psi(x_0) \in Z_a(x_0),$$

for any $x_0 \in M$, any C^1 supertangent ψ to u at x_0 .

The previous definition can be equivalently rephrased by taking test functions of class C^k , $1 < k \leq +\infty$, or simply differentiable, instead of C^1 .

Third definition of weak subsolution A function u is a *viscosity subsolution of second type* if it satisfies the previous definition with subtangent in place of supertangent.

Fourth definition of weak subsolution A function u is a subsolution in the sense of Clarke if

$$\partial u(x) \subset Z_a(x) \quad \text{for all } x \in M.$$

Note that this last definition is unique based on a condition holding at any point of M , and not just at the differentiability points of u or at points where some test function exists. This fact will be exploited in the forthcoming definition of strict subsolution.

Proposition 3 *The previous four definitions are equivalent.*

We first show that the viscosity subsolutions of both types are locally Lipschitz-continuous. It is exploited that Z_a , being upper semicontinuous, is also locally bounded, namely for any bounded subset B of M there is a positive r with

$$Z_a(x) \subset B(0, r) \quad \text{for } x \in B, \quad (9)$$

where $B(0, r)$ is the Euclidean ball centered at 0 with radius r . The minimum r for which (9) holds true will be indicated by $|Z_a|_{\infty, B}$. The argument is given for a viscosity subsolution of first type, say u ; the proof for the others is similar.

Assume by contradiction that there is an open bounded domain B_1 where u is not Lipschitz-continuous, then consider an open bounded domain B_2 containing B_1 such that

$$\alpha := \inf\{|x - y|, x \in B_1, y \in \partial B_2\} > 0,$$

and choose an l_0 such that

$$|Z_a|_{\infty, B_2} < l_0 \quad (10)$$

$$\sup_{B_2} u - \inf_{B_1} u - l_0 \alpha < 0 \quad (11)$$

Since u is not Lipschitz-continuous on B_1 , a pair of points x_0, x_1 in B_1 can be found satisfying

$$u(x_1) - u(x_0) > l_0 |x_1 - x_0|, \quad (12)$$

which shows that the function $x \mapsto u(x) - u(x_0) - l_0 |x - x_0|$, has a positive supremum in B_2 . On the other hand such a function is negative on ∂B_2 by (11), and so it attains its maximum in B_2 at a point $\bar{x} \in B_2$, or, in other terms, $x \mapsto u(x_0) + l_0 |x - x_0|$ is supertangent to u at $\bar{x} \neq x_0$. Consequently

$$l_0 \frac{x_1 - x_0}{|x_1 - x_0|} \in Z_a(x_1),$$

in contradiction with (10).

Since at every differentiability point u can be taken as a test function of itself, then any viscosity subsolution of both types is also an a.e. subsolution. The a.e. subsolutions, in turn, satisfy the fourth definition above, thanks to the definition of generalized gradient, Remark 2, and the fact that H is convex in p and continuous in both arguments.

Finally, exploit (7) and (8) to see that the differential of any viscosity test function to u at some point x_0 is contained in $\partial u(x_0)$. This shows that any subsolution in the Clarke sense is also a viscosity subsolution of the first and second type.

In view of Proposition 3, from now on any element of this class of functions will be called a subsolution of (1) without any further specification, similarly the notion of a subsolution in an open subset of M can be given. Note that for any subsolution u , any bounded open domain B , the quantity $|Z_a|_{\infty, B}$ is a Lipschitz constant in B for every subsolution, consequently the family of all subsolutions to (1) is locally equiLipschitz-continuous.

A conjugate Hamiltonian \check{H} can be associated to H , it is defined by

$$\check{H}(x, p) = H(x, -p) \quad \text{for any } x, p. \quad (13)$$

Note that \check{H} satisfies, as H does, assumptions (3)–(5). The two corresponding conjugate Hamilton–Jacobi equations have the same family of subsolutions, up to a change of sign, as is apparent looking at the first definition of subsolution.

Next we will have a closer look at the family of subsolutions to (1), denoted from now on by S_a ; the properties deduced will be exploited in the next sections. Advantage is taken of this to illustrate a couple of basic arguments coming from viscosity solutions theory.

Proposition 4 *The family S_a is stable with respect to the local uniform convergence.*

The key point in the proof of this result is to use the same C^1 function, at different points, for testing the limit as well as the approximating functions. This is indeed the primary trick to obtaining stability properties in the framework of viscosity solutions theory.

Let u_n be a sequence in S_a and $u_n \rightarrow u$ locally uniformly in M . Let ψ be a supertangent to u at some point x_0 , it can be assumed, without loss of generality, that ψ is a strict subgradient, by adding a suitable quadratic term. Therefore, there is a compact neighborhood U of x_0 where x_0 itself is the unique maximizer of $u - \psi$. Any sequence x_n of maximizers of $u_n - \psi$ in U converges to a maximizer of $u - \psi$, and so $x_n \rightarrow x_0$, and consequently lies in the interior of U for n large enough. In other terms ψ is supertangent to u_n at x_n , when n is sufficiently large.

Consequently $H(x_n, D\psi(x_n)) \leq a$, which implies, exploiting the continuity of the Hamiltonian and passing at the limit, $H(x_0, D\psi(x_0)) \leq a$, as desired.

Take into account the equiLipschitz character of subsolutions to (1), and the fact that the subsolution property is not affected by addition of a constant, to obtain by slightly adjusting the previous argument, and using the Ascoli Theorem:

Proposition 5 *Let $u_n \in S_{a_n}$, with a_n converging to some a . Then the sequence u_n converges to $u \in S_a$, up to addition of constants and extraction of a subsequence.*

Before ending the section, a notion which will have some relevance in what follows will be introduced.

A subsolution u is said to be *strict* in some open subset $\Omega \subset M$ if

$$\partial u(x) \subset \text{int } Z_a(x) \quad \text{for any } x \in \Omega,$$

where int stands for interior. Since the multivalued map $x \mapsto \partial u(x)$ is upper semicontinuous, this is equivalent to

$$\begin{aligned} \text{ess sup}_{\Omega'} H(x, Du(x)) &< a \\ &\text{for any } \Omega' \text{ compactly contained in } \Omega, \end{aligned}$$

where the expression *compactly contained* means that the closure of Ω' is compact and is contained in Ω . Accordingly, the maximal (possibly empty) open subset W_u where u is strict is given by the formula

$$W_u := \{x: \partial u(x) \subset \text{int } Z_a(x)\}. \quad (14)$$

Solutions

Unfortunately the relevant stability properties pointed out in the previous section for the family of subsolutions, do not hold for the a.e. *solutions*, namely the locally Lipschitz-continuous functions satisfying the equation up to a subset of M with vanishing measure. Take, for instance, the sequence u_n in \mathbb{T}^1 , obtained by linear interpolation of

$$u_n \left(\frac{k}{2n} \right) = 0 \quad \text{for } k \text{ even}, 0 \leq k \leq 2n$$

$$u_n \left(\frac{k}{2n} \right) = \frac{1}{2n} \quad \text{for } k \text{ odd}, 0 \leq k \leq 2n,$$

then it comprises a.e. solutions of (1), with $H(x, p) = |p|$ and $a = 1$. But its uniform limit is the null function, which is an a.e. subsolution of the same equation, according to Proposition 4, but fails to be an a.e. solution. This lack of stability motivates the search for a stronger notion of weak

solution. The idea is to look at the properties of S_a with respect to the operations of sup and inf.

Proposition 6 *Let $\tilde{S} \subset S_a$ be a family of locally equibounded functions, then the function defined as the pointwise supremum, or infimum, of the elements of \tilde{S} is a subsolution to (1).*

Set $u(x) = \inf\{v(x): v \in \tilde{S}\}$. Let ψ, u_n be a C^1 subgradient to u at a point x_0 , and a sequence of functions in \tilde{S} with $u_n(x_0) \rightarrow u(x_0)$, respectively.

Since the sequence u_n is made up of locally equibounded and locally equiLipschitz-continuous functions, it locally uniformly converges, up to a subsequence, by Ascoli Theorem, to a function w which belongs, in force of Proposition 4, to S_a . In addition $w(x_0) = u(x_0)$, and w is supergradient to u at x_0 by the very definition of u . Therefore, ψ is also subgradient to v at x_0 and so $H(x_0, D\psi(x_0)) \leq a$, which shows the assertion. The same proof, with obvious adaptations, allows us to handle the case of the pointwise supremum.

Encouraged by the previous result, consider the subsolutions of (1) enjoying some extremality properties. A definition is preliminary. A family \tilde{S} of locally equibounded subsolutions to (1) is said to be *complete* at some point x_0 if there exists ε_{x_0} such that if two subsolutions u_1, u_2 agree outside some neighborhood of x_0 with radius less than ε_{x_0} and $u_1 \in \tilde{S}$ then $u_2 \in \tilde{S}$.

The interesting point in the next proposition is that the subsolutions which are extremal with respect to a complete family, possess an additional property involving the viscosity test functions.

Proposition 7 *Let u be the pointwise supremum (infimum) of a locally equibounded family $\tilde{S} \subset S_a$ complete at a point x_0 , and let ψ be a C^1 subgradient (supergradient) to u at x_0 . Then $H(x_0, D\psi(x_0)) = a$.*

Only the case where u is a pointwise supremum will be discussed. The proof is based on a *push up* method that will be again used in the sequel.

If, in fact, the assertion were not true, there should be a C^1 strict subgradient ψ at x_0 , with $\psi(x_0) = u(x_0)$, such that $H(x_0, D\psi(x_0)) < a$. The function ψ , being C^1 , is a (classical) subsolution of (1) in a neighborhood U of x_0 . Push up a bit the test function to define

$$v = \begin{cases} \max\{\psi + \varepsilon, u\} & \text{in } B(x_0, \varepsilon) \\ u & \text{otherwise} \end{cases} \quad (15)$$

with the positive constant ε chosen so that $B(x_0, \varepsilon) \subset U$ and $\varepsilon < \varepsilon_{x_0}$, where ε_{x_0} is the quantity appearing in the definition of a complete family of subsolutions at x_0 . By the

Proposition 6, the function v belongs to S_a , and is equal to $u \in \tilde{S}$ outside $B(x_0, \varepsilon)$. Therefore, $v \in \tilde{S}$, which is in contrast with the maximality of u because $v(x_0) > u(x_0)$.

Proposition 7 suggests the following two definitions of weak solution in the viscosity sense, or viscosity solutions for Eq. (1).

The function u is a *viscosity solution of the first type* if it is a subsolution and for any x_0 , any C^1 subgradient to u at x_0 one has $H(x_0, D\psi(x_0)) = a$. The *viscosity solutions of the second type* are definite by replacing the subgradient with supergradient. Such functions are clearly a.e. solutions.

The same proof of Proposition 4, applied to C^1 subgradients as well as supergradients, gives:

Proposition 8 *The family of viscosity solutions (of both types) to (1) is stable with respect to the local uniform convergence.*

Moreover, the argument of Proposition 6, with obvious adaptations, shows:

Proposition 9 *The pointwise infimum (supremum) of a family of locally equibounded viscosity solutions of the first (second) type is a viscosity solution of the first (second) type to (1).*

Morally, it can be said that the solutions of the first type enjoy some maximality properties, and some minimality properties hold for the others. Using the notion of the strict subsolution, introduced in the previous section, the following partial converse of Proposition 7 can be obtained:

Proposition 10 *Let Ω , u , φ be a bounded open subset of M , a viscosity solution to (1) of the first (second) type, and a strict subsolution in Ω coincident with u on $\partial\Omega$, respectively. Then $u \geq \varphi$ ($u \leq \varphi$) in Ω .*

The proof rests on a regularization procedure of φ by mollification. Assume that u is a viscosity solution of the first type, the other case can be treated similarly. The argument is by contradiction, then admit that the minimizers of $u - \varphi$ in $\overline{\Omega}$ (the closure of Ω) are in an open subset Ω' compactly contained in Ω . Define for $x \in \Omega'$, $\delta > 0$,

$$\varphi_\delta(x) = \int \zeta_\delta(y - x) \varphi(y) dy,$$

where ζ_δ is a standard C^∞ mollifier supported in $B(0, \delta)$. By using the convex character of the Hamiltonian and Jensen Lemma, it can be found

$$H(x, D\varphi_\delta(x)) \leq \int \zeta_\delta(y - x) H(x, D\varphi(y)) dy.$$

Therefore, taking into account the stability of the set of minimizers under the uniform convergence, and that φ

is a strict subsolution, δ can be chosen so small that φ_δ is a C^∞ strict subsolution of (1) in Ω' and, in addition, is subgradient to u at some point of Ω' . This is in contrast with the very definition of viscosity solution of the first type. The above argument will be used again, and explained with some more detail, in the next section.

The family of viscosity solutions of first and second type coincide for conjugate Hamilton–Jacobi equations, with Hamiltonian H and \check{H} , up to a change of sign. More precisely:

Proposition 11 *A function u is a viscosity solution of the first (second) type to*

$$\check{H}(x, Du) = a \quad \text{in } M \quad (16)$$

if and only if $-u$ is a viscosity solution of the second (first) type to (1).

In fact, if u , ψ are a viscosity solution of the first type to (16) and a C^1 supergradient to $-u$ at a point x_0 , respectively, then $-u$ is a subsolution to (1), and $-\psi$ is supergradient to u at x_0 so that

$$a = \check{H}(x_0, -D\psi(x_0)) = H(x_0, -D\psi(x_0)),$$

which shows that $-u$ is indeed a viscosity solution of the second type to (1). The other implications can be derived analogously.

The choice between the two types of viscosity solutions is just a matter of taste, since they give rise to two completely equivalent theories. In this article those of the first type are selected, and they are referred to from now on as (viscosity) solutions, without any further specification.

Next a notion of regularity is introduced, called semiconcavity (semiconvexity), which fits the viscosity solutions framework, and that will be used in a crucial way in Sect. “Main Regularity Result”. The starting remark is that even if the notion of viscosity solution of the first (second) type is more stringent than that of the a.e. solution, as proved above, the two notions are nevertheless equivalent for concave (convex) functions. In fact a function of this type, say u , which is locally Lipschitz-continuous, satisfies the inequality

$$u(y) \leq (\geq) u(x_0) + p(y - x),$$

for any $x_0, y, p \in \partial u(x_0)$. It is therefore apparent that it admits (global) linear supergradients (subgradients) at any point x_0 . If there were also a C^1 subgradient (supergradient), say ψ , at x_0 , then u should be differentiable at x_0 by Proposition 1, and $Du(x_0) = D\psi(x_0)$, so that if u were an a.e. solution then $H(x_0, D\psi(x_0)) = a$, as announced.

In the above argument the concave (convex) character of u was not exploited to its full extent, but it was just used to show the existence of C^1 super(sub)tangents at any point x and that the differentials of such test functions make up ∂u . Clearly such a property is still valid for a larger class of functions. This is the case for the family of so-called *strongly semiconcave* (*semiconvex*) functions, which are concave (convex) up to the subtraction (addition) of a quadratic term. This point will now be outlined for strongly semiconcave functions, a parallel analysis could be performed in the strongly semiconvex case.

A function u is said to be strongly semiconcave if $u(x) - k|x - x_0|^2$ is concave for some positive constant k , some $x_0 \in M$. Starting from the inequality

$$\begin{aligned} u(\lambda x_1 + (1 - \lambda)x_2) - k|\lambda x_1 + (1 - \lambda)x_2 - x_0|^2 &\geq \\ \lambda(u(x_1) - k|x_1 - x_0|^2) + (1 - \lambda)(u(x_2) - k|x_2 - x_0|^2) \end{aligned}$$

which holds for any $x_1, x_2, \lambda \in [0, 1]$, it is derived through straightforward calculations

$$\begin{aligned} u(\lambda x_1 + (1 - \lambda)x_2) - \lambda u(x_1) - (1 - \lambda)u(x_2) \\ \geq -k\lambda(1 - \lambda)|x_1 - x_2|^2, \quad (17) \end{aligned}$$

which is actually a property equivalent to strong semiconcavity. This shows that for such functions the subtraction of $k|x - x_0|^2$, for any $x_0 \in M$, yields the concavity property. Therefore, given any x_0 , and taking into account (8), one has

$$u(x) - k|x - x_0|^2 \leq u(x_0) + p(x - x_0),$$

for any x , any $p \in \partial u(x_0)$ which proves that the generalized gradient of u at x_0 is made up by the differentials of the C^1 supertangents to u at x_0 . The outcome of the previous discussion is summarized in the following statement.

Proposition 12 *Let u be a strongly semiconcave (semiconvex) function. For any x , $p \in \partial u(x)$ if and only if it is the differential of a C^1 supertangent (subtangent) to u at x . Consequently, the notions of a.e. solution to (1) and viscosity solution of the first (second) type coincide for this class of functions.*

In Sect. “[Main Regularity Result](#)” a weaker notion of semiconcavity will be introduced, obtained by requiring a milder version of (17), which will be crucial to proving the existence of C^1 subsolutions to (1).

Even if the previous analysis, and in particular the part about the equivalent notions of subsolutions, is only valid for (1), one can define a notion of viscosity solution for a wider class of Hamilton–Jacobi equations than (1), and

even for some second-order equations. To be more precise, given a Hamilton–Jacobi equation $G(x, u, Du) = 0$, with G nondecreasing with respect to the second argument, a continuous function u is called the viscosity solution of it, if for any x_0 , any C^1 supertangent (subtangent) ψ to u at x_0 the inequality

$$G(x_0, u(x_0), D\psi(x_0)) \leq (\geq) 0$$

holds true.

Loosely speaking the existence of comparison principles in this context is related to the strict monotonicity properties of the Hamiltonian with respect to u or the presence in the equation of the time derivative of the unknown. For instance such principles hold for (2), see Sect. “[Long-Time Behavior of Solutions to the Time-Dependent Equation](#)”.

Obtaining uniqueness properties for viscosity solutions to (1) is a more delicate matter. Such properties are actually related to the existence of strict subsolutions, since this, in turn, allows one to slightly perturb any solution obtaining a strict subsolution. To exemplify this issue, Proposition 10 is exploited to show:

Proposition 13 *Let Ω , g be an open-bounded subset of M and a continuous function defined on $\partial\Omega$, respectively. Assume that there is a strict subsolution φ to (1) in Ω . Then there is at most one viscosity solution to (1) in Ω taking the datum g on the boundary.*

Assume by contradiction the existence of two viscosity solutions u and v with $v > u + \varepsilon$ at some point of Ω , where ε is a positive constant. The function $v_\lambda := \lambda\varphi + (1 - \lambda)v$ is a strict subsolution to (1) for any $\lambda \in]0, 1]$, by the convexity assumption on H . Further, λ can be taken so small that the points of Ω for which $v_\lambda > u + \frac{\varepsilon}{2}$ make up a non-empty set, say Ω' , are compactly contained in Ω . This goes against Proposition 10, because u and $v_\lambda + \frac{\varepsilon}{2}$ agree on $\partial\Omega'$ and the strict subsolution $v_\lambda + \frac{\varepsilon}{2}$ exceeds u in Ω' .

First Regularity Results for Subsolutions

A natural question is when does a classical subsolution to (1) exist? The surprising answer is that it happens whenever there is a (locally Lipschitz) subsolution, provided the assumptions on H introduced in the previous section are strengthened a little. Furthermore, any subsolution can be approximated by regular subsolutions in the topology of locally uniform convergence.

This theorem is postponed to Sect. “[Main Regularity Result](#)”. Some preliminary results of regularity for subsolution, holding under the assumptions (3)–(6), are presented below.

Firstly the discussion concerns the existence of subsolutions to (1) that are regular, say C^1 , at least on some distinguished subset of M . More precisely an attempt is made to determine when such functions can be obtained by mollification of subsolutions. The essential output is that this smoothing technique works if the subsolution one starts from is strict, so that, loosely speaking, some room is left to perturb it locally, still obtaining a subsolution. A similar argument has already been used in the proof of Proposition 10.

In this analysis it is relevant the critical level of the Hamiltonian which is defined as the one for which the corresponding Hamilton–Jacobi equation possesses a subsolution, but none of them are strict on the whole ground space. It is also important subset of M , named after Aubry and indicated by \mathcal{A} , made up of points around which no critical subsolution (i. e. subsolution to (1) with $a = c$) is strict.

According to what was previously outlined, to smooth up a critical subsolution around the Aubry set, seems particularly hard, if not hopeless. This difficulty will be overcome by performing a detailed qualitative study of the behavior of critical subsolutions on \mathcal{A} .

The simple setting to be first examined is when there is a strict subsolution, say u , to (1) satisfying

$$H(x, Du(x)) \leq a - \varepsilon \quad \text{for a.e. } x \in M, \text{ and some } \varepsilon > 0, \quad (18)$$

and, in addition, H is uniformly continuous on $M \times B \subset T^*M$, whenever B is a bounded subset of \mathbb{R}^N . In this case the mollification procedure plainly works to supply a regular subsolution. The argument of Proposition 10 can be adapted to show this. Define, for any x , any $\delta > 0$

$$u_\delta(x) = \int \zeta_\delta(y - x)u(y) \, dy,$$

where ζ_δ is a standard C^∞ mollifier supported in $B(0, \delta)$, and by using the convex character of the Hamiltonian and Jensen Lemma, get

$$H(x, Du_\delta(x)) \leq \int \zeta_\delta(y - x)H(x, Du(y)) \, dy,$$

so that if $o(\cdot)$ is a continuity modulus of H in $M \times B(0, r)$, with r denoting a Lipschitz constant for u , a δ can be selected in such a way that $o(\delta) \leq \frac{\varepsilon}{2}$, and consequently u_δ is the desired smooth subsolution, and is, in addition, still strict on M .

Even if condition (18) does not hold on the whole underlying space, the previous argument can be applied locally, to provide a smoothing of any subsolution u , at least

in the open subset W_u where it is strict (see (14) for the definition of this set), by introducing countable open coverings and associated C^∞ partition of the unity. The uniform continuity assumption on H as well as the global Lipschitz-continuity of u can be bypassed as well. It can be proved:

Proposition 14 *Given $u \in S_a$ with W_u nonempty, there exists $v \in S_a$, which is strict and of class C^∞ on W_u .*

Note that the function v appearing in the statement is required to be a subsolution on the whole M . In the proof of Proposition 14 an extension principle for subsolutions will be used that will be explained later.

Extension principle *Let v and C be a subsolution to (1) and a closed subset of M , respectively. Any continuous extension of $v|_C$ which is a subsolution on $M \setminus C$ is also a subsolution in the whole M .*

The argument for showing Proposition 14 will also provide the proof, with some adjustments, of the main regularity result, i. e. Theorem 35 in Sect. “Main Regularity Result”.

By the very definition of W_u an open neighborhood U'_x , compactly contained in W_u , can be found, for all $x \in W_u$, in such a way that

$$H(y, Du(y)) < a - \varepsilon_x \quad \text{for a.e. } y \in U'_x \text{ and some } \varepsilon_x > 0. \quad (19)$$

Through regularization of u by means of a C^∞ mollifier ζ_δ supported in $B(0, \delta)$, for $\delta > 0$ suitably small, a smooth function can be then constructed still satisfying (19) in a neighborhood of x slightly smaller than U'_x , say U_x . The next step is to extract from $\{U_x\}$, $x \in W_u$, a countable locally finite cover of W_u , say $\{U_{x_i}\}$, $i \in \mathbb{N}$. In the sequel the notations U_i , ε_i are adopted in place of U_{x_i} , ε_{x_i} , respectively. The regularized function is denoted by u_i .

Note that such functions are not, in general, subsolutions to (1) on M , since their behavior outside U_i cannot be controlled. To overcome this difficulty a C^∞ partition of the unity β_i subordinated to U_i is introduced.

The crucial point here is that the mollification parameters, denoted by δ_i , can be adjusted in such a way that the uniform distance $|u - u_i|_{\infty, U_i}$ is as small as desired. This quantity, more precisely, is required to be small with respect to $\frac{1}{|D\beta_i|_\infty}$, $\frac{1}{2^i}$ and the ε_j corresponding to indices j such that $U_j \cap U_i \neq \emptyset$. In place of $\frac{1}{2^i}$ one could take the terms of any positive convergent series with sum 1. Define v via the formula

$$v = \begin{cases} \sum \beta_i u_i & \text{in } W_u \\ u & \text{otherwise.} \end{cases} \quad (20)$$

Note that a finite number of terms are involved in the sum defining v in W_u since the cover $\{U_i\}$ is locally finite. It can be surprising at first sight that the quantity $\sum \beta_i u_i$, with u_i subsolution in U_i and β_i supported in U_i , represents a subsolution to (1), since, by differentiating, one gets

$$D\left(\sum \beta_i u_i\right) = \sum \beta_i Du_i + \sum D\beta_i u_i,$$

and the latter term does not seem easy to handle. The trick is to express it through the formula

$$\sum D\beta_i u_i = \sum D\beta_i (u_i - u), \quad (21)$$

which holds true because $\sum \beta_i \equiv 1$, by the very definition of partition of unity, and so $\sum D\beta_i \equiv 0$. From (21) deduce

$$\left|\sum D\beta_i u_i\right| \leq \sum |D\beta_i|_\infty |u - u_i|_{\infty, U_i},$$

and consequently, recalling that $|u - u_i|_{\infty, U_i}$ is small with respect to $\frac{1}{|D\beta_i|_\infty}$

$$D\left(\sum \beta_i u_i\right) \sim \sum \beta_i Du_i.$$

Since the Hamiltonian is convex in p , β_i is supported in U_i and u_i is a strict subsolution to (1) in U_i , we finally discover that v , defined by (20), is a strict subsolution in W_u .

Taking into account the extension principle for subsolutions, it is left to show, for proving Proposition 14, that v is continuous. For this, first observe that for any $n \in \mathbb{N}$ the set

$$\overline{\cup_{i \leq n} U_i}$$

is compact and disjoint from ∂W_u , and consequently

$$\min\{i: x \in U_i\} \rightarrow +\infty \quad \text{when } x \in W_u \text{ approaches } \partial W_u.$$

This, in turn, implies, since $|u - u_i|_{\infty, U_i}$ is small compared to $\frac{1}{2^i}$,

$$\begin{aligned} \left|\sum \beta_i(x) u_i(x) - u(x)\right| &\leq \sum_{\{i: x \in U_i\}} \beta_i(x) |u - u_i|_{\infty, U_i} \\ &\leq \sum_{\{i: x \in U_i\}} \beta_i(x) \frac{1}{2^i} \rightarrow 0, \end{aligned}$$

whenever $x \in W_u$ approaches ∂W_u . This shows the assertion.

The next step is to look for subsolutions that are strict in a subset of M as large as possible. In particular a strict

subsolution to (1) on the whole M does apparently exist at any level a of the Hamiltonian with

$$a > \inf\{b: H(x, Du) = b \text{ has a subsolution}\}. \quad (22)$$

The infimum on the right-hand side of the previous formula is the critical value of H ; it will be denoted from now on by c . Accordingly the values $a > c$ (resp. $a < c$) will be qualified as *supercritical* (resp. *subcritical*). The inf in (22) is actually a minimum in view of Proposition 5. By the coercivity properties of H , the quantity $\min_p H(x, p)$ is finite for any x , and clearly

$$c \geq \sup_x \min_p H(x, p),$$

which shows that $c > -\infty$, but it can be equal to $+\infty$ if M is noncompact. In this case no subsolutions to (1) should exist for any a . In what follows it is assumed that c is finite. Note that the critical value for the conjugate Hamiltonian \check{H} does not change, since, as already noticed in Sect. “Subsolutions”, the family of subsolutions of the two corresponding Hamilton–Jacobi equations are equal up to a change of sign.

From Proposition 14 can be derived:

Theorem 15 *There exists a smooth strict subsolution to (1) for any supercritical value a .*

Critical Equation and Aubry Set

Here the attention is focused on the critical equation

$$H(x, Du) = c. \quad (23)$$

A significant progress in the analysis is achieved by showing that there is a critical subsolution v with W_v , see (14) for the definition, enjoying a maximality property. More precisely the following statement holds:

Proposition 16 *There exists $v \in S_c$ with*

$$W_v = W_0 := \bigcup \{W_u: u \text{ is a critical subsolution}\}.$$

This result, combined with Proposition 14, gives the

Proposition 17 *There exists a subsolution to (23) that is strict and of class C^∞ on W_0 .*

To construct v appearing in the statement of Proposition 16 a covering technique to W_0 , as in Proposition 14, is applied and then the convex character of H is exploited. Since no regularity issues are involved, there is no need to introduce smoothing procedures and partitions of unity, so the argument is altogether quite simple.

Any point $y \in W_0$ possesses a neighborhood U_y where some critical subsolution v_y satisfies

$$H(x, Dv_y(x)) \leq c - \varepsilon_y \quad \text{for a.e. } x \in U_y, \text{ some positive } \varepsilon_y.$$

A locally finite countable subcover $\{U_{y_i}\}$, $i \in \mathbb{N}$, can be extracted, the notations U_i , v_i , ε_i are used in place of U_{y_i} , v_{y_i} , ε_{y_i} . The function v is defined as an infinite convex combination of u_i , more precisely $v = \sum \frac{1}{2^i} v_i$.

To show that v has the properties asserted in the statement, note that the functions v_i are locally equiLipschitz-continuous, being critical subsolutions, and can be taken, in addition, locally equibounded, up to addition of a constant. The series $\sum \lambda_i u_i$, $\sum \lambda_i Du_i$ are therefore locally uniformly convergent by the Weierstrass M -test. This shows that the function v is well defined and is Lipschitz-continuous, in addition

$$Dv(x) = \sum \lambda_i Du_i(x) \quad \text{for a.e. } x.$$

If x belongs to the full measure set where v and all the v_i are differentiable, one finds, by exploiting the convex character of H

$$\begin{aligned} H\left(x, \sum_{i \leq n} \lambda_i Du_i(x)\right) &\leq \sum_{i \leq n} \lambda_i H(x, Du_i(x)) \\ &\quad + \left(1 - \sum_{i \leq n} \lambda_i\right) H(x, 0), \end{aligned}$$

for any fixed n . This implies, passing to the limit for $n \rightarrow +\infty$

$$\begin{aligned} H(x, Dv(x)) &= H\left(x, \sum_{i=1}^{\infty} \lambda_i Du_i(x)\right) \\ &\leq \sum_{i=1}^{\infty} \lambda_i H(x, Du_i(x)). \end{aligned}$$

The function v is thus a critical subsolution, and, in addition, one has

$$\begin{aligned} H(x, Dv(x)) &\leq \sum_{i \neq j} H(x, Dv_i(x)) + \lambda_j H(x, Dv_j(x)) \\ &\leq c - \lambda_j \varepsilon_j, \end{aligned} \quad (24)$$

for any j and a.e. $x \in U_j$. This yields Proposition 16 since $\{U_j\}$, $j \in \mathbb{N}$, is a locally finite open cover of W_0 , and so it comes from (24) that the essential sup of v , on any open set compactly contained in W_0 , is strictly less than c .

The Aubry set \mathcal{A} is defined as $M \setminus W_0$. According to Propositions 14, 16 it is made up by the *bad* points around

which no function of S_c can be regularized through mollification still remaining a critical subsolution. The points of \mathcal{A} are actually characterized by the fact that no critical subsolution is strict around them. Note that a local as well as a global aspect is involved in such a property, for the subsolutions under investigation must be subsolutions on the whole space. Note further that \mathcal{A} is also the Aubry set for the conjugate critical equation with Hamiltonian \tilde{H} .

A qualitative analysis of \mathcal{A} is the main subject of what follows. Notice that the Aubry set must be nonempty if M is compact, since, otherwise, one could repeat the argument used for the proof of Proposition 16 to get a *finite* open cover $\{U_i\}$ of M and a *finite* family u_i of critical subsolutions satisfying

$$H(x, Du_i(x)) \leq c - \varepsilon_i \quad \text{for a.e. } x \in U_i \text{ and some } \varepsilon_i > 0,$$

and to have for a *finite* convex combination $u = \sum_i \lambda_i u_i$

$$H(x, Du(x)) \leq c - \min_i \{\lambda_i \varepsilon_i\},$$

in contrast with the very definition of critical value. If, on the contrary, M is noncompact Hamiltonian such that the corresponding Aubry set is empty, then it can be easily exhibited.

One example is given by $H(x, p) = |p| - f(x)$, in the case where the potential f has no minimizers. It is easily seen that the critical level is given by $-\inf_M f$, since, for a less than this value, the sublevels $Z_a(\cdot)$ are empty at some $x \in M$ and consequently the corresponding Hamilton–Jacobi equation does not have any subsolution; on the other side

$$H(x, 0) = -f(x) < -\inf_M f,$$

which shows that any constant function is a strict critical subsolution on M . This, in turn, implies the emptiness of \mathcal{A} .

In view of Proposition 14, one has:

Proposition 18 *Assume that M is noncompact and the Aubry set is empty, then there exists a smooth strict critical subsolution.*

The points y of \mathcal{A} are divided into two categories according to whether the sublevel $Z_c(y)$ has an empty or nonempty interior. It is clear that a point y with $\text{int } Z_c(y) = \emptyset$ must belong to \mathcal{A} because for such a point

$$H(y, p) = c \quad \text{for all } p \in Z_c(y), \quad (25)$$

and, since any critical subsolution u must satisfy $\partial u(y) \subset Z_c(y)$, it cannot be strict around y . These points

are called equilibria, and \mathcal{E} indicates the set of all equilibria. The reason for this terminology is that if the regularity assumptions on H are enough to write the Hamilton's equations on T^*M , then (y, p_0) is an equilibrium of the related flow with $H(y, p_0) = c$ if and only if $y \in \mathcal{E}$ and $Z_c(y) = \{p_0\}$. This point of view will not be developed further herein. From now on the subscript c will be omitted to ease notations.

Next the behavior of viscosity test functions of any critical subsolution at points belonging to the Aubry set is investigated. The following assertion holds true:

Proposition 19 *Let u, y, ψ be a critical subsolution, a point of the Aubry set and a viscosity test function to u at y , respectively. Then $H(y, D\psi(y)) = c$.*

Note that the content of the proposition is an immediate consequence of (25) if, in addition, y is an equilibrium. In the general case it is not restrictive to prove the statement when ψ is a strict subgradient. Actually, if the inequality $H(y, D\psi(y)) < c$ takes place, a contradiction is reached by constructing a subsolution v strict around y by means of the push-up argument introduced in Sect. “Subsolutions” for proving Proposition 7.

By using the previous proposition, the issue of the existence of (viscosity) solutions to (23) or, more generally, to (1) can be tackled. The starting idea is to fix a point y in M , to consider the family

$$\tilde{S}_a^y = \{u \in S_a : u(y) = 0\}, \quad (26)$$

and to define

$$w_a^y(x) = \sup_{\tilde{S}_a^y} u(x), \quad (27)$$

Since \tilde{S}_a^y is complete (this terminology was introduced in Sect. “Solutions”) at any $x \neq y$, the function w_a^y is a subsolution to (1) on M , and a viscosity solution to $M \setminus \{y\}$, by Propositions 6, 7.

If $a = c$, and the point y belongs to \mathcal{A} then, in view of Proposition 19, w_c^y is a critical solution on the whole M . On the contrary, the fact that $y \notin \mathcal{A}$ i.e. $y \in W_0$, prevents this function from being a global solution. In fact in this case, according to Propositions 14, 16, there is a critical subsolution φ , which is smooth and strict around y , and it can be also assumed, without any loss of generality, to vanish at y . Therefore, φ is subgradient to w_c^y at y for the maximality property of w_c^y and $H(y, D\varphi(y)) < c$. A characterization of the Aubry set then follows:

First characterization of \mathcal{A} *A point y belongs to the Aubry set if and only if the function w_c^y , defined in (27) with $a = c$, is a critical solution on the whole M .*

If $\mathcal{A} \neq \emptyset$, which is true when M is compact, then the existence of a critical solution can be derived. Actually in the compact case the critical level is the unique one for which a viscosity solution to (1) does exist. If, in fact $a > c$, then, by Theorem 15, the equation possesses a smooth strict critical subsolution, say φ , which is subgradient to any other function f defined on M at the minimizers of $f - u$, which do exist since M is assumed to be compact. This rules out the possibility of having a solution of (1) since $H(x, D\varphi(x)) < a$ for any x .

Next is discussed the issue that in the noncompact case a solution does exist at the critical as well as at any supercritical level.

Let a be supercritical. The idea is to exploit the noncompact setting, and to throw away the points where the property of being a solution fails, by letting them go to infinity.

Let $w_n := w_a^{y_n}$ be a sequence of subsolutions given by (27), with $|y_n| \rightarrow +\infty$. The w_n are equiLipschitz-continuous, being subsolutions to (1), and locally equibounded, up to the addition of a constant. One then gets, using Ascoli Theorem and arguing along subsequences, a limit function w . Since the w_n are solutions around any fixed point, for n suitably large, then, in view of the stability properties of viscosity solutions, see Proposition 8, w is a solution to (1) around any point of M , which means that w is a viscosity solution on the whole M , as announced. The above outlined properties are summarized in the next statement.

Proposition 20

- (i) *If M is compact then a solution to (1) does exist if and only if $a = c$.*
- (ii) *If M is noncompact then (1) can be solved in the viscosity sense if and only if $a \geq c$.*

An Intrinsic Metric

Formula (27) gives rise to a nonsymmetric semidistance $S_a(\cdot, \cdot)$ by simply putting

$$S_a(y, x) = w_a^y(x).$$

This metric viewpoint will allow us to attain a deeper insight into the structure of the subsolutions to (1) as well as of the geometric properties of the Aubry set.

It is clear that S_a satisfies the triangle inequality and $S_a(y, y) = 0$ for any y . It fails, in general to be symmetric and non-negative. It will be nevertheless called, from now on, distance to ease terminology.

An important point to be discussed is that S_a is a length distance, in the sense that a suitable length functional ℓ_a

can be introduced in the class of Lipschitz-continuous curves of M in such a way that, for any pairs of points x and y , $S_a(y, x)$ is the infimum of the lengths of curves joining them. Such a length, will be qualified from now on as *intrinsic* to distinguish it from the natural length on the ground space, denoted by ℓ . It only depends on the corresponding sublevels of the Hamiltonian. More precisely one defines for a (Lipschitz-continuous) curve ξ parametrized in an interval I

$$\ell_a(\xi) = \int_I \sigma_a(\xi, \dot{\xi}) dt, \quad (28)$$

where σ_a stands for the support function of the a -sublevel of H . More precisely, the function σ_a is defined, for any $(x, q) \in TM$ as

$$\sigma_a(x, q) = \max\{p \cdot q : p \in Z_a(x)\},$$

it is accordingly convex and positively homogeneous in p , upper semicontinuous in x , and, in addition, continuous at any point possessing a sublevel with a nonempty interior. The positive homogeneity property implies that the line integral in (28) is invariant under change of parameter preserving the orientation. The intrinsic length ℓ_a is moreover lower semicontinuous for the uniform convergence of a equiLipschitz-continuous sequence of curves, by standard variational argument, see [7]. Let \bar{S}_a denote the length distance associated to ℓ_a , namely

$$\bar{S}_a(y, x) = \inf\{\ell_a(\xi) : \xi \text{ connects } y \text{ to } x\}.$$

The following result holds true:

Proposition 21 \bar{S}_a and S_a coincide.

Note that by the coercivity of the Hamiltonian $\ell_a(\xi) \leq r\ell(\xi)$ for some positive r . Taking into account that the Euclidean segment is an admissible junction between any pair of points, deduce the inequality

$$|\bar{S}_a(y, x)| \leq r|y - x| \quad \text{for any } y, x,$$

which, combined with the triangle inequality, implies that the function $x \mapsto \bar{S}_a(y, x)$ is locally Lipschitz-continuous, for any fixed y . Let y_0, x_0 be a pair of points in M . Since $u := S_a(y_0, \cdot)$ is locally Lipschitz-continuous, one has

$$S_a(y_0, x_0) = u(x_0) - u(y_0) = \int_I \frac{d}{dt} u(\xi(t)) dt$$

for any curve ξ connecting y_0 to x_0 , defined in some interval I . It is well known from [8] that

$$\begin{aligned} \frac{d}{dt} u(\xi(t)) &= p(t) \dot{\xi}(t) \\ &\text{for a.e. } t \in I, \text{ some } p(t) \in \partial u(\xi(t)), \end{aligned}$$

and, since $\partial u(x) \subset Z_a(x)$ for any x , derive

$$S_a(y_0, x_0) \leq \ell_a(\xi) \quad \text{for all every curve } \xi \text{ joining } y_0 \text{ to } x_0,$$

which, in turn, yields the inequality $S_a(y_0, x_0) \leq \bar{S}_a(x_0, y_0)$. The converse inequality is obtained by showing that the function $w := \bar{S}_a(y_0, \cdot)$ is a subsolution to (1), see [20,29].

From now on the subscript from Z_a , S_a and σ_a will be omitted in the case where $a = c$.

It is clear that, in general, the intrinsic length of curves can have any sign. However, if the curve is a cycle such a length must be non-negative, according to Proposition 21, otherwise going several times through the same loop the identity $S_a \equiv -\infty$ would be obtained. This remark will have some relevance in what follows.

Proposition 21 allows us to determine the intrinsic metric related to the a -sublevel of the conjugate Hamiltonian \check{H} , denoted by $\check{Z}_a(\cdot)$. Since $\check{Z}_a(x) = -Z_a(x)$, for any x , because of the very definition of \check{H} , the corresponding support function $\check{\sigma}$ satisfies

$$\check{\sigma}_a(x, q) = \sigma_a(x, -q) \quad \text{for any } x, q.$$

Therefore, the intrinsic lengths ℓ_a and $\check{\ell}_a$ coincide up to a change of orientation. In fact, given ξ , defined in $[0, 1]$, and denoted by $\gamma(s) = \xi(1 - s)$ the curve with opposite orientation, one has

$$\check{\ell}_a(\xi) = \int_0^1 \sigma_a(\xi, -\dot{\xi}) ds,$$

and using $r = 1 - t$ as a new integration variable one obtains

$$\int_0^1 \sigma_a(\xi, -\dot{\xi}) ds = \int_0^1 \sigma_a(\gamma, \dot{\gamma}) dr = \ell_a(\gamma).$$

This yields

$$\check{S}_a(x, y) = S_a(y, x) \quad \text{for any } x, y,$$

where \check{S}_a stands for the conjugate distance. The function $S_a(\cdot, y)$ is thus the pointwise supremum of the family

$$\{v : v \text{ is a subsolution to (16) and } v(y) = 0\},$$

and accordingly $-S_a(\cdot, y)$ the pointwise infimum of $\{u \in S_a : u(y) = 0\}$. Summing up:

Proposition 22 Given $u \in S_a$, $y \in M$, the functions $S_a(y, \cdot)$, $-S_a(\cdot, y)$ are supertangent and sub-tangent, respectively, to u at y .

The Extension Principle for subsolutions can now be proved. Preliminarily the fifth characterization of the family of subsolutions is given.

Proposition 23 *A continuous function u is a subsolution to (1) if and only if*

$$u(x) - u(y) \leq S_a(y, x) \quad \text{for any } x, y. \quad (29)$$

It is an immediate consequence that any subsolution satisfies the inequality in the statement. Conversely, let ψ be a C^1 sub-tangent to u at some point y . By the inequality (29) the subsolution $x \mapsto S_a(y, x)$ is supertangent to u at y , for any y . Therefore, ψ is sub-tangent to $x \mapsto S_a(y, x)$ at the same point, and so one has $H(y, D\psi(y)) \leq a$, which shows the assertion taking into account the third definition of subsolution given in Sect. “Subsolutions”.

To prove the Extension Lemma, one has to show that a function w coincident with some subsolution to (1) on a closed set C , and being a subsolution on $M \setminus C$, is a subsolution on the whole M . The intrinsic length will play a main role here. Two facts are exploited:

- (i) if a curve connects two points belonging to C then the corresponding variation of u is estimated from above by the its intrinsic length because of Proposition 23, and since u coincides with a subsolution to (1) on C ,
- (ii) the same estimate holds true for any pair of points if the curve joining them lies outside C in force of the property that u is a subsolution in $M \setminus C$.

Let ε be a positive constant, x, y a pair of points and ξ a curve joining them, whose intrinsic length approximates $S_a(y, x)$ up to ε . The interval of definition of ξ can be partitioned in such a way that the portion of the curve corresponding to each subinterval satisfies the setting of one of the previous items (i) and (ii). By exploiting the additivity of the intrinsic length one finds

$$u(x) - u(y) \leq \ell_a(\xi) \leq S_a(y, x) + \varepsilon,$$

and the conclusion is reached taking into account the characterization of a subsolution given by Proposition 23, and the fact that ε is arbitrary.

To carry on the analysis, it is in order to discuss an apparent contradiction regarding the Aubry set. Let $y_0 \in \mathcal{A} \setminus \mathcal{E}$ and $p_0 \in \text{int } Z(y)$, then p_0 is also in the interior of the c -sublevels at points suitably close to y , say belonging to a neighborhood U of y , since Z is continuous at y . This implies

$$p(x - y_0) \leq \ell_c(\xi), \quad (30)$$

for any $x \in U$, any curve ξ joining y to X and lying in U . However, this inequality does not imply by any means that the function $x \mapsto p(x - y)$ is sub-tangent to $x \mapsto S(y, x)$

at y . This, in fact, should go against Proposition 19, since $H(y, p_0) < c$.

The unique way to overcome the contradiction is to admit that, even for points very close to y , the critical distance from y is realized by the intrinsic lengths of curves going out of U . In this way one could not deduce from the inequality (30) the previously indicated subtangency property. This means that S is not localizable with respect to the natural distance, and the behavior of the Hamiltonian in points far from y in the Euclidean sense can affect it.

There thus exist a sequence of points x_n converging to y and a sequence of curves joining y to x_n with intrinsic length approximating $S(y, x_n)$ up to $\frac{1}{n}$ and going out U . By juxtaposition of ξ_n and the Euclidean segment from x_n to y , a sequence of cycles γ_n can be constructed based on y (i. e. passing through y) satisfying

$$\ell_c(\gamma_n) \rightarrow 0, \quad \inf_n \ell(\gamma_n) > 0. \quad (31)$$

This is a threshold situation, since the critical length of any cycle must be non-negative. Next it is shown that (31) is indeed a metric characterization of the Aubry set.

Metric characterization of the Aubry set *A point y belongs to \mathcal{A} if and only if there is a sequence γ_n of cycles based on y and satisfying (31).*

What remains is to prove that the condition (31) holds at any equilibrium point and, conversely, that if it is true at some y , then such a point belongs to \mathcal{A} .

If $y \in \mathcal{E}$ then this can be directly proved exploiting that $\text{int } Z(y)$ is empty and, consequently the sublevel, being convex, is contained in the orthogonal of some element, see [20].

Conversely, let $y \in M \setminus \mathcal{A}$, according to Proposition 17, there is a critical subsolution u which is of class C^∞ and strict in a neighborhood U of y . One can therefore find a positive constant δ such that

$$Du(x)q \leq \sigma(x, q) - \delta \quad \text{for any } x \in U, \text{ any unit vector } q. \quad (32)$$

Let now ξ be a cycle based on y and parametrized by the Euclidean arc-length in $[0, \ell(\xi)]$, then $\xi(t) \in U$, for t belonging to an interval that can be assumed without loss of generality of the form $[0, t_1]$ for some $t_1 \geq \text{dist}(y, \partial U)$ (where dist indicates the Euclidean distance of a point from a set). This implies, taking into account (32) and that ξ is a cycle

$$\begin{aligned} \ell_c(\xi) &= \ell_c(\xi|_{[0, t_1]}) + \ell_c(\xi|_{[t_1, T]}) \\ &\geq (u(\xi(t_1)) - u(\xi(0)) + \delta t_1 + (u(\xi(T)) - u(\xi(t_1))) \\ &\geq \delta \text{dist}(y, \partial U)). \end{aligned}$$

This shows that the condition (31) cannot hold for sequences of cycles passing through y . By slightly adapting the previous argument, a further property of the intrinsic critical length, to be used later in Sect. “Long-Time Behavior of Solutions to the Time-Dependent Equation”, can be deduced.

Proposition 24 *Let M be compact. Given $\delta > 0$, there are two positive constants α, β such that any curve ξ lying at a distance greater than δ from \mathcal{A} satisfies*

$$\ell_c(\xi) \geq -\alpha + \beta \ell(\xi).$$

An important property of the Aubry set is that it is a *uniqueness set* for the critical equation, at least when the ground space is compact. This means that two critical solutions coinciding on \mathcal{A} must coincide on M . More precisely it holds:

Proposition 25 *Let M be compact. Given an admissible trace g on \mathcal{A} , i. e. satisfying the compatibility condition*

$$g(y_2) - g(y_1) \leq S(y_1, y_2),$$

the unique viscosity solution taking the value g on \mathcal{A} is given by

$$\min\{g(y) + S(y, \cdot) : y \in \mathcal{A}\}.$$

The representation formula yields indeed a critical solution thanks to the first characterization of the Aubry set and Proposition 9. The uniqueness can be obtained taking into account that there is a critical subsolution which is strict and C^∞ in the complement of \mathcal{A} (see Proposition 17), and arguing as in Proposition 13.

Some information on the Aubry set in the one-dimensional case can be deduced from both the characterizations of \mathcal{A}

Proposition 26 *Assume M to have dimension 1, then*

- (i) *if M is compact then either $\mathcal{A} = \mathcal{E}$ or $\mathcal{A} = M$,*
- (ii) *if M is noncompact then $\mathcal{A} = \mathcal{E}$, and, in particular, $\mathcal{A} = \emptyset$ if $\mathcal{E} = \emptyset$.*

In the one-dimensional case the c -sublevels are compact intervals. Set $Z(x) = [\alpha(x), \beta(x)]$ with α, β continuous, and consider the Hamiltonian

$$\bar{H}(x, p) = H(x, p - \alpha(x)).$$

It is apparent that u is a critical (sub)solution for H if and only if $u + F$, where F is any antiderivative of α , is

a (sub)solution to $\bar{H} = c$, and in addition, u is strict as a subsolution in some $\Omega \subset M$ if and only if $u + F$ is strict in the same subset. This proves that c is also the critical value for \bar{H} .

Further, $u \in \tilde{S}_c^y$ for some y , with \tilde{S}_c^y defined as in (26), if and only if $u + F_0$, where F_0 is the antiderivative of α vanishing at y , is in the corresponding family of subsolutions to $\bar{H} = c$. Bearing in mind the first characterization of \mathcal{A} , it comes that the Aubry sets of the two Hamiltonians H and \bar{H} coincide.

The advantage of using \bar{H} is that the corresponding critical sublevels $\bar{Z}(x)$ equal $[0, \beta(x) - \alpha(x)]$, for any x , and accordingly the support function, denoted by $\bar{\sigma}$, satisfies

$$\bar{\sigma}(x, q) = \begin{cases} q(\beta(x) - \alpha(x)) & \text{if } q > 0 \\ 0 & \text{if } q \leq 0 \end{cases}$$

for any x, q . This implies that the intrinsic critical length related to \bar{H} , say $\bar{\ell}_c$, is non-negative for all curves. Now, assume M to be noncompact and take $y \notin \mathcal{E}$, the claim is that $y \notin \mathcal{A}$. In fact, let $\varepsilon > 0$ be such that

$$m := \inf\{\beta(x) - \alpha(x) : x \in I_\varepsilon :=]y - \varepsilon, y + \varepsilon[\} > 0,$$

given a cycle ξ based on y , there are two possibilities: either ξ intersects ∂I_ε or is entirely contained in I_ε . In the first case

$$\bar{\ell}_c(\xi) \geq m \varepsilon, \quad (33)$$

in the second case ξ can be assumed, without losing generality, to be parametrized by the Euclidean arc-length; since it is a cycle one has

$$\int_0^{\ell(\xi)} \dot{\xi} \, ds = 0,$$

so that $\dot{\xi}(t) = 1$ for t belonging to a set of one-dimensional measure $\frac{\ell(\xi)}{2}$. One therefore has

$$\bar{\ell}_c(\xi) \geq m \frac{\ell(\xi)}{2}. \quad (34)$$

Inequalities (33), (34) show that $\bar{\ell}_c(\xi)$ cannot be infinitesimal unless $\ell(\xi)$ is infinitesimal. Hence item (ii) of Proposition 26 is proved. The rest of the assertion is obtained by suitably adapting the previous argument.

Dynamical Properties of the Aubry Set

In this section the convexity and the coercivity assumptions on H are strengthened and it is required, in addition

to (3),

$$H \text{ is convex in } p \quad (35)$$

$$\lim_{|p| \rightarrow +\infty} \frac{H(x, p)}{|p|} = +\infty \quad \text{uniformly in } x. \quad (36)$$

The Lagrangian L can be therefore defined through the formula

$$L(x, q) = \max\{p q - H(x, p) : p \in \mathbb{R}^N\}.$$

A curve γ , defined in some interval I , is said to be *critical* provided that

$$S(\gamma(t_1), \gamma(t_2)) = \int_{t_1}^{t_2} L(\gamma, \dot{\gamma}) + c \, ds = -S(\gamma(t_2), \gamma(t_1)), \quad (37)$$

for any t_1, t_2 in I . It comes from the metric characterization of \mathcal{A} , given in the previous section, that any critical curve is contained in \mathcal{A} . In fact if such a curve is supported on a point, say x_0 , then $L(x_0, 0) = -c$ and, consequently, the critical value is the minimum of $p \mapsto H(x, p)$. This, in turn, implies, in view of (35), that the sublevel $Z(x_0)$ has an empty interior so that $x_0 \in \mathcal{E} \subset \mathcal{A}$.

If, on the contrary, a critical curve is nonconstant and x_1, x_2 are a pair of different points lying in its support, then $S(x_1, x_2) + S(x_2, x_1) = 0$ and there exist two sequences of curves, ξ_n and η_n whose intrinsic length approximates the $S(x_1, x_2)$ and $S(x_2, x_1)$, respectively. Hence the trajectories obtained through juxtaposition of ξ_n and η_n are cycles with critical length infinitesimal and natural length estimated from below by a positive constant, since they contain x_1 and x_2 , with $x_1 \neq x_2$. This at last implies that such points, and so the whole support of the critical curve, are contained in the Aubry set.

Next the feature of the parametrization of a critical curve γ is investigated, since it apparently matters for a curve to be critical. For this purpose let $x_0, q_0 \neq 0$, and $p_0 \in Z(x_0)$ with $\sigma(x_0, q_0) = p_0 q_0$, it comes from the definition of the Lagrangian

$$\sigma(x_0, q_0) - c = p_0 q_0 - H(x_0, p_0) \leq L(x_0, q_0),$$

by combining this formula with (37), and recalling the relationship between intrinsic length and distance, one gets

$$L(\gamma, \dot{\gamma}) + c = \sigma(\gamma, \dot{\gamma}) \quad \text{for a.e. } t. \quad (38)$$

A parametrization is called *Lagrangian* if it satisfies the above equality. As a matter of fact it is possible to prove that any curve η , which stays far from \mathcal{E} , can be endowed with such a parametrization, see [10].

A relevant result to be discussed next is that the Aubry set is fully covered by critical curves. This property allows us to obtain precious information on the behavior of critical subsolution on \mathcal{A} , and will be exploited in the next sections in the study of long-time behavior of solutions to (2). More precisely the following result can be shown:

Theorem 27 *Given $y_0 \in \mathcal{A}$, there is a critical curve, defined in \mathbb{R} , taking the value y_0 at $t = 0$.*

If $y_0 \in \mathcal{E}$ then the constant curve $\xi(t) \equiv y_0$ is critical, as pointed out above. It can therefore be assumed $y_0 \notin \mathcal{E}$. It is first shown that a critical curve taking the value y_0 at 0, and defined in a bounded interval can be constructed.

For this purpose start from a sequence of cycles ξ_n , based on y_0 , satisfying the properties involved in the metric characterization of \mathcal{A} , and parametrized by (Euclidean) arc length in $[0, T_n]$, with $T_n = \ell(\xi_n)$. By exploiting Ascoli Theorem, and arguing along subsequences, one obtains a uniform limit curve ξ of the ξ_n , with $\xi(0) = y_0$, in an interval $[0, T]$, where T is strictly less than $\inf_n T_n$. It is moreover possible to show that ξ is nonconstant.

A new sequence of cycles γ_n can be defined through juxtaposition of ξ , the Euclidean segment joining $\xi(T)$ to $\xi_n(T)$ and $\xi_n|_{[T, T_n]}$. By the lower semicontinuity of the intrinsic length, the fact that $\xi_n(T)$ converges to $\xi(T)$, and consequently the segment between them has infinitesimal critical length, one gets

$$\lim_n \ell_c(\gamma_n) = 0. \quad (39)$$

The important thing is that all the γ_n coincide with ξ in $[0, T]$, so that if $t_1 < t_2 < T$, $S(\xi(t_1), \xi(t_2))$ is estimated from above by

$$\ell_c(\xi|_{[t_1, t_2]}) = \ell_c(\gamma_n|_{[t_1, t_2]}),$$

and $S(\xi(t_2), \xi(t_1))$ by the intrinsic length of the portion of γ_n joining $\xi(t_2)$ to $\xi(t_1)$. Taking into account (39) one gets

$$0 = \lim_n \ell_c(\gamma_n) \geq S(\xi(t_1), \xi(t_2)) + S(\xi(t_2), \xi(t_1)),$$

which yields the crucial identity

$$S(\xi(t_2), \xi(t_1)) = -S(\xi(t_1), \xi(t_2)). \quad (40)$$

In addition the previous two formulae imply that $\xi|_{[t_1, t_2]}$ is a minimal geodesic whose intrinsic length realizes $S(\xi(t_1), \xi(t_2))$, so that (40) can be completed as follows:

$$S(\xi(t_2), \xi(t_1)) = \int_{t_1}^{t_2} \sigma(\xi, \dot{\xi}) \, ds = -S(\xi(t_1), \xi(t_2)). \quad (41)$$

Finally, ξ has a Lagrangian parametrization, up to a change of parameter, so that it is obtained in the end

$$\int_{t_1}^{t_2} \sigma(\xi, \dot{\xi}) \, ds = \int_{t_1}^{t_2} (L(\xi, \dot{\xi}) + c) \, ds.$$

This shows that ξ is a critical curve. By applying Zorn lemma ξ can be extended to a critical curve defined in \mathbb{R} , which concludes the proof of Theorem 27.

As a consequence a first, perhaps surprising, result on the behavior of a critical subsolution on the Aubry set is obtained.

Proposition 28 *All critical subsolutions coincide on any critical curve, up to an additive constant.*

If u is such a subsolution and ξ any curve, one has

$$S(\xi(t_1), \xi(t_2)) \geq u(\xi(t_2)) - u(\xi(t_1)) \geq -S(\xi(t_2), \xi(t_1)),$$

by Proposition 22. If in addition ξ is critical then the previous formula holds with equality, which proves Proposition 28.

Next is presented a further result on the behavior of critical subsolutions on \mathcal{A} . From this it appears that, even in the broad setting presently under investigation (the Hamiltonian is supposed to be just continuous), such subsolutions enjoy some extra regularity properties on \mathcal{A} .

Proposition 29 *Let ξ be a critical curve, there is a negligible set E in \mathbb{R} such that, for any critical subsolution u the function $u \circ \xi$ is differentiable whenever in $\mathbb{R} \setminus E$ and*

$$\frac{d}{dt} u(\xi(t)) = \sigma(\xi(t), \dot{\xi}(t)).$$

More precisely E is the complement in \mathbb{R} of the set of Lebesgue points of $\sigma(\xi, \dot{\xi})$ where, in addition, ξ is differentiable. E has a vanishing measure thanks to Rademacher and Lebesgue differentiability theorem. See [10] for a complete proof of the proposition.

The section ends with the statement of a result, that will be used for proving the forthcoming Theorem 34. The proof can be obtained by performing Lagrangian reparametrizations.

Proposition 30 *Let ξ be a curve defined in $[0, 1]$. Denote by \mathcal{E} the set of curves obtained through reparametrization of ξ in intervals with right endpoint 0, and for $\gamma \in \mathcal{E}$ indicate by $[0, T(\gamma)]$ its interval of definition. One has*

$$\ell_c(\xi) = \inf \left\{ \int_0^{T(\gamma)} (L(\gamma, \dot{\gamma}) + c) \, ds : \gamma \in \mathcal{E} \right\}.$$

Long-Time Behavior of Solutions to the Time-Dependent Equation

In this section it is assumed, in addition to (3), (36),

$$M \text{ is compact} \quad (42)$$

$$H \text{ is strictly convex in } p. \quad (43)$$

A solution of the time-dependent Eq. (2) is said to be *stationary* if it has the variable-separated form

$$u_0(x) - at, \quad (44)$$

for some constant a . Note that if ψ is a supertangent (subtangent) to u_0 at some point x_0 , then $\psi - at$ is supertangent (subtangent) to $u_0 - at$ at (x_0, t) for any t , so that the inequality

$$-a + H(x_0, D\psi(x_0)) \leq (\geq) 0$$

holds true. Therefore, u_0 is a solution to (1) in M . Since such a solution does exist only when $a = c$, see Proposition 20, it is the case that in (44) u_0 is a critical solution and a is equal to c . The scope of this section is to show that any solution to the time-dependent equation uniformly converge to a stationary solution, as t goes to $+\infty$.

In our setting there is a comparison principle for (2), stating that two solutions v, w , issued from initial data v_0, w_0 , with $v_0 \geq w_0$, satisfies $v \geq w$, from any $x \in M, t > 0$. In addition there exists a viscosity solution v , for any continuous initial datum v_0 , which is, accordingly, unique, and is given by the Lax–Oleinik representation formula:

$$v(x, t) = \inf \left\{ v_0(\xi(0)) + \int_0^t L(\xi, \dot{\xi}) \, ds : \right. \\ \left. \xi \text{ is a curve with } \xi(t) = x \right\}. \quad (45)$$

This shows that Eq. (2) enjoys the semigroup property, namely if w and v are two solutions with $w(\cdot, 0) = v(\cdot, t_0)$, for some $t_0 > 0$, then

$$w(x, t) = v(x, t_0 + t).$$

It is clear that the solution of (2), taking a critical solution u_0 as initial datum, is stationary and is given by (44). For any continuous initial datum v_0 it can be found, since the underlying manifold is compact, a critical solution u_0 and a pair of constants $\alpha > \beta$ such that

$$u_0 + \alpha > v_0 > u_0 + \beta,$$

and consequently by the comparison principle for (2)

$$u_0 + \alpha > v(\cdot, t) + ct > u_0 + \beta \quad \text{for any } t.$$

This shows that the family of functions $x \mapsto v(x, t) + ct$, for $t \geq 0$, is equibounded. It can also be proved, see [10], that it is also equicontinuous, so that, by Ascoli Theorem, every sequence $v(\cdot, t_n) + ct_n$, for $t_n \rightarrow +\infty$, is uniformly convergent in M , up to extraction of a subsequence. The limits obtained in this way will be called ω -limit of $v + ct$. The first step of the analysis is to show:

Proposition 31 *Let v be a solution to (2). The pointwise supremum and infimum of the ω -limit of $v + ct$ are critical subsolutions.*

The trick is to introduce a parameter ε small and consider the functions

$$v^\varepsilon(x, t) = v(x, t/\varepsilon) \quad \text{for } \varepsilon > 0.$$

Arguing as above, it can be seen that the family $v^\varepsilon + ct$ is equibounded, moreover $v^\varepsilon + ct$ is apparently the solution to

$$\varepsilon w_t + H(x, Dw) = c,$$

for any ε . Hence, we exploit the stability properties that have been illustrated in Sects “Subsolutions”, “Solutions”, to prove the claim.

The following inequality will be used

$$L(x, q) + c \geq \sigma(x, q) \quad \text{for any } x, q,$$

which yields, by performing a line integration

$$\int_0^t (L(\gamma, \dot{\gamma}) + c) \, ds \geq \ell_c(\gamma), \quad (46)$$

for any $t > 0$, any curve γ defined in $[0, t]$, moreover, taking into account Lax–Oleinik formula and that M is assumed in this section to be compact

$$v(x, t) + ct \geq v_0(y) + S(y, x) \quad \text{for some } y \text{ depending on } x, \quad (47)$$

for a solution v of (2) taking a function v_0 as initial datum. If, in addition, v_0 is a critical subsolution, invoke Proposition 23 to derive from (47)

$$v(x, t) \geq v_0(x) - ct \quad \text{for any } x, t. \quad (48)$$

A crucial point to be exploited, is that, for such a v_0 , the evolution induced by (2) on the Aubry set takes place on the critical curves. Given $t > 0$ and $x \in \mathcal{A}$, pick a critical curve ξ with $\xi(t) = x_0$, whose existence is guaranteed by Theorem 27, and then employ Proposition 29, about the

behavior of the subsolution to (23) on critical curves, to obtain

$$v_0(x) - ct = v_0(\xi(0)) + \int_0^t L(\xi, \dot{\xi}) \, ds \geq v(x, t). \quad (49)$$

By combining (49) and (48), one finally has

$$v(x, t) = v_0(\xi(0)) + \int_0^t L(\xi, \dot{\xi}) \, ds = v_0(x) - ct,$$

which actually shows the announced optimal character of critical curves with respect to the Lax–Oleinik formula, and, at the same time, the following

Proposition 32 *Let v be a solution to (2) taking a critical subsolution v_0 as initial datum at $t = 0$. Then*

$$v(x, t) = v_0(x) - ct \quad \text{for any } x \in \mathcal{A}.$$

Summing up: stationary solutions are derived by taking as initial datum solutions to (23); more generally, solutions issued from a critical subsolution are stationary at least on the Aubry set. The next step is to examine the long-time behavior of such solutions on the whole M .

Proposition 33 *Let v be a solution to (2) taking a critical subsolution v_0 as initial datum at $t = 0$. One has*

$$\lim_{t \rightarrow +\infty} v(x, t) + ct = u_0(x) \quad \text{uniformly in } x,$$

where u_0 is the critical solution with trace v_0 on \mathcal{A} .

The starting remark for getting the assertion is that, for any given x_0 , an ε -optimal curve for $v(x_0, t_0)$, say ξ , must be close to \mathcal{A} for some $t \in [0, t_0]$, provided ε is sufficiently small and t_0 large enough.

If in fact ξ stayed far from \mathcal{A} for any $t \in [0, t_0]$ then $L(\xi(s), 0) + c$ could be estimated from below by a positive constant, since $\mathcal{E} \subset \mathcal{A}$, and the same should hold true, by continuity, for $L(\xi(s), q) + c$ if $|q|$ is small. One should then deduce that $\ell(\xi)$, and consequently (in view of Proposition 24) $\ell_c(\xi)$ were large. On the other side

$$u_0(x_0) \geq v(x_0, t_0) + ct_0 \geq v_0(\xi(0)) + \ell_c(\xi) - \varepsilon, \quad (50)$$

by (46) and the comparison principle for (2), which shows that the critical length of ξ is bounded from above, yielding a contradiction.

It can be therefore assumed that, up to a slight modification, the curve ξ intersects \mathcal{A} at a time $s_0 \in [0, t_0]$ and satisfies

$$\begin{aligned} v(x_0, t_0) &\geq v_0(\xi(0)) + \int_0^{t_0} L(\xi, \dot{\xi}) \, ds - \varepsilon \\ &\geq v(\xi(s_0), s_0) + \int_{s_0}^{t_0} L(\xi, \dot{\xi}) \, ds - \varepsilon. \end{aligned}$$

It is known from Proposition 32 that $v(\xi(s_0), s_0) = v_0(\xi(s_0)) - c s_0$, so we have from the previous inequality, in view of (46)

$$\begin{aligned} v(x_0, t_0) &\geq v_0(\xi(s_0)) - c t_0 + \ell_c \left(\xi|_{[s_0, t_0]} \right) - \varepsilon \\ &\geq v_0(\xi(s_0)) - c t_0 + S(\xi(s_0), x_0) - \varepsilon. \end{aligned}$$

Bearing in mind the representation formula for u_0 given in Proposition 25, we obtain in the end

$$v(x_0, t_0) \geq u_0(x_0) - c t_0 - \varepsilon,$$

and conclude exploiting $u_0 \geq v_0$ and the comparison principle for (2).

The previous statement can be suitably generalized by removing the requirement of v_0 being a critical subsolution. One more precisely has:

Theorem 34 *Let v be a viscosity solution to (2) taking a continuous function v_0 as initial datum for $t = 0$, then*

$$\lim_{t \rightarrow +\infty} v(x, t) + c t = u_0(x) \quad \text{uniformly in } x,$$

where u_0 is the critical solution given by the formula

$$u_0(x) = \inf_{y \in \mathcal{A}} \inf_{z \in M} (v_0(z) + S(z, y) + S(y, x)). \quad (51)$$

The claim is that u_0 , as defined in (51), is the critical solution with trace

$$w_0 := \inf_{z \in M} v_0(z) + S(z, \cdot) \quad (52)$$

on the Aubry set. This can indeed be deduced from the representation formula given in Proposition 25, once it is proved that w_0 is a critical subsolution. This property, in turn, comes from the characterization of critical subsolutions in terms of critical distance, presented in Proposition 23, the triangle inequality for S , and the inequalities

$$\begin{aligned} w_0(x_1) - w_0(x_2) &\leq v_0(z_2) + S(z_2, x_1) - v_0(z_2) - S(z_2, x_2) \leq S(x_2, x_1), \end{aligned}$$

which hold true if z_2 is a point realizing the infimum for $w_0(x_2)$. If, in particular, v_0 itself is a critical subsolution, then it coincides with w_0 , so that, as announced, Theorem 34 includes Proposition 33.

In the general case, it is apparent that $w_0 \leq v_0$, moreover if $z \in M$ and \bar{w}_0 is a critical subsolution with $\bar{w}_0 \leq v_0$, one deduces from Proposition 23

$$\bar{w}_0(x) \leq v_0(z) + S(z, x) \quad \text{for any } z,$$

which tells that $\bar{w}_0 \leq w_0$, therefore w_0 is the maximal critical subsolution not exceeding v_0 .

A complete proof of Theorem 34 is beyond the scope of this presentation. To give an idea, we consider the simplified case where the equilibria set \mathcal{E} is a uniqueness set for the critical equation.

Given $x_0 \in \mathcal{E}$, $\varepsilon > 0$, take a z_0 realizing the infimum for $w_0(x_0)$, and a curve η , connecting z_0 to x_0 , whose intrinsic length approximates $S(z_0, x_0)$ up to ε . By invoking Proposition 30 one deduces that, up to a change of the parameter, such a curve, defined in $[0, T]$, for some $T > 0$, satisfies

$$\int_0^T (L(\eta, \dot{\eta}) + c) ds < \ell_c(\eta) + \varepsilon.$$

Therefore, taking into account the Lax–Oleinik formula, one discovers

$$\begin{aligned} w_0(x_0) &\geq v_0(z_0) + \int_0^T (L(\eta, \dot{\eta}) + c) ds - 2\varepsilon \\ &\geq v(x_0, T) + cT - 2\varepsilon. \end{aligned} \quad (53)$$

Since $L(x_0, 0) + c = 0$, by the very definition of equilibrium, it can be further derived from Lax–Oleinik formula that $t \mapsto v(x_0, t)$ is nonincreasing, so that the inequality (53) still holds if T is replaced by every $t > T$. This, together with the fact that ε in (53) is taken arbitrarily, shows in the end, that any ω -limit ψ of $v + ct$ satisfies

$$w_0(x_0) \geq \psi(x_0) \quad \text{for any } x_0 \in \mathcal{E}. \quad (54)$$

On the other side, the initial datum v_0 is greater than or equal to w_0 , and the solution to (2) with initial datum w_0 , say w , has as unique ω -limit the critical solution u_0 with trace w_0 on \mathcal{E} [recall that \mathcal{E} is assumed to be a uniqueness set for (23)]. Consequently, by the comparison principle for (2), one obtains

$$u_0 \leq \psi \quad \text{in } M, \quad (55)$$

which, combined with (54), implies that ψ and w_0 coincide on \mathcal{E} . Further, by Proposition 31, the maximal and minimal ψ are critical subsolutions, and u_0 is the maximal critical subsolution taking the value w_0 on \mathcal{E} . This finally yields $\psi = u_0$ for any ψ , and proves the assertion of Theorem 34.

In the general case the property of the set of ω -limits of critical curves (i. e. limit points for $t \rightarrow +\infty$) of being a uniqueness set for the critical equation must be exploited. In this setting the strict convexity of H is essential, in [4,10] there are examples showing that Theorem 34 does not hold for H just convex in p .

Main Regularity Result

This section is devoted to the discussion of

Theorem 35 *If the Eq. (1) has a subsolution then it also admits a C^1 subsolution. Moreover, the C^1 subsolutions are dense in S_a with respect to the local uniform convergence.*

Just to sum up: it is known that a C^1 subsolution does exist when a is supercritical, see Theorem 15, and if the Aubry set is empty, see Proposition 18. So the case where $a = c$ and $\mathcal{A} \neq \emptyset$ is left. The starting point is the investigation of the regularity properties of *any* critical subsolution on \mathcal{A} .

For this, and for proving Theorem 35 conditions (43), (35) on H are assumed, and (3) is strengthened by requiring

$$H \text{ is locally Lipschitz-continuous in both arguments.} \quad (56)$$

This regularity condition seems unavoidable to show that the functions $S_a(y, \cdot)$ and $\check{S}_a(y, \cdot) = S_a(\cdot, y)$ enjoy a weak form of semiconcavity in $M \setminus \{y\}$, for all y , namely, if v is any function of this family, x_1, x_2 are different from y and $\lambda \in [0, 1]$, then

$$v(\lambda x_1 + (1 - \lambda)x_2) - \lambda v(x_1) - (1 - \lambda)v(x_2)$$

can be estimated from below by a quantity of the same type as that appearing on the right-hand side of (17), with $|x_1 - x_2|^2$ replaced by a more general term which is still infinitesimal for $|x_1 - x_2| \rightarrow 0$. Of course the fundamental property of possessing C^1 supertangents at any point different from y and that the set made up by their differentials coincides with the generalized gradient is maintained in this setting.

To show the validity of this semiconcavity property, say for $S_a(y, \cdot)$, at some point x it is crucial that for any neighborhood U of x suitably small there are curves joining y to x and approximating $S_a(y, x)$ which stays in U for a (natural) length greater than a fixed constant depending on U . This is clearly true if $x \neq y$ and explains the reason why the initial point y has been excluded. However, if $a = c$ and $y \in \mathcal{A}$, exploiting the metric characterization of the Aubry set, it appears that this restriction on y can be removed, so that the following holds:

Proposition 36 *Let $y \in \mathcal{A}$, then the functions $S(y, \cdot)$ and $S(\cdot, y)$ are semiconcave (in the sense roughly explained above) on the whole M . This, in particular, implies that both functions possess C^1 supertangents at y and their differentials comprise the generalized gradient.*

From this the main regularity property of critical subsolutions on \mathcal{A} are deduced. This reinforces the results given in Propositions 28, 29 under less stringent assumptions.

Theorem 37 *Every critical subsolution is differentiable on \mathcal{A} . All have the same differential, denoted by $p(y)$, at any point $y \in \mathcal{A}$, and*

$$H(y, p(y)) = c.$$

Furthermore, the function $y \mapsto p(y)$ is continuous on \mathcal{A} .

It is known from Proposition 22 that $S(y, \cdot)$, $-S(\cdot, y)$ are supertangent and subgradient, respectively, to every critical subsolution u at any y . If, in particular, $y \in \mathcal{A}$ then $S(y, \cdot)$ and $-S(\cdot, y)$ admit C^1 supertangents and subgradients, respectively, thanks to Proposition 36. This, in turn, implies that u is differentiable at y by Proposition 1 and, in addition, that all the differentials of supertangents to $S(y, \cdot)$ coincide with $Du(y)$, which shows that its generalized gradient reduces to a singleton and so $S(y, \cdot)$ is strictly differentiable at y by Proposition 36. If $p(y)$ denotes the differential of $S(y, \cdot)$ at y , then $Du(y) = p(y)$ for any critical subsolution and, in addition, $H(y, p(y)) = c$ by Proposition 19. Finally, the strict differentiability of $S(y, \cdot)$ at y gives that $p(\cdot)$ is continuous on \mathcal{A} .

The first application of Theorem 37 is relative to critical curves, and confirm their nature of generalized characteristics. If ξ is such a curve (contained in \mathcal{A} by the results of Sect. “[Dynamical Properties of the Aubry Set](#)”) and u a critical subsolution, it is known from Proposition 29 that

$$\begin{aligned} \frac{d}{dt}u(\xi(t)) &= p(\xi(t))\dot{\xi}(t) = \sigma(\xi(t), \dot{\xi}(t)) \\ &= L(\xi(t), \dot{\xi}(t)) + c, \end{aligned}$$

for a.e. $t \in \mathbb{R}$. Bearing in mind the definition of L , it is deduced that $p(\xi(t))$ is a maximizer of $p \mapsto p\dot{\xi}(t) - H(\xi(t), p)$, then by invoking (7) one obtains:

Proposition 38 *Any critical curve ξ satisfies the differential inclusion*

$$\dot{\xi} \in \partial_p H(\xi, p(\xi)) \quad \text{for a.e. } t \in \mathbb{R}.$$

In the statement ∂_p denotes the generalized gradient with respect to the variable p .

The proof of Theorem 35 is now attacked. Combining Proposition 17 and Theorem 37, it is shown that there exists a critical subsolution, say w , differentiable at any point of M , strict and of class C^∞ outside the Aubry set, and with $Dw|_{\mathcal{A}} = p(\cdot)$ continuous. The problem is therefore to adjust the proof of Proposition 14 in order to have continuity of the differential on the whole M .

The first step is to show a stronger version of the Proposition 16 asserting that it is possible to find a critical subsolution u , which is not only strict on $W_0 = M \setminus \mathcal{A}$,

but also strictly differentiable on \mathcal{A} . Recall that this means that if $y \in \mathcal{A}$ and x_n are differentiability points of u with $x_n \rightarrow y$, then $Du(x_n) \rightarrow Du(y) = p(y)$. This implies that if $\bar{p}(\cdot)$ is a continuous extension of $p(\cdot)$ in M , then $\bar{p}(x)$ and $Du(x)$ are close at every differentiability point of u close to \mathcal{A} .

Starting from a subsolution u enjoying the previous property, the idea is then to use for defining the sought C^1 subsolution, say v , the same formula (20) given in Proposition 14, i. e.

$$v = \begin{cases} \sum \beta_i u_i & \text{in } W_0 \\ u & \text{in } \mathcal{A} \end{cases}$$

where β_i is a C^∞ partition of unity subordinated to a countable locally finite open covering U_i of W_0 , and the u_i are obtained from u through suitable regularization in U_i by mollification. Look at the sketch of the proof of Proposition 16 for the precise properties of these objects. It must be shown

$$D\left(\sum \beta_i u_i(x_n)\right) \rightarrow Du(y) = p(y),$$

for any sequence x_n of elements of W_0 converging to $y \in \mathcal{A}$, or equivalently

$$\left| \bar{p}(x_n) - D\left(\sum \beta_i u_i(x_n)\right) \right| \rightarrow 0.$$

One has

$$\begin{aligned} \left| \bar{p}(x_n) - D\left(\sum \beta_i u_i(x_n)\right) \right| &\leq \sum_i \beta_i(x_n) |Du_i(x_n) \\ &\quad - \bar{p}(x_n)| D\beta_i(x_n) |u_i(x_n) - u(x_n)|. \end{aligned}$$

The estimates given in the proof of Proposition 14 show that the second term of the right-hand side of the formula is small. To estimate the first term calculate

$$\begin{aligned} |Du_i(x_n) - \bar{p}(x_n)| &\leq \int \zeta_{\delta_i}(z - x_n) (|Du(z) - \bar{p}(z)| \\ &\quad + |\bar{p}(z) - \bar{p}(x_n)|) dz, \end{aligned}$$

and observe first that $|Du(z) - \bar{p}(z)|$ is small, as previously explained, since $x_n \rightarrow y \in \mathcal{A}$ and z is close to x_n , second, that the mollification parameter δ_i can be chosen in such a way that $|\bar{p}(z) - \bar{p}(x_n)|$ is also small.

What is left is to discuss the density issue of the C^1 subsolutions. This is done still assuming $a = c$ and \mathcal{A} non-empty. The proof in the other cases is simpler and goes along the same lines.

It is clear from what was previously outlined that the initial subsolution u and v obtained as a result of the regularization procedure are close in the local uniform topology. It is then enough to show that any critical subsolution w can be approximated in the same topology by a subsolution enjoying the same property of u , namely being strict in W_0 and strictly differentiable on the Aubry set.

The first property is easy to obtain by simply performing a convex combination of w with a C^1 subsolution, strict in W_0 , whose existence has been proved above. It can be, in turn, suitably modified in a neighborhood of \mathcal{A} in order to obtain the strict differentiability property, see [20].

Future Directions

This line of research seems still capable of relevant developments. In particular to provide exact and approximate correctors for the homogenization of Hamilton–Jacobi equations in a stationary ergodic environment, see Lions and Souganidis [26] and Davini and Siconolfi [11,12], or in the direction of extending the results about long-time behavior of solutions of time-dependent problems to the noncompact setting, see Ishii [22,23]. Another promising field of utilization is in mass transportation theory, see Bernard [5], Bernard and Buffoni [6], and Villani [30]. The generalization of the model in the case where the Hamiltonian presents singularities should also make it possible to tackle through these techniques the N -body problem. With regard to applications, the theory outlined in the paper could be useful for dealing with topics such as the analysis of dielectric breakdown as well as other models in fracture mechanics.

Bibliography

1. Arnold WI, Kozlov WW, Neishtadt AI (1988) Mathematical aspects of classical and celestial mechanics. In: Encyclopedia of Mathematical Sciences: Dynamical Systems III. Springer, New York
2. Bardi M, Capuzzo Dolcetta I (1997) Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman equations. Birkhäuser, Boston
3. Barles G (1994) Solutions de viscosité des équations de Hamilton–Jacobi. Springer, Paris
4. Barles G, Souganidis PE (2000) On the large time behavior of solutions of Hamilton–Jacobi equations. SIAM J Math Anal 31:925–939
5. Bernard P (2007) Smooth critical subsolutions of the Hamilton–Jacobi equation. Math Res Lett 14:503–511
6. Bernard P, Buffoni B (2007) Optimal mass transportation and Mather theory. J Eur Math Soc 9:85–121
7. Buttazzo G, Giaquinta M, Hildebrandt S (1998) One-dimensional Variational Problems. In: Oxford Lecture Series in Mathematics and its Applications, 15. Clarendon Press, Oxford

8. Clarke F (1983) Optimization and nonsmooth analysis. Wiley, New York
9. Contreras G, Iturriaga R (1999) Global Minimizers of Autonomous Lagrangians. In: 22nd Brazilian Mathematics Colloquium, IMPA, Rio de Janeiro
10. Davini A, Siconolfi A (2006) A generalized dynamical approach to the large time behavior of solutions of Hamilton–Jacobi equations. *SIAM J Math Anal* 38:478–502
11. Davini A, Siconolfi A (2007) Exact and approximate correctors for stochastic Hamiltonians: the 1-dimensional case. to appear in *Mathematische Annalen*
12. Davini A, Siconolfi A (2007) Hamilton–Jacobi equations in the stationary ergodic setting: existence of correctors and Aubry set. (preprint)
13. Evans LC (2004) A survey of partial differential methods in weak KAM theory. *Commun Pure Appl Math* 57
14. Evans LC, Gomes D (2001) Effective Hamiltonians and averaging for Hamilton dynamics I. *Arch Ration Mech Anal* 157: 1–33
15. Evans LC, Gomes D (2002) Effective Hamiltonians and averaging for Hamilton dynamics II. *Arch Ration Mech Anal* 161: 271–305
16. Fathi A (1997) Solutions KAM faibles et barrières de Peierls. *C R Acad Sci Paris* 325:649–652
17. Fathi A (1998) Sur la convergence du semi-groupe de Lax–Oleinik. *C R Acad Sci Paris* 327:267–270
18. Fathi A () Weak Kam Theorem in Lagrangian Dynamics. Cambridge University Press (to appear)
19. Fathi A, Siconolfi A (2004) Existence of C^1 critical subsolutions of the Hamilton–Jacobi equations. *Invent Math* 155:363–388
20. Fathi A, Siconolfi A (2005) PDE aspects of Aubry–Mather theory for quasiconvex Hamiltonians. *Calc Var* 22:185–228
21. Forni G, Mather J (1994) Action minimizing orbits in Hamiltonian systems. In: Graffi S (ed) *Transition to Chaos in Classical and Quantum Mechanics*. Lecture Notes in Mathematics, vol 1589. Springer, Berlin
22. Ishii H (2006) Asymptotic solutions for large time Hamilton–Jacobi equations. In: *International Congress of Mathematicians*, vol III. Eur Math Soc, Zürich, pp 213–227
23. Ishii H () Asymptotic solutions of Hamilton–Jacobi equations in Euclidean n space. *Anal Non Linéaire, Ann Inst H Poincaré* (to appear)
24. Koike S (2004) A beginner’s guide to the theory of viscosity solutions. In: *MSJ Memoirs*, vol 13. Tokyo
25. Lions PL (1987) Papanicolaou G, Varadhan SRS, Homogenization of Hamilton–Jacobi equations. Unpublished preprint
26. Lions PL, Souganidis T (2003) Correctors for the homogenization of Hamilton–Jacobi equations in the stationary ergodic setting. *Commun Pure Appl Math* 56:1501–1524
27. Roquejoffre JM (2001) Convergence to Steady States of Periodic Solutions in a Class of Hamilton–Jacobi Equations. *J Math Pures Appl* 80:85–104
28. Roquejoffre JM (2006) Propriétés qualitatives des solutions des équations de Hamilton–Jacobi et applications. *Séminaire Bourbaki* 975, 59ème année. Société Mathématique de France, Paris
29. Siconolfi A (2006) Variational aspects of Hamilton–Jacobi equations and dynamical systems. In: *Encyclopedia of Mathematical Physics*. Academic Press, New York
30. Villani C () Optimal transport, old and new. <http://www.umpa.ens-lyon.fr/cvillani/>. Accessed 28 Aug 2008
31. Weinan E (1999) Aubry–Mather theory and periodic solutions of the forced Burgers equation. *Commun Pure and Appl Math* 52:811–828

Health Care in the United Kingdom and Europe, System Dynamics Applications to

ERIC WOLSTENHOLME^{1,2}

¹ South Bank University, London, UK

² Symmetric SD, Brighton, UK

Article Outline

Glossary

Definition of the Subject

Introduction

The History of System Dynamics

The Need for System Dynamics

The Components of System Dynamics

An Overview of Health and Social Care

in the UK and Europe

A Case Study: Using System Dynamics to Influence

Health and Social Care Policy Nationally

in the UK – Delayed Hospital Discharges

Review of System Dynamics Studies

in Epidemiology in Europe

Review of System Dynamics Studies in Health

and Social Care Management in Europe

System Dynamics Workforce Planning Models

to Support Health Management

Future Directions

Acknowledgments

Bibliography

This paper describes the application of system dynamics to health and social care in Europe.

Systems thinking and the simulation tool set of system dynamics are introduced together with an overview of current strategic health issues and responses in the UK and Europe. A case study is then presented to demonstrate how effective and apposite system dynamics studies can be. This is followed by a pan-European review of applications of system dynamics in epidemiology and in health treatment and diagnosis in different sectors of health and social care, based on an extensive bibliography. Reference is also made to health workforce planning studies. Lastly, a review of future directions is described.

The knowledge base of this paper is located in published work by internal and external consultants and Universities, but it should also be said that there is far more

work in system dynamics in health than is referred to in these sources. Many internal and external consultancies undertake studies which remain unpublished.

The description of the subject and the applications described are comprehensive, but the review is a personal interpretation of the current state of a fast-moving field by the author and apologies are made in advance for any unintended omissions.

The case study in Sect. “[A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges](#)” is extracted from material published by Springer-Verlag, US and published with their permission.

Glossary

System dynamics

System A collection of elements brought together for a purpose and whose sum is greater than the parts.

Systems thinking The process of interpreting the world as a complex, self regulating and adaptive system.

System dynamics A method based on quantitative computer simulation to enhance learning and policy design in complex systems.

Qualitative system dynamics The application of systems thinking and system dynamics principles, without formal simulation.

Dynamic complexity The number of interacting elements contained in a system and the consequences of their interactions over time.

Human activity system Any system created and regulated by human intervention.

Reductionism The opposite of systemic – seeing the world only in its constituent parts.

Feedback Feedback refers to the interaction of the elements of the system where a system element, X, affects another system element, Y, and Y in turn affects X perhaps through a chain of causes and effects. Feedback thus controls the performance of the system. Feedback can be either natural or behavioral (created by human intervention) (System Dynamics Society).

Unintended consequences Undesirable consequences arising well intended action – or vice versa.

Continuous simulation The aggregate method of computer simulation used in system dynamics based on a continuous time analogy with fluid dynamics and used to test out patterns of behavior over time.

System structure The term used in system dynamics to refer to the total structure of a system (composing processes, organization boundaries, information feedback, policy and delays).

System behavior The term used in system dynamics to refer to the behavior over time of a particular structure.

Reference mode of behavior An observed past trend and future projected trends used to assist defining model scope and time frame.

Discrete entity simulation A method of simulation based on the movement of individual entities through systems over time either as processes or as interactions between entities.

Health and Social Care

Epidemiology The study of factors affecting the health and the incidence and prevalence of illness of populations.

Health treatment The application of drugs, therapies, and medical/surgical interventions to treat illness.

National health service (NHS) The organization in the UK responsible for the delivery of health care.

Primary care trusts (PCTs) The local operating agencies of the NHS, which both commission (buy) and deliver health services.

General practitioners (GPs) Locally-based general clinicians who deliver primary care services and control access to specialist health services.

Social services In England, care services which provide non-health related care, mainly for children and older people, located within local government in the UK.

Nursing/residential home care In England, private and public residential establishments for the care of older people.

Domiciliary care In England, care for older people in their own homes.

Acute hospitals Hospital dealing with short term conditions requiring mainly one-off treatment.

Outliers Patients located in hospital in wards not related to their condition, due to bed capacity issues.

Intermediate care Short term care to expedite the treatment of non-complex conditions.

Definition of the Subject

All too often complexity issues are ignored in decision making simply because they are just too difficult to represent. Managers feel that to expand the boundaries of the decision domain to include intricate, cross-boundary interconnections and feedback will detract from the clarity of the issue at stake. This is particularly true when the interconnections are behavioral and hard to quantify. Hence, the focus of decision making is either very subjective or based on simple, linear, easy to quantify components. However, such a reductionist stance, which ig-

nore information feedback (for example, the effects of health supply on health demand management) and multiple-ownership of issues can result in unsustainable, short term benefits with major unintended consequences.

System dynamics is a particular way of thinking and analyzing situations, which makes visible the dynamic complexity of human activity systems for decision support.

It is particularly important in the health and social care field where there are major issues of complexity associated with the incidence and prevalence of disease, an aging population, a profusion of new technologies and multiple agencies responsible for the prevention and treatment of illness along very long patient pathways. Health is also linked at every stage to all facets of life and health policy has a strong political dimension in most countries.

Introduction

This paper describes and reviews work in applying system dynamics to issues of health and social care in the UK and Europe. Although the fundamental issues in health and social care and many of the strategies adopted are similar the world over, there are differences in culture, operational policies and funding even over short geographical distances. Additionally, the health field can be dissected in many different ways both internally and between countries.

There is, moreover, a fundamental dilemma at the center of health that determines both its structure and emphasis. Although the real long term and systemic solution to better health lies in the prevention of illness, the health field focuses on the study of the incidence and prevalence of disease (Epidemiology) and on the 'health service' issues of how to manage ill health (Health Diagnosis and Treatment).

There are many reasons for this, not the least being that illness prevention is in fact the province of a field much bigger than health, which includes economics, social deprivation, drugs, poverty, power and politics.

The field of system dynamics in health reflects this dilemma. Whilst all studies would conclude that prevention is better than the cure, the majority of applications focus on illness. Whilst more studies are required on the truly systemic goal of moving attention away from the status quo, for example, modeling the German system of health care and drug addicts [52], the major focus and impact of system dynamics in Europe in recent years has been in terms of Epidemiology and Health Treatment. Hence, it is these categories that will be the focus of this paper. However, work often transcends the two and models often in-

clude both disease and treatment states. For example, work on AIDS covers both prevalence and drug treatment and work on long term conditions, particularly mental health conditions, covers condition progression as well as alternative therapies.

It is important to emphasize what this paper does not cover. By definition system dynamics is a strategic approach aimed at assisting with the understanding of high level feedback effects at work in organizations. It is therefore separate from the many applications of spreadsheets and discrete entity simulation methods applied to answer short term operational level issues in health [9,21,29].

It is also important to note where the knowledge base of this paper is located. System dynamics applications in health in Europe began in the 1980s and are expanding rapidly. However, as will be seen from the bibliography to this paper, much of the work is applied by internal and external consultants and Universities for health care managers and reported in management, operational research and system dynamics journals. Little of the work so far has been addressed directly at clinicians or published in the health literature. It should also be said that there is far more work in system dynamics in health than is referred to in this publication. Many internal and external consultancies undertake studies which remain unpublished.

Initially the fundamentals of system dynamics will be described followed by an overview of current health issues and responses in the UK and Europe. This is followed by a case study to demonstrate how effective and apposite system dynamics studies can be. There then follows a review of applications in epidemiology and in both physical and mental health diagnosis and treatment. Mention is also made of health workforce planning studies. Lastly, a review of future directions is described.

The History of System Dynamics

System dynamics was conceived at MIT, Boston in the late 60s and has now grown into a major discipline [25,47] which was formally celebrated and reviewed in 2008 [48]. It is widely used in the private business sector in production, marketing, oil, asset management, financial services, pharmaceuticals and consultancy. It is also used in the public sector in defense, health and criminal justice.

System dynamics has a long history in the UK and Europe. The first formal university group was established at the University of Bradford in England 1970. Today there are at least a dozen university departments and business schools offering courses in system dynamics and numerous consultancies of all types using the method in one form or another. Thousands of people have attended pri-

vate and university courses in system dynamics and, additionally, there are almost one hundred UK members of the System Dynamics Society, which is the largest national grouping outside the US.

The Need for System Dynamics

Most private and public organizations are large and complex. They exhibit both ‘detailed’ complexity (the number of elements they contain), but more importantly ‘dynamic’ complexity (the number of interconnections and interactions they embrace). They have long processes which transcend many sectors, each with their own accounting and performance measures. In the case of health and social care organizations this translates into long patient pathways across many agencies. Complexity and decision making in the public sector is also compounded by a multitude of planning time horizons and the political dimension.

Long processes mean that there are many opportunities for intervention, but that the best levers for overall improvement are often well away from symptoms of problems. Such interventions may benefit sectors other than those making the investments and require an open approach to improving patient outcomes, rather than single agency advantage.

The management of complex organizations is complicated by the fact that human beings have limited cognitive ability to understand interconnections and consequently have limited mental models about the structure and dynamics of organizations.

A characteristic of complex organizations is a tendency for management to be risk averse, policy resistant and quick to blame. This usually means they prefer to stick to traditional solutions and reactive, short term gains. In doing this managers ignore the response of other sectors and levels of the organization. In particular, they underestimate the role and effect of behavioral feedback.

Such oversight can result in unintended consequences in the medium term that undermine well-intended actions. Self organizing and adaptive responses in organizations can lead to many types of informal coping actions, which in turn, inhibit the realization of improvement attempts and distort data. A good example of these phenomena, arising from studies described here, is the use of ‘length of stay’ in health and social care services as a policy lever to compensate for capacity shortages.

Planning within complex organization reflects the above characteristics. The core of current planning tends to be static in nature, sector-based and reliant on data and financial spreadsheets with limited transparency of assumptions. For example the planning of new acute hos-

pitals can quickly progress to detailed levels without assessment of trends in primary and post acute care; that is, where hospital patients come from and go to.

In contrast, sustainable solutions to problems in complex organizations often require novel and balanced interventions over whole processes, which seem to defy logic and may even be counterintuitive.

However, in order to realize such solutions requires a leap beyond both the thinking and planning tools commonly used today. In order to make significant changes in complex organizations it is necessary to think differently and test ideas before use. System dynamics provides such a method.

The Components of System Dynamics

System dynamics is based on the idea of resisting the temptation to be over reactive to events, learning instead to view patterns of behavior in organizations and ground these in the structure (operational processes and policies) of organizations. It uses purpose-built software to map processes and policies at a strategic level, to populate these maps with data and to simulate the evolution of the processes under transparent assumptions, policies and scenarios.

System dynamics is founded upon:

- Non linear dynamics and feedback control developed in mathematics, physics and engineering,
- Human, group and organizational behavior developed in cognitive and social psychology and economics,
- Problem solving and facilitation developed in operational research and statistics.

System dynamics provides a set of *thinking* skills and a set of *modeling* tools which underpin the current trend of ‘whole systems thinking’ in health and social care.

System Dynamics Thinking Skills for the Management of Complex Organizations

In order to understand and operate in complex organizations it is necessary to develop a wide range of thinking skills [45]. The following are summarized after Richmond [42].

- **Dynamic thinking** – The ability to conceptualize how organizations behave over time and how we would like them to behave.
- **System-as-cause thinking** – The ability to determine plausible explanations for the behavior of the organization over time in terms of past actions.
- **Forest thinking** – The ability to see the “big picture” (transcending organizational boundaries).

- **Operational thinking** – The ability to analyze the contribution made to the overall behavior by the interaction of processes, information feedback, delays and organizational boundaries.
- **Closed-loop thinking** – The ability to analyze feedback loops, including the way that results can feedback to influence causes.
- **Quantitative thinking** – The ability to determine the mathematical relationships needed to model cause and effect.
- **Scientific thinking** – The ability to construct and test hypotheses through modeling.

System Dynamics Modeling Tools for Planning in Complex Organizations

A useful way to appreciate the tool set of system dynamics is by a brief comparison with other computer based management tools for decision support.

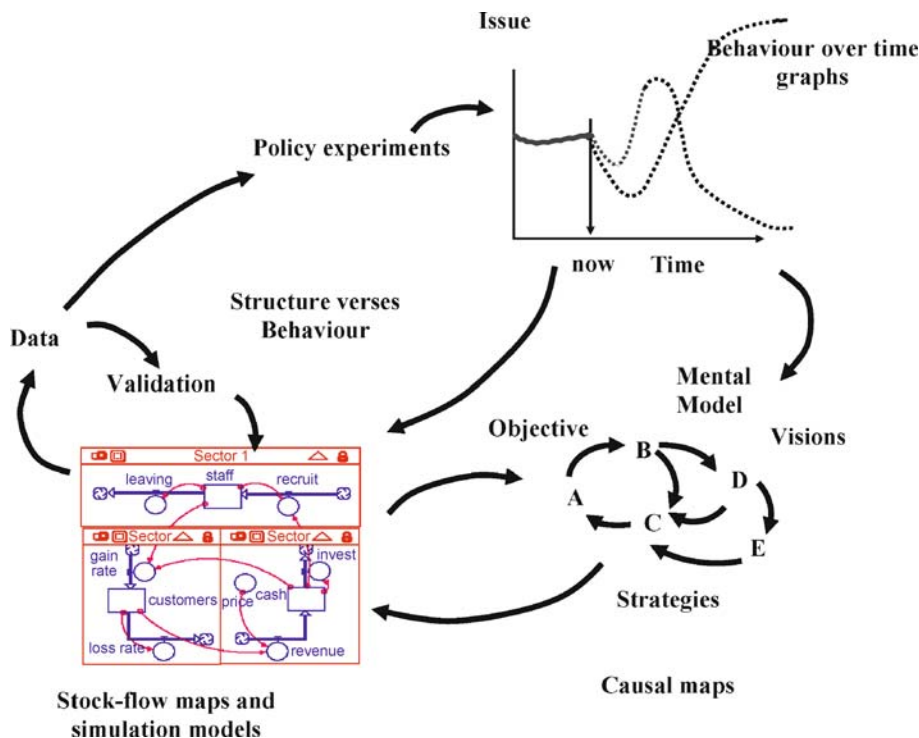
System dynamics is, by definition, a strategic rather than operational tool. It can be used in a detailed operational role, but is first and foremost a *strategic* tool aimed at integrating policies across organizations, where behavioral feedback is important. It is unique in its ability to address the strategic domain and this places it apart from more op-

erational toolsets such as process mapping, spreadsheets, data analysis, discrete entity simulation and agent-based simulation.

System dynamics is based on representing process flows by 'stock' and 'rate' variables. Stocks are important measurable accumulations of physical (and non-physical) resources in the world. They are built and depleted over time as input and output rates to them change under the influence of feedback from the stocks and outside factors. Recognizing the difference between stocks and rates is fundamental to understanding the world as a system. The superimposition of organizational sectors and boundaries on the processes is also fundamental to understanding the impact of culture and power on the flows. System dynamics also makes extensive use of causal maps to both help conceptualize models and to highlight feedback processes within models.

Applying System Dynamics with Management Teams

However, the success of system dynamics lies as much in its process of application as in the tool set and hence demands greater skill in conceptualization and use than spreadsheets.



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 1
The systems thinking/system dynamics method

Figure 1 shows the overall process of applying system dynamics. A key starting point is the definition of an initial significant issue of managerial concern and the establishment of a set of committed and consistent management teams from all agencies involved in the issue. Another requirement is a set of facilitators experienced in both conceptualizing and formulating system dynamics models. The models created must be shared extensions of the mental models of the management teams, not the facilitators and, importantly owned by the team.

The next step is the analysis of existing trends in major performance measures of the organizations and of their future trajectories, desired and undesired. This is referred to as the reference model of behavior of the issue and helps with the establishment of the time scale of the analysis. The key contribution of system dynamics is then to formulate a high level process map, at an appropriate level of aggregation, linking operations across organizations and to populate this with the best data available. Once validated against past data, the mental models of the management team and shown capable of reproducing the reference mode of behavior of the issue ('what is'), the model is used to design policies to realize desired futures ('what might be'). Maps and models are constructed in relatively inexpensive purpose-built software (for example *ithink*, *Vensim* and *Powersim*) with very transparent graphical interfaces.

The key is to produce the simplest model possible consistent with maintaining its transparency and having confidence in its ability to cast new light on the issue of concern. This means keeping the resolution of the model at the highest possible level and this distinguishes it from most spreadsheets and process maps.

An Overview of Health and Social Care in the UK and Europe

Ensuring that all residents have access to health and social care services is an important goal in all EU countries and all have universal or almost universal health care coverage (European Observatory 'Healthcare in Transition' profiles and OECD Health Data 2004). Even in the Netherlands, where only 65% of the population are covered by a compulsory scheme, with voluntary private insurance available to the remainder, only 1.6% of the population are without health insurance.

At the present time, most care in the EU is publicly financed, with taxation and social insurance provide the main sources of funding. Taxation is collected at either the national level or local level, or both and social insurance contributions are generally made by both employees and

employers. The role of private insurance varies between countries and generally private insurance is as a supplement to, rather than as a substitute for, the main care system. The exceptions to this are Germany and the Netherlands. Further, people are increasingly required to pay part of the cost of medical care.

The delivery of health and social care is a mixture of public and private with only 10 countries not having any private delivery sector at all.

This paper is primarily concerned with health and social care supply issues. Although the structure and terminology associated with supply varies across the EU the underlying issues tend to be similar between countries. Hence the major issues will be described for England.

Health in England is primarily managed and delivered by the National Health Service (NHS) and is at the center of a modernization agenda, whereby the government sets out a program of change and targets against which the public may judge improved services.

A major mechanism for reform tends to be via frequent changes to organizational structure. The current structure consist of large primary care trusts (PCTs), which both deliver services such as General Practitioner Services (GPs), but also purchase (commission) more specialist services from other agencies, both public and private. A key driver of structural change is to enhance primary care and to take the pressure off acute hospitals (acute is a word used to differentiate short term hospitals from long stay ones). Initiatives here center on providing new services, such as diagnostic and treatment centers and shorter term 'intermediate' care. Emphasis is on bringing the services to the users, patient choice, payment by results (rather than through block contracts) and service efficiency, the latter being driven by target setting and achievement. The government has made reform of public services a key plank in its legislative program and pressure to achieve a broad range of often conflicting targets is therefore immense. However, despite continual increases in funding new initiatives are slow to take effect and the performance and viability of the service is problematic with money often being used to clear deficits rather than generate new solutions.

Social care in England is delivered both by a public sector located with Local Government Social Services Directorates and a private sector. It consists of numerous services to support children and older people. The latter consisting of care homes, nursing homes and domiciliary (at home) care.

Many patient processes, particularly for older people, transcend health and social care boundaries and hence create a serious conflict of process structure and organi-

zational structure, where the relative power of the different agencies is a major determinant of resource allocation [64]. Consequently, emphasis in this paper will be on joint health and social care work.

A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges

In order to give a flavor of the relevance and impact of applying system dynamics to health and social care issues a concise case study will be presented [65,67,70].

Issue

Delayed hospital discharge was an issue which first came onto the UK legislative agenda in late 2001. The ‘reference mode’ of behavior over time for this situation was that of increasing numbers of patients occupying hospital beds, although they had been declared “medically fit”. In March 2002, 4,258 people were “stuck” in hospital and some were staying a long time, pushing up the number of bed days and constituting significant lost capacity.

The government’s approach to this issue was to find out who was supposed to “get the patients out” of acute hospitals and threaten them with ‘fines’ if they did not improve performance. This organization proved to be social services for older people, who are located within the local government sector and who are responsible for a small, but significant, number of older people needing ex-hospital (‘post-acute’) care packages. Such patients are assessed and packages organized by hospital social workers. There was also pressure on the government from hospitals claiming that some of the problem was due to lack of hospital capacity.

The idea of fines was challenged by the Local Government Association (LGA), which represents the interests of all local government agencies at the national level) who suggested that a ‘system’ approach should be undertaken to look at the complex interaction of factors affecting delayed hospital discharges. This organization, together with the NHS Confederation (the partner organization representing the interests of the National Health Service organizations at a national level) then commissioned a system dynamics study to support their stance.

The remit was for consultants working with the representatives of the two organizations to create a system dynamics model of the ‘whole patient pathway’ extending upstream and downstream from the stock of people delayed in hospital, to identify and test other interventions affecting the issue.

Model

A system dynamics model was developed interactively with managers from the LGA and NHS, using national data to simulate pressures in a sample health economy covering primary, acute and post acute care over a 3 year period. The model was driven by variable demand including three winter pressure “peaks” when capacity in each sector was stretched to the limit. Figure 2 shows an overview of the sectors of the model.

The patient flows through the model were broken down into medical flows and surgical with access to the medical and surgical stocks of beds being constrained by bed capacity. The medical flows were mainly emergency patients and the surgical flows mainly non-emergency ‘elective’ patients, who came via referral processes and wait lists.

Further, medical patients were broken down into ‘fast’ and ‘slow’ streams. The former were the normal patients who had a short stay in hospital and needed few post acute services and the latter the more complex cases (mainly older people), who require a longer stay and hospital and complex onward care packages from social services. This split was because although the slow patients were few in number they constituted most of the people who caused delayed discharges.

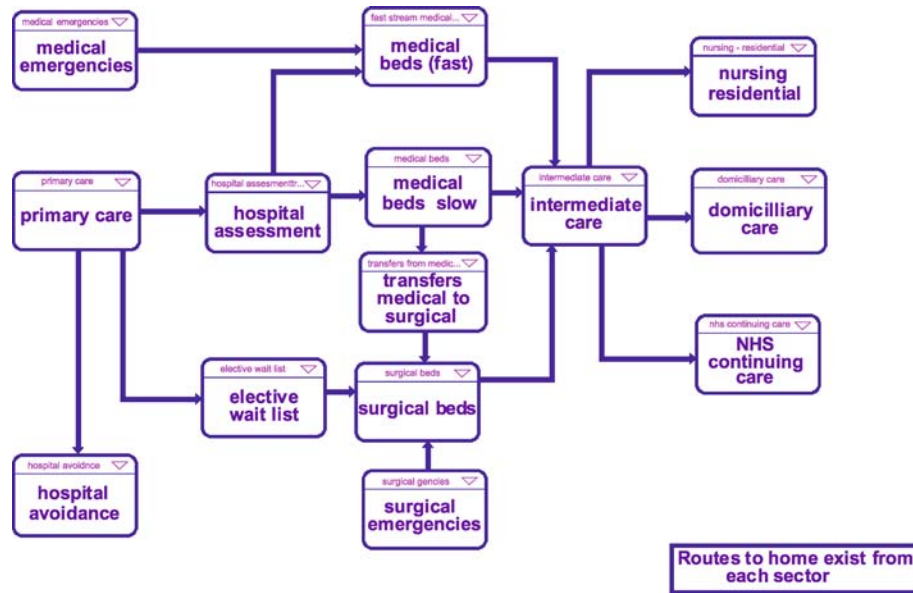
The post hospital health and social care services of intermediate care, nursing/residential home care, and domiciliary care were included in the model and were also capacity constrained in terms of the number of care packages they could provide.

The model incorporated a number of mechanisms by which hospitals coped during periods of high demand, for example, moving medical patients to surgical beds (outliers) and early discharges with allowance for readmissions.

Configuration of the Model

The model was set up to simulate a typical sample health economy over a 3 year period when driven by a variable demand (including three winter “peaks”). The capacity constrained sectors of the model were given barely sufficient capacity to cope. This situation was designed to create shocks against which to test alternative policies for performance improvement. Major performance measures in use in the various agencies were incorporated. These included:

1. Cumulative episodes of elective surgery.
2. Elective wait list size and wait time.



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 2
An overview of the sectors of the delayed discharge model

3. Numbers of patients in hospital having completed treatment and assessment, but not yet discharged (delayed discharges).
4. Number of 'outliers'.

The model was initially set up with a number of fixed experiments, to introduce people to the range of experiments that yielded useful insights into the behavior of the whole system. From there, they were encouraged to devise their own experiments and develop their own theories of useful interventions and commissioning strategies.

The three main policies tested in the fixed runs were:

1. Adding additional acute hospital bed capacity. This is the classic response used over many years by governments throughout the world to solve any patient pathway capacity problems and was a favorite 'solution' here.
2. Adding additional post acute capacity, both nursing and residential home beds but also more domiciliary capacity.
3. Diverting more people away from hospital admission by use of pre-hospital intermediate capacity and also expansion of treatment in primary care GP surgeries.

Example Results from the Delayed Hospital Discharge Model

Figures 3, 4 and 5 show some typical outputs for the delayed hospital discharge model. Figure 3 captures the way

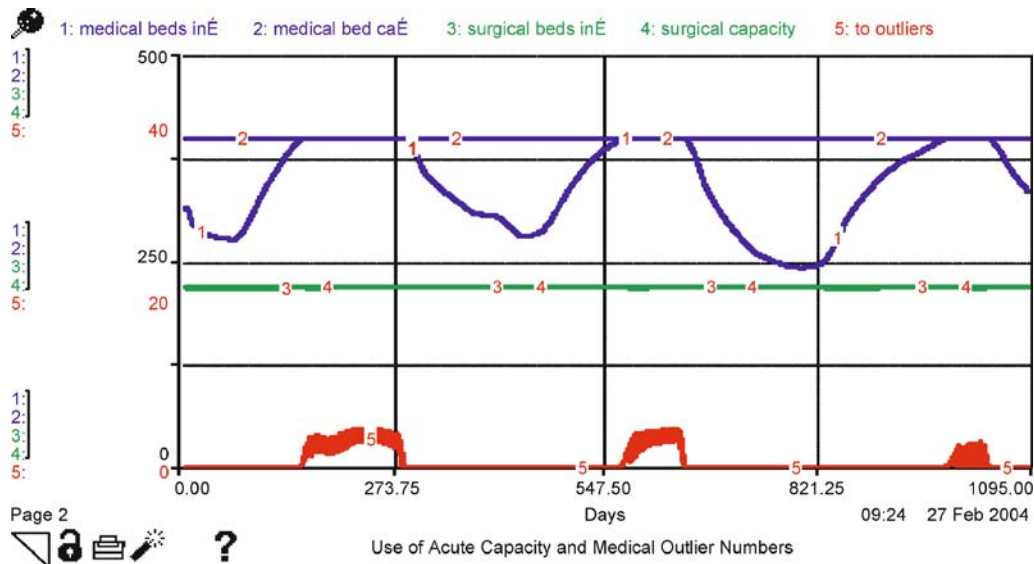
capacity utilization was displayed (actual beds occupied v total available for both medical and surgical sectors of the hospital) and shows the occurrence of 'outliers' (transfers of patients from medical to surgical beds) whenever medical capacity was reached.

Figures 4 and 5 show comparative graphs of 3 policy runs for 2 major performance measures for 2 sectors of the patient pathway – delayed discharges for post acute social services and cumulative elective procedures for acute hospitals. In each case the base run is line 1. Line 2 shows the effect of increasing hospital beds by 10% and line 3 shows the effect of increasing post acute capacity by 10%.

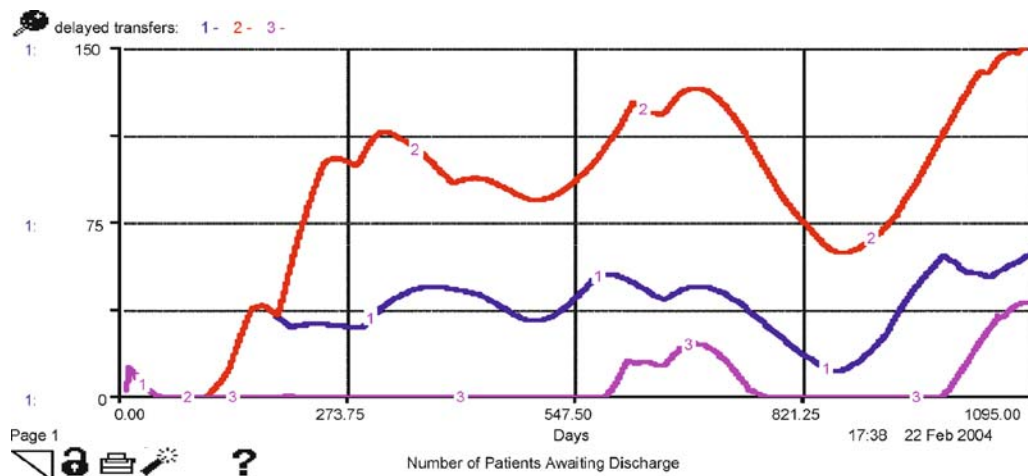
The interesting feature of this example output is that the cheaper option of increasing post acute capacity gives lower delayed discharges and higher elective operations whereas the more expensive option of increasing acute hospital beds benefits the hospital but makes delayed discharges worse. The key to this counter intuitive effect is that increasing post acute capacity results in higher hospital discharges which in turn reduces the need for the 'outlier' coping policy in the hospital, hence freeing up surgical capacity for elective operations.

Outcomes

Common Sense Solutions Can Be Misleading The obvious unilateral solution of adding more acute capacity was shown to exacerbate the delayed discharge situation. Increasing hospital capacity means facilitating more hos-



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 3
Medical and surgical bed utilization's in hospital and 'outliers'



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 4
Delayed hospital discharges for 3 policy runs of the model

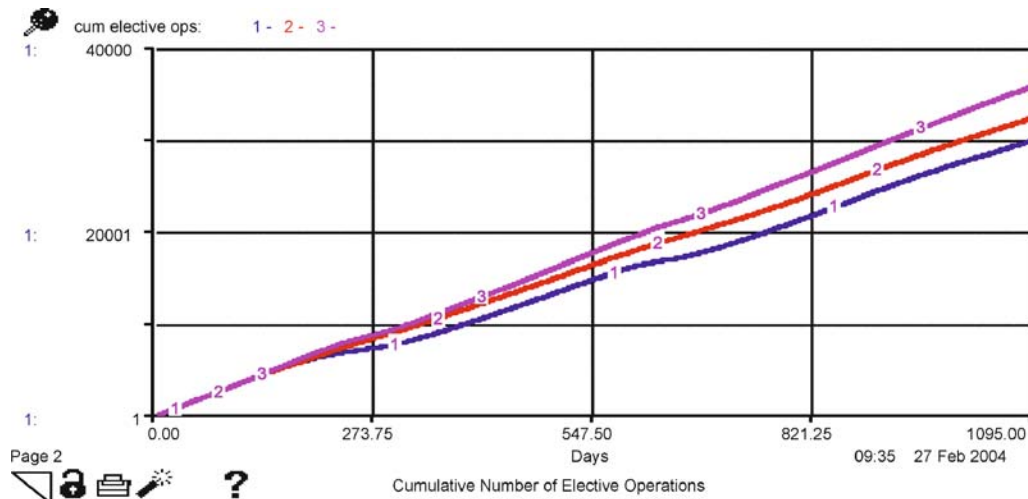
pital admissions, but with no corresponding increase in hospital discharges. Hence, the new capacity will simply fill up and then more early discharges and outliers will be needed.

Fines May Have Unintended Consequences This solution was shown to depend on where the money raised by fines was spent. If the money levied from social services was given to the acute sector to finance additional capacity it was clearly demonstrated that this would make delayed discharges worse. It would be worse still if it causes the post-acute sector to cut services. The effects of service cuts

may also then spill over into other areas of local government including housing and education.

It was demonstrated that there were some interventions that could help:

1. Increasing post acute capacity gives a win-win solution to both health and social care because it increases all acute and post acute sector performance measures. Such action allows hospital discharges to directly increase, and eliminates the need for the hospitals to apply coping policies, which in turn increases elective operations and reduces elective wait times. Further,



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 5
Cumulative elective operations for 3 policy runs of the model

counter intuitively, increasing medical capacity in hospital is more effective than increasing surgical capacity for reducing elective wait times.

2. Reducing assessment times and lengths of stay in all sectors is beneficial to all performance measures, as is reducing variation in flows, particularly reinforcing feedback loops like re-admission rates.
3. Increasing diversion from hospitals into pre-admission intermediate care was almost as beneficial as increasing post acute capacity.
4. If fines are levied they need to be re-invested from a whole systems perspective. This means re-balancing resources across all the sectors (NOT just adding to hospital capacity).
5. In general the model showed that keeping people out of hospital is more effective than trying to get them out faster. This is compounded by the fact that in-patients are more prone to infections so the longer patients are in hospital, the longer they will be in hospital.
6. Improving the quality of data was shown to be paramount to realizing the benefits of all policies. This is an interesting conclusion associated with many system dynamics studies, where explicit representation of the structure of the organization can lead to a review and redesign of the information needed systems to really manage the organization.

An interesting generalization of the findings was that increasing stock variables where demand is rising (such as adding capacity) is an expensive and unsustainable solution. Whereas increasing rate variables, by reducing delays and lengths of stay, is cheaper and sustainable.

Impact

This model was shown at the Political Conferences of 2002 and generated considerable interest. It was instrumental in causing re-thinking of the intended legislation, so that social services was provided with investment funding to address capacity issues, and the implementation of fines was delayed for a year. Reference to the model was made in the House of Lords.

Moving the main amendment, Liberal Democrat health spokesperson Lord Clement-Jones asked the House to agree that the Bill failed to tackle the causes of delayed discharges and would create perverse incentives which would undermine joint working between local authorities and the NHS and distort priorities for care of elderly people by placing the requirement to meet discharge targets ahead of measures to avoid hospital admission ... **He referred to "ithink", the whole systems approach being put forward by the Local Government Association, health service managers and social services directors involving joint local protocols and local action plans prepared in co-operation.**

Postscript

This case study demonstrates the ability of system dynamics to be applied quickly and purposefully to shed rigor and insight on an important issue. The study enabled the development of a very articulate and compelling case for the government to move from a reactive position of blaming social services to one of understanding and acting on

a systemic basis. The whole project including modeling and communication of the outcomes was completed in 6 weeks.

Review of System Dynamics Studies in Epidemiology in Europe

The potential for system dynamics in population health and disease control began in the UK in the late eighties/early nineties with the extensive studies carried out on AIDS modeling. The majority of these studies were by Prof. Brian Dangerfield and Carole Roberts and were ongoing until 2000 [14,15,16,17,18,19,20].

The earlier studies [16] used a transition model to portray the nature of the disease and to better specify the types of data collection required for further developments of the model. The model was then developed further over the years [18,19] and was fed with time-series data of actual cases. This enabled projections of future occurrence to be forecast. The latter models were more concerned with examining the resource and cost implications of treatments given to HIV positive individuals and at their varying stages up until the ensuing onset of AIDS.

A recent study by Dangerfield et al. [20] saw further development of the original model with parameter optimization and recent data on the spread of AIDS in the UK was also integrated. The rationale for the update of the model was to investigate the recent dramatic decrease in diagnosed Aids cases in the West. The model assesses the effects of relatively new emergent triple antiretroviral therapy given to HIV patients causing this reduction and examines the possibility of continuity of the effectiveness of this therapy.

Dangerfield explains some of the reasons [13] why system dynamics acts as an excellent tool for epidemiological modeling. The positive and negative feed-back loops help imitate the natural disposition of the spread and containment of diseases amongst the general population. Further, system dynamics allows delays associated with the incubation predisposition of infectious diseases to be accurately and easily modeled without the need for complicated mathematical representation.

The work in the UK was complemented by work in Holland on simulation as a tool in the decision-making process to prevent HIV incidence among homosexual men [23] and on models for analysis and evaluation of strategies for preventing AIDS [32]. Further epidemiological studies in system dynamics in the UK related to the outbreak out of BSE and the subsequent infection of humans with its human form nvCJD [12].

These models are all characterized by modeling the flow of people through different stocks over time representing the different stages of the disease progression. The purpose of the model is then to test the effects of interventions aimed at slowing down the rate of progression of the condition or indeed moving people 'upstream' to less severe states of the condition.

Review of System Dynamics Studies in Health and Social Care Management in Europe

By far the greatest number of studies and publications in the use of system dynamics in health and social care is associated with patient flow modeling for health care planning. That is, the flow of patients through multiple service delivery channels. Patient pathway definition has been an area of health modernization and these pathways lend themselves to representation as stock/flow resource flows in system dynamics. The purpose of this type of modeling is to identify bottlenecks, plan capacity, reduce wait lists, improve the efficiency of patient assessments and times and the design of alternative pathways with shorter treatment times, (for example, intermediate care facilities both pre and post hospital treatment).

A characteristic of patient flows is that they are long and pass through multiple agencies and hence confront the major health issues of working across boundaries and designing integrated policies. Studies in this area have examined the flow of many different populations of patients and often resulted in arrayed models to represent the flow of different 'populations' or 'needs groups' through several parallel service channels.

The studies have covered both physical and mental conditions and have sometimes combined both the dynamic progression of people through undiagnosed and untreated disease states and the dynamic progression of diagnosed people through treatment pathways.

The Modeling of the Diagnosis and Treatment of Physical Conditions

Here the most common set of models are associated with the flow of patients from primary care, through acute hospitals and onwards into post acute care such as social services provisions for home care, nursing care and residential care. The populations have often been split between the simple everyday cases and the complex cases associated with older people needing greater degrees of care. They have also involves medical and surgical splits. There are a number of review papers which supplement the work described below [1,18,19].

In addition to work in the 1990s on the interface between health and social care [59,60,61] and the national level UK work on older people flows through hospitals [65,67,71,72], Wolstenholme has reported that system dynamics applications are currently underway by the authors in 10 local health communities around the UK with the objectives of modeling patient flows across agency boundaries to provide a visual and quantitative stimulus to strategic multi-agency planning [65].

Lane has reported work in Accident and Emergency Departments [33] and in mapping acute patient flows [34] whilst Royston worked with the NHS to help develop and implement policies and programs in health care in England [43]. Taylor has undertaken award winning modeling of the feedback effects of reconfiguring health services [49,50,51], whilst Lacey has reported numerous UK studies to support the strategic and performance management roles of health service management, including provision of intermediate care and reduction of delayed hospital discharges [31]. Other intermediate care and social care delivery studies are described by Bayer [6,7] and further hospital capacity studies by Coyle [11]. Elsewhere, there have been specific studies on bed-blocking [24] and screening [37].

In Norway system dynamics-based studies have focused on mapping the flows of patients in elderly non-acute care settings [10]. The purpose of this study according to Chen is to differentiate between acute and non-acute settings and thereby increase understanding of the complexity and dynamics caused by influencing elements in the system. Also it is to provide a tool for local communities in Norway for their long term budget planning in the non-acute health sector for the elderly.

Work on reducing waiting lists has been reported in Holland [30,53,54,57]). Also in Holland Vennix has reported comprehensive work on modeling a regional Dutch health care system [56].

Work has been undertaken to balance capacities in individual hospitals in Italy [44] and in Norway [38,41]. Whilst normally the realm of more operational types of simulation system dynamics has proved very effective here. There has also been work to assess the impact on health and social care of technological innovation, particularly telecare [5,8]. Additionally, system thinking has been undertaken by doctors to examine the European time directive [40].

Given the similar nature of a lot of these studies further detail here will focus on the work of Vennix in participative model building and Wolstenholme in extracting insights from numerous studies.

Participative Model Building

A characteristic of all Vennix's work has been group model building [55]. The main objectives of this [27] are communication and learning and integration of multiple perspectives where the process of model building is frequently more important than the resulting model itself [56]. Vennix brought together strategic managers and important stakeholders to participate in the process of building a system dynamics model of the Dutch healthcare system. The policy problem which is modeled in Vennix's 1992 study is related to the gradual, but persistent, rise in health care costs in the Netherlands. Vennix [56] attempts to find the underlying causes of those increases that emanate from within the health care system itself rather than focusing on exogenous factors. By doing so Vennix stands to identify potential levers within the health care system that can be practically and appropriately be adjusted to reduce cost increases.

Vennix attempts to extract important assumptions from the key players by posing three straight forward questions;

- a) What factors have been responsible for the increase in health care costs?
- b) How will health care costs develop in the future?
- c) What are the potential effects of several policy options to reduce these costs?

Participants are asked if they agreed or disagreed with the statements and why they thought the statements were true or not. The most frequently given reasons for the verbal statements were then incorporated in to the statements to create causal arguments from the participant's mental models.

Similar methods were adopted to identify policies which represent the aggregate of many individual actions. For example, why a GP may decide on such matters as frequency of patients appointments, drugs choice, referral to other medical specialist or a combination of all these. Vennix's model was subsequently formalized and quantified and converted into a computer-based learning environment for use by a wider range of health personnel.

The idea of using system dynamics as a means of participative modeling for learning is also inherent in other work [35].

Offering Insights into Managing the Demand for Health Care

Wolstenholme reports the insights from many applications of his own and other work. He suggests a hypothesis that the 'normal' mode of operation for many health and

social care organizations today is often well beyond their safe design capacity. This situation arises from having to cope with whatever demand arrives at their door irrespective of their supply capability. Risk levels can be high in these organizations and the consequences could be catastrophic for patients [71,72].

Evidence for the hypothesis has emerged at many points along patient pathways in health and social care from a number of studies carried out using system dynamics simulation to identify and promote systemic practice in local health communities. The rigor involved in knowledge-capture and quantitative simulation model construction and running has identified mismatches between how managers claim their organizations work and the observed data and behavior. The discrepancies can only be explained by surfacing informal coping policies. For example, transferring medical emergency patients to surgical wards, resulting in canceled elective procedures, also reported by Lane [35]. Indeed, the data itself becomes questionable as it reflects more the actions of managers than the true characteristics of patients.

The result of capacity pressure can mean that managers are unable, physically and financially, to break out from a fire-fighting mode to implement better resource investment and development policies for systemic and sustainable improvement. The insights reported are important for Health and Social Care management, the meaning of data and for modeling. The key message here is that much-needed systemic solutions and whole system thinking can never be successfully implemented until organizations are allowed to articulate and dismantle their worst coping strategies and return to working within best practice capacities.

The Modeling of the Treatment of Mental Health Diagnosis and Treatments in the UK

Modeling to assist mental health reform has recently developed as a separate strand of health work in the UK [46, 69,72].

Mental health services in the UK over the past 50 years have undergone numerous major reforms. The National Institute for Clinical Excellence [36] has recently published extensive research-based guidelines on the way stepped care might be best achieved. These involved moves towards a balanced, mixed community/institutional provision of services set within a range of significant reforms to the National Health Service. The latest and perhaps most significant reform is that associated with the introduction of 'stepped care'. Stepped care is aimed at bringing help to more patients more cheaply by devel-

oping intermediate staff, services and treatments between GPs and the specialist health hospitals.

Having decided on the new treatments at each step and having designed the basic patient pathways, modeling has been used in the North West of England to help with communication of the benefits and to overcome anticipated problems with resource reallocation issues [69]. Further work in Lincolnshire UK [58] reports the increasing use of 'matrix' modeling in mental health to capture the dynamics of both patient needs and treatments. This work also demonstrates the dangers of over-investment in situations where much demand is in accrued backlogs and incidence is reducing due to better and more successful interventions.

The depression work has also led to work at the Department of Health in the UK to help analyze the national impact of stepped services for mental health on the totality of the labor market and unemployment [72]. This work is an example of the value that system dynamics can add to conventional cost benefit analysis. A static cost benefit analysis was developed into a system dynamics model. By developing a bigger picture of the issue, both upstream to where patients go after treatment and downstream from where patients originate in the labor market, and by simulation of the enhanced vision, the dynamic cost benefit analysis is shown to advance understanding of the issue and plans.

The work questions the magnitude of the potential benefits, introduces phasing issues, surfaces structural insights, takes account of the dynamics of the labor market and forces linkages between the plan and other initiatives to get people back to work. The paper suggests that cost benefit analysis and system dynamics are very complementary and should be used together in strategic planning.

Other mental health capacity planning studies have been carried out for individual mental health hospitals and trusts. One such study [71] describes the application of system dynamics to assist decision making in the reallocation of resources within a specialist mental health trust in south London. Mental health service providers in the UK are under increasing pressure to both reduce their own costs and to move resources upstream in mental health patient pathways to facilitate treating more people, whilst not compromising service quality.

The investigation here focused on the consequences of converting an existing specialist service ward in a mental health hospital into a 'triage' ward, where patients are assessed and prioritized during a short stay for either discharge or onward admission to a normal ward. Various policies for the transition were studied together with the

implications for those patients needing post hospital services and relocation within the community. The model suggested that the introduction of a triage ward could meet the strategic requirement of a 10% shift away from institutional care and into community services. The paper includes a number of statements from the management team involved on the benefits of system dynamics and the impact of its application on their thinking.

System Dynamics Workforce Planning Models to Support Health Management

It is also important to mention that work has been carried out in a number of countries in the field of workforce planning related to health. In the UK the NHS has deployed sophisticated workforce planning models to determine the training and staffing needs associated with numerous alternative service configurations. In the Spanish Health system modeling has been used to determine the number of doctors required for a number of specialists services and to attempt to explore solutions for the current imbalance among supply and demand of physicians [2,4,5]. Elsewhere the factors affecting staff retention has been studied [28] and in the Netherlands, an advisory body of the Dutch government was given the responsibility of implying a new standard for the number of rheumatologists [39]. One of the main factors that were studied in the scenario analysis stage was the influences of changing demographics on the demand of manpower in the health system. Other studies have covered time reduction legislation on doctor training [22].

Future Directions

System dynamics has already made a significant impact on health and social care thinking across the EU. Many policy insights have been generated and the organizations are increasingly being recognized as complex adaptive systems. However, true understanding and implementation of the messages requires much more work and too many organizations are still locked into a pattern of short-termism which leads them to focus on the things they feel able to control – usually variables within their own individual spheres of control. There are also some aspects of system reform in some countries that are producing perverse incentives which encourage organizations to apply short-term policies.

Wider communication of existing studies and further studies are necessary to demonstrate the advantages of sustainable, systemic solutions. The key challenge lies in demonstrating to a wider audience of managers and clinicians that they can add value to the whole whilst remain-

ing autonomous. An important element is to train more people capable of modeling and facilitating studies and to simplify the process and software of system dynamics.

Acknowledgments

The author would like to acknowledge the many practitioners of system dynamics in health and social care throughout Europe for their dedication to using these methods to improve health and wellbeing. Particular thanks are due to my colleagues in the International Systems Dynamics community and its Society and those in the UK Chapter of the Society, without whom I would not have the knowledge base to undertake this review. Special thanks are also due to my colleagues in Symmetric SD without whom much of the personal work reported here would not have been carried out at all.

Bibliography

Primary Literature

1. Abdul-Salam O (2006) An overview of system dynamics applications. In: A dissertation submitted to the University of Salford Centre for Operational Research and Applied Statistics for the degree of MSc Centre for Operational Research and Applied Statistics, University of Salford
2. Alonso Magdaleno MI (2002) Administrative policies and MIR vacancies: Impact on the Spanish Health System. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
3. Alonso Magdaleno MI (2002) Dynamic analysis of some proposals for the management of the number of physicians in Spain. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
4. Alonso Magdaleno MI (2002) Elaboration of a model for the management of the number of specialized doctors in the Spanish health system. In: Proceedings of the 20th International Conference of the System Dynamics Society, System Dynamics Society, Palermo, 2002
5. Bayer S (2001) Planning the implementation of telecare services. In: The 19th International Conference of the System Dynamics Society, System Dynamics Society, Atlanta, 2001
6. Bayer S (2002) Post-hospital intermediate care: Examining assumptions and systemic consequences of a health policy prescription. In: Proceedings of the 20th International Conference of the System Dynamics Society, System Dynamics Society, Palermo, 2002
7. Bayer S, Barlow J (2003) Simulating health and social care delivery. In: Proceedings of the 21st International Conference of the System Dynamics Society, System Dynamics Society, New York, 2003
8. Bayer S, Barlow J (2004) Assessing the impact of a care innovation: Telecare. In: 22nd International Conference of the System Dynamics Society, System Dynamics Society, Oxford, 2004
9. Brailsford SC, Lattimer VA (2004) Emergency and on-demand health care: Modelling a large complex system. *J Operat Res Soc* 55:34–42

10. Chen Y (2003) A system dynamics-based study on elderly non-acute service in norway. In: Proceedings of the 21st International Conference of the System Dynamics Society, System Dynamics Society, New York, 2003
11. Coyle RG (1996) A systems approach to the management of a hospital for short-term patients. *Socio Econ Plan Sci* 18(4):219–226
12. Curram S, Coyle JM (2003) Are you vMad to go for surgery? Risk assessment for transmission of vCJD via surgical instruments: The contribution of system dynamics. In: Proceedings of the 21st International Conference of the System Dynamics Society, New York, 2003
13. Dangerfield BC (1999) System dynamics applications to european health care issues. *System dynamics for policy, strategy and management education. J Operat Res Soc* 50(4):345–353
14. Dangerfield BC, Roberts CA (1989) A role for system dynamics in modelling the spread of AIDS. *Trans Inst Meas Control* 11(4):187–195
15. Dangerfield BC, Roberts CA (1989) Understanding the epidemiology of HIV infection and AIDS: Experiences with a system dynamics. In: Murray-Smith D, Stephenson J, Zobel RN (eds) Proceedings of the 3rd European Simulation Congress. Simulation Councils Inc, San Diego, pp 241–247
16. Dangerfield BC, Roberts CA (1990) Modelling the epidemiological consequences of HIV infection and AIDS: A contribution from operational research. *J Operat Res Soc* 41(4):273–289
17. Dangerfield BC, Roberts CA (1992) Estimating the parameters of an AIDS spread model using optimisation software: Results for two countries compared. In: Vennix JAM, Faber J, Scheper WJ, Takkenberg CA (eds) System Dynamics. System Dynamics Society, Cambridge, pp 605–617
18. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five european countries. In: Dangerfield BC, Roberts CA (eds) O.R. Work in HIV/AIDS 2nd edn. Operational Research Society, Birmingham, pp 7–13
19. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: Some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
20. Dangerfield BC, Roberts CA, Fang Y (2001) Model-based scenarios for the epidemiology of HIV/AIDS: The consequences of highly active antiretroviral therapy. *Syst Dyn Rev* 17(2):119–150
21. Davies R (1985) An assessment of models in a health system. *J Operat Res Soc* 36:679–687
22. Derrick S, Winch GW, Badger B, Chandler J, Lovett J, Nokes T (2005) Evaluating the impacts of time-reduction legislation on junior doctor training and service. In: Proceedings of the 23rd International Conference of the System Dynamics Society, Boston, 2005
23. Dijkgraaf MGW, van Greenstein GJP, Gourds JLA (1998) Interactive simulation as a tool in the decision-making process to prevent HIV incidence among homosexual men in the Netherlands: A proposal. In: Jager JC, Rotenberg EJ (eds) Statistical Analysis and Mathematical Modelling of AIDS. OUP, Oxford, pp 112–122
24. El-Darzi E, Vasilakis C (1998) A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in hospitals. *Health Care Manag Sci* 1(2):143–149
25. Forrester JW (1961) Industrial Dynamics. MIT Press
26. Gonzalez B, Garcia R (1999) Waiting lists in spanish public hospitals: A system dynamics approach. *Syst Dyn Rev* 15(3):201–224
27. Heyne G, Geurts JL (1994) DIAGNOST: A microworld in the healthcare for elderly people. In: 1994 International System Dynamics Conference. System Dynamics Society, Sterling
28. Holmstroem P, Elf M (2004) Staff retention and job satisfaction at a hospital clinic: A case study. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
29. Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. *J Operat Res Soc* 50(2):109–123
30. Kim DH, Gogi J (2003) System dynamics modeling for long term care policy. Proceedings of the 21st International Conference of the System Dynamics Society. System Dynamics Society, New York
31. Lacey P (2005) Futures through the eyes of a health system simulator, Paper presented to the System Dynamics Conference, Boston 2005
32. Lagergren M (1992) A family of models for analysis and evaluation of strategies for preventing AIDS. In: Jager JC (eds) Scenario Analysis. Elsevier, Amsterdam, pp 117–145
33. Lane DC (2000) Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *J Operat Res Soc* 51(5):518
34. Lane DC, Husemann E (2008) System dynamics mapping of acute patient flows. *J Oper Res Soc* 59:213–224
35. Lane DC, Monefeldt C et al (2003) Client involvement in simulation model building: Hints and insights from a case study in a London hospital. *Health Care Manag Sci* 6(2):105–116
36. National Institute for Clinical Excellence (2004) Depression: Management of depression in primary and secondary care – NICE guidance. National Clinical Practice Guideline 23
37. Osipenko L (2006) System dynamics model of a new prenatal screening technology (Screening pregnant women). In: 24th International Conference of the System Dynamics Society, Nijmegen, The Netherlands, 23–27 July 2006
38. Petersen LO (2000) How should the capacity for treating heart disease be expanded? In: 18th International Conference of the System Dynamics Society, Bergen, 2000
39. Posmta TJBM, Smits MT (1992) Personnel planning in health care: An example in the field of rheumatology. In: Proceedings of the 1992 International System Dynamics Conference of the System Dynamics Society. System Dynamics Society, Utrecht
40. Ratnarajah M (2005) European union working time directive. In: 7th Annual Gathering. System Dynamics Society, Harrogate, February 2005
41. Ravn H, Petersen LO (2007) Balancing the surgical capacity in a hospital. *Int J Healthcare Technol Manag* 14:4023–4089
42. Richmond B (1994) ISEE Systems Inc, Hanover
43. Royston G, Dost A (1999) Using system dynamics to help develop and implement policies and programmes in health care in England. *Syst Dyn Rev* 15(3):293–315
44. Sedehi H (2001) HDS: Health department simulator. In: The 19th International Conference of the System Dynamics Society. System Dynamics Society, Atlanta, 2001
45. Senge P (1990) The fifth discipline doubleday. New York
46. Smith G, Wolstenholme EF (2004) Using system dynamics in modeling health issues in the UK. In: 22nd International Conference of the System Dynamics Society. The System Dynamics Society, Oxford, 2004

47. Sterman J (2000) Business dynamics: System thinking and modelling for a complex world. McGraw-Hill, Boston
48. Sterman T (ed) (2008) Exploring the next frontier: System Dynamics at 50. *Syst Dyn Rev* 23:89–93
49. Taylor KS (2002) A system dynamics model for planning and evaluating shifts in health services: The case of cardiac catheterisation procedures in the NHS London, London School of Economics and Political Science. *J Oper Res Soc* 56:659–1229
50. Taylor KS, Dangerfield BC (2004) Modelling the feedback effects of reconfiguring health services. *J Oper Res Soc* 56:659–675 (Published on-line Sept 2004)
51. Taylor KS, Dangerfield BC, LeGrand J (2005) Simulation analysis of the consequences of shifting the balance of health care: A system dynamics approach. *J Health Services Res Policy* 10(4):196–202
52. Tretter F (2002) Modeling the system of health care and drug addicts. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
53. Van Ackere A, PC Smith (1999) Towards a macro model of national health service waiting lists. *Syst Dyn Rev* 15(3):225
54. Van Dijkum C, Kuijk E (1998) Experiments with a non-linear model of health-related actions. In: 16th International Conference of the System Dynamics Society. System Dynamics Society, Quebec, 1998
55. Vennix AM (1996) Group Model Building. Wiley, Chichester
56. Vennix JAM, JW Gubbels (1992) Knowledge elicitation in conceptual model building: A case study in modeling a regional dutch health care system. *Europ J Oper Res* 59(1):85–101
57. Verburgh LD, Gubbels JW (1990) Model-based analyses of the dutch health care system. In: System Dynamics '90: Proceedings of the 1990 International Systems Dynamics Conference, International System Dynamics Society, Chestnut Hill, 1990
58. Wolstenholme E, McKelvie D, Monk D, Todd D, Brad C (2008) Emerging opportunities for system dynamics in UK health and social care – the market-pull for systemic thinking. Paper submitted to the 2008 System Dynamics Conference, Athens 2008
59. Wolstenholme EF (1993) A case study in community care using systems thinking. *J Oper Res Soc* 44(9):925–934
60. Wolstenholme EF (1996) A management flight simulator for community care. In: Cropper S (ed) Enhancing decision making in the NHS. Open University Press, Milton Keynes
61. Wolstenholme EF (1999) A patient flow perspective of UK health services: Exploring the case for new intermediate care initiatives. *Syst Dyn Rev* 15(3):253–273
62. Wolstenholme EF (2004) Using generic system archetypes to support thinking and learning. *Syst Dyn Rev* 20(2):341–356
63. Wolstenholme EF, McKelvie D (2004) Using system dynamics in modeling health and social care commissioning in the UK. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
64. Wolstenholme EF, Monk D (2004) Using system dynamics to influence and interpret health and social care policy in the UK. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
65. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics in modelling health and social care commissioning in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
66. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics in modelling mental health issues in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
67. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics to influence and interpret health and social care policy in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
68. Wolstenholme EF, Arnold S, Monk D, Todd D, McKelvie D (2005) Coping but not coping in health and social care – masking the reality of running organisations well beyond safe design capacity. *Syst Dyn Rev* 23(4):371–389
69. Wolstenholme EF, Arnold S, Monk D, Todd D, McKelvie D (2006) Reforming mental health services in the UK – using system dynamics to support the design and implementation of a stepped care approach to depression in north west england. In: Proceedings of the 2006 System Dynamics Conference, Nijmegen, 2006
70. Wolstenholme EF, Monk D, McKelvie D (2007) Influencing and interpreting health and social care policy in the UK In: Qudrat-Ullah H, Spector MJ, Davidsen PI (eds) Complex decision making: Theory and practice. Springer, Berlin
71. Wolstenholme EF, Monk D, McKelvie D, Gillespie P, O'Rourke D, Todd D (2007) Reallocating mental health resources in the borough of Lambeth, London, UK. In: Proceedings of the 2007 System Dynamics Conference, Boston, 2007
72. Wolstenholme EF, Monk D, McKelvie D, Todd D (2007) The contribution of system dynamics to cost benefit analysis – a case study in planning new mental health services in the UK. In: Proceedings of the 2007 System Dynamics Conference, Boston

Books and Reviews

- Dangerfield BC (1999) System dynamics applications to european health care issues. *J Oper Res Soc* 50(4):345–353
- Dangerfield BC, Roberts CA (eds) (1994) Health and health care dynamics. Special issue of *Syst Dyn Rev* 15(3)
- Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. *Syst Dyn Rev* 19(1):7–26

Health Care in the United States, System Dynamics Applications to

GARY HIRSCH¹, JACK HOMER²

¹ Independent Consultant, Wayland, USA

² Independent Consultant, Voorhees, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Four Applications](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Chronic illness A disease or adverse health state that persists over time and cannot in general be cured, although its symptoms may be treatable.

Stock An accumulation or state variable, such as the size of a population.

Flow A rate-of-change variable affecting a stock, such as births flowing into a population or deaths flowing out.

Feedback loop A closed loop of causality that acts to counterbalance or reinforce prior change in a system state.

Definition of the Subject

Health care involves a complex system of interactions among patients, providers, payers, and other stakeholders. This system is difficult to manage in the United States because of its free market approach and relative lack of regulation. System Dynamics simulation modeling is an effective method for understanding and explaining causes of dysfunction in U.S. health care and for suggesting approaches to improving health outcomes and slowing rising costs. Applications since the 1970s have covered diverse areas in health care including the epidemiology of diseases and substance abuse, as well as the dynamics of health care capacity and delivery and their impacts on health. Many of these applications have dealt with the mounting burden of chronic illnesses, such as diabetes. In this article four such applications are described.

Introduction

Despite remarkable successes in some areas, the health enterprise in the United States faces difficult challenges in meeting its primary goal of reducing the burden of disease and injury. These challenges include the growth of the underinsured population, epidemics of obesity and asthma, the rise of drug-resistant infectious diseases, ineffective management of chronic illness [33], long-standing racial and ethnic health disparities [32], and an overall decline in the health-related quality of life [64]. Many of these complex problems have persisted for decades, often proving resistant to attempts to solve them [36].

It has been argued that these interventions fail because they are made in piecemeal fashion, rather than comprehensively and from a whole-system perspective [15]. This compartmentalized approach is engrained in the financial structures, intervention designs, and evaluation methods of most health agencies. Conventional analytic methods are generally unable to satisfactorily address situations in which population needs change over time (often in re-

sponse to the interventions themselves), and in which risk factors, diseases, and health resources are in a continuous state of interaction and flux [52].

The term *dynamic complexity* has been used to describe such evolving situations [56]. Dynamically complex problems are often characterized by long delays between causes and effects, and by multiple goals and interests that may in some ways conflict with one another. In such situations, it is difficult to know how, where, and when to intervene, because most interventions will have unintended consequences and will tend to be resisted or undermined by opposing interests or as a result of limited resources or capacities.

The systems modeling methodology of System Dynamics (SD) is well suited to addressing the challenges of dynamic complexity in public health. The methodology involves the development of causal diagrams and policy-oriented computer simulation models that are unique to each problem setting. The approach was developed by computer pioneer Jay W. Forrester in the mid-1950s and first described at length in his book *Industrial Dynamics* [11] with some additional principles presented in later works [8,9,10,12]. The International System Dynamics Society was established in 1983, and within the Society a special interest group on health issues was organized in 2003.

SD modeling has been applied to health and health care issues in the U.S. since the 1970s. Topic areas have included:

- Disease epidemiology including work in heart disease [24,40], diabetes [24,34,43], obesity [25], HIV/AIDS [29], polio [57] and drug-resistant pneumococcal infections [28];
- Substance abuse epidemiology covering heroin addiction [37], cocaine prevalence [30], and tobacco reduction policy [50,58];
- Health care capacity and delivery in such areas as population-based HMO planning [21], dental care [20,38], and mental health [38], and as affected by natural disasters or terrorist acts [16,22,41]; and
- Interactions between health care or public health capacity and disease epidemiology [17,18,19,23,27].

Most of these modeling efforts have been done with the close involvement of clinicians and policymakers who have a direct stake in the problem being modeled. Established SD techniques for group model building [60] can help to harness the insights and involvement of those who deal with public health problems on a day-to-day basis.

It is useful to consider how SD models compare with those of other simulation methods that have been applied to public health issues, particularly in epidemio-

logical modeling. One may characterize any population health model in terms of its degree of aggregation, that is, the extent to which individuals in the population are combined together in categories of disease, risk, or age and other demographic attributes. At the most aggregate end of the scale are lumped contagion models [3,35]; more disaggregated are Markov models [13,31,44]; and the most disaggregated are microsimulations at the level of individuals [14,51,63].

The great majority of SD population health models are high or moderately high in aggregation. This is related to the fact that most SD models have a broad model boundary sufficient to include a variety of realistic causal factors, policy levers, and feedback loops. Although it is possible to build models that are both broad in scope and highly disaggregated, experience suggests that such very large models nearly always suffer in terms of their ability to be easily and fully tested, understood, and maintained. In choosing between broader scope and finer disaggregation, SD modelers tend to opt for the former, because a broad scope is generally needed for diagnosing and finding effective solutions to dynamically complex problems [55,56].

The remainder of this article describes four of the System Dynamics modeling applications cited above, with a focus on issues related to chronic illnesses and their care and prevention. The U.S. Centers for Disease Control and Prevention (CDC) estimates that chronic illness is responsible for 70% of all deaths and 75% of all health care costs in the U.S. [5]. The applications discussed below address:

- Diabetes and heart failure management at the community level;
- Diabetes prevention and management from an epidemiological perspective;
- General chronic illness care and prevention at a community level; and
- General chronic illness care and prevention at the national level.

The article concludes with a discussion of promising areas for future work.

Four Applications

Diabetes and Heart Failure Management at the Community Level

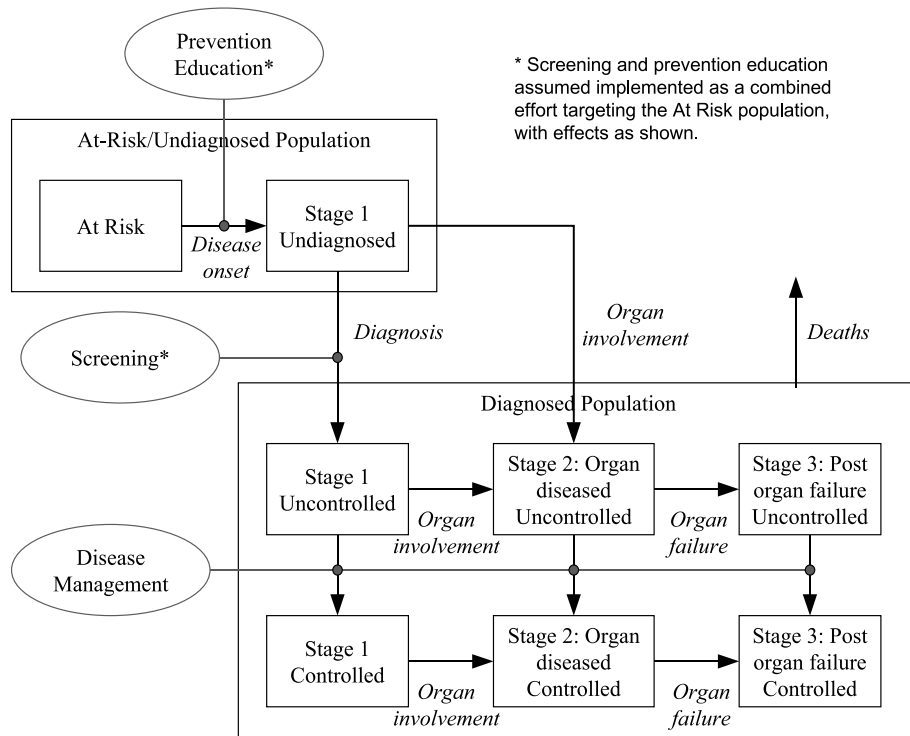
Two hours north of Seattle in the state of Washington lies Whatcom County, with a population of about 170 thousand. The county embarked on a major effort to address chronic illness care and was selected by the Robert Wood Johnson Foundation as one of seven sites in a larger

chronic care program called Pursuing Perfection [24]. The program initially concentrated on two chronic illnesses as prototypes for improved care: diabetes and congestive heart failure. Both of these illnesses affect millions of people in the U.S. and other countries and exact a heavy toll in terms of direct medical expenditures as well as indirect costs due to disability and premature mortality [2,45,47]. The prevalence of both diseases is growing rapidly as the numbers of people above age 65 increase, and also due to the epidemic rise in obesity, which is a risk factor for both diabetes and heart disease [7,46].

Leaders of the Whatcom County program had two critical needs for making decisions about potential interventions for improving the care of chronic illnesses such as diabetes and heart failure. First, they wanted to get a sense of the overall impact of these interventions on incidence and prevalence of diabetes and heart failure, health care utilization and cost, and mortality and disability rates in the community. Second, they wanted to understand the impact of the various interventions on individual health care providers in the community and on those who pay for care—insurers, employers, and patients themselves. There was a concern that the costs and benefits of the program be shared equitably and that providers who helped produce savings should not suffer a resulting loss of revenue to their businesses.

These analytic needs could not be met with spreadsheet and other models that project impacts in a simple, linear fashion. Interventions in chronic illness do not have simple direct impacts. The aging of the population, incidence of new cases, progression of disease, deaths, and the interventions themselves all create a constantly changing situation. Interventions ideally reduce mortality rates, but this leaves more people with the disease alive and requiring care for years to come.

Figure 1 presents a simplified view of the stock-and-flow structure used in modeling non-insulin-dependent (Type 2) diabetes. The actual model has two separate structures like those shown in Fig. 1, one for the 18-to-64 age group and one for the 65-and-older age group, which are linked by flows of patients turning 65. The model also calculates an inflow of population turning 18, death outflows from each stock based on patient age and stage of illness, and flows of migration into and out of the county. The rectangular boxes in Fig. 1 represent sub-populations with particular characteristics. The arrows signify flows of people from one population group to another (e.g., from uncontrolled to controlled diabetes at a particular stage). Lines from ovals (programmatic interventions such as disease management) to population flows indicate control of or influence on those flows.



Health Care in the United States, System Dynamics Applications to, Figure 1
Disease stages and intervention points in the Whatcom County Diabetes Model

The three stages of diabetes portrayed in this figure were identified through discussions with clinicians in Whatcom County. The population At Risk includes those with family history, the obese, and, most directly, those with a condition of moderate blood sugar known as pre-diabetes. Further increases in blood sugar lead to Stage 1 diabetes, in which blood vessels suffer degradation, but there is not yet any damage to organs of the body, nor typically any symptoms of the encroaching disease. More than half of Stage 1 diabetics are undiagnosed. If Stage 1 diabetics go untreated, most will eventually progress to Stage 2, marked by organ disease. In Stage 2 diabetes, blood flow disturbances impair the functioning of organ systems and potentially lead to irreversible damage. A patient who has suffered irreversible organ damage, or organ failure, is said to be in Stage 3; this would include diabetics who suffer heart attacks, strokes, blindness, amputations, or endstage renal disease. These patients are at the greatest risk of further complications leading to death.

Several studies have demonstrated that the incidence, progression, complications, and costs of diabetes can be reduced significantly through concerted intervention [1,4,6,59,61]. Such intervention may include primary prevention or disease management. As indicated in Fig. 1,

primary prevention would consist of efforts to screen the at-risk population and educate them about the diet and activity changes they need to prevent progression to diabetes. Disease management, on the other hand, addresses existing diabetics. A comprehensive disease management approach, such as that employed by the Whatcom County program, can increase the fraction of patients who are able to keep their blood sugar under effective control from the 40% or less typically seen without a program up to perhaps 80% or more.

The SD model of diabetes in Whatcom County was first used to produce a 20-year status quo or baseline projection, which assumes that no intervention program is implemented. In this projection, the prevalence of diabetes among all adults gradually increases from 6.5% to 7.5%, because of a growing elderly population; the prevalence of diabetes among the elderly is 17%, compared with 5% among the non-elderly. Total costs of diabetes, including direct costs for health care and pharmaceuticals and indirect economic losses due to disability, grow substantially in this baseline projection.

The next step was to use the model to examine the impact of various program options. These included: (1) a partial approach enhancing disease management but not pri-

mary prevention, (2) a full implementation approach combining enhancement of both disease management and primary prevention, and (3) an approach that goes beyond full implementation by also providing greater financial assistance to the elderly for purchasing drugs needed for the control of diabetes.

Simulations of these options projected results in terms of various outcome variables, including deaths from complications of diabetes and total costs of diabetes. Figure 2 shows typical simulation results obtained by projecting these options, in this case, the numbers of deaths over time that might be expected due to complications of diabetes. “Full VCTIS” refers to the complete program of primary prevention and disease management. Under the status quo projection, the number of diabetes-related deaths grows continuously along with the size of the diabetic population. The partial (disease management only) approach is effective at reducing deaths early on, but becomes increasingly less effective over time. The full program approach (including primary prevention) overcomes this shortcoming and by the end of the 20 year simulation reduces diabetes-related deaths by 40% relative to the status quo. Addition of a drug purchase plan for the elderly does even better, facilitating greater disease control and thereby reducing diabetes related deaths by 54% relative to the status quo.

With regard to total costs of diabetes, the simulations indicate that the full program approach can achieve net savings only two years after the program is launched. Four years after program launch, a drug plan for the elderly generates further reductions in disability costs beyond those provided by the program absent such a plan. The partial program approach, in contrast, achieves rapid net savings initially, but gives back most of these savings over time as diabetes prevalence grows. By the end of 20 years, the full program approach results in a net savings amounting to 7% of the status quo costs, two-thirds of that savings coming from reduction in disability-related costs. The model suggests that these anticipated net savings are the result of keeping people in the less severe stages of the diseases for a longer period of time and reducing the number of diabetes-related hospitalizations.

The simulations provided important information and ideas to the Whatcom County program planners, as well as supporting detailed discussions of how various costs and benefits could be equitably distributed among the participants. This helped to reassure participants that none of them would be unfairly affected by the proposed chronic illness program. Perhaps the most important contribution of modeling to the program planning process was its ability to demonstrate that the program, if implemented in its

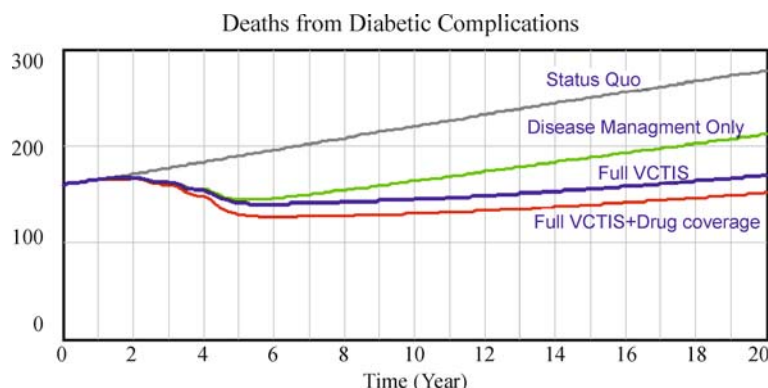
full form, would likely reduce total costs, even though it would extend the longevity of many diabetics requiring costly care. Given the sensitivity of payers who were already bearing high costs, this finding helped to motivate their continued participation in the program.

Diabetes Prevention and Management from an Epidemiological Perspective

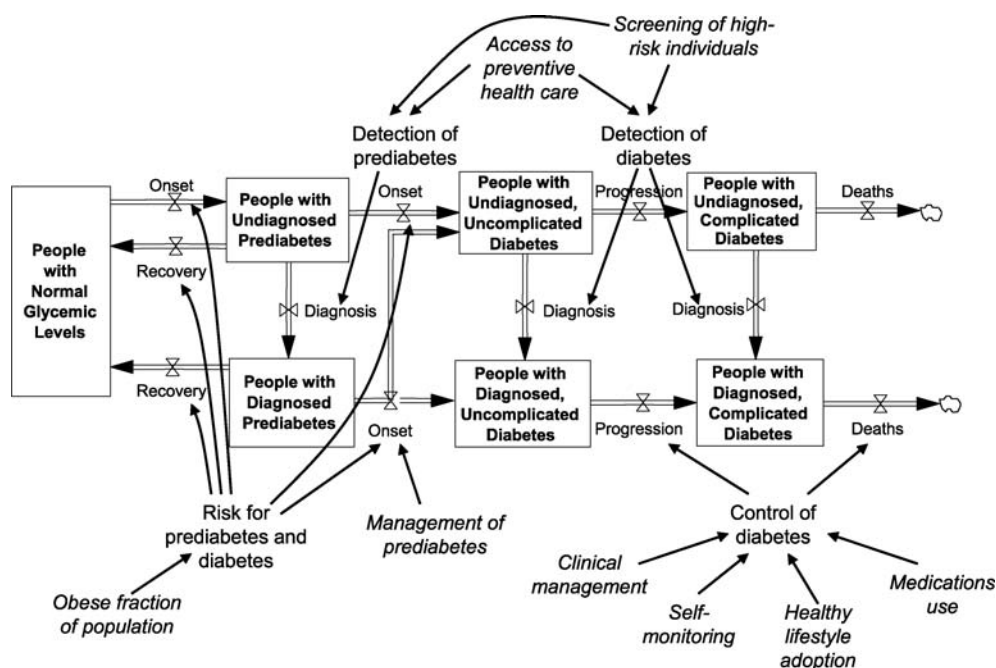
Another SD model of diabetes in the population was developed for the CDC’s Division of Diabetes Translation [34]. This model, a structural overview of which is presented in Fig. 3, builds upon the Whatcom County work but looks more closely at the drivers of diabetes onset, including the roles of prediabetes and obesity. The core of the CDC model is a chain of population stocks and flows portraying the movement of people among the stages of normal blood glucose, prediabetes, uncomplicated diabetes, and complicated diabetes. The prediabetes and diabetes stages are further divided among stocks of people whose conditions are diagnosed or undiagnosed. Also shown in Fig. 3 are the potentially modifiable influences in the model that affect the rates of population flow. These flow-rate drivers include obesity and the detection and management of prediabetes and of diabetes.

The model’s parameters were calibrated based on historical data available for the U.S. adult population, as well as estimates from the scientific literature. The model is able to reproduce historical time series, some going as far back as 1980, on diagnosed diabetes prevalence, the diagnosed fraction of diabetes, prediabetes prevalence, the obese fractions of people with prediabetes and diabetes, and the health burden (specifically, the mortality, morbidity, and costs) attributable to diabetes. The model suggests that two forces worked in opposition to affect the diabetes health burden from 1980 to 2004. The first force is a rise in the prevalence of obesity, which led to a greater incidence and prevalence of prediabetes and diabetes through the chain of causation seen in Fig. 3. The second and opposing force is a significant improvement in the control of diabetes, achieved through greater efforts to detect and manage the disease. The second force managed to hold the health burden of diabetes more or less flat during 1980 to 2004.

Looking toward the future, a baseline scenario assumes that no further changes occur in obesity prevalence after 2006, and that inputs affecting the detection and management of prediabetes and diabetes remain fixed at their 2004 values through 2050. This fixed-inputs assumption for the baseline scenario is not meant to represent a forecast of what is most likely to in the future but does pro-



Health Care in the United States, System Dynamics Applications to, Figure 2
Typical results from policy simulations with Whatcom County Diabetes Model



Health Care in the United States, System Dynamics Applications to, Figure 3 Structure of the CDC Diabetes Model

vide a useful and easily-understood starting point for policy analysis.

The baseline simulation indicates a future for diabetes burden outcomes for the period 2004–2050 quite different from the past. With obesity prevalence fixed, by assumption, at a high point of 37% from 2006 onward, the diabetes onset rate remains at a high point as well, and diabetes prevalence consequently continues to grow through 2050, becoming more level (after about 2025) only when the outflow of deaths starts to catch up with the inflow of onset.

The CDC model has been used to examine a variety of future scenarios involving policy interventions (singly or in combination) intended to limit growth in the burden of diabetes. These include scenarios improving the management of diabetes, increasing the management of prediabetes, or reducing the prevalence of general population obesity over time. Enhanced diabetes management can significantly reduce the burden of diabetes in the short term, but does not prevent the growth of greater burden in the longer term due to the growth of diabetes prevalence. Indeed, the effect of enhanced dia-

betes management on diabetes prevalence is not to decrease it at all, but rather to increase it somewhat by increasing the longevity of people with diabetes. Increased management of prediabetes does, in contrast, reduce diabetes onset and the growth of diabetes prevalence. However, it does not have as much impact as one might expect; this is because many people with prediabetes are not diagnosed, and also because the policy does nothing to reduce the growth of prediabetes prevalence due to obesity in the general population. A reduction in prediabetes can be achieved only by reducing population obesity. Significant obesity reduction may take 20 years or more to accomplish fully, but the model suggests that such a policy can be quite a powerful one in halting the growth of diabetes prevalence and burden even before those 20 years are through.

Overall, the CDC model suggests that no single type of intervention is sufficient to limit the growth of the diabetes burden in both the short term and the long term. Rather, what is needed is a combination of disease management for the short term and primary prevention for the longer term. The model also suggests that effective primary prevention may require obesity reduction in the general population a focus on managing diagnosed prediabetes.

At the state and regional level, the CDC model has become the basis for a model-based workshop called the “Diabetes Action Lab”. Participants have included state and local public health officials along with non-governmental stakeholders including health care professionals, leaders of not-for-profit agencies, and advocates for people living with diabetes. The workshops have helped the participants improve their intervention strategies and goals and become more hopeful and determined about seeing their actions yield positive results in the future.

The CDC diabetes model has led to other SD modeling efforts at the CDC emphasizing disease prevention, including studies of obesity [25] and cardiovascular risk. The obesity study involved the careful analysis of population survey data to identify patterns of weight gain over the entire course of life from childhood to old age. It explored likely impacts decades into the future of interventions to reduce or prevent obesity that may be targeted at specific age categories. Tentative findings included (1) that obesity in the U.S. should be expected to grow at a much slower pace in the future than it did in the 1980s and 1990s; (2) that the average amount of caloric reduction necessary to reverse the growth of obesity in the population is less than 100 calories per day; (3) that the current trend of focusing intervention efforts on school-age children will likely have only a small impact on future obesity in the adult population; and (4) that it may take decades to see the full

impacts of interventions to reduce obesity in the overall population.

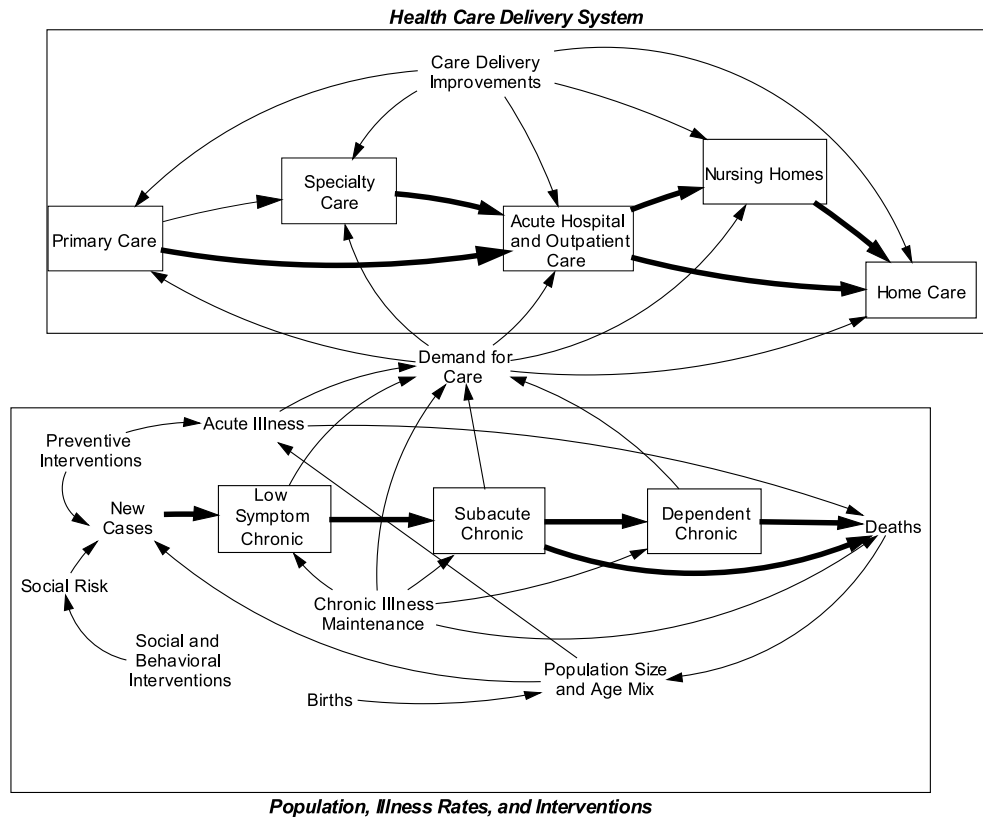
General Health Care and Illness Prevention at a Community Level

Hirsch and Immediato [19] describe a comprehensive view of health at the level of a community. Their “Health Care Microworld”, depicted in highly simplified form in Fig. 4, simulates the health status and health care delivery for people in the community. The Microworld was created for a consortium of health care providers who were facing a wide range of changes in the mid-1990s and needed a means for their staffs to understand the implications of those changes for how they managed. The underlying SD model consists of many hundreds of equations and was designed to reflect with realistic detail a typical American community and its providers, with data taken from public sources as well as proprietary surveys. Users of the Microworld have a wide array of options for expanding the capacity and performance of the community’s health care delivery system such as adding personnel and facilities, investing in clinical information systems, and process redesign. They have a similar range of alternatives for improving health status and changing the demand for care including screening for and enhanced maintenance care of people with chronic illnesses, programs to reduce behavioral risks such as smoking and alcohol abuse, environmental protection, and longer-term risk reduction strategies such as providing social services, remedial education, and job training.

The Microworld’s comprehensive view of health status and health care delivery can provide insights not available from approaches that focus on one component of the system at a time. For example, users can play roles of different providers in the community and get a better understanding of why many attempts at creating integrated delivery systems have failed because participating providers care more about their own bottom lines and prerogatives than about creating a viable system. When examining strategies for improving health status, users can get a better sense of how a focus on enhanced care of people with chronic illnesses provides short-term benefits in terms of reduced deaths, hospital admissions, and costs, but how better long-term results can be obtained by also investing in programs that reduce social and behavioral health risks.

General Health Care and Illness Prevention at the National Level

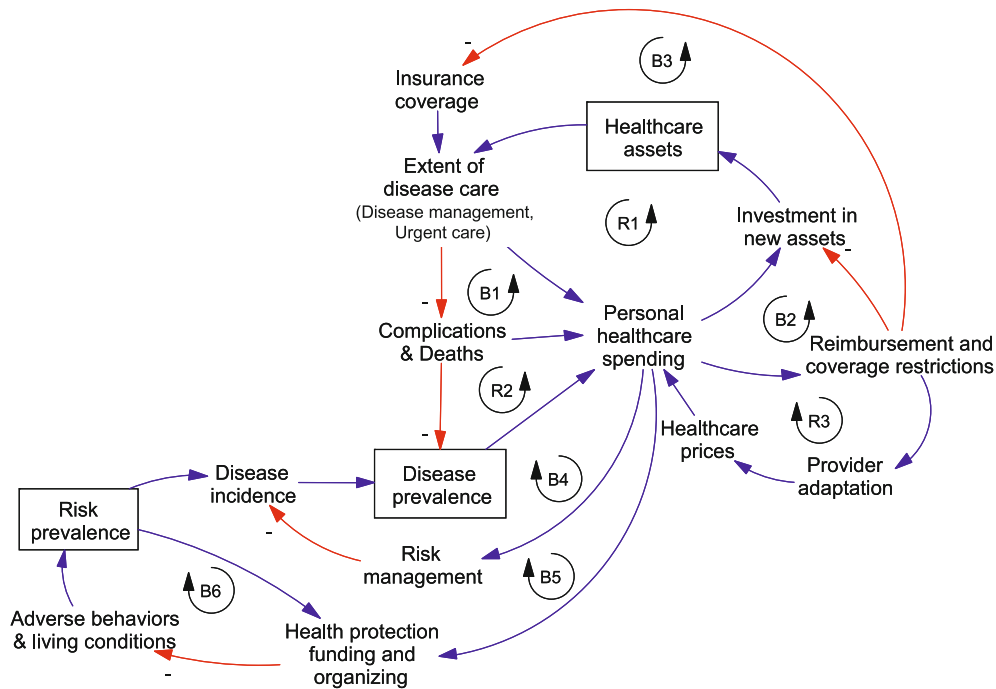
Despite rapid growth in health care spending in the U.S. in recent decades, the health of Americans has not noticeably



Health Care in the United States, System Dynamics Applications to, Figure 4
Overview of the Health Care Microworld

improved. A recent SD model [23] addresses the question of why the U.S. has not been more successful in preventing and controlling chronic illness. This model can faithfully reproduce patterns of change in disease prevalence and mortality in the U.S., but its structure is a generic one and should be applicable to other countries. The model examines the growing prevalence of disease and responses to it, responses which include the treatment of complications as well as disease management activities designed to slow the progression of illness and reduce the occurrence of future complications. The model shows how progress in complications treatment and disease management has slowed since 1980 in the U.S., largely due to a behavioral tug-of-war between health care payers and providers that has resulted in price inflation and an unstable climate for health care investments. The model is also used to demonstrate the impact of moving “upstream” by managing known risk factors to prevent illness onset, and moving even further upstream by addressing adverse behaviors and living conditions linked to the development of these risk factors in the first place.

An overview of the model’s causal structure is presented in Fig. 5. The population stock of disease prevalence is increased by disease incidence and decreased by deaths. The death rate can be reduced by a greater extent of disease care, including urgent care and disease management. Disease incidence draws from a stock of risk prevalence, where risk refers to physical or psychological conditions or individual behaviors that may lead to disease. Effective risk management can reduce the flow of people from risk to disease, and may also in some cases allow people to return to a condition of being no longer at risk. Such management may include changes in nutrition or physical activity, stress management, or the use of medications. The risk prevalence stock is increased by adverse behaviors and living conditions. Adverse behaviors may include poor diet, lack of physical activity, or substance abuse. Adverse living conditions can encompass many factors, including crime, lack of access to healthy foods, inadequate regulation of smoking, weak social networks, substandard housing, poverty, or poor educational opportunities.



Health Care in the United States, System Dynamics Applications to, Figure 5

Overview of a National-Level Model of Health Care and Illness Prevention. Key to feedback loops ("R" denotes self-reinforcing, "B" denotes counterbalancing):

R1 Health care revenues are reinvested for further growth

B1 Disease management reduces need for urgent care

R2 Disease care prolongs life and further increases need for care

B2 Reimbursement restriction limits spending growth

B3 Insurance denial limits spending growth

R3 Providers circumvent reimbursement restrictions, leading to a tug-of-war with payers

B4 Risk management proportional to downstream spending can help limit it

B5 Health protection proportional to downstream spending can help limit it

B6 Health protection (via sin taxes) proportional to risk prevalence can help limit it

The extent of care is explained in the model by two key factors: the abundance of health care assets, and insurance coverage. Health care assets are the structures and fixed equipment used directly for health care or for the production of health care products, as well as the human capital of personnel involved. Insurance coverage refers to the fraction of the population with some form of health care insurance, either with a private insurer or through a government plan. The uninsured are less likely than the insured to receive health care services, especially disease management services, something which most of the uninsured cannot afford whereas in most cases they can get urgent care at a hospital emergency department.

The stock of assets is increased by investments, which may be viewed as the reinvestment of some fraction of health care revenues. Such reinvestment drives further growth of care and revenue, and the resulting exponential growth process is identified as loop R1 in Fig. 5.

The data indicate, however, that the reinvestment process has slowed significantly since 1980. It is hypothesized that this decline in the reinvestment rate has been the response by potential investors to various forms of cost control, including the restriction of insurance reimbursements, which affect the providers of health care goods and services. With increasing controls and restrictions, these potential investors face greater risk and uncertainty about the future return on their investments, and the result is a greater reluctance to build a new hospital wing, or to purchase an expensive new piece of equipment, or even, at an individual level, to devote a decade or more of one's life to the hardship of medical education and training. Health care costs and cost controls have also led to elimination of private health insurance coverage by some employers, although some of the lost coverage has been replaced by publicly-funded insurance.

One additional part of the downstream health care story portrayed in Fig. 5 is the growth of health care prices. Health care prices are measured in terms of a medical care consumer price index (CPI), which since 1980 has grown much more rapidly than the general CPI for the overall economy. For the period 1980–2004, inflation in medical care prices averaged 6.1% versus general inflation of 3.5%. Why has health care inflation exceeded that of the general economy? Several different phenomena have contributed to health care inflation, but not all have contributed with sufficient magnitude or with the timing necessary to explain the historical pattern. One phenomenon that does appear to have such explanatory power is shown in Fig. 5 as “provider adaptation”. This is the idea that, in response to cost containment efforts, providers may “increase fees, prescribe more services, prescribe more complex services (or simply bill for them), order more follow-up visits, or do a combination of these. . .” [49] Many tests and procedures are performed that contribute little or no diagnostic or therapeutic value, thereby inflating the cost per quality of care delivered. By one estimate, unnecessary and inflationary expense may have represented 29% of all personal health care spending in the year 1989 [23].

The dynamics involving the extent of disease care are portrayed in Fig. 5 in the feedback loops labeled R1, B1, R2, B2, B3, and R3. Taken together, one may view these loops—with the exception of Loop R3—as the story of a “rational” downstream health care system that favors growth and investment until the resulting costs get to a point where further increases are perceived to be no longer worth the expected incremental improvements in health and productivity. Loop R3, however, introduces dysfunction into this otherwise rational system. The loop describes a tug-of-war between payers restricting reimbursement in response to high health care costs, and providers adapting to these restrictions by effectively raising health care prices in an attempt to circumvent the restrictions and maintain their incomes. Because this loop persistently drives up health care costs, it ends up hurting health care investments and insurance coverage (through Loops B2 and B3, respectively), thus dampening growth in the extent of care.

Simulations of the model suggest that there are no easy downstream fixes to the problem of an underperforming and expensive health care system in the U.S. mold. The simulations seem to suggest—perhaps counterintuitively—that health insurance should be stable and non-restrictive in its reimbursements, so as to avoid behavioral backlashes that can trigger health care inflation and underinvestment. Although a broad mandate of this sort would likely be politically infeasible in the U.S., movement in this

direction could perhaps start with the government’s own Medicare and Medicaid insurance programs, and then diffuse naturally to private insurers over time. It is interesting to consider whether a more generous and stable approach to reimbursement could not only combat illness better than the current restrictive approach, but do it more efficiently and perhaps even at lower cost.

The model also includes structure for evaluating the upstream prevention of disease incidence. There are two broad categories of such efforts described in the literature: Risk management for people already at risk, and health protection for the population at large to change adverse behaviors and mitigate unhealthy living conditions. While spending on population-based health protection and risk management programs has grown somewhat, it still represents a small fraction of total U.S. health care spending, on the order of 5% in 2004 [23].

Figure 5 includes three balancing loops to indicate how, in general terms, efforts in risk management and health protection might be funded or resourced more systematically and in proportion to indicators of capability or relative need. Loop B4 suggests that funding for programs promoting risk management could be made proportional to spending on downstream care, so that when downstream care grows funding for risk management would grow as well. Loop B5 suggests something similar for health protection, supposing that government budgets and philanthropic investments for health protection could be set in proportion to recent health care spending. Loop B6 takes a different approach to the funding of health protection, linking it not to health care spending but to risk prevalence, the stock which health protection most directly seeks to reduce. The linkage to risk prevalence can be made fiscally through “sin taxes” on unhealthy items, such as cigarettes (already taxed throughout the U.S. to varying extents [39]) and fatty foods [42]. In theory, the optimal magnitude of such taxes may be rather large in some cases, as the taxes can be used both to discourage unhealthy activities and promote healthier ones [48].

Simulations of the model suggest that whether the approach to upstream action is risk management or health protection, such actions can reduce illness prevalence and ultimately save money. However, the payback time, in terms of reduced downstream health care costs, may be a relatively long one, perhaps on the order of 20 years. It should be noted, however, that the model does not include losses in productivity to employers and society at large. The Whatcom County models described above suggest that when these losses are taken into account, the payback on upstream action may shrink to a much shorter time period that may be acceptable to the public as well

as to those decision makers in a position to put upstream efforts into effect [24].

Future Directions

As long as there are dynamically complex health issues in search of answers, the SD approach will have a place in the analytic armamentarium. There is still much to be learned about the population dynamics of individual chronic conditions like hypertension and risk factors like obesity. SD models could also address multiple interacting diseases and risks, giving a more realistic picture of their overall epidemiology and policy implications, particularly where the diseases and risks are mutually reinforcing. For example, it has been found that substance abuse, violence, and AIDS often cluster in the same urban subpopulations, and that such “syndemics” are resistant to narrow policy interventions [53,54,62]. This idea could also be extended to the case of mental depression, which is often exacerbated by other chronic illnesses, and may, in turn, interfere with the proper management of those illnesses. An exploratory simulation model has indicated that SD can usefully address the concept of syndemics [26].

There is also more to be learned about health care delivery systems and capacities, with the inclusion of characteristics specific to selected real-world cases. Models combining delivery systems and risk and disease epidemiology could help policymakers and health care providers understand the nature of coordination required to put ambitious public health and risk reduction programs in place without overwhelming delivery capacities. Such models could reach beyond the health care delivery system per se to examine the potential roles of other delivery systems, such as schools and social service agencies, in health risk reduction.

The more complete view of population health dynamics advocated here may also be extended to address persistent challenges in the U.S. that will likely require policy changes at a national and state level, and not only at the level of local communities. Examples include the large underinsured population, persistent racial and ethnic health disparities, and the persistent shortage of nurses. SD modeling can help to identify the feedback structures responsible for these problems, and point the way to policies that can make a lasting difference.

Bibliography

1. American Diabetes Association and National Institute of Diabetes and Digestive and Kidney Diseases (2002) The prevention or delay of Type 2 diabetes. *Diabetes Care* 25:742–749
2. American Heart Association (2000) 2001 Heart and Stroke Statistical Update. AHA, Dallas
3. Anderson R (1994) Populations, infectious disease and immunity: A very nonlinear world. *Phil Trans R Soc Lond B* 346:457–505
4. Bowman BA, Gregg EW, Williams DE, Engelgau MM, Jack Jr L (2003) Translating the science of primary, secondary, and tertiary prevention to inform the public health response to diabetes. *J Public Health Mgmt Pract* November (Suppl):S8–S14
5. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion (2007) Chronic Disease Overview. Available at <http://www.cdc.gov/nccdphp/overview.htm>
6. Diabetes Prevention Program Research Group (2002) Reduction in the incidence of Type 2 diabetes with lifestyle intervention or metformin. *New Engl J Med* 346:393–403
7. Flegal KM, Carroll MD, Ogden CL, Johnson CL (2002) Prevalence and trends in obesity among US adults, 1999–2000. *JAMA* 288:1723–1727
8. Forrester JW, Senge PM (1980) Tests for building confidence in system dynamics models. In: *System Dynamics, TIMS Studies in the Management Sciences*. North-Holland, New York, pp 209–228
9. Forrester JW (1971) Counterintuitive behavior of social systems. *Technol Rev* 73:53–68
10. Forrester JW (1980) Information sources for modeling the national economy. *J Amer Stat Assoc* 75:555–574
11. Forrester JW (1961) *Industrial Dynamics*. MIT Press, Cambridge
12. Forrester JW (1969) *Urban Dynamics*. MIT Press, Cambridge
13. Gunning-Schepers LJ (1989) The health benefits of prevention: A simulation approach. *Health Policy Spec Issue* 12:1–255
14. Halloran EM, Longini IM, Nizam A, Yang Y (2002) Containing bioterrorist smallpox. *Science* 298:1428–1432
15. Heirich M (1999) *Rethinking Health Care: Innovation and Change in America*. Westview Press, Boulder
16. Hirsch GB (2004) Modeling the consequences of major incidents for health care systems. In: *22nd International System Dynamics Conference*. System Dynamics Society, Oxford. Available from: <http://www.systemdynamics.org/publications.htm>
17. Hirsch G, Homer J (2004) Integrating chronic illness management, improved access to care, and idealized clinical practice design in health care organizations: A systems thinking approach. In: *International Conference on Systems Thinking in Management*. AFEI/University of Pennsylvania, Philadelphia. Available from: <http://www.afei.org/documents/CDRomOrderForm.pdf>
18. Hirsch G, Homer J (2004) Modeling the dynamics of health care services for improved chronic illness management. In: *22nd International System Dynamics Conference*. System Dynamics Society, Oxford. Available from: <http://www.systemdynamics.org/publications.htm>
19. Hirsch GB, Immediato CS (1999) Microworlds and generic structures as resources for integrating care and improving health. *Syst Dyn Rev* 15:315–330
20. Hirsch GB, Killingsworth WR (1975) A new framework for projecting dental manpower requirements. *Inquiry* 12:126–142
21. Hirsch G, Miller S (1974) Evaluating HMO policies with a computer simulation model. *Med Care* 12:668–681
22. Hoard M, Homer J, Manley W, Furbie P, Haque A, Helmkamp J (2005) Systems modeling in support of evidence-based

- disaster planning for rural areas. *Int J Hyg Environ Health* 208:117–125
23. Homer J, Hirsch G, Milstein B (2007) Chronic illness in a complex health economy: The perils and promises of downstream and upstream reforms. *Syst Dyn Rev* 23(2–3):313–334
 24. Homer J, Hirsch G, Minniti M, Pierson M (2004) Models for collaboration: How system dynamics helped a community organize cost-effective care for chronic illness. *Syst Dyn Rev* 20:199–222
 25. Homer J, Milstein B, Dietz W, Buchner D, Majestic E (2006) Obesity population dynamics: Exploring historical growth and plausible futures in the U.S. In: 24th International System Dynamics Conference. System Dynamics Society, Nijmegen. Available from: <http://www.systemdynamics.org/publications.htm>
 26. Homer J, Milstein B (2002) Communities with multiple afflictions: A system dynamics approach to the study and prevention of syndemics. In: 20th International System Dynamics Conference. System Dynamics Society, Palermo. Available from: <http://www.systemdynamics.org/publications.htm>
 27. Homer J, Milstein B (2004) Optimal decision making in a dynamic model of community health. In: 37th Hawaii International Conference on System Sciences. IEEE, Waikoloa. Available from: <http://csdl.computer.org/comp/proceedings/hicss/2004/2056/03/2056toc.htm>
 28. Homer J, Ritchie-Dunham J, Rabbino H, Puente LM, Jorgensen J, Hendricks K (2000) Toward a dynamic theory of antibiotic resistance. *Syst Dyn Rev* 16:287–319
 29. Homer JB, St. Clair CL (1991) A model of HIV transmission through needle sharing. *Interfaces* 21:26–49
 30. Homer JB (1993) A system dynamics model of national cocaine prevalence. *Syst Dyn Rev* 9:49–78
 31. Honeycutt AA, Boyle JP, Broglio KR et al (2003) A dynamic Markov model for forecasting diabetes prevalence in the United States through 2050. *Health Care Mgmt Sci* 6:155–164
 32. Institute of Medicine (Board on Health Sciences Policy) (2003) *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press, Washington, DC
 33. Institute of Medicine (Committee on Quality of Health Care in America) (2001) *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academies Press, Washington, DC
 34. Jones AP, Homer JB, Murphy DL, Essien JDK, Milstein B, Seville DA (2006) Understanding diabetes population dynamics through simulation modeling and experimentation. *Am J Public Health* 96(3):488–494
 35. Kaplan EH, Craft DL, Wein LM (2002) Emergency response to a smallpox attack: The case for mass vaccination. In: *Proceedings of the Natl Acad of Sciences* 99:10935–10940
 36. Lee P, Paxman D (1997) Reinventing public health. *Annu Rev Public Health* 18:1–35
 37. Levin G, Roberts EB, Hirsch GB (1975) *The Persistent Poppy*. Ballinger, Cambridge
 38. Levin G, Roberts EB, Hirsch GB, Kligler DS, Wilder JF, Roberts N (1976) *The Dynamics of Human Service Delivery*. Ballinger, Cambridge
 39. Lindblom E (2006) State cigarette excise tax rates and rankings. Campaign for Tobacco-Free Kids, Washington, DC. Available from: <http://www.tobaccofreekids.org/research/factsheets/pdf/0097.pdf>
 40. Luginbuhl W, Forsyth B, Hirsch G, Goodman M (1981) Prevention and rehabilitation as a means of cost-containment: The example of myocardial infarction. *J Public Health Policy* 2:1103–1115
 41. Manley W, Homer J et al (2005) A dynamic model to support surge capacity planning in a rural hospital. In: 23rd International System Dynamics Conference, Boston. Available from: <http://www.systemdynamics.org/publications.htm>
 42. Marshall T (2000) Exploring a fiscal food policy: The case of diet and ischaemic heart disease. *BMJ* 320:301–305
 43. Milstein B, Jones A, Homer J, Murphy D, Essien J, Seville D (2007) Charting Plausible Futures for Diabetes Prevalence in the United States: A Role for System Dynamics Simulation Modeling. *Preventing Chronic Disease*, 4(3), July 2007. Available at: http://www.ignorespaces.cdc.gov/pcd/issues/2007/jul/06_0070.htm
 44. Naidoo B, Thorogood M, McPherson K, Gunning-Schepers LJ (1997) Modeling the effects of increased physical activity on coronary heart disease in England and Wales. *J Epidemiol Community Health* 51:144–150
 45. National Institute of Diabetes and Digestive and Kidney Diseases (2004) National Diabetes Statistics. Available from: <http://diabetes.niddk.nih.gov/dm/pubs/statistics/index.htm>
 46. National Institute of Diabetes and Digestive and Kidney Diseases (2004) Statistics Related to Overweight and Obesity. Available from: <http://www.niddk.nih.gov/health/nutrit/pubs/statobes.htm#preval>
 47. O'Connell JB, Bristow MR (1993) Economic impact of heart failure in the United States: Time for a different approach. *J Heart Lung Transplant* 13(suppl):S107–S112
 48. O'Donoghue T, Rabin M (2006) Optimal sin taxes. *J Public Econ* 90:1825–1849
 49. Ratanawijitrasin S (1993) *The dynamics of health care finance: A feedback view of system behavior*. Ph.D. Dissertation, SUNY Albany
 50. Roberts EB, Homer J, Kasabian A, Varrell M (1982) A systems view of the smoking problem: Perspective and limitations of the role of science in decision-making. *Int J Biomed Comput* 13:69–86
 51. Schlessinger L, Eddy DM (2002) Archimedes: A new model for simulating health care systems—the mathematical formulation. *J Biomed Inform* 35:37–50
 52. Schorr LB (1997) *Common Purpose: Strengthening Families and Neighborhoods to Rebuild America*. Doubleday/Anchor Books, New York
 53. Singer M, Clair S (2003) Syndemics and public health: Reconceptualizing disease in bio-social context. *Med Anthropol Q* 17:423–441
 54. Singer M (1996) A dose of drugs, a touch of violence, a case of AIDS: Conceptualizing the SAVA syndemic. *Free Inquiry* 24:99–110
 55. Sterman JD (1988) A skeptic's guide to computer models. In: Grant L (ed) *Foresight and National Decisions*. University Press of America, Lanham, pp 133–169
 56. Sterman JD (2000) *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill, Boston
 57. Tompson KM, Tebbens RJD (2007) Eradication versus control for poliomyelitis: An economic analysis. *Lancet* 369:1363–1371. doi:10.1016/S0140-6736(07)60532-7
 58. Tengs TO, Osgood ND, Chen LL (2001) The cost-effectiveness of intensive national school-based anti-tobacco education: Results from the tobacco policy model. *Prev Med* 33:558–70

59. U.K. Prospective Diabetes Study Group (1998) Tight blood pressure control and risk of macrovascular and microvascular complications in Type 2 diabetes. *The Lancet* 352:703–713
60. Vennix JAM (1996) *Group Model-building: Facilitating Team Learning Using System Dynamics*. Wiley, Chichester
61. Wagner EH, Sandhu N, Newton KM et al (2001) Effect of improved glycemic control on health care costs and utilization. *JAMA* 285(2):182–189
62. Wallace R (1988) A synergism of plagues. *Environ Res* 47:1–33
63. Wolfson MC (1994) POHEM: A framework for understanding and modeling the health of human populations. *World Health Stat Q* 47:157–176
64. Zack MM, Moriarty DG, Stroup DF, Ford ES, Mokdad AH (2004) Worsening trends in adult health-related quality of life and self-rated health—United States, 1993–2001. *Public Health Rep* 119:493–505

Hierarchical Dynamics

MARTIN NILSSON JACOBI

Complex Systems Group, Department of Energy and Environment, Chalmers University of Technology, Gothenburg, Sweden

Article Outline

[Glossary](#)

[Definition](#)

[Introduction](#)

[Overview](#)

[Temporal Hierarchies: Separation of Time Scales](#)

[Structural Hierarchies: Foliations](#)

[Conclusion](#)

[Future Directions](#)

[Acknowledgment](#)

[Bibliography](#)

Glossary

In this section some definitions and results that are important as background for the later exposition are summarized. Out of necessity the definitions are very short and the reader might want to consult standard textbooks on the subjects. This is especially true in the section on group theory, which is dense and only meant to recapitulate main results to a reader already familiar with the basics of the subject. It would have been appropriate to also include a brief background on differential geometry. Unfortunately this subject is so large that even a minimal introduction must span several pages. The reader is advised to consult a standard book on differential geometry and Lie group theory, e. g. [4], to find the necessary definitions and results.

Liouville's Theorem, Conservative and Dissipative Systems

Consider a dynamical system $\dot{x} = f(x)$ with a phase space $x \in \mathbb{R}^n$. Let $\rho(t, x)$ be a probability density on the phase space, defined so that $\rho(t, x)d^n x$ is the probability of finding the system in the phase space volume $d^n x$ at time t . Given an initial value, the differential equation has a unique solution. This observation results in the following continuity equation

$$\frac{\partial \rho}{\partial t} + \sum_i \frac{\partial (f_i \rho)}{\partial x_i} = 0, \quad (1)$$

for the density under the flow f . In general, we may write the time evolution of a probability density as

$$\frac{\partial \rho}{\partial t} = - \sum_i \frac{\partial (f_i \rho)}{\partial x_i} = -\mathcal{L}\rho, \quad (2)$$

where \mathcal{L} is called the Liouville operator. In quantum mechanics the convention $\frac{\partial \rho}{\partial t} = -i\mathcal{L}\rho$ is often used to ensure that \mathcal{L} is a Hermitian operator. We do not use this convention here. The evolution of the probability density along a trajectory is given by the total time derivative:

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_i f_i \frac{\partial \rho}{\partial x_i} = -(\nabla \cdot f)\rho, \quad (3)$$

where Eq. (1) is used in the last step. The factor $\nabla \cdot f$ measures the phase space contraction under the flow of the dynamical system. If the system is Hamiltonian, then the degrees of freedom are given by x_i and p_i and there exist a function $H(x, p, t)$ (the Hamiltonian) such that the dynamics can be written on the form:

$$\dot{x}_i = f_i(x, p) = -\frac{\partial H}{\partial p_i}, \quad \dot{p}_i = f_{i+d}(x, p) = \frac{\partial H}{\partial x_i},$$

$i = 1, \dots, d$. It then follows that

$$\begin{aligned} \nabla \cdot f &= \sum_{i=1}^d \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^d \frac{\partial f_{i+d}}{\partial p_i} \\ &= - \sum_{i=1}^d \frac{\partial^2 H}{\partial x_i \partial p_i} + \sum_{i=1}^d \frac{\partial^2 H}{\partial p_i \partial x_i} = 0. \end{aligned}$$

We conclude that $\frac{d\rho}{dt} = 0$ for a Hamiltonian system, i. e. the phase space volume is conserved. This result is called Liouville's theorem and it is valid for closed Hamiltonian systems. We also note that, with the standard scalar product $(f, g) = \int dx f(x)g(x)$, the Liouville operator is anti-symmetric ($\mathcal{L}^\dagger = -\mathcal{L}$) for Hamiltonian systems.

If $\nabla \cdot f < 0$ in some region of the phase space the phase space volume is contracting, we say that the system is dissipative. A consequence of a contracting phase space is that there exists an attractor. An attractor is a subset to which trajectories from some regions of initial values evolve asymptotically. For chaotic dissipative systems, the attractor is usually not smooth and can have non-integer Hausdorff dimension. By contrast, a conservative system do not have any attractors.

The Langevin and Fokker–Planck Equations

One of the simplest stochastic differential equations is the Langevin equation

$$\dot{v} = -\gamma v + \zeta(t), \quad (4)$$

where $\zeta(t)$ represents white noise with zero mean $\langle \zeta(t) \rangle_t = 0$ and variance defined as $\langle \zeta(t)\zeta(t + dt) \rangle_t = 2B\delta(dt)$ (δ denotes a Dirac delta function). The first moments of v can be shown to be [55]:

$$\langle v(t) \rangle = 0 \quad \langle v(t)^2 \rangle = B/\gamma.$$

If v is interpreted as velocity, then by the definition of temperature $\langle v(t)^2 \rangle = k_B T$ (assuming unit mass) and the fluctuation-dissipation theorem follows from the variance:

$$B = \gamma k_B T. \quad (5)$$

For a deterministic dynamical system, the time evolution of a probability density on the phase space is given by Liouville's equation (2). For a stochastic differential equation, such as the Langevin equation, the time evolution of a probability density is given by a Fokker–Planck equation:

$$\frac{\partial \rho}{\partial t} = -\gamma \nabla \cdot \rho + B \Delta \rho.$$

Note that the stochastic part of the dynamics shows up as a diffusion term. In general, for a stochastic differential equation on the form

$$\dot{x}_i = f_i(x) + \zeta_i(t), \quad (6)$$

with $\langle \zeta_i(t) \rangle = 0$ and

$$\langle \zeta_i(t_1)\zeta_j(t_2) \rangle = 2B_{ij}(x)\delta(t_1 - t_2),$$

the corresponding Fokker–Planck equation reads

$$\frac{\partial \rho}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} (f_i \rho) + \sum_{ij} \left(\frac{\partial}{\partial x_i} B_{ij} \frac{\partial}{\partial x_j} \right) \rho. \quad (7)$$

It is important to be aware of subtleties with the interpretation of stochastic differential equations. The fundamental problem lies in the ambiguous representation of the noise $\zeta(t)$, which is not a regular function (it is not even continuous). There are essentially two different interpretations of (6), called Itô and Stratonovich. They lead to different forms of the diffusion term in Eq. (7). This issue are not of central concern to the current presentation and we do not discuss it further. The interested reader is recommended to read two standard references on the Fokker–Planck equation: Risken & Frank and Gardiner [13,41].

Ergodicity and Mixing

Let (X, Σ, μ) be a probability space, where X is the space (or set), Σ is a σ -algebra, and μ is a probability measure on Σ . Consider a map T that is measure preserving, $\mu(T^{-1}(E)) = \mu(E)$ for any $E \in \Sigma$. The map T is called ergodic if, for almost all $x \in X$ and any Lebesgue measurable function f ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = \int f d\mu,$$

where μ is an invariant measure. Alternatively one may define ergodic as follows: whenever $T(E) = E$, i. e. E is an invariant measure, then it follows that either $\mu(E) = 0$ or $\mu(E) = 1$. Intuitively ergodicity means that we can study the map's properties by looking at a single trajectory. The definition of ergodicity for continuous dynamical system is analogous.

A measure preserving map T is called (strongly) mixing if

$$\lim_{n \rightarrow \infty} \mu(T^n(A) \cap B) = \mu(A)\mu(B),$$

where A, B are measurable sets. Intuitively, mixing means that different sets of initial distributions become intertwined with each other as the system evolves. The generic metaphor is a drop of dye in a glass of water. Mixing systems are always ergodic but the converse is not necessarily true. For example, an irrational flow on a torus is ergodic but not mixing. The mixing property is strongly connected with chaotic motion. The exponential stretching and folding of the phase space associated with nonzero Lyapunov exponents results in a mixing behavior for chaotic dynamics.

Some Concepts from Group Theory

Group A group is a set G endowed with a binary operation $*$: $G \times G \rightarrow G$. We often denote the group

by $(G, *)$, or if there is no risk for confusion just G . The operator $*$ fulfills three group axioms: Associativity $a * (b * c) = (a * b) * c$; Identity element, there exist a unique element $e \in G$ such that $\forall a \in G \ a * e = e * a = a$; Inverse, for each element $a \in G$ there exist a unique element $a^{-1} \in G$ such that $a * a^{-1} = a^{-1} * a = e$.

Subgroup A subset H of G is called a subgroup if it is closed under the group operation ($\forall a, b \in H$ it follows that $a * b \in H$), and if H is a group in itself ($e \in H$ and $\forall a \in H$ it follows that $a^{-1} \in H$).

Classes We say that a and a' are said to be conjugates if $\exists b \in G$ such that $a' = b^{-1} * a * b$. Conjugacy is an equivalence relation and each element $a \in G$ belong to exactly one class. We call the resulting group a direct product of G_1 and G_2 , denoted by $G_1 \otimes G_2$. Note that G_1 and G_2 are both normal subgroups of $G_1 \otimes G_2$.

Normal subgroup If N is a subgroup of G and furthermore $\forall b \in N$ and $\forall a \in G$ it is true that $a * b = b' * a$ for some $b' \in N$, then N is called a normal subgroup of G . A normal subgroup can be identified with a union of classes.

Direct product Let $(G_1, *)$ and (G_2, \cdot) be two groups. We can form ordered pairs (a_1, a_2) where $a_1 \in G_1$ and $a_2 \in G_2$. We can then define a group product on the ordered pair as $(a_1, a_2) \circ (a'_1, a'_2) = (a_1 * a'_1, a_2 \cdot a'_2)$.

Coset and quotient group If H is a subgroup of G , a left (right) coset of an element $a \in G$ is defined as $aH = \{a * b : b \in H\}$ ($Ha = \{b * a : b \in H\}$). The collection of aH (Ha) $\forall a \in G$ are called the left (right) cosets of H in G . The left and right cosets of H in G coincide ($aH = Ha$) if and only if H is a normal subgroup of G . In fact the importance of normal subgroups comes from a possibility to define a quotient group (or factor group) G/N . The elements in the quotient group G/N can be identified with the cosets of N in G . The group operation \cdot on G/N is defined as $(aN) \cdot (bN) \doteq (a * b)N$ (note that this definition makes sense since $Nb = bN$).

Homomorphism, isomorphism, and automorphism

Let $(G, *)$ and (F, \cdot) be two groups. A map $\phi: G \rightarrow F$ is called a homomorphism if $\phi(a * b) = \phi(a) \cdot \phi(b)$ for all $a, b \in G$. It follows directly that $\phi(e_G) = e_F$ and $\phi(a^{-1}) = \phi(a)^{-1}$. If the map ϕ is also bijective, it is called an isomorphism and in this case G and F are considered “essentially the same”, denoted $G \simeq F$. If $\Psi: G \rightarrow G$ is an isomorphism from G to itself, then Ψ it is called an automorphism. The kernel of an homomorphism $\phi: G \rightarrow F$, denoted $\ker(\phi)$ is defined as the pre-image of the identity element in F , i.e. $\ker(\phi) = \{a \in G : \phi(a) = e_F\}$. In general, the

pre-image of any subgroup in F is a subgroup in G . The pre-image of the trivial subgroup e_F , i.e. $\ker(\phi)$, is also a normal subgroup in G (this follows from $a \in \ker(\phi)$ gives $\phi(b * a * b^{-1}) = \phi(b) \cdot \phi(a) \cdot \phi(b^{-1}) = \phi(b) \cdot e_F \cdot \phi(b^{-1}) = \phi(e_G) = e_F$). With these definitions we can state the fundamental theorem of isomorphisms: $\phi(G) \simeq G / \ker(\phi)$.

Semi-direct product Let $(N, *)$ and (H, \cdot) be groups and $\Phi: H \rightarrow \text{Aut}(N)$ map from H to the set of automorphisms on N , we use the notation $\Psi(h)(\cdot) \doteq \Psi_h(\cdot)$. A semi-direct product of N and H with respect to Ψ , denoted $N \rtimes_{\Psi} H$, is defined by the set of pairs (n, h) where $n \in N$ and $h \in H$, and the group operator defined by $(n_1, h_1) \circ (n_2, h_2) = (n_1 * \Psi_{h_1}(n_2), h_1 \cdot h_2)$. N is a normal subgroup of $N \rtimes_{\Psi} H$ (if H is also a normal subgroup then $N \rtimes_{\Psi} H \simeq N \otimes H$). Conversely, if N is a normal subgroup and H a subgroup of G such that for each $g \in G$, $g = n \circ h$ for some $n \in N$ and $h \in H$ and $N \cap H = e_G$, then $G \simeq N \rtimes_{\Psi} H$ with $\Psi_h n = h \circ n \circ h^{-1}$. The normal subgroup can be eliminated by the quotient group construction $(N \rtimes_{\Psi} H) / N \simeq H$.

Semigroup A semigroup is a set S with an associative binary operation \circ . A semigroup can trivially always be extended with an identity element e by considering $e \cup S$, and let $e \circ a = a \circ e = a \forall a \in S$. It is the absence of inverse elements that makes semigroups fundamentally different from groups. A dynamical system can be viewed as a transformation semigroup acting on a state space (see below). Semigroups are also important in the study of finite automata (e.g. Krohn–Rhodes decomposability theorem [24]) and theoretical computer science in general.

Lie group and Lie algebra A Lie group is a d dimensional differentiable manifold with a group structure. Both the group operation $*$: $G \times G \rightarrow G$ and the inversion map $i: G \rightarrow G$, $i(a) = a^{-1}$ are smooth maps between manifolds. The vector fields spanning the tangent space at the identity element in a Lie group spans an algebra, called a Lie algebra $\mathfrak{g} \simeq TG|_e$. The vector fields in the Lie algebra are characterized by their invariance under group multiplication (from the right or left). The Lie algebra is then a vector space together with a bilinear operator $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$, called the Lie bracket. The Lie bracket is defined as the commutator of two vector fields $[\mathbf{v}, \mathbf{w}] = \mathbf{v}(\mathbf{w}) - \mathbf{w}(\mathbf{v})$. It is clear that the Lie algebra is closed under the Lie bracket since the vector fields span a tangent space (they must form an involution). Any group element $g \in G$, connected to the identity element, can be expressed in terms of a finite succession of

exponentials of the vector fields in the Lie algebra $g = e^{\epsilon_1 v_{i_1}} e^{\epsilon_2 v_{i_2}} \dots e^{\epsilon_d v_{i_d}}$, for some ϵ_j . We therefore view the Lie algebra as a generator of the Lie group. Most properties of the Lie group can be analyzed in terms of the Lie algebra. These concepts are similar in spirit to how all global information about an analytic function is contained in a local Taylor expansion.

Transformation group A group G with an action ψ such that G can act on a set M , $\psi: G \times M \rightarrow M$ is called a transformation group. Examples of transformation groups are rotations acting on \mathbb{R}^n . Let $a, b \in G$ and $x \in M$. A transformation group action then fulfills the following properties: $\psi(a, \psi(b, x)) = \psi(a * b, x)$ and $\psi(e, x) = x$. An orbit of a point $x \in M$ is a set in M that is invariant under the action ψ . Let \mathcal{O} denote an orbit then if $x \in \mathcal{O}$ it follows that $\psi(a, x) \in \mathcal{O} \forall a \in G$. It is clear that an element $x \in M$ belongs to exactly one orbit, i. e. the orbits form a partition of M through their associated equivalence classes. A group action is called transitive if there is only one orbit covering the entire set M . The set of all orbits is defined as a quotient set (or quotient manifold if M is a manifold), denoted by M/G . In the case when G is a Lie group and M is a smooth manifold, the regularity of the quotient manifold M/G is not guaranteed unless the following extra conditions are assumed for the group action: ψ is regular. Semi-regular means that all orbits have the same dimension as submanifolds in M . Regular means that the action is semi-regular, and in addition for each $x \in M$ there exists a neighborhood U of x such that each orbit intersecting U is a pathwise connected subset.

Finite state automata A finite state automaton is defined as a triplet $\mathcal{A} = (A, X, \delta)$ where A is a finite (nonempty) state set, X is an input alphabet, and $\delta: X \times A \rightarrow A$ is a transition function. There is an optional output alphabet but this could be identified with A . The transition function can be extended to act on words: $\delta(uv, a) = \delta(v, \delta(u, a))$ etc. The set of all words formed out of the alphabet X is denoted by X^* , and the set of nonempty finite words are denoted by X^+ . There is a natural equivalence defined on X^* : $U \sim V$ if $\delta(U, a) = \delta(V, a)$ for all $a \in A$ (where $U, V \in X^*$). The equivalence classes X^+ / \sim together with the concatenation operation form the characteristic semigroup of the automaton $S(\mathcal{A})$. The characteristic semigroup $S(\mathcal{A})$ acting on A with the action δ defines a transformation semigroup. In this way any finite automaton can be analyzed as a transformation semigroup.

Definition

The concept of hierarchies is often used in our descriptions of the world. Sometimes the hierarchical structure is primarily a construct of our mind, and can therefore be considered subjective. Examples are the Linnean taxonomy, Chomsky's linguistic hierarchies, hierarchies in object oriented programming, and hierarchical structures within human organizations. In other situations, one may argue that the physical world, external to the human mind (as far as something of that nature actually exists), is objectively organized in a hierarchical way. Examples of the latter situation can be physical hierarchies of particles and length/time scales, evolutionary taxonomies in biology, hierarchical organization within organisms (cells versus organs), ecosystems (food chains), and computational complexity classes in theoretical computer science. The distinction between subjective and objective is provocative. Not only is it unclear if the distinction actually makes sense in all relevant cases, it is in fact problematic to argue sharply in any of the examples given. The Linnean taxonomy can, for example, be argued for as objective since it is based on important morphological differences and similarities between the organisms. The characteristics used are also central for the function, which is often reproduction. Is it then fair to say that the result is just a reflection of the human visual cortex combined with higher cognitive functions in the human brain? On the other side, is not our description of the universe in terms of physical law, expressed as mathematical relations, subjective *in absurdum*? However, the important difference between the subjective and objective hierarchies listed above lies in whether or not the hierarchical structure is pre-assumed a priori to the model building, and included as a central part of the description of the system; or, as in the hierarchies listed as objective, a posteriori result derived from a model of reality that does not pre-assume the existence of a hierarchical organization (we may say that the discovery of a hierarchy includes some element of surprise).

While the idea of using hierarchies to describe the world goes back to Aristotle, Simon was one of the first to discuss hierarchical organization as a key ingredient in evolvable and/or controllable complex systems [51]. The important insight was to acknowledge how hard it is to make a complex system robust and adaptable if it is not organized hierarchically. For a system to remain robust as the overall complexity increases, the internal complexity must be confined by recursively "hiding" internal degrees of freedom into modules. The modules then communicate with each other only through narrow channels. In this way, a combinatorial explosion of complexity stemming from

“everyone talking to everyone about everything” can be avoided. In conclusion, hierarchies through modularity is a central theme when trying to combine complexity with robustness, controllability, and evolvability.

Understanding hierarchal organization in dynamical systems may also lead to insights into mechanisms behind emergence in systems that are, at least in principle, reductionistic. Emergence is one of the central themes in complex systems science. Methods and ideas used for analyzing hierarchical dynamics are likely to play a central role in any theoretical framework addressing emergence. Finally, many of the techniques that we discuss in the context of hierarchical dynamics originated in the related fields of model reduction and multi-scale simulation. In model reduction one typically seeks systematic methods for reducing the complexity of a specific model, or a class of models, thereby making them more manageable in terms of simulation or analytic studies. As we will see, this is a special case of hierarchical dynamics (a case of great practical and conceptual importance).

Introduction

Technically, a definition of a hierarchy must contain some measure that defines the levels, i. e. the objects in the hierarchy must be ordered according to their complexity, or some other measure that defines the hierarchy. In this presentation we focus on hierarchies in dynamical systems. The most natural measure to form a hierarchy around in this case is the dimensionality of the phase space, or in the discrete case the cardinality of the state space. Furthermore, it is natural to assume an upward causality on the hierarchy based on projections from a higher dimensional phase space (or if the system is discrete state space of higher cardinality) to a lower dimensional phase space (state space of lower cardinality). An interesting exception is the renormalizing projections used to analyze critical phenomena. It builds a self-similar hierarchy without reducing the cardinality of the state space. This is possible in an infinite system (the thermodynamic limit) since the cardinality of e. g. the natural numbers does not change when half of the states are eliminated. Returning to the discussion on projections that do reduce the dimensionality of the phase space. To avoid arbitrary projections from defining new hierarchical levels, we must also require the resulting dynamics to be closed. Closure means that the dynamics on the higher level (lower dimensionality) is self-contained, i. e. the evolution of the system can be effectively described using only its internal degrees of freedom, not information from the lower levels in the hierarchy. However, this definition is still not strict enough. We must also

require that the resulting system is a first order ordinary differential equation (it is a continuous time system) and Markovian, which means that the future evolution of the system depends only on the current state, not on the history of the system. To demonstrate the importance of the Markov requirement, we consider a dynamical system on the form

$$\begin{aligned}\dot{x} &= f(x, y) \\ \dot{y} &= g(x)y.\end{aligned}$$

Formally we can solve the second equation for y and put the result into the first equation. The result reads

$$\frac{\partial x}{\partial t} = f\left(x, y_0 e^{\int_0^t ds g(x(s))}\right).$$

We end up with a one-dimensional equation but with an infinite memory term. This reduction corresponds to using the projection $\pi(x, y) = x$ (detailed explanation of the meaning of this projection is given in Sect. “[Structural Hierarchies: Foliations](#)”). The exact form of the right hand side of \dot{y} was chosen for simplicity and clarity. In principle the result generalizes, at least locally, to generic functional forms $\dot{y} = g(x, y)$. The projection in this example was arbitrary and did not reflect any structure in the dynamical system. It is clear that such projections are not interesting when defining a hierarchy. By requiring the projected system to be Markovian we restrict attention to projections that do reflect hierarchical organization of the dynamics.

The fact that no information is lost by a generic projection of a dynamical system is useful in practice. Delay-time embedding for attractor reconstruction is for example based on this observation [38,45,53]. In our context we are interested in projections that actually do hide some information on the lower levels in the hierarchy. Takens’ embedding theorem states that such projections, if they exist, constitute singular (with regards to a some natural measure) points in the space of all possible projections. The purpose here is to identify the constraints that have these singular projections as solutions. We conclude the introduction by repeating the central theme that will be our guide throughout this presentation:

Definition 1 Each level in a hierarchy should be a self-contained Markovian dynamical system.

For a more extensive discussion on objectivity in hierarchical dynamics and the Markov property see [49].

Overview

There are mainly two types of hierarchies in dynamical systems: structural and temporal hierarchies. Temporal hi-

erarchies are defined by separation of time scales between the different levels. The local nature of physical interactions connects time scales to length scales. As a consequence, the levels in temporal hierarchies are also often associated with separation in length scales. One may turn the argument around and claim that our choice of metric is a reflection of how the interactions between objects in the universe behave. In any case the result is the same, there is a tight coupling between time and length scales. Structural hierarchies are derived from geometric properties of the dynamics. These geometric properties stems from decomposability, or skew-product decomposability (to be defined later), of the vector field defining the flow. Simple examples of structural hierarchies are non-interacting subsystems or systems with constants of motion.

Temporal hierarchies are discussed extensively in the literature. The mechanisms behind structural hierarchies are also very carefully analyzed in the context of classical mechanics and quantum mechanics. In classical mechanics, the connection between symmetries of the Hamiltonian and invariants of the motion was clarified by Noether's theorem. In quantum mechanics one is often interested in the result of composing multiple particles with certain symmetries into a single object with a larger symmetry group. The general mathematical setting for modern gauge theory is fiber bundles. As we discuss in Sect. "Structural Hierarchies: Foliations" this is also the natural framework for working with structural hierarchies. In this presentation however, our aim is not to clarify the connection between modern theoretical physics and hierarchical dynamics, but rather to show the connection between temporal and structural hierarchies in dynamical systems. This connection is usually not emphasized in the literature.

Temporal Hierarchies: Separation of Time Scales

Dynamical systems with high dimensionality often display dynamics on vastly different time scales. As a consequence, when the system is analyzed at a specific time scale, some degrees of freedom evolve so slowly that they can effectively be treated as (adiabatic) constants, while other degrees of freedom evolve very fast compared to the time scale of interest. We wish to systematically eliminate both the very fast and the very slow degrees of freedom. The fast dynamics must be treated differently depending on its characteristics: it can be represented by its average influence; it can be treated as white or colored noise; or its dynamics is dissipative and the dynamics quickly relaxes to an adiabatic fixed point. For a review on model reduc-

tion in dynamical systems with time scale separation, see Givon et al. [14].

Elimination of Slaved Degrees of Freedom

Self-organization can be defined as the tendency for a system to increase internal order without influence from the outside. From a more technical perspective, self-organization is a result of a collapse of the phase space volume. As a result the effective dimensionality of the system's phase space is reduced. As was discussed in Sect. "Liouville's Theorem, Conservative and Dissipative Systems", a shrinking volume of phase space elements in the Liouville equation is a result of energy dissipation, or "friction". It is clear that a closed physical system cannot be self-organizing since this would break the first and second law of thermodynamics, as well as Liouville's theorem that ensures conserved phase space volume for Hamiltonian systems. Self-organizing systems are open, and often kept in a out-of-equilibrium, but stationary, state by external energetic driving. A simple example of a driven dissipative systems is a forced damped pendulum:

$$\ddot{\theta} + \gamma \dot{\theta} + \sin(\theta) = A \sin(\omega t + \phi). \quad (8)$$

If we re-write this system as a first order differential equation we can show that $\nabla \cdot f = -\gamma$. Equation (3) then implies that the phase space volume shrinks exponentially. The attractor for Eq. (8) is a limit cycle. Note that the shrinking of the phase space volume only depend on the dissipation, not on the driving on the right hand side. Naively one could have expected that the driving would tend to expand the phase space volume, but this is not the case. Liouville's theorem holds true also for mechanical systems with time dependent Hamiltonians.

A generic feature of driven dissipative systems is that fast degrees of freedom, due to large negative exponent associated with dissipation, often relaxes to an adiabatic fixed point, i. e. a point in the phase space that appears effectively fixed on the time scale of the fast dynamics but that changes on the time scale set by the slow degrees of freedom. The overall dynamics is therefore slaved to a slow positively invariant, or inertial, manifold (or more correctly attractor) and the resulting dimensionality is reduced. It should be noted that the geometry of the reduced system is often very complicated, taking e. g. the form of a strange attractor [42]. This picture of self-organization has been advanced by Haken in his work on synergetics [18]. Lately the same idea has also been revitalized in the turbulence community, primarily by a proof of existence of inertial manifolds in a class of hyperbolic dynamical systems [11]. Positive invariant manifolds are also used

in model reduction schemes in chemical kinetics [17]. In general it is hard to strictly prove the existence of global inertial manifolds even though it is often suspected that they exist and lead to spontaneous dimensional reduction in driven dissipative systems.

We now present the generic setup for inertial manifolds. Consider a dissipative dynamical system for a function $u(t, x)$ on the form

$$\dot{u} = Au + F(u), \quad (9)$$

where A is a symmetric linear operator with compact resolvent, defined on a Hilbert space H and F is a nonlinear function. The operator A contains a spatial coupling through space derivatives. The operator F is nonlinear but contains no derivatives (sometimes F is also defined to include derivatives of lower order than A , but here we assume that this is not the case). For systems of the type (9) the main result is [29]:

Theorem 2 *Assume that F is Lipschitz continuous with constant c . If there is a gap in the spectrum of A such that $|\lambda_n - \lambda_{n+1}| > c$, then Eq. (9) then has an inertial manifold of dimension n .*

Consider the special case when the operator A is the Laplacian ($A = \Delta$). Common examples of this situation are reaction-diffusion equations [28]. If $A = \Delta$ in d -dimensions, the spectrum scales as $\lambda_k \sim -k^{2/d}$, and it follows that

$$\begin{aligned} \lambda_k - \lambda_{k-1} &\sim k & d = 1 \\ \lambda_k - \lambda_{k-1} &\sim 1 & d = 2. \end{aligned}$$

We conclude that the spectral gap becomes arbitrarily large for $d = 1$, which shows that inertial manifolds do exist. For $d = 2$ the spectral gap is constant so the existence of an inertial manifold is not clear from the general scaling argument. However, in $d = 2$ on finite domains there are more advanced arguments showing that under certain general conditions the spectral gap can still become large [28]. Reaction-diffusion equations in two dimensions do possess inertial manifolds.

Now let us assume the existence of an inertial manifold of dimension n . We proceed by defining a spectral projection operator \mathcal{P} onto the first n eigenfunctions of A

$$\mathcal{P}u = \sum_{i=1}^n (u, v_i) v_i,$$

where v_i are normalized eigenfunctions of A and (\cdot, \cdot) denotes a scalar product on H . The operator \mathcal{P} is idempotent, i. e. $\mathcal{P}^2 = \mathcal{P}$, and the complement $\mathcal{Q} = \mathcal{I} - \mathcal{P}$ is also

a projection operator and $\mathcal{P}\mathcal{Q} = 0$. Furthermore, \mathcal{P} commutes with A , $\mathcal{P}A = A\mathcal{P}$. We split the Eq. (9) according to

$$\frac{\partial \mathcal{P}u}{\partial t} = \mathcal{P}A\mathcal{P}u + \mathcal{P}F(\mathcal{P}u + \mathcal{Q}u) \quad (10)$$

$$\frac{\partial \mathcal{Q}u}{\partial t} = \mathcal{Q}A\mathcal{Q}u + \mathcal{Q}F(\mathcal{P}u + \mathcal{Q}u). \quad (11)$$

The inertial manifold can be expressed implicitly as a graph $\Phi: \mathcal{P}H \rightarrow \mathcal{Q}H$. Inserting $\mathcal{Q}u = \Phi(\mathcal{P}u)$ into Eq. (10) gives

$$\frac{\partial \mathcal{P}u}{\partial t} = \mathcal{P}A\mathcal{P}u + \mathcal{P}F(\mathcal{P}u + \Phi(\mathcal{P}u)),$$

which is a closed evolution equation for the slow dynamics $\mathcal{P}u$. The problem is to calculate Φ . The crudest approximation is to set $\Phi = 0$. This is referred to as the linear Galerkin method. In contrast, there are many numerical schemes for nonlinear Galerkin methods that provides more nontrivial approximations of Φ , see e.g. [10,39]. In general one assumes that the fast dynamics relaxes on a time scale τ and use some method for solving Eq. (11) under this condition. Implicit Euler gives $\Phi(\mathcal{P}u)$ as a solution to the fixed point map

$$\mathcal{Q}u \rightarrow -\tau(I + \tau A\mathcal{Q})^{-1} \mathcal{Q}F(\mathcal{P}u + \mathcal{Q}u). \quad (12)$$

This map is a contraction for small enough τ . Due to the time scale separation we can chose $\tau \sim \lambda_{n+1}^{-1}$. In [10] the fixed point is approximated by one application of (12) on the initial linear Galerkin guess $\mathcal{Q}u = 0$.

There exist many methods for calculating invariant manifolds in dissipative systems on more general forms than Eq. (9). Chemical kinetics and transport theory have been especially active areas [17].

Example 1 As a simple explicit example of slaving we look at the system (from [14])

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_3 \\ \dot{x}_2 &= x_1 + x_2/5 \\ \dot{x}_3 &= 1/5 - 5x_3 + y \\ \dot{y} &= \epsilon^{-1}(x_1x_3 - y), \end{aligned}$$

where $\epsilon \ll 1$ defines the spectral gap ($|\lambda_x - \lambda_y| \sim \epsilon^{-1}$). We assume that the y variable has time to relax to its adiabatic fixed point $\dot{y} = \mathcal{O}(\epsilon)$. The graph that defines the inertial manifold is in this case given by $\Phi(y) = x_1x_3 + \mathcal{O}(\epsilon)$. The resulting reduced equations read

$$\begin{aligned} \dot{X}_1 &= -X_2 - X_3 \\ \dot{X}_2 &= X_1 + X_2/5 \\ \dot{X}_3 &= 1/5 - 5X_3 + X_1X_3, \end{aligned}$$

which is recognized as the Rössler system.

Averaging

Averaging is a phenomenon whose effect is reminiscent of the slaving described in the last section. However, in an averaging system the fast degrees of freedom typically do not evolve to an adiabatic fixed point. They stay active and continue to evolve on the fast time scales. Their effect on the slow degrees of freedom can be described by their average influence. As a result it is possible to derive closed equations for the slow degrees of freedom, while only taking the average affect of the fast dynamics into account. As an example, consider a system on the form

$$\begin{aligned}\dot{x} &= f(x, y) \\ \dot{y} &= \epsilon^{-1} g(x, y),\end{aligned}\quad (13)$$

where $\epsilon \ll 1$ and f and g are assumed to be of order unity. If, for a fixed x the y -dynamics is ergodic, Anasov's theorem states that the slow dynamics will on some finite time interval converge uniformly to an average equation ([44]):

$$\dot{X} = F(X),$$

where

$$\begin{aligned}F(\xi) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T d\tau f(\xi, \eta[\tau, y; x]) \\ &= \int f(\xi, y) \mu_\xi(dy),\end{aligned}$$

where $\eta[\tau/\epsilon, y; x]$ is the solution to the differential equation (fast dynamics with x fixed and initial value y , which will not matter asymptotically due to ergodicity):

$$\frac{\partial \eta[\tau, y; x]}{\partial \tau} = \epsilon^{-1} g(x, \eta[\tau, y; x]) \quad \eta[\tau = 0, y; x] = y, \quad (14)$$

and μ_ξ is the invariant ergodic measure of y for a fixed $x = \xi$. The main result is that $|x(t) - X(t)|$ is of order $\mathcal{O}(\epsilon)$ for $\epsilon \ll 1$. In practice, the time average is taken over an interval proportional to the separation in time scale, i. e. (using a slight abuse of notation) $\epsilon \ll T \ll 1$. There are also many results on averaging in situations when the ergodicity assumption is not valid, see [14] and references therein. For extensive expositions on averaging and systematic perturbation methods see [25,44]. At the end of the next section we return to the issue of averaging as a limiting case of a noise approximation of the fast degrees of freedom.

White Noise Reduction: Mori–Zwanzig Projections

In non-equilibrium statistical physics dimensional reduction often means going from a deterministic high dimensional

model to a reduced Langevin-type model that includes noise. The randomness stems from fast motion that, on the time scale of the relevant (slow) degrees of freedom, can be approximated as white noise. The result is a Markovian dynamics for the slow degrees of freedom. This idea was first formalized by Zwanzig [54] and has later matured into the general framework described e. g. in [13,41]. The classic example of bulk degrees of freedom behaving like noise is a heath bath in contact with a heavy particle, see e. g. [55] for details. More modern studies are often focused on low dimensional chaotic fast subsystems as noise generators [3,8,43]. In the current presentation we review some of the later findings. Many of the results in this section are based on a series of papers by Just et al., e. g. [21,22,40].

Assume we have a dynamical system whose corresponding Liouville equation reads

$$\frac{\partial \rho_t}{\partial t} = -\mathcal{L}\rho_t. \quad (15)$$

We like to project away some part of the system (the fast degrees of freedom). To this end we define a projection operator \mathcal{P} , which splits the phase space density according to $\rho_t = \mathcal{P}\rho_t + \mathcal{Q}\rho_t$, where $\mathcal{Q} = 1 - \mathcal{P}$ and $\mathcal{P}\rho$ is the subsystem we are interested in (the slow dynamics). The projection operator is idempotent, i. e. $\mathcal{P}^2 = \mathcal{P}$, which also implies that $\mathcal{P}\mathcal{Q} = 0$. Equation (15) splits into

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}(\mathcal{P}\rho_t + \mathcal{Q}\rho_t) \quad (16)$$

$$\frac{\partial \mathcal{Q}\rho_t}{\partial t} = -\mathcal{Q}\mathcal{L}(\mathcal{P}\rho_t + \mathcal{Q}\rho_t). \quad (17)$$

(Compare these equations to Eqs. (10)–(11).) We want to find a closed equation for the time evolution of $\mathcal{P}\rho_t$. In general, if $\mathcal{P}\mathcal{L}\mathcal{Q}\rho_t = 0$, then Eq. (16) is closed. This is fulfilled, for example, if the projection operator commutes with the Liouville operator, $\mathcal{P}\mathcal{L} = \mathcal{L}\mathcal{P}$, then the closure follows immediately from $\mathcal{P}\mathcal{Q} = 0$. This is an interesting special case, connected to symmetries of the dynamical system, that will be discussed extensively in Sect. “Structural Hierarchies: Foliations”. To achieve this one may naively consider projections on the form [26] (based on Cauchy's formula):

$$\mathcal{P} = \frac{1}{2\pi i} \int_{\Gamma} d\zeta (1 - \zeta \mathcal{L})^{-1}, \quad (18)$$

where Γ is a closed curve in the complex plane. This is a spectral projection, i. e. a generalization of the projections used in Eq. (10). The operator (18) projects onto the space spanned by the eigenfunctions corresponding to the

eigenvalues contained within the closed curve Γ . Projections defined as in (18) commutes with \mathcal{L} by definition. However, projections on the density space do not in general correspond to dimensional reduction on the phase space. Therefore we cannot simply follow the same line of manipulations as in Eq. (10). We will discuss explicit forms of projections that do correspond to dimensional reductions shortly, but for now it is enough to accept that we often want to consider projections where $\mathcal{P}\mathcal{L} \neq \mathcal{L}\mathcal{P}$. In this case we proceed by formally solving Eq. (17) for $\mathcal{Q}\rho_t$ as

$$\mathcal{Q}\rho_t = \mathcal{Q}\rho_0 - \int_0^t d\tau e^{\mathcal{Q}\mathcal{L}(\tau-t)} \mathcal{Q}\mathcal{L}\mathcal{P}\rho_\tau.$$

(Note that this approach does not work for Eq. (10) since F is a nonlinear function.) We are interested in the asymptotic solutions, and we can therefore ignore the term that comes from the initial distribution. Inserting the solution for $\mathcal{Q}\rho_t$ into Eq. (16) gives a closed equation for $\mathcal{P}\rho_t$:

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}\mathcal{P}\rho_t + \mathcal{P}\mathcal{L} \int_0^t d\tau e^{-\tau\mathcal{Q}\mathcal{L}} \mathcal{Q}\mathcal{L}\mathcal{P}\rho_{t-\tau}. \quad (19)$$

The structure of Eq. (19) is interesting. As discussed in the Introduction, the dimensional reduction comes at a cost: the system has (infinite) memory, i.e. it is not Markovian. It should be noted that Eq. (19) is not simpler than Eq. (15). So far we have achieved nothing. The central idea is that, if there is a clear separation of time scales between $\mathcal{P}\rho_t$ and $\mathcal{Q}\rho_t$, then Eq. (19) can be simplified by using a Markovian approximation. To proceed we use a dynamical system where the fast and the slow degrees of freedom have been separated explicitly. Let x and y denote the slow and the fast degrees of freedom respectively. The dynamical system is given by

$$\begin{aligned} \dot{x} &= f(x, y) \\ \dot{y} &= \epsilon^{-1} g(x, y), \end{aligned} \quad (20)$$

with $0 < \epsilon \ll 1$ measuring the scale separation, f , and g are on the order of unity. There is a hidden subtlety in Eq. (20). As it stands, it is identical to the situation that was analyzed in connection to averaging, Eq. (13). In fact, in the $\epsilon \rightarrow 0^+$ limit, the y -dependent part of the back-coupling in $f(x, y)$ must scale as $\mathcal{O}(\epsilon^{-1/2})$. The intuitive rationale for this scaling is that the y -term in Eq. (20) should behave as noise.

The corresponding Liouville operator splits into a fast and a slow part:

$$-\mathcal{L} = \underbrace{\sum_i \frac{\partial}{\partial x_i} f_i(x, y)}_{-\mathcal{L}_s} + \frac{1}{\epsilon} \underbrace{\sum_i \frac{\partial}{\partial y_i} g_i(x, y)}_{-\mathcal{L}_f}. \quad (21)$$

The idea is now to find a closed evolution for the slow degrees of freedom x . We proceed by assuming that the fast variables have time to relax to a stationary state before x changes. We define the projection operator as

$$\mathcal{P}\rho_t(x, y) = \rho_{\text{ad}}(y|x) \int dy' \rho_t(x, y'), \quad (22)$$

where $\rho_{\text{ad}}(y|x)$ denotes the adiabatic equilibrium of the fast degrees of freedom, y , for a given value of x , i.e.

$$\rho_{\text{ad}}(y|x) = \lim_{\tau \rightarrow \infty} e^{\tau \mathcal{L}_f} \rho_t(x, y) \quad (23)$$

$$\mathcal{L}_f \rho_{\text{ad}}(y|x) = 0. \quad (24)$$

Equation (24) means that $\rho_{\text{ad}}(y|x)$ is a zero eigenfunction to \mathcal{L}_f and Eq. (23) is based on the assumption that the zero eigenfunction is unique for all fixed x , which is often true for a mixing system.

At this point it is interesting to make a connection to a result called Trotter's theorem. Assume that we have an operator that can be decomposed as $\mathcal{L} = \mathcal{L}_s + \epsilon^{-1} \mathcal{L}_f$. Then the following limit is well defined and converges [46]:

$$e^{t(\mathcal{L}_s + \epsilon^{-1} \mathcal{L}_f)} = \lim_{N \rightarrow \infty} (e^{t\mathcal{L}_f/2\epsilon N} e^{t\mathcal{L}_s/N} e^{t\mathcal{L}_f/2\epsilon N})^N, \quad (25)$$

where \mathcal{L}_s and \mathcal{L}_f do not commute in general, $\mathcal{L}_f \mathcal{L}_s \neq \mathcal{L}_s \mathcal{L}_f$ (if the operators do commute, the relation is trivial). The system can be accurately integrated from $\tau = 0$ to $\tau = t$ using N steps, where $1 \ll N/t \ll 1/\epsilon$. The intuitive rationale for the step size, and behind defining the projection as in Eq. (22), is the assumption that the fast degrees of freedom has time to effectively equilibrate before the slow degrees of freedom change significantly. In Eq. (25) the scheme is explicit: $e^{t\mathcal{L}_f/\epsilon N}$ relaxes the fast dynamics, whereas $e^{t\mathcal{L}_s/N}$ evolves the slow dynamics. Technically this means that the limit in Eq. (23) actually converge as $\tau \sim \mathcal{O}(\epsilon) \ll 1$. The (unique) adiabatic equilibrium distribution of the fast variables is used to generate noise, which drives the slow degrees of freedom.

To proceed with this scheme we need to re-write Eq. (19) as an expansion in ϵ . From (22), (23) and (24) follows the relation:

$$\mathcal{P}\mathcal{L}_f = \mathcal{L}_f \mathcal{P} = 0. \quad (26)$$

Using this fundamental relation we can formally close Eq. (16). Under the assumption that the fast variables relaxes to a stationary state, we set

$$\frac{\partial \mathcal{Q}\rho_t}{\partial t} = 0,$$

and use this in Eq. (17) to solve for $\mathcal{L}\mathcal{Q}\rho_t$. Then we use this result to close Eq. (16) (formally identifying the pseudo-inverse relation $\mathcal{Q}^{-1}\mathcal{Q} = \mathcal{Q}$, which follows from $\mathcal{Q}\mathcal{Q} = \mathcal{Q}$):

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}_s\mathcal{P}\rho_t + \mathcal{P}\mathcal{L}_s \left(\mathcal{L}_s + \frac{1}{\epsilon}\mathcal{L}_f \right)^{-1} \mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_t,$$

where we have used (26) to set $\mathcal{P}\mathcal{L} = \mathcal{P}\mathcal{L}_s$ and $\mathcal{L}\mathcal{P} = \mathcal{L}_s\mathcal{P}$. Expanding to first order in ϵ gives (under the assumption that $\alpha(\epsilon) \ll 1/\epsilon$):

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}_s\mathcal{P}\rho_t + \epsilon\mathcal{P}\mathcal{L}_s\mathcal{L}_f^{-1}\mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_t, \quad (27)$$

where \mathcal{L}_f^{-1} is well defined since its null-space is projected out by \mathcal{Q} . Note that for the second term in Eq. (27) to be non-zero in the $\epsilon \rightarrow 0$ limit, \mathcal{L}_s must be of order $\mathcal{O}(1/\sqrt{\epsilon})$. As discussed earlier, this scaling must come from the y -dependent part of the coupling $f(x, y)$. We will see this effect explicitly in an example below. To use Eq. (27) in practice the eigenfunctions of \mathcal{L}_f must be found. Solutions of (27) can then be expressed in terms of series expansions in these eigenfunction basis. In most examples where this approach is successful the fast variables have a stochastic dynamics, i. e. \mathcal{L}_f is a Fokker–Planck operator, see [13,41] for derivation of e. g. the Smoluchowski equation using a Hermitian polynomial basis. However, Eq. (27) is not the most convenient form for deriving a closed dynamics for the slow variables if the eigenfunctions of \mathcal{L}_f are not easy to find (as is usually the case when the fast variables have deterministic dynamics). One can often introduce a trick by adding a small amount of noise to the fast degrees of freedom to make the stationary distribution smooth and well defined, i. e. to “fatten” the usually fractal attractor. Here we discuss an alternative, more direct, approach. We start by using Dyson’s operator identity,

$$\begin{aligned} e^{-\tau\mathcal{Q}\mathcal{L}} &= e^{-\tau\mathcal{Q}\mathcal{L}_f} + \int_0^\tau d\sigma \frac{\partial}{\partial \sigma} \left(e^{-\sigma\mathcal{Q}\mathcal{L}} e^{-(\tau-\sigma)\mathcal{Q}\mathcal{L}_f} \right) \\ &= e^{-\tau\mathcal{Q}\mathcal{L}_f} - \int_0^\tau d\sigma e^{-\sigma\mathcal{Q}\mathcal{L}} \mathcal{Q}\mathcal{L}_s e^{(\sigma-\tau)\mathcal{Q}\mathcal{L}_f}, \end{aligned} \quad (28)$$

to write a perturbation expansion of Eq. (19) in terms of ϵ (using Eq. (26)):

$$\begin{aligned} \frac{\partial \mathcal{P}\rho_t}{\partial t} &= -\mathcal{P}\mathcal{L}_s\mathcal{P}\rho_t + \mathcal{P}\mathcal{L}_s \int_0^t d\tau e^{\mathcal{Q}\mathcal{L}_f(\tau-t)/\epsilon} \mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_\tau \\ &\quad - \epsilon^2 \mathcal{P}\mathcal{L}_s \int_0^{t/\epsilon} d\tau \int_0^{\tau/\epsilon} d\sigma e^{\mathcal{Q}\mathcal{L}_f(\tau-\sigma)/\epsilon} \mathcal{Q}\mathcal{L}_s \\ &\quad \cdot e^{-\mathcal{Q}(\mathcal{L}_f + \epsilon\mathcal{L}_s)\tau} \mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_{t-\epsilon\sigma}. \end{aligned}$$

The last term is of order ϵ^2 will from now on be dropped. We now define adiabatic averages as (the last equality assumes ergodicity in the fast degrees of freedom):

$$\begin{aligned} \langle h \rangle_{\text{ad}}(x) &\doteq \int dy h(x, y) \rho_{\text{ad}}(y|x) \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T d\tau h(x, \eta[\tau/\epsilon, y; x]), \end{aligned} \quad (29)$$

where $\eta[\tau/\epsilon, y; x]$ is defined as in Eq. (14). Using this formalism we can write Eq. (28) as:

$$\frac{\partial \bar{\rho}_t}{\partial t} = -\langle \mathcal{L}_s \rangle_{\text{ad}} \bar{\rho}_t + \int_0^t d\tau \langle \mathcal{L}_s e^{-\tau\mathcal{L}_f/\epsilon} \mathcal{Q}\mathcal{L}_s \rangle_{\text{ad}} \bar{\rho}_{t-\tau}, \quad (30)$$

where $\bar{\rho}_t(x) = \int dy \rho_t(x, y)$ is the density for the slow degrees of freedom only. It remains to analyze the kernel in the last term. For the dynamics to become approximately Markovian we need to assume that the fast degrees of freedom are exponentially mixing. Under this assumption, the correlation functions captured by the kernel decay rapidly on the time scale ϵ , and we write:

$$\begin{aligned} &\int_0^t d\tau \langle \mathcal{L}_s e^{-\tau\mathcal{L}_f/\epsilon} \mathcal{Q}\mathcal{L}_s \rangle_{\text{ad}} \bar{\rho}_{t-\tau} \\ &= \int_0^\infty d\tau \langle \mathcal{L}_s e^{-\tau\mathcal{L}_f/\epsilon} \mathcal{Q}\mathcal{L}_s \rangle_{\text{ad}} \bar{\rho}_t. \end{aligned}$$

After some further algebraic manipulations we arrive at a Fokker–Planck equation for the slow degrees of freedom:

$$\begin{aligned} \frac{\partial \bar{\rho}_t}{\partial t} &= - \sum_i \frac{\partial}{\partial x_i} D_i^{(1)}(x) \bar{\rho}_t(x) \\ &\quad + \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} D_{ij}^{(2)}(x) \bar{\rho}_t(x). \end{aligned} \quad (31)$$

The drift term is defined as:

$$\begin{aligned} D_i^{(1)}(x) &= \langle f_i \rangle_{\text{ad}}(x) + \sum_j \int_0^\infty d\tau \left\langle f_j(x, y) \right. \\ &\quad \left. \frac{\partial}{\partial x_j} \delta f_i(x, \eta[\tau/\epsilon, y; x]) \right\rangle_{\text{ad}}, \end{aligned} \quad (32)$$

where we use the notation

$$\delta f_i(x, y) = f_i(x, y) - \langle f_i \rangle_{\text{ad}}(x)$$

as an abbreviation for the fluctuations around the adiabatic equilibrium. The diffusion term is defined as:

$$D_{ij}^{(2)}(x) = \int_0^\infty d\tau \langle \delta f_i(x, \eta[\tau/\epsilon, y; x]) \delta f_j(x, y) \rangle_{\text{ad}}. \quad (33)$$

Example 2 As a simple example of the method we consider a circle map coupled to a fast evolving Lorenz system in the chaotic regime.

$$\begin{aligned}\dot{x}_1 &= x_2 + x_1 \left(\frac{\gamma + \alpha y_1}{\sqrt{x_1^2 + x_2^2}} - 1 \right) \\ \dot{x}_2 &= -x_1 + x_2 \left(\frac{\gamma + \alpha y_1}{\sqrt{x_1^2 + x_2^2}} - 1 \right) \\ \dot{y}_1 &= -\frac{3(y_1 - y_2)}{\epsilon} \\ \dot{y}_2 &= -\frac{y_1 y_3 + 26.5 y_1 - y_2}{\epsilon} \\ \dot{y}_3 &= \frac{y_1 y_2 - y_3}{\epsilon}.\end{aligned}\quad (34)$$

Changing to cylindrical variables for the slow degrees of freedom, their dynamics simplifies as

$$\begin{aligned}\dot{r} &= -\gamma(r - 1) + \alpha y_1 \\ \dot{\theta} &= 1.\end{aligned}\quad (35)$$

The derivation of the terms in (31) is now quite straightforward:

$$\begin{aligned}\langle f_r \rangle_{\text{ad}}(r, \theta) &= -\gamma(1 - r) \\ D_{rr}^{(1)}(r, \theta) &= D_{\theta\theta}^{(1)}(r, \theta) = D_{r\theta}^{(1)}(r, \theta) = 0 \\ \langle f_\theta \rangle_{\text{ad}}(r, \theta) &= 1 \\ D_{rr}^{(2)}(r, \theta) &= \alpha^2 \int_0^\infty d\tau \langle y_1(\tau) |_{y_1(0)=y_1} y_1 \rangle_{\text{ad}} \\ \delta f_r(r, \theta, y_1) &= y_1 \\ D_{\theta\theta}^{(2)}(r, \theta) &= D_{r\theta}^{(2)}(r, \theta) = 0 \\ \delta f_\theta(r, \theta, y_1) &= 0\end{aligned}$$

and the resulting approximative Langevin equation (4) for r is

$$\dot{r} = -\gamma(r - 1) + \zeta(t), \quad (36)$$

with ζ representing white noise: $\langle \zeta(t) \rangle = 0$ and $\langle \zeta(t_1) \zeta(t_2) \rangle = 2D_{rr}^{(2)} \delta(t_1 - t_2)$ (see e.g. [13,41,55] for details on how to relate a Langevin dynamics to a Fokker–Planck equation). The mixing time for y_1 is $\mathcal{O}(\epsilon)$ and this limits the effective support in the integral defining $D_{rr}^{(2)}$ so that $D_{rr}^{(2)} = \mathcal{O}(\alpha^2 \epsilon)$. From this we conclude that, in the limit $\epsilon \rightarrow 0$, $D_{rr}^{(2)} \rightarrow 0$ unless $\alpha \sim 1/\sqrt{\epsilon}$. This is in agreement with earlier remarks on the size of the back-coupling from the fast degrees of freedom to the slow dynamics.

The resulting dynamics is shown in Fig. 1. A numerical simulation with parameters $\gamma = 5$, $\alpha = 1$, and $\epsilon = 0.005$ is shown. From the trajectory of the fast dynamics we measure $D_{rr}^{(2)} = 0.082$. The Langevin equation (36) with white noise predicts

$$\langle (r(t) - 1)^2 \rangle = \frac{D_{rr}^{(2)}}{\gamma},$$

from the fluctuation-dissipation theorem. We measure $\langle (r(t) - 1)^2 \rangle = 0.076/\gamma$ from the simulation, which can be considered in good agreement with expectations with an error of order $\mathcal{O}(\epsilon)$.

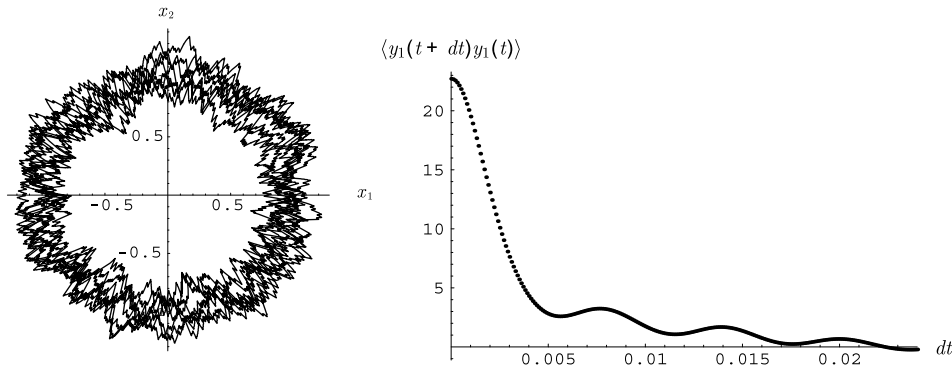
Separation of Time-Scales in Discrete Markov Chains

We now look at the discrete equivalent of separation of time scales discussed in Sect. “Temporal Hierarchies: Separation of Time Scales”. The model problem is a Markov process with a transition matrix that is approximately block-diagonal:

$$T = \epsilon \begin{pmatrix} \frac{1}{\epsilon^{(1)}} T^{(1)} & Q^{(12)} & \dots & Q^{(1n)} \\ Q^{(21)} & \frac{1}{\epsilon^{(2)}} T^{(2)} & \dots & Q^{(2n)} \\ \vdots & \vdots & \ddots & \vdots \\ Q^{(n1)} & Q^{(n2)} & \dots & \frac{1}{\epsilon^{(n)}} T^{(n)} \end{pmatrix}, \quad (37)$$

where the elements in the Q -matrices are of order unity or smaller and the ϵ s represent small numbers. Intuitively this system consists of n subsystems with internal dynamics $T^{(i)}$. On the time scale of $\mathcal{O}(1)$ the dynamics typically remains within one of these subsystems, but on a time scale of the order $\mathcal{O}(1/\epsilon)$ the systems switch from being in one subsystem to being in another. The slow dynamics therefore consists of n states whereas the fast dynamics have a varying number of states depending on the dimensionality of $T^{(i)}$. Note that the setup is analogous to the continuous case with $\mathcal{L} = \mathcal{L}_s + \epsilon^{-1} \mathcal{L}_f$, where the Liouville operator is viewed as a linear transition operator.

Normally, the transition matrix is not given on the near block diagonal form as in (37), but has the columns and rows mixed in random order just like in Sect. “Hierarchies in Markov Chains Through Aggregation of States”. Finding a permutation matrix that transforms the transition matrix to the near block diagonal form is however a much simpler problem than finding a hidden tensor decomposition. Effective algorithms for finding optimal permutations are based on the observation that a block diagonal matrix has a degeneracy of the stationary distribution corresponding to the number of blocks, i.e. there is a n -fold degeneracy of the eigenvalue 1 (the Perron roots).



Hierarchical Dynamics, Figure 1

To the *left*: The trajectory in the $x_1 - x_2$ -plane of the dynamical system defined in (34). The parameters used: $\alpha = 1$, $\epsilon = 0.005$ and $\gamma = 5$. To the *right*: the auto-correlation function $\langle y_1(t + dt)y_1(t) \rangle$ used to calculate $D_{rr}^{(2)}$

This also reflects the broken ergodicity for a Markov process with a block diagonal transition matrix. For a system where the transition matrix can be transformed into near block diagonal form, there is a set of eigenvalues close to 1, separated from the rest of the eigenvalues by a spectral gap. The aggregates of states building up the slow dynamics are identified by the approximately identical sign structure in the corresponding (right) eigenvectors. See [7] for details.

It is worth mentioning that Markov processes of the type given in (37) are often used as approximations of stochastic differential equations of Langevin type (4):

$$\ddot{x} = -\nabla U(x) - \gamma \dot{x} + \zeta(t).$$

If the free energy potential U has multiple minima, then the jumping between the basins around these minima can sometimes be described as a Markov process of the type in (37). The spectral gap $\sim 1/\epsilon$ is then depending on the height and width of the potential barriers between the local minima, as well as the temperature. In the crudest approximation one only considers the second derivatives at the minima to estimate the transition times. For more details see e. g. [5,13,19].

Computational Mechanics

We end the presentation with a pointer to a framework that is useful for complementing the hierarchical decomposition of discrete systems; computational mechanics. Computational mechanics [6,47,48] is a technique for deriving optimal predictors for stochastic processes. The predictors, called ϵ -machines, are automata whose nodes are equivalence classes, causal states, of observed histories of states. All the states in a causal state must have the same probability distribution of future observed states. An ϵ -machine is the minimal and maximally efficient

model of the observed process [48]. In practice an ϵ -machine can be acquired approximately from generated time series or other statistics [50]. In the context of hierarchical dynamics, the ϵ -machines can be used to find an optimal Markovian dynamics for a discrete system (with finite memory), on which we can apply the methods presented below to infer the hierarchical structure.

Invariants of the Motion

Invariants of the motion play a central role in the analysis of mechanical systems. The oldest roots in this tradition can be traced to the systematic study of continuous symmetries, advanced by Lagrange, Poisson, Jacobi, Lie, and Noether. The resulting reduction schemes eliminate inactive, i. e. constant, degrees of freedom. The most elegant product of this line of thought is Noethers' theorem, which directly connects continuous symmetries of the Hamiltonian to conserved quantities such as momentum, energy, angular momentum (spin) etc. Reduction of mechanical systems with symmetries is still a very active field, see for example the recent article [15] on reduction by stages in the Kepler problem. The literature in this field is voluminous. For a brief introduction, look in Arnold [1] (and also in Smale's article on topology in classical mechanics [52]). For modern reviews see Marsden et al. [31,32] and Marmo et al. [30]. In the current context it is interesting to note that invariants of the motion can be classified as both temporal and structural: a constant has an infinitely slow dynamics, and at the same time the dynamics is trivially self-contained (see Sect. "Structural Hierarchies: Foliations" for definition of structural hierarchies). Alternatively, one may argue that projections that eliminates, or projects onto, invariants of the motion do not qualify to define hierarchies since they merely reflect that the origi-

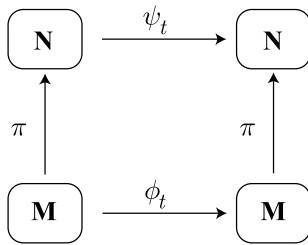
nal description of the system is “over-determined”. Which way we choose to interpret inactive degrees of freedom is not critical for the definition of hierarchies, so we leave it for the reader to decide for her or himself.

Structural Hierarchies: Foliations

Hierarchies in dynamical systems is a more general concept than dimensional reduction. To define a hierarchy in a more general setting, we focus on projections of the dynamics that constitutes a new “self-contained” (Markovian) dynamical system. The levels in the hierarchy do not necessarily evolve on different time scales. The general idea of autonomy naturally leads to the concept of preserved fibrations and foliations, as we define shortly.

Let a dynamical system be defined on a manifold M through a flow ϕ_t generated by a vector field \mathbf{v} . We define a new hierarchical level through a map π from M onto a lower dimensional manifold N . For N to define a new level of description we require π to induce a well defined flow on N , i. e., the differential $\pi_*(\mathbf{v})$ should be a well defined vector field. The hierarchical organization of dynamical systems is most naturally expressed in the language of fiber bundles. Roughly, a fiber bundle consist of a map $\pi: M \rightarrow N$ of a total space M onto a base space N such that all pre-image spaces $\pi^{-1}(x)$, where $x \in N$, are considered equivalent. The pre-image spaces $\pi^{-1}(x)$ are called fibers over a base point x . We say that a fibration π is preserved by a flow ϕ_t if fibers are carried over to fibers. In other words, if x and y belongs to the same fiber, i. e. $\pi(x) = \pi(y)$, then $\phi_t(x)$ and $\phi_t(y)$ also belong to the same fiber: $\pi(\phi_t(x)) = \pi(\phi_t(y))$. It should be clear that a preserved fibration also defines a new level in the dynamical hierarchy as defined in Fig. 2.

Foliation are geometric constructs, closely related to fibration. A foliation of an open subset of \mathbb{R}^n is a union of disjoint subsets called leaves. We say that a foliation is pre-



Hierarchical Dynamics, Figure 2

The manifold M is the original phase space, N is a lower dimensional phase space, and ϕ_t and ψ_t denotes flows on M respectively N . The projective map π describes a new level of description if the diagram commutes

served by a flow ϕ_t if leaves are carried over to leaves, i. e., if x and y belong to the same leaf L_1 then $\phi_t(x)$ and $\phi_t(y)$ also belong to the same leaf L_2 . Note that we do not require $L_1 = L_2$. If $L_1 = L_2$, then we say that the foliation is invariant under the flow. Further, if L_1 has the structure of a manifold, then it is an invariant manifold of the flow. If the set of leaves is taken as the base space then a foliation becomes a fibration, with the projection defined by collapsing all points on a leaf to the same point on N . Hierarchical dynamics can be expressed either in terms of preserved fibration or preserved foliations.

If we let \mathbf{v} denote the vector field that generates the flow ϕ_t , the criterion for π to be a preserved fibration under the flow generated by the vector field \mathbf{v} can be expressed as:

$$\pi(x) = \pi(y) \Rightarrow \pi_*(\mathbf{v}|_x) = \pi_*(\mathbf{v}|_y), \quad (38)$$

where $\pi_*: TM \rightarrow TN$ denotes the differential of the map $\pi: M \rightarrow N$. The foliations we are considering in this paper are such that the leaves L are immersed integral submanifolds of M . It then follows from Frobenius' theorem that the vector fields spanning the tangent space of the leaves TL form an involution [4], i. e., if $\mathbf{w}_k \in TL$ then

$$[\mathbf{w}_k, \mathbf{w}_l] = \sum_m g_{kl}^m(x) \mathbf{w}_m, \quad (39)$$

for some smooth functions g_{kl}^m . Expressed as a fibration, π is invariant under the translations along the vector fields in TL . The invariance of the fibration π is then expressed infinitesimally as

$$\mathbf{w}_k(\pi) = 0, \quad (40)$$

for all $\mathbf{w}_k \in TL$. The space of leaves, or the base space, $N(\pi: M \rightarrow N)$ carries the structure of a quotient manifold, $N = M/L$. We have the following central result:

Theorem 3 Let M be a manifold of dimension m . Consider a (singular) foliation F of M through a partition of M into connected immersed integral submanifolds (leaves). Let the tangent space of a leaf at a point x , $TL|_x$, be spanned by an involution of vector fields \mathbf{w}_k at x . The foliation is preserved under the flow generated by the vector field \mathbf{v} if and only if

$$[\mathbf{v}, \mathbf{w}_k] = \sum_l f_k^l(x) \mathbf{w}_l. \quad (41)$$

Proof The proof of this theorem is given e. g. in [35], see also [33,34] for a discussion on how to use this result for constructing geometric integrators. We sketch the proof idea as follows. For π to be a preserved fibration, we require

$$\pi(x) = \pi(y) \Rightarrow \pi(\exp(t\mathbf{v})x) = \pi(\exp(t\mathbf{v})y).$$

Now, if $\pi(x) = \pi(y)$ then $x = \exp(\mathbf{w})y$ for some $\mathbf{w} \in TL$. So

$$\begin{aligned}\pi(\exp(\mathbf{t}\mathbf{v})x) &= \pi(\exp(\mathbf{t}\mathbf{v})\exp(\mathbf{w})y) \\ &= \pi(\exp(\widetilde{\mathbf{w}})\exp(\mathbf{t}\mathbf{v})y) \\ &= \pi(\exp(\mathbf{t}\mathbf{v})y),\end{aligned}$$

for some vector field $\widetilde{\mathbf{w}} \in TL$, given explicitly by the Baker–Campbell–Hausdorff formula. \square

An important special case of Eq. (41) is $f_k^l(x) = 0$, i. e., when

$$[\mathbf{v}, \mathbf{w}_k] = 0 \quad \forall k. \quad (42)$$

The vector field \mathbf{w}_k is then a symmetry of the dynamics generated by \mathbf{v} (and vice versa), see [37] for details. In practice, it is much easier to search for solutions to Eq. (42) than Eq. (41). Equation (42) is a closed partial differential equation whereas Eq. (41) contains the unknown functions $f_k^l(x)$ which can be difficult to handle.

Quotient Manifold Projection

Given a set of vector fields that generates a preserved foliation, we want to construct the corresponding reduced dynamical system. The most straightforward approach is to find the projection π by solving Eq. (40). In local coordinates, if $\mathbf{w}_k = \sum_i \eta_k^i(x) \frac{\partial}{\partial x^i}$, Eq. (40) can be written as a set of quasi-linear first order partial differential equations:

$$\mathbf{w}_k(\pi) = 0, \quad (43)$$

or in local coordinates

$$\sum_j \eta_k^j(x) \frac{\partial \pi}{\partial x^j} = 0,$$

for all k and j . To find an explicit expression for π we need to recursively solve this system, using e.g. the method of characteristics.

There are technical complications with the quotient manifold construction presented above. Regularity is not guaranteed. The resulting quotient manifold may not even be Hausdorff. However, if we assume that the involution \mathbf{w}_i form regular submanifolds, it follows that the vector fields can be defined so that the functions $f_k^l(x)$ in (41) and g_{kl}^m in (39) are independent of x and \mathbf{w}_k form a Lie algebra [35]. The corresponding Lie group has a regular action on M . In this case the quotient manifold M/G is smooth and well defined.

Example 3 (Linear dynamics projected onto the real projective plane.) In general, any linear system $\dot{x} = Ax$ has

two trivial symmetries: $\mathbf{w}_1 = \sum_{ij} A_{ij} x_i \frac{\partial}{\partial x_j}$ and $\mathbf{w}_2 = \sum_i x_i \frac{\partial}{\partial x_i}$. The first symmetry is just the dynamics itself and the corresponding projection maps onto an invariant of the motion. The latter symmetry comes from the trivial observation that the identity matrix commutes with A , but it actually gives a non-trivial foliation. According to Eq. (43), the projection must fulfill

$$\sum_j x_j \frac{\partial \pi_i(x)}{\partial x_j} = 0,$$

for all components i . The general solution reads

$$\pi_i(x) = F_i\left(\frac{x_i}{x_j}\right),$$

for some arbitrary coordinate x_j and general functions F_i . Note that $\pi_j(x)$ is a constant. This reflects the reduction of dimensionality by the projective map. Just as in the case of the circle there is no single projective map valid over the entire phase space. For different choices of coordinate x_β , π provides “coordinate charts” valid in regions where $x_j \neq 0$. The resulting manifold can be identified as the real projective plane, $P\mathbb{R}^{d-1}$, if the original dynamics was in \mathbb{R}^d .

The circle is a special case, $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ where it is natural to choose

$$\pi(x, y) = \arctan(x/y),$$

in the coordinate chart where $y \neq 0$, and

$$\pi(x, y) = \pi/2 - \arctan(y/x),$$

in the coordinate chart where $x \neq 0$. The projection is onto the angular coordinate in the cylindric coordinate system. Note that $P\mathbb{R}^1 \simeq S^1$.

The other projections of a linear system follows directly from the Jordan form of A that fully resolves all invariant subspaces of the dynamics. Each invariant subspace can then be projected out to form a new level in the hierarchy. A set of non-redundant eigenvalues λ_i can be projected immediately using

$$\pi(x) = \left(\prod_i (A - \lambda_i \mathbf{I}) \right) x, \quad (44)$$

or the more abstract formula in Eq. (18). Note that Eq. (44) is not strictly a projection since $P^2 \neq P$ while Eq. (18) is a projection. If it is important, this can be fixed by using a pseudo-inverse, see Sect. “Hierarchies in Markov Chains

Through Aggregation of States". Equation (44) is perhaps the most natural projection of a linear system. The special case in the example above is focused on projecting onto one degree of freedom corresponding to one half of a complex conjugate pair. This is the reason why the topology of the resulting reduced manifold becomes non-trivial.

Example 4 (Skew-product systems) Consider a dynamical system with the following nonlinear skew-product structure:

$$\begin{aligned}\dot{x} &= f(x) \\ \dot{y} &= g(x, y),\end{aligned}\quad (45)$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^{m-n}$. The family of vector fields $\mathbf{w}_1 = \frac{\partial}{\partial y_1}, \dots, \mathbf{w}_{m-n} = \frac{\partial}{\partial y_{m-n}}$ forms an involution. The corresponding foliation $\pi(x, y) = F(x)$ for any function F is in accordance with $\dot{x} = f(x)$ being a self-contained subsystem. Note that this result is also consistent with the discussion in Sect. "White Noise Reduction: Mori-Zwanzig Projections" since if $f(x, y)$ in Eq. (20) is independent of y it follows that $\delta f(x, y) = 0$ and therefore $D^{(1)} = D^{(2)} = 0$.

Non-autonomous Dynamical Systems

It is straight forward to generalize the foliation framework to non-autonomous dynamical systems:

$$\dot{x} = f(t, x). \quad (46)$$

To analyze Eq. (46) we need to find an involution of vector fields \mathbf{v}_k that can generate the dynamics, i. e.

$$\begin{aligned}f(t, x) &= \text{span}\{\mathbf{v}_i\} \quad \forall t \\ [\mathbf{v}_k, \mathbf{v}_l] &= \sum_m g_{kl}^m(x) \mathbf{v}_m.\end{aligned}$$

Not that if f is free to take any form, then \mathbf{v}_i must span the entire tangent space. The condition (41) is generalized to:

$$[\mathbf{v}_k, \mathbf{w}_l] = \sum_m f_{kl}^m(x) \mathbf{w}_m \quad \forall k. \quad (47)$$

Hierarchies in Discrete Dynamical System Through Normal Subgroup Extensions

In this section we present the analogy of non-autonomous foliation (Sect. "Non-autonomous Dynamical Systems") for discrete systems. Let H be a semigroup with a finite number of generators acting on a finite state space Σ . The semigroup H together with the state space Σ define a dynamical system. We define a new finite group N , also acting on Σ . Further, we define the joint group $F = \text{span}(H, N)$ generated by the elements in both H and

N : $f = \prod_i f_i$ where $f_i \in H \cup N$. The action ψ of F on Σ is well defined through the action of H and N individually. Two elements f_1 and f_2 are identical if and only if $\psi(f_1, \sigma) = \psi(f_2, \sigma)$, $\forall \sigma \in \Sigma$. Note that both H and N are subgroups of F .

Assume that N has been carefully chosen so that, for each $h \in H$, $\sigma \in \Sigma$, and $n \in N$, there exist a $n' \in N$ such that

$$\psi(h * n, \sigma) = \psi(n' * h, \sigma). \quad (48)$$

Note that the element n' may depend on the state σ , i. e. $\psi(h * n, \sigma_1) = \psi(n' * h, \sigma_1)$ and $\psi(h * n, \sigma_2) = \psi(n'' * h, \sigma_2)$ do not generally imply $n' = n''$. Relation (48) ensures that N is a normal subgroup of F (for each fixed $\sigma \in \Sigma$). Furthermore, the commutation relation in Eq. (48) ensures that, for each $\sigma \in \Sigma$, the elements in the product $f = \prod_i f_i$ can be re-arranged so that $f = \prod_i n_i \prod_j h_j$, i. e. for each $f \in F$ there exist $n \in N$ and $h \in H$ such that $f^{(\sigma)} = n^{(\sigma)} * h^{(\sigma)}$ (the superscript indicates that the decomposition may be different for different σ). It then follows that (see Sect. "Some Concepts from Group Theory")

$$F^{(\sigma)} \simeq N \rtimes_{\Psi^{(\sigma)}} H, \quad (49)$$

with the automorphism mapping $\Psi_h^{(\sigma)}(n) = n'$, where $\psi(h * n, \sigma) = \psi(n' * h, \sigma)$.

Equations (48) and (49) are the equivalent of the ideal relation for vector fields used in Eq. (41) or more generally Eq. (47). It is therefore the central relation in the reduction of discrete dynamical systems. Note that the definition of the automorphism map $\Psi^{(\sigma)}$ plays the same role as the commutator of the generating vector fields $\exp(t\mathbf{v}) \exp(\mathbf{w}) = \exp(\widetilde{\mathbf{w}}) \exp(t\mathbf{v})$ in the proof of reduction in continuous systems using preserved foliations. In the continuous case, the solution to the equation $\psi(h * n, \sigma) = \psi(n' * h, \sigma)$ is given explicitly by the Baker–Campbell–Hausdorff formula. In the discrete case there is no such explicit approach. However, as we shall see there is no need for deriving the explicit expression for $\Psi^{(\sigma)}$.

To form the quotient projection in the discrete case, we form equivalence classes on Σ : $\sigma_1 \sim \sigma_2$ if there exist a $n \in N$ such that $\sigma_1 = \psi(n, \sigma_2)$, i. e. the orbits of N form the equivalence classes. The equivalence classes are preserved under the action of H . To see this, let $\sigma_1 = \psi(n, \sigma_2)$. Then when H acts on σ_1 and σ_2 we have

$$\psi(h, \sigma_1) = \psi(h * n, \sigma_2) = \psi(n' * h, \sigma_2) \sim \psi(h, \sigma_2)$$

for some $n' \in N$. Thus $\psi(h, \sigma_1) \sim \psi(h, \sigma_2)$, so the equivalence classes are preserved. This implies that the action

of H is well defined on the set of equivalence classes, i. e., the quotient set Σ/N . We say that the resulting dynamical system, i. e. H acting on Σ/N , is reduced.

The algebraic structure presented in this section makes a clear connection to reduction through foliations for continuous dynamical systems. Moreover, since any automaton can be represented as a semigroup (see Sect. “Some Concepts from Group Theory”), it also shows how to make hierarchical decompositions of automata. The algebraic structure is also connected to Wreath products and Krohn–Rhodes theory for finite automata [9,24]. Furthermore, automata theory can also be applied to Markov processes. The Markov process must then be decomposed into a Bernoulli shift, combined with a set of deterministic transition matrices [27]. Intuitively this means that at each time step the system is updated using a randomly chosen deterministic transition matrix. The semigroup H is composed of all the deterministic transition matrices that generate the Markov process. However, for a Markov process the approach introduced in this section is unnecessarily complicated. In the next section we discuss a more useful technique.

Hierarchies in Markov Chains Through Aggregation of States

Consider a stochastic Markov process with transition matrix T and state space consisting of symbols in an alphabet Σ . If T can be decomposed as a tensor, or Kronecker, product

$$T = T^{(1)} \otimes T^{(2)} \otimes \dots \otimes T^{(N)}, \quad (50)$$

then $T^{(i)}$ are transition matrices for independent processes. Hence, a projection onto a state space representing one, or many, subsystems results in a new Markov process over a state space with reduced cardinality. Let $T^{(2:N)} = T^{(2)} \otimes \dots \otimes T^{(N)}$ and $T_{i,j}^{(1)}$ denote the elements $T^{(1)}$. From the definition of a tensor product it then follows that

$$T = \begin{pmatrix} T_{1,1}^{(1)} T^{(2:N)} & T_{1,2}^{(1)} T^{(2:N)} & \dots & T_{1,K}^{(1)} T^{(2:N)} \\ T_{2,1}^{(1)} T^{(2:N)} & T_{2,2}^{(1)} T^{(2:N)} & \dots & T_{2,K}^{(1)} T^{(2:N)} \\ \vdots & \vdots & \ddots & \vdots \\ T_{K,1}^{(1)} T^{(2:N)} & T_{K,2}^{(1)} T^{(2:N)} & \dots & T_{K,K}^{(1)} T^{(2:N)} \end{pmatrix},$$

where $T_{ij}^{(1)} T^{(2:N)}$ represents sub-matrices in T . Since $T^{(1)}$ is a transition matrix, its column sum is normalized. It follows that

$$T^{(2:N)} = \sum_{j=1}^K T_{i,j}^{(1)} T^{(2:N)} \quad i = 1, \dots, K,$$

or explicitly

$$T_{k,l}^{(2:N)} = \sum_{j=0}^{K-1} T_{k+iK,l+jK} \quad i = 1, \dots, K. \quad (51)$$

On the one hand, under the assumption that T can be decomposed as in the ansatz (50), then Eq. (51) gives a Markov process on the reduced state space. On the other hand, Eq. (51) also tests the ansatz since, for each element $T_{k,l}^{(2:N)}$ the right hand side should evaluate identically for all choices of the index i . The last observation can be used to identify hierarchies in Markov processes.

When a subsystem $T^{(1)}$ is projected out according to Eq. (51), the corresponding reduction on the state space can, as usual, be defined in terms of equivalence classes. Assume that the state vector is ordered in the default manner, so that the element in position i represents the probability that the system is in a state denoted by $\sigma_i \in \Sigma$. The reduction described by (51) is then equivalent to forming equivalence classes according to $\sigma_i \sim \sigma_j$ iff $(i - j) \bmod K = 0$. It is clear that the composition of the equivalence classes depends on the ordering in the state vector. As a result, to find a tensor decomposition of the transition matrix presented with the rows and columns in random ordering, the validity of Eq. (51) should be tested for all permutations (excluding those not affecting the equivalence classes). Formally we express this as:

Theorem 4 *An $N \times N$ transition matrix T can be decomposed as a tensor product of an $K \times K$ matrix and an $N/K \times N/K$ matrix if and only if there exists a $P \in \text{Sym}(\Sigma)$ such that*

$$\sum_{j=0}^{K-1} (P^T T P)_{k+iK,l+jK} \quad (52)$$

gives the same result independent of the index i .

A major problem with this result is that the cardinality of the symmetric group increase extremely fast with the number of states in the state space (actually $|\text{Sym}(\Sigma)| = |\Sigma|!$). The practical usefulness of the summation condition is of limited. Below we present a criterion that is more useful for identify projections onto a Markov process with reduced state space. Before this however, we generalize from direct (Kronecker) product to semi-direct products.

The decomposition in (51) is not the most general form permitting projections onto well-defined Markov subsystems. In (51) the subsystems are completely decoupled. In a more general case it may happen that a subsystem can influence the dynamics of another but not vice

versa. In this situation it is still possible to project onto the first subsystem, but not onto the second. This situation is equivalent with the one presented in Eq. (45) for the continuous case. In the discrete setting a decomposition of this type uses a semi-direct product of submatrices. It is straightforward to define consistency equations on the form in Eq. (51) in this more general case. The size of the different equivalence classes in the partitioning of the state space may vary, and this needs to be taken into account in the summation. We start by introducing some notation.

We consider a Markov chain X_t , $t = 0, 1, \dots$, with a finite state space $\Sigma = \{1, \dots, N\}$ and transition probability matrix $P = [p_{ij}]$. The transition matrix operates from the right $x_{t+1} = x_t P$. A reduction of the state space is a lumping, or state aggregation, by a partition of the state space $\tilde{\Sigma} = \{L_1, \dots, L_M\}$ where L_k is a nonempty subset of Σ , $L_k \cap L_l = \emptyset$ if $k \neq l$ and $\bigcup_k L_k = \Sigma$. Clearly $M \leq N$. The reduction can be defined by an $N \times M$ matrix $\Pi = [\pi_{ik}]$ where $\pi_{ik} = 1$ if $i \in L_k$ and $\pi_{ik} = 0$ if $i \notin L_k$. We call the reduction, or the aggregation, a Lumping $\tilde{\Sigma}$. A lumping induces a quotient process $\tilde{x}_t = x_t \Pi$. If the process \tilde{x}_t is a Markov process (which is, as always, not usually the case), then we say that the Markov chain is *strongly lumpable* with respect to Π (or $\tilde{\Sigma}$). The following criterion is a generalization of the summation condition above and is necessary and sufficient for a Markov chain with transition matrix P to be lumpable with respect to Π or $\tilde{\Sigma}$ [23,26]:

1. $\Pi \tilde{P} = P \Pi$, where $\tilde{P} = \Pi^+ P \Pi$ is the transition matrix on the reduced phase space ($\Pi^+ \doteq (\Pi^T \Pi)^{-1} \Pi^T$ is the left pseudo-inverse of Π , $\Pi^+ \Pi = I$).
2. $\ker(\Pi)$ is P -invariant, i.e. $y \Pi = 0 \Rightarrow y P \Pi = 0$.
3. For any $L_k, L_l \in \tilde{\Sigma}$, the total probability of going from any state $i \in L_k$ to L_l , i.e. $\sum_{j \in L_l} p_{ij}$, is independent of i .

Note that condition 2 is equivalent to Eq. (43) and condition 1 is related to Eq. (18). If the Markov chain is lumpable, then the probability distribution over the reduced state space is updated according to $\tilde{x}_{t+1} = \tilde{x}_t \tilde{P}$, where \tilde{P} is defined in Criterion 1. Furthermore, the transition matrix for the reduced Markov chain is given by

$$\tilde{p}_{kl} = \sum_{j \in L_l} p_{ij} \quad i \in L_k, \quad (53)$$

where we note that $\tilde{P} = [\tilde{p}_{kl}]$ is well defined since the sum is independent of which representative $i \in L_k$ we chose according to criterion 3.

As mentioned in connection with the summation criterion in Eq. (52), criteria 1–3 not immediately useful

for identifying lumpings of a Markov chain. Barr and Thomas [2] presented a necessary lumpability criterion involving the left eigenvectors of the transition matrix. It was first noted that the spectrum of \tilde{P} must be a subset of the spectrum of P (this is also discussed in detail in a more general setting in [26]). It also follows that if $vP = \lambda v$ then $\tilde{v}\tilde{P} = \lambda \tilde{v}$, with $\tilde{v} = v\Pi$. It follows that if λ is an eigenvalue of both P and \tilde{P} , then \tilde{v} is an eigenvector of \tilde{P} , but if λ is not an eigenvalue of \tilde{P} then $\tilde{v} = v\Pi = 0$. Intuitively this observation can be understood as Π eliminating a subset of the eigenvectors of P . This is also clear from criterion 1 and 2, as well as from our previous discussion on linear systems in general. Equation (44).

Barr and Thomas' result suggests a search for lumpings defined by Π such that $v^\alpha \Pi = 0$ for some subset of the left eigenvectors of P , $\{v^\alpha\}_{\alpha \in J}$, $J \subseteq \Sigma$. Since Π should be a matrix with zeros and ones, $v^\alpha \Pi = 0$ essentially means searching for eigenvectors with subsets of elements that sums to zero. For lumpings only involving agglomeration of two states this is straightforward since the eigenvector(s) must have pairs of elements $v_i^\alpha = -v_j^\alpha$. However, agglomeration of k states means searching for partial sums evaluating to zero and involving k elements. This leads back to the combinatorial explosion of possibilities discussed before. As Barr and Thomas point out, there is no obvious algorithm to generate Π based on their result.

In a recent study a more useful method for identifying possible partitions of the state space leading to a reduced Markov process is presented [36]. The key is an observation that the dual of the probability space can be used to identify lumpings. The criterion $v^\alpha \Pi = 0$ is viewed as an orthogonality condition between the left eigenvectors $\{v^\alpha\}_{\alpha \in J}$ and the column space of Π . The orthogonal complement of a set of left eigenvectors is spanned by complementary right eigenvectors (defined naturally in the dual vector space). These complementary eigenvectors span the column space of Π . Requiring that Π consists of zeros and ones does corresponds to a criterion of repeated elements within each complementary right eigenvector. Clearly, identifying repeated elements in the right eigenvectors is algorithmically straight forward. The precise result reads as follows (proof is to be published [36]):

Theorem 5 Assume that P is a diagonalizable transition matrix with full rank describing a Markov process $x_{t+1} = x_t P$. Consider a set of linearly independent right eigenvectors of P , $P u^\alpha = \lambda^\alpha u^\alpha$. Let $I \subseteq \Sigma$ be the set of indices for the eigenvectors. Form state equivalence classes defined by states with identical elements in all eigenvectors u^α , i.e. $i \sim j$ iff $u_i^\alpha = u_j^\alpha \forall \alpha \in I$. The equivalence classes define a partitioning $\tilde{\Sigma}$ of the state space. This partitioning

is a lumping of the Markov chain if the number of partition elements equals the number of eigenvectors, i. e. $|\tilde{\Sigma}| = |I|$.

Conversely, if $\tilde{\Sigma}$ is a lumping then there exist $|\tilde{\Sigma}|$ linearly independent right eigenvectors that are invariant under permutations within the lumps.

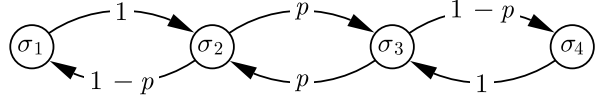
There are two important points to make in connection to this result. First, the result can be viewed as an extension of the result in Sect. “Separation of Time-Scales in Discrete Markov Chains”. A block diagonal matrix can be constructed as a semi-direct product of the identity matrix and a set of matrices appearing in the blocks. The second point is that the criterion in general provides a relatively efficient method for generating projections of a given Markov chain. The most time consuming part of the algorithm is to diagonalize the transition matrix. This problem scales approximately cubic in the number of states, i. e. not exponential like the naive approaches. However, there are several subtleties to take into consideration. Degenerate eigenvalues can cause difficulties since the corresponding eigenvectors are in this situation no longer uniquely defined. The criterion is still valid but one must chose which linear combination of eigenvectors to use. The partitions suggested by the eigenvectors can also be nested to produce an exponential number of possible reductions. This problem is intrinsic, no algorithm can find all possible reductions in linear time since there might be an exponential number of possible projections. The identity matrix is a good example where any projection is clearly an acceptable lumping.

Example 5 (Kronecker product system) As a simple example we use the method on a process with transition matrix $T_{\mathcal{A}}$ described in Fig. 3. If we sort the state vector according to the default ordering: $\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$, then the corresponding transition matrix reads

$$T_{\mathcal{A}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1-p & 0 & p & 0 \\ 0 & p & 0 & 1-p \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

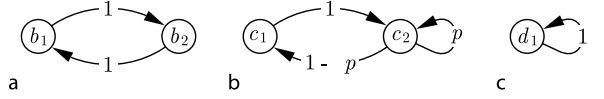
As it stands, $T_{\mathcal{A}}$ is not directly decomposable but if we change the ordering of the states to $\{\sigma_1, \sigma_4, \sigma_3, \sigma_2\}$, we end up with a decomposable transition matrix:

$$\begin{aligned} P^T T_{\mathcal{A}} P &= \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1-p & 0 & p \\ 1-p & 0 & p & 0 \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}}_{T_C} \otimes \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{T_B}, \end{aligned}$$



Hierarchical Dynamics, Figure 3

Example process $T_{\mathcal{A}}$ over the state space $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$. The edges are labeled with transition probabilities



Hierarchical Dynamics, Figure 4

Dynamics (a) T_B , (b) T_C and (c) T_D (a trivial process with one state) resulting from projections of the example process $T_{\mathcal{A}}$ in Fig. 3. The states are defined in terms of equivalence classes as follows: $b_1 = \{\sigma_1, \sigma_3\}$, $b_2 = \{\sigma_2, \sigma_4\}$, $c_1 = \{\sigma_1, \sigma_4\}$, $c_2 = \{\sigma_2, \sigma_3\}$, $d_1 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$

where P is the permutation matrix providing the re-ordering: $\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\} \rightarrow \{\sigma_1, \sigma_4, \sigma_3, \sigma_2\}$. From this it follows that the process $T_{\mathcal{A}}$ can be projected onto T_B , T_C , or the trivial process with one state T_D , all described in Fig. 4.

The eigenvalues of $P_{\mathcal{A}}$ are $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = 1 - p$ and $\lambda_4 = p - 1$. The corresponding right eigenvectors are $u_1 = (1, 1, 1, 1)^T$, $u_2 = (-1, 1, -1, 1)^T$, $u_3 = (-1, p-1, 1-p, 1)^T$, and $u_4 = (1, p-1, p-1, 1)^T$. By the condition on repeated elements, the two pairs of eigenvectors $\{u_1, u_2\}$ and $\{u_1, u_4\}$ implies the two possible lumpings found above: $b_1 = \{\sigma_1, \sigma_3\}$, $b_2 = \{\sigma_2, \sigma_4\}$, $c_1 = \{\sigma_1, \sigma_4\}$, $c_2 = \{\sigma_2, \sigma_3\}$, $d_1 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$.

Example 6 (Semi-direct product system) Consider the transition matrix

$$P = \begin{pmatrix} a+b+(c-1)/2 & 1-a-b & (1-c)/2 \\ -a+(c+1)/2 & a & (1-c)/2 \\ 1-b-c & b & c \end{pmatrix},$$

with $0 \leq a, b, c \leq 1$. P has the eigenvalues $\lambda^1 = 1$, $\lambda^2 = 2a + b - 1$ and $\lambda^3 = (3c - 1)/2$, and the eigenvectors

$$\begin{aligned} u^1 &= (1, 1, 1)^T, \\ u^2 &= (1 + c - 2a - 2b, 2(a - c), 2b + c - 1)^T \text{ and} \\ u^3 &= (-1, -1, 2)^T. \end{aligned}$$

There are three possible lumpings of P :

$$\begin{aligned} \tilde{\Sigma}_1 &= \{\{1, 2, 3\}\} && \text{from } \{u^1\}, \\ \tilde{\Sigma}_2 &= \{\{1, 2\}, \{3\}\} && \text{from } \{u^1, u^3\} \text{ and} \\ \tilde{\Sigma}_3 &= \{\{1\}, \{2\}, \{3\}\} && \text{from } \{u^1, u^2, u^3\}, \end{aligned}$$

where $\tilde{\Sigma}_2$ is valid if $2a + b - 1 \neq 0$ and $3c - 1 \neq 0$.

Conclusion

Analysis of hierarchical dynamics contains two separate steps. Initially the states, or degrees of freedom, that should be eliminated in each hierarchical transition need to be identified. After the identification, the reduced dynamics can be derived. The first step is by far the most computationally expensive. This is especially transparent in the analysis of decomposition of transition matrices for Markov processes (see Sect. “[Hierarchies in Markov Chains Through Aggregation of States](#)”), but the conclusion holds in general. In the discussion on temporal hierarchies (Sect. “[Temporal Hierarchies: Separation of Time Scales](#)”), the problem of separating the slow and the fast degrees of freedom was tactically assumed to be given a priori (with the exception of inertial manifolds where numerical algorithms do exist [10,17,39]), and the nearly decomposable Markov chains discussed in Sect. “[Separation of Time-Scales in Discrete Markov Chains](#)”, see also [7]. In the section on structural hierarchies (Sect. “[Structural Hierarchies: Foliations](#)”), the conditions on the projections are clearly stated. However the practical applicability of the methods presented depends critically on efficient algorithms for finding the projections. Naive approaches typically lead to algorithms that are exponentially slower than solving the original system. There are in fact reasons to believe that the Markov chains are typical. As we have seen, naive approaches to finding reductions lead to exponentially slow algorithms. There are more efficient methods that work most of the time, like to eigenvector criterion presented in Sect. “[Hierarchies in Markov Chains Through Aggregation of States](#)”, but in worst case scenarios these methods also become exponentially slow. In conclusion, methods for deriving appropriate projections for reduction is an area that is in need of more extensive study.

Future Directions

In the physics community, one of the main areas where hierarchical dynamics is discussed is renormalization of systems with critical behavior. The most prominent example is the Ising model in two or three dimensions [16]. The main idea in renormalization theory is that the state of the system, drawn from an ensemble of possible states with probability defined by the Hamiltonian, is structurally self-similar under the projection operator. Clearly, since the projection in general reduces the dimensionality of the phase space, self-similarity can only be well defined if the dimensionality of the system is infinite. For certain parameter values (often the temperature is the parameter), the dynamics is not only structurally self-similar but

the interaction strength between the agglomerated states does not decay, i. e. the projection operator has a non-trivial fixed point on the space of Hamiltonians (or rather a parametrization of the Hamiltonians). In this situation the system is said to be in a critical state, characterized by self-similarity and non-vanishing fluctuations on all length scales. It seems obvious that renormalization theory and hierarchical dynamics are closely connected. In fact it is relatively straight forward to formulate the renormalization group in the framework of structural hierarchies as presented in Sect. “[Structural Hierarchies: Foliations](#)”. A detailed investigation of the relation remains to be done. Furthermore, the more interesting aspect of this connection would be to explore possible generalizations where the constraint of self-similarity is relaxed and we focus on non-vanishing interactions with different structure on different levels in the hierarchy. The result of such an effort should, if successful, be of central interest in complex systems, especially for studying emergence (the author would like to acknowledge private discussions with Nils Baas on this subject).

The remarkable ingenuity in the renormalization group is the step of abstraction when the projection is viewed as generator of a discrete dynamical system on the space of Hamiltonians. This leap enables us to understand the origin of universality, as the details of the physical system become unimportant for the analysis of the behavior (the fixed points) of the dynamical system generated by the projection. It is exciting to speculate about the possibility of formulating similar “meta-models” for the dynamics generated by the projections used to define general hierarchical dynamics.

Finally, one of the central problems with defining emergence in complex dynamical systems is the lack of a framework for describing creation and destruction of objects. On the basic level, the number of degrees of freedom in a dynamical system does not change during the time evolution. From this perspective one may argue that dynamical systems are not suited for describing systems consisting of entities that can be created, destroyed and change internal properties [12]. These features are central in many complex systems, especially in biological and social systems. Methods used to analyze hierarchical dynamics can possibly be used to define the emergence of objects and organizations in dynamical systems. Near decomposability into independent subsystems (as defined in Sect. “[Hierarchies in Markov Chains Through Aggregation of States](#)”) can for example be used to define objects, see [20] for an alternative definition. Emergence, creation, destruction and change can be understood as different regions of the phase space permitting different

hierarchical structures. A “meta-model”, as described in the previous paragraph, could possibly be used to understand emergent higher-order organization in dynamical systems.

Acknowledgment

The author would like to thank Olof Görnerup and Kolbjørn Tunstrøm for discussions and comments on the manuscript. The author would also like to acknowledge support from the EU integrated project FP6-IST-FET PACE, by EMBIO, a European Project in the EU FP6 NEST initiative, and by MORPHEX, a European Project in the EU FP6 NEST initiative.

Bibliography

Primary Literature

- Arnold VI (1989) *Mathematical Methods of Classical Mechanics*, Graduate Texts in Mathematics, 2nd edn. Springer, New York
- Barr DR, Thomas MU (1977) An eigenvector condition for Markov chain lumpability. *Oper Res* 25(6):1028–1031
- Beck C, Schlögl F (1993) *Thermodynamics of chaotic systems*, Cambridge Nonlinear Science Series, vol 4. Cambridge University Press, Cambridge
- Boothby WM (2002) An introduction to differentiable manifolds and Riemannian geometry, *Pure and Applied Mathematics*, vol 120, 2nd edn. Academic Press
- Caroli B, Caroli C, Roulet B (1979) Diffusion in a bistable potential: a systematic wkb treatment. *J Stat Phys* 21:415–536
- Crutchfield J, Young K (1989) Inferring statistical complexity. *Phys Rev Lett* 63:105
- Deuffhard P, Huisinga W, Fischer A, Schütte C (2000) Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Linear Algebra Appl* 315:39–59
- Dorfman JR (1998) *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*, Cambridge Lecture Notes in Physics, vol 14. Cambridge University Press, Cambridge
- Egri-Nagy A (2005) Algebraic hierarchical decompositions of finite state automata a computational approach. Ph D thesis, University of Hertfordshire
- Foias C, Jolly MS, Kevrekidis IG, Sell GR, Titi ES (1988) On the computation of inertial manifolds. *Phys Lett A* 131(7–8): 433–436
- Foias C, Sell GR, Temam R (1988) Inertial manifolds for non-linear evolutionary equations. *J Differ Equ* 73:309–353
- Fontana W, Buss LW (1996) The barrier of objects: From dynamical systems to bounded organizations. In: Casti J, Karlqvist A (eds) *Barriers and Boundaries*. Addison-Wesley, Reading, pp 56–116
- Gardiner C (2004) *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer Series in Synergetics, vol 13, 3rd edn. Springer, Berlin
- Givon D, Kupferman R, Stuart A (2004) Extracting macroscopic dynamics: model problems and algorithms. *Nonlinear-ity* 17:55–127
- Godfreyt SE, Princet GE (1991) A canonical reduction of order for the kepler problem. *J Phys A: Math Theor* 24: 5465–5475
- Goldenfeld N (1992) *Lectures on Phase Transitions and the Renormalization Group*. Perseus Books, Oxford
- Gorban A, Karlin I (2005) *Invariant Manifolds for Physical and Chemical Kinetics*. Lecture Notes in Physics. Springer, Berlin
- Haken H (1983) *Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*, 3rd edn. Springer, New York
- Hänggi P, Talkner P, Borkovec M (1990) Reaction-rate theory: fifty years after kramers. *Rev Mod Phys* 62(2):251–341
- Jost J, Bertschinger N, Olbrich E, Aya N, Frankela S (2007) An information theoretic approach to system differentiation on the basis of statistical dependencies between subsystems. *Physica A* 378(1):1–10
- Just W, Kantz H, Rödenbeck C, Helm M (2001) Stochastic modelling: Replacing fast degrees of freedom by noise. *J Phys A: Math Theor* 34:3199–3213
- Just W, Gelfert K, Baba N, Riebert A, Kantz H (2003) Elimination of fast chaotic degrees of freedom: On the accuracy of the born approximation. *J Stat Phys* 112:277–292
- Kemeny JG, Snell JL (1976) *Finite Markov Chains*, 2nd edn. Springer, New York
- Krohn K, Rhodes J (1965) Algebraic theory of machines. i. prime decomposition theorem for finite semigroups and machines. *Trans Am Math Soc* 116:450–464
- Lichtenberg AJ, Lieberman MA (1983) *Regular and Stochastic Motion*, Applied Mathematical Sciences, vol 38. Springer, New York
- Lorch E (1962) *Spectral theory*. Oxford University Press, New York
- Maler O (1995) A decomposition theorem for probabilistic transition systems. *Theor Comput Sci* 145(1–2):391–396
- Mallet-Paret J, Sell GS (1988) Inertial manifolds for reaction diffusion equations in higher space dimensions. *J Am Math Soc* 1(4):805–866
- Mane R (1977) Reduction of semilinear parabolic equations of finite dimensional c^1 flows. In: *Geometry and Topology*, no. 597 in *Lecture Notes in Mathematics*. Springer, New York, pp 361–378
- Marmo G, Saletan EJ, Simoni A, Vitale B (1985) *Dynamical Systems: a differential geometric approach to symmetry and reduction*. Wiley, New York
- Marsden JE, Ratiu TS (2002) *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*, Texts in Applied Mathematics, 2nd edn. Springer, New York
- Marsden J, Misiolek G, Ortega JP, Perlmutter M, Ratiu T (2007) *Hamiltonian Reduction by Stages*. Lecture Notes in Mathematics. Springer, New York
- McLachlan RI, Perlmutter M, Quispel GRW (1998) Numerical integrators that preserve symmetries and reversing symmetries. *SIAM J Numer Anal* 35(2):586–599
- McLachlan RI, Perlmutter M, Quispel GRW (2003) Lie group foliations: dynamical systems and integrators. *Futur Gener Comput Syst* 19(7):1207–1219
- Molino P (1988) *Riemannian Foliations*. Birkhäuser, Boston
- Nilsson Jacobi M, Görnerup O A dual eigenvector condition on lumpability in markov chains. To be published
- Olver P (2000) Applications of Lie Groups to Differential Equa-

tions, Graduate Texts in Mathematics, 2nd edn. Springer, New York

38. Packard N, Crutchfield J, Farmer D, Shaw R (1980) Geometry from a time series. *Phys Rev Lett* 45:712–716
39. Rega G, Troger H (2005) Dimension reduction of dynamical systems: Methods, models. *Nonlinear Dyn* 41:1–15
40. Riebert A, Baba N, Gelfert K, Just W, Kantz H (2005) Hamiltonian chaos acts like a finite energy reservoir: Accuracy of the Fokker–Planck approximation. *Phys Rev Lett* 94:54–103
41. Risken H, Frank T (1996) *The Fokker–Planck Equation: Methods of Solutions and Applications*, 2nd edn. Springer Series in Synergetics. Springer, Berlin
42. Ruelle D (1989) Chaotic evolution and strange attractors. Cambridge University Press, Cambridge
43. Ruelle D (1999) Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. *J Stat Phys* 95(1–2):393–468
44. Sanders JA, Verhulst F (1985) *Averaging Methods in Nonlinear Dynamical Systems*, Applied Mathematical Sciences, vol 59. Springer, New York
45. Sauer T, Yorke JA, Casdagli M (1991) Embedology. *J Stat Phys* 65:579–616
46. Schulman LS (1996) *Techniques and applications of path integration*. John Wiley, New York
47. Shalizi C (2001) Causal architecture, complexity and self-organization in time series and cellular automata. Ph D thesis, University of Wisconsin
48. Shalizi C, Crutchfield J (2001) Computational mechanics: Pattern and prediction, structure and simplicity. *J Stat Phys* 104:816
49. Shalizi C, Moore C (2003) What is a macrostate? Subjective observations and objective dynamics. <http://arxiv.org/abs/cond-mat/0303625>
50. Shalizi C, Shalizi K (2004) Blind construction of optimal nonlinear recursive predictors for discrete sequences. In: AUA1 '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUA1 Press, Arlington, pp 504–511
51. Simon H (1962) The architecture of complexity. *Proc Am Philos Soc* 106(6):467–482
52. Smale S (1970) Topology and mechanics. *Invent Math* 10(4):305–331
53. Takens F (1981) Detecting strange attractors in turbulence. In: Rand D, Young L (eds) *Dynamical Systems and Turbulence*, Warwick 1980. Springer, Berlin, p 366
54. Zwanzig R (1960) Ensemble methods in the theory of irreversibility. *J Chem Phys* 33:1338
55. Zwanzig R (2001) *Nonequilibrium statistical mechanics*. Oxford University Press, New York

Books and Reviews

- Anderson PW (1972) More is Different. *Science* 177(4047):393–396
- Badii R, Politi A (1997) *Complexity: Hierarchical structures and scaling in physics*. Cambridge Nonlinear Science Series, vol 6. Cambridge University Press, Cambridge
- van Kampen NG (1985) Elimination of fast variables. *Phys Rep* 124(2):69–160
- Laughlin RB, Pines D (2000) The theory of everything *Proc Natl Acad Sci* 97(1):28–31
- Laughlin RB, Pines D, Schmalian J, Stojkovic BP, Wolynes P (2000) The middle way. *Proc Natl Acad Sci* 97(1):32–37

Human Behavior, Dynamics of

DOUGLAS R. WHITE^{1,2}

¹ University of California, Irvine, USA

² Santa Fe Institute, Santa Fe, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Networks and Cohesion in HB Dynamics

Cooperation, Connectivity- k and “Critical Mass” in Collective Action

Transition Models with Thresholds

Aggregate (“Sufficient Unit”) Equation-Based Modeling

Institutions, Network, Economic Models and Experiments: Testing Causality

Future Directions

Acknowledgment

Bibliography

Glossary

Connectivity- k (k -connected, k -cohesive, structural cohesion, cohesive.blocks) refer to the Menger (1927) theorem for structure/traversal isomorphism in graph theory, as explained in the text, where k -components are the largest possible expansion (maximal group) that preserve structural k -cohesion. Computation is provided by cohesive.blocks in the igraph R package.

Scale-free network where the probability that a node i in the network connects with k other nodes is inversely proportional to the number of k 's links (see: power law), more generally, $p_i(k) \sim k^{-\lambda}$, with $\lambda = 1$ for scale-free.

Nonindependence is characteristic of complex phenomena with built-in interdependencies, where distributions of attributes or relations should not be not directly subject to statistical inference using the null hypothesis of independence, as in structural measures sampled from networks, and autocorrelated time series or autocorrelated spatial distributions.

Sufficient statistic a *sufficient statistic* for a statistical model is one that captures the information relevant to statistical inference within the context of the model, including the size and composition of the units of study. Let X_1, \dots, X_M be a random sample, governed by the density or probability mass function $f(x|\theta)$. The statistic $T(x)$ is sufficient for θ if the conditional distribu-

tion of x , given $T(x) = t$, is independent of θ . Equivalently, the functional form of $f_{\theta|x}(x)$ does not involve θ , and the Fisher–Neyman factorization theorem may be used to help spot sufficient statistics. The likelihood ratio test can often be reduced to a sufficient statistic of the data for hypothesis testing. The minimum variance unbiased estimator of a parameter θ can be characterized in parameter estimation by sufficient statistics and the Rao–Blackwell Theorem. See Scharf *Statistical Signal Processing* [107]. A sufficient unit is one for which a random sample of aggregate statistics are sufficient.

Aggregate (“sufficient unit”) equation modeling

assumes that causality can be found with quantitative equation models that use **sufficient statistics**, which implies that the aggregate units studied have cohesive mass or entitivity for causal interactions to act on their aggregate characteristics. See Sect. “Aggregate (“Sufficient Unit”) Equation-based Modeling”.

NP-complete (NPC) algorithms require an order of non-deterministic polynomial time (NP) but are exceptionally difficult: if a deterministic polynomial time solution can be found for any of them, it would provide a solution to every other problem in NP and empty out the class of NPC.

Dictator game where the first player proposes a split of some endowment and the second, entirely passive, receives the remainder. Not formally a game at all (as the term is used in game theory, where every player’s outcome must depend on the actions of others), it is used in decision theory to test the *homo economicus* model of individual behavior, where selfishness would dictate allocation entirely to oneself. Henrich et al. [50] discovered in a 15-society cross cultural study that people do allocate a share of the endowment to others. Skyrms [113] gives the dynamics of an evolutionary game theory variant.

Concentration indices such as the Laakso–Taagepera Index $1/\sum_i p_i$, where p_i is an effective proportion weighting for each unit, are used for problems such as “what are the effective numbers of political parties self-weighted by their membership (for politics: by their population or area)”, e.g., US party proportions { .49, .49, .02} would have an effective number of 2.08 while France with 101 parties (each weighted by its number of members) might have effective party number of 22.1.

Power law is a Pareto distribution where probability $p(x) \sim x^{-\alpha}$, as for example: “multifractals have tails that follow a power law” (p. 209 in [75]) in how the frequency of similar units at different scales varies

with the scale; see multifractal. Power laws tend to become ubiquitous when what is studied involves dimensional constraints. Power-law *growth* is expressed as $N = K/(t_0 - t)^k$ where K is an initial constant, t is calendrical time, and t_0 is the calendrical singularity date at which $K/(t_0 - t) = K/0$, where division by zero produces dynamical instability as $K/(t_0 - t) \rightarrow \infty$.

Fractal is a pattern or object (e.g. geometrical) whose parts echo the whole, only scaled down, i.e., scale invariant; invariant at any scale of magnification or reduction. Fractal prices occur when positive and negative changes in prices (daily, weekly, monthly, yearly) follow a power law. “To improve almost any fractal model it is a good idea to replace it with a multifractal one” (p. 209 in [75]). A multifractal (with root and generator) is a composite pattern that begins with an initial root (e.g., a straight line) that is successively replaced with a generator (e.g., a zagged line) that replaces every instance of the initial element. See power law.

Causality is a relation holding between two variables such that manipulation of one of the variables (the potential cause) is reliably associated with variation in the other (the response), for some configuration of the values of other potential causes of the response. Estimation includes classical structural equations approaches [74], the treatment effects framework [102,103], the directed acyclic graph (DAG) probabilistic approach [95], and the settable system probabilistic approach that unifies all three [141]. Another aspect of causation is probabilistic evaluation and decision theory, in which case the effect of evidence in revising beliefs about causation can be studied in a Bayesian framework [28,112]. Probability of causation is not causation of probability, although there are probabilistic causative models.

Definition of the Subject

Dynamics of human behavior (abbreviations DHB, HB, HD) deals with the effects of multiple causal forces in human behavior, including network interactions, groups, social movements, and historical transitions, among many other concerns. Description of movement and change distinguishes kinematics from statics, while dynamics considers causes of movement and change. Pearl [95] summarizes issues of causality with two fundamental questions: (1) What empirical evidence is required for legitimate inference of cause-effect relationships? (2) Given that we are willing to accept causal information about a phenomenon, what inferences can we draw from such information, and how? Policy issues entail beliefs about causation and open

a second framework for evaluating beliefs about causality [28,112]. HB dynamics is a field replete with new discoveries—and applications of methods derived from problems and principles that apply across disciplines. Insights transfer across disciplinary boundaries. This is because research strategies for studying causality in a hierarchy of sciences are typically not a reductionism of one level to another but involve recognition of emergence at different levels. Common principles that apply are often shared but with different detailed applications more finely tuned to irreducible aspects of concurrent phenomena. Mathematics and physical principles apply at various levels in the scientific disciplines, but principles discovered in the human and evolutionary sciences are increasingly found to apply and generalize as well.

DHB takes into account the distinctive behaviors of humans and the range of their sociopsychocultural variations. Focusing on causes, HB dynamics may refer, for different levels of social entities, to spatial and temporal, local and long-distance interactions, growth and decline, oscillations, changes in distributional properties, and synchronous or time-lagged causality in dynamical evolution. Examples of precursors in DHB include Ibn-Khaldun's (c.1379) dynamical characterizations of the oscillations of Muslim and Berber political dynasties and charismatic tribal initiatives [60]. Ibn-Khaldun's work was an extraordinary early precursor of the empirical study of oscillatory sociopolitical dynamics (as contrasted with beliefs in cycles of renewal, for example, derived from experience with cycles in nature) and is incorporated into contemporary DHB modeling. Similarly, Richardson's *Statistics of deadly quarrels* (1960) searched for causality of war and posed behavioral dynamic equation-based decision models with basins of attraction for stability, disarmament, or the arms race [108,109]. Schelling's "focal point" solution in the study of strategic behavior and bargaining ("each person's expectation of what the other expects him to expect to be expected to do") advanced the game theoretic policy sciences while his *Micromotives and Macrobehavior* [101] was seminal for modeling complex causal feedbacks. Interest in lower-level processes and how they link to higher levels motivates much of HB dynamical modeling. This is the case as well in biological modeling, as in SFI researcher David Krakauer's statement of research on "the evolutionary history of information processing mechanisms in biology, with an emphasis on robust information transmission, signaling dynamics and their role in constructing novel, higher level features. The research spans several levels of organization finding analogous processes in genetics, cell biology, microbiology and in organismal behavior" [68].

Introduction

HB dynamics is grounded within an evolutionary framework and interacts well with research in biology and primate and human ethology. Fundamental problems in new and old approaches to HB dynamics include general approaches to identify and model (1) units of analysis, (2) interaction equations and structures, and (3) levels of analysis, with (4) sufficient statistics. Many problems concerned with the "units" of investigation, organized into systems, are multifractal, and are explored through detailed study of social organization, biological reproduction, evolutionary phylogeny, and developmental ontology. A focus on networks recognizes the fluidity of dynamical interactions in living systems (i. e., recognizing the limits of hard-unit and hard-wired modeling). Network analysis also links to hydrodynamics, nonlinear synchronization, percolation, and other physical processes as well as models derived from the study of graphs and lattices. Generalizations of entropy measures may also provide approaches for testing general principles in physics that are more useful than mechanics, solid state physics, or conventional models of entropy. While many principles of complexity sciences will apply across many disciplines, how they apply varies with subject matter.

Formal approaches to HB dynamics—where *formal* means theories have been stated in a formalized language, usually mathematical, that does not allow for variable readings [70,120,121]—require construction on the basis of careful descriptive, qualitative, and quantitative research about human behavior and institutions such as are independently carried out in the disciplines (history, sociology, economics, psychology, cognitive science, political science, linguistics, and anthropology, including ethnography, archeology and other domains) as well as in cross-disciplinary fields including those of complexity sciences.

The modeling of human behavior is still in its infancy, and there are likely to be widespread advances in many different areas in coming years. The examples here show a range of concepts and practices but are not intended to cover all of the definitive techniques for modeling human behavior. Among the formal and complexity science approaches in HB dynamics, some of the examples include network modeling, aggregate equation-based modeling, and simulation modeling (equation or agent-based, or both), and how these deal with problems of non-independence. Network modeling depends on finding means of bounding and measuring fields of interaction where particular kinds of units and their causal interrelations can be specified. "Sufficient unit" modeling looks for aggregates at particular scales that represent relative closures of

systems in which causality from internal dynamics can be studied for certain types of relatively well-bounded units that occur within limited ranges of scale. Briefly exemplified are institutional studies of the evolution of market systems extended by experiments in network economics. Not covered are generalized “open system” entropy maximization [119], fields such as fractal dynamics that have challenged fundamental economic axioms and start with the notion that “units” of behavior operate with memory compressed through repetitions of structure that are not dependent on scale. The topics and examples presented form an overall outline about structural k -cohesion and resistance as measurable social forces in human behavior; what enhances or limits scalability in cohesion; what produces and inhibits resistance; and the multiple ways that these two social forces, very different from physical forces, interact dynamically.

Networks and Cohesion in HB Dynamics

The interconnected theme of these illustrative examples will vary from basic measurement to exploratory models to findings built on the mathematics of universality classes, focusing on two features of human ethology that make for unusual dynamics. One is an open-field bonding ability, like chimpanzees, gorillas, and orangutans, which involves recognition of community organized by weak rather than strong ties [39,78]. Humans are additionally equipped with a huge range of social and cultural abilities that derive from our use of symbols, which can widen community and cohesion and enable scalable networks of trust through strong ties as well [135]. These emergents can alter the scale and especially the dynamics of human social organizations. One foundational base for a theory of such emergents are the scalable cohesive groups whose boundaries are identified with the concept of *structural k -cohesion* in sociology [83,134], with new parallels recently discovered in the signaling properties of human and biological networks [104]. Another is the role of *k -cohesive resistance* in human ethology. Taken together, the scalability of cohesive human groups, which allow the scale-up of group sizes that contribute greatly to political expansion and warfare, and the role of decentralized cohesive resistance in pushing back political aggression, exhibit some of the properties of laws of momentum and of proportional reaction, not atypical of complex systems with complex interiors.

To understand the potential for such regularities in phenomena as irregular as human sociopolitical histories (ones that were not lost on the pre-Einsteinian Henry Adams [53]), the concepts underlying indefinite extensi-

bility of scalably emergent cohesive human groups need to be carefully drawn. Rather than harking back to Ibn Khaldun, they draw on Menger’s 1927 theorem [80] for graphs or networks, which is now in use in sociology [83,97,134,138] and anthropology [17] even if rarely used in physics or chemistry, although applications are beginning in graph-theoretic formalization of biological signaling network models [104]. In a network of connected elements, a maximal group (one that cannot be expanded further without losing the property) with structural cohesion k is one that (a) cannot be disconnected without removal of at least k elements, and which, as proved by Menger [80], is equivalent to its having (b) at least k element-disjoint paths between every pair of elements. Property (a) provides *external resistance* to complete disruption (i. e., removing fewer than k elements leaves the structurally k -cohesive group connected), and property (b) proves the existence of a measure k of *internal cohesive traversal* through concomitant existence of at least k redundant paths of transmission or potential communication between every pair of elements. Neither the internal nor the external *cohesive* properties can be surpassed by extending its boundary to include others, whereby each structural cohesion k -group has a unique social boundary. Perfect *scalability* occurs for the numeric size of the *intensive* variable k by any scale-up *extensive* multiplier m because while a structurally k -cohesive group of size n requires only $k < n$ links per element, the same is true at size nm . Note that while dying or migrating might be due to external forces or attractions that remove people from groups, sometimes group members themselves decide to leave, or are expelled. This raises the point that cohesion models and measures are appropriate where the inter-element or interpersonal ties are positive, not antagonistic or negative, by restriction on what should be included in such a model.

Broad problem areas of HB dynamics can be understood from the pairing of (1) the indefinite extensibility of scalably emergent structurally cohesive groups (which have an indefinite supportive potential for scale-up in size of cooperative groups and community) with (2) the contending abilities to form both emergent centralized social structures and (3) cohesive resistance to invasion or centralized authority. HB dynamical processes that can be phrased in terms of symbolic and social interactions of types (1)–(3) are discussed in Sect. “[Cooperation, Connectivity- \$k\$ and “Critical Mass” in Collective Action](#)”. Central to these issues, John Turner’s (2002) *Face to Face: Toward a Sociological Theory of Interpersonal Behavior* [128] presents evidence for the deeply rooted ethological twosidedness of humans as a species pitting *cohesion* against

resistance. A reviewer's summary is worth quoting:

Turner forcefully argues that we are not the solidarity-seeking emotional animals that theorists like Durkheim, Goffman, and Mead would have us to be. Nor are we normally the tortured beings of the Freudian perspective. Reflecting our origins among the great apes, we are a deeply ambivalent species of two minds, craving strong emotional attachments and at the same time, bridle against the constraints in closed social circles of even strong interpersonal ties. Turner argues that this two-sidedness is rooted deeply in our biology, and is not simply the product of historically specific ideologies and social structures. Clearly this viewpoint has enormous implications for the study of face-to-face interactions, as well as many other aspects of sociology. However, in his deep respect for the traditional perspectives in this field, these implications are often obscured and hidden in Turner's exegesis of the general problems and principles in this area of study. None of the other theorists analyzed here have created a better model of ambivalence. Capturing the two-sided nature of social linkages was not a key part of theorists such as Mead, Goffman, and Schutz. Freud made ambivalence central to his model, but locked it into a narrow sexual model. As Neil Smelser has argued, the future of sociological theory will depend in large part on its ability to deal with ambivalence, and Turner's model goes a long way in this regard [45].

Issues of two-sidedness, through a number of steps in logic and measurement, are not unrelated to those of scalability in structural cohesion. To clarify the first three steps in this logic, we can refer to the number of elements in a maximally-sized *k-cohesive* group as its *k-cohsize* or extension and so state, for clarity, that *k-cohesion* and *k-cohsize* = n ($> k$ by definition) of such a group can vary independently for a given level of *k-cohesion* that defines the boundaries of a particular subgroup in a network:

Step 1. *Intensive versus extensive aspects of structural k-cohesion are independent*. Evidence of the causal effect of *k-cohesion* is found in empirical studies and is unrelated to *k-cohsize*. There are three major tests of this to date, one where the major variance in student attachment to high school [83] (as measured by a half-dozen validated questions) was consistently predicted, in multiple tests (ten American high schools randomly selected from the 100-school sample of US Adolescent Health network surveys [10]), by level of *k-cohesion* in which each student was embedded in the school's network of friendships.

With complete data on students and networks in each school, replication of this result was achieved in logistic regressions where all other attribute and pertinent network measures competed in accounting for variance. The influence coefficients for *k-cohesion* replicated in each of the ten independent populations [83]; and the *k-cohsize* of the friendship groups for individual students did not account for school attachment. Since these groups varied in size for each level of *k-cohesion*, this is evidence that the causal effect of *k-cohesion* is not diluted by size, that is, it is an intensive predictive property independent of its scalability in size.

In a second major study (Powell et al. 2006 [97]), Attraction to *k-cohesion* along with recruitment of diversity were the major predictors in a 12-year time-series analysis of variables accounting for tie-formation probabilities proportional to *k* in the biotech industry. Because of the recruitment of new entrants, with fewer ties the overall industry levels of cohesion varied relatively little and oscillated in alternation with 3-year to 4-year waves of variation in attracting new recruits. While *k* did vary slightly for the maximally cohesive core of the industry, it grew neither uniformly nor uniformly decreased over time. There is a consistency here with the finding that *greater cohesion* was the attractor in tie formation and not greater network centrality as hypothesized in the Barabási scale-free network model [8]. The tie-preference attractor was a *sufficient level of k-cohesion that is scalable by addition of members to the structurally cohesive group*, as is shown to occur over time in the biotech industry study.

A third study, of cohesive decay (White and Harary 2001 [134]), tested predictions of how a single 4-cohesive group disintegrated into two competing and eventually disconnected groups. With leaders in opposing groups, order of dissolution of ties followed the pattern predicted, as individuals dissolved their ties successively on the side of the leader with whom they had less *k-cohesion*, and if cohesion was equal, dissolved these ties to the opposing side that had the longer path lengths. While the larger 4-cohesive group dissolved, ties redistributed to the two smaller 4-cohesive groups that formed around the disputant leaders.

Step 2. *Further evidence for scalability* is that *k-cohesion* and *k-cohsize* measures find cohesive groups on much larger scales than do density-based measures called *community detection* [84,87] that split networks into mutually exclusive groups such that higher densities are within rather than between them. Calling these density groups "communities" ignores the fact that such groups overlap and form *k-cohesive* groups on much larger scales. Community detection lacks the scalability of structural *k-co-*

hesion. White et al. [138] demonstrate how the boundaries of these much larger cohesive groups can be approximated in extremely large networks, which is needed because cohesive.blocks computation is NP-complete. Further, they show how density-based and row-column correlation-based algorithms fail in 30 out of 31 methodological studies in a meta-analysis of a classical small-network dataset. They also show analytically, as do Harary and White [134], how k -cohesive components of networks stack hierarchically for successive values of k , providing core-group centralization in addition to horizontal cohesion within a group. The analytical properties of multiconnectivity (*aka* structural cohesion and also k -cohesion) as a precisely measurable and scalable concept (connectivity- k in graph theory) for hierarchical and overlapping group-cohesion boundaries make it an ideal construct for studying the relation between micro (small group and local network properties) and macro properties of social networks, those of political and other social units, and the social construction of roles and institutions [59]. A large number of studies show cohesive scalability in the ways in which symbols and attachments are deployed in human groups and networks, as will be discussed in Sect. “Cooperation, Connectivity- k and “Critical Mass” in Collective Action”, although some effects are preserved only up to certain scale-up thresholds in group size.

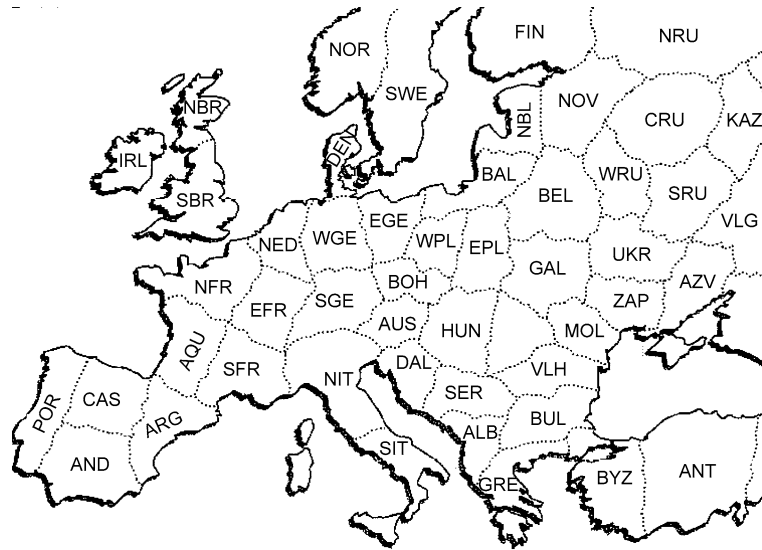
Step 3. *Evidence discussed in Sect. “Cooperation, Connectivity- k and “Critical Mass” in Collective Action” supports the hypothesis that k -cohesive components of human social networks amplify transmission quality and the utility of information that can be cross-checked from multiple independent paths.* For example, distinguishing carefully between dominance (force or force threat) and prestige (freely conferred deference), generalized prestige rankings are scalable along with the transmission quality of multiple channels in k -cohesive groups, while dyadic dominance hierarchies are not. Henrich and Gil-White [47] tested and found support from data across the social sciences for the predictions of a prestige model of social learning as opposed to dominance imprinting. This supports their argument that “natural selection favored social learners who could evaluate potential models and copy the most successful among them,” and that prestige rankings were an emergent product of psychological adaptations that evolved to improve the quality of information acquired via cultural transmission. Finally, studies of networks where utility is gained from long-range interactions [21,56,117] show a variety of network topologies that may combine the benefits of centralized hubs (which are often thought in network economics to maximize efficiency by minimiz-

ing redundancy) with those of redundancy in k -cohesive components.

The approach to cohesion taken here—also contrasting to methods for the partitioning of roles [100]—is not that of trying to specify analytical boundaries using matrix-based methods (Newman [85,86]), which are insufficient as tools to capture the precise boundaries and overlaps in the concept of k -connectivity. The analogy between physical forces and social cohesion or repulsion breaks down because the latter do not involve the kinds of hard-body (Hamiltonian) equations used to describe simple systems such as a bouncing ball, billiard balls, a pendulum, or an oscillating spring. The algorithmic complexity identifying k -cohesive units given an arbitrary graph is NP-complete and not susceptible to matrix-analytic detection, although humans are often better at perceiving accessible and simple but algorithmically complex patterns than are computers.

Because a great many fundamental issues in HB dynamics can be framed in the context of the pairing of cohesion and resistance—similar to but much more complex than the concept of attractive and repulsive force—this pairing is used to organize many of the research questions and findings presented here, not the least of which is related to the problem of the units of analysis needed for tests of HB dynamics, and how these units interact or embed in one another.

How, for example, do scalability and resistance play out in terms of HB dynamics on the larger historical scale? Peter Turchin’s [121] examination of 50 cases (Fig. 1) in the historical military expansion of agrarian states in European history over the last two millennia is illustrative as a test of historical DHB theories that engage concepts of social cohesion. What happens when agrarian states or empires invade a sizeable group that differs in major metaethnic markers (multiple cultural differentiations in *religion, language, and ethnicity*) that are internally cohesive for the group invaded? The framing of this problem is given initially by comparison of dynamical equation-based models for ordinary differential equations of zero-order (unbounded growth or decline), first-order (bounded growth or decline), and second-order (oscillatory growth and decline) [120,121]. Empires show growth and collapse that fail to conform to the first two types of dynamical equations, but could be governed by a second-order dynamics in which there are time lags and negative feedback. The next steps in this study engage the ethological issues that will also be examined here. For example: *What accounts for the resistive capabilities of human social groups, e.g., against outside invasion?* (It is useful to recall that this research was finished before 11 September



Human Behavior, Dynamics of, Figure 1

Turchin's [121] 50 cultural regions used as geographical units in the statistical analysis of the relationship between metaethnic frontiers and polity size (courtesy of the author)

2001). The study draws parallels with the dynamical theory of Ibn Khaldun, who used the term *asabiya* for collective solidarity:

Ibn Khaldun was clearly aware of the nested nature of ethnic groups, and that each level has its own *asabiya* associated with it. . . . [T]he leading or ruling element within a group must be vested in a family or lineage that has the strongest and most natural claim to the control of the available *asabiya* (Ibn Khaldun [60]). Only the leader who controls an *asabiya* of sufficient strength may succeed in founding a dynasty (pp. 38–39 in [47]).

Ibn Khaldun is widely credited as a thoroughly modern sociological scientist of culture, knowledge, politics, and urban life and in his theory of oscillations of Arab and Berber polities. His theory and historical analysis is framed in terms of second-order dynamics:

It [the theory] is held together by his central concept of “*asabiyyah*”, or “social cohesion.” It is this cohesion, which arises spontaneously in tribes and other small kinship groups, but which can be intensified and enlarged by a religious ideology, that provides the motive force that carries ruling groups to power. Its inevitable weakening, due to a complex combination of psychological, sociological, economic, and political factors, which Ibn Khaldun analyzes with consummate skill, heralds the decline of a dynasty

or empire and prepares the way for a new one, based on a group bound by a stronger cohesive force [29].

The thesis of the 50-case study of European military expansion is that “areas where imperial and metaethnic frontiers coincide act as *asabiya* incubators” (p. 56 [121]), areas where new ethnies (i. e., ethnicities, nationalities) are born in the growth of collective resistance. These solidary groups with high *asabiya* have the attributes of *k*-connectivity: “An important element of the theory is the ability of ethnic groups to scale up without splintering into sub-groups” (p. 57 [121]). Examples of integrative mechanisms in this particular context of differing ethnies are *religion*, *society-wide mechanisms of male socialization*, and *rulership with primogeniture*.

External conflict has long been seen to stimulate cohesion on both sides of the conflict boundaries [24,111], as exemplified in the fault line frontiers in history [79,127] and in the marcher state [9] conflicts along these frontiers. A remarkable display of the dynamics of history for the 50-case study is provided by the maps constructed to show, for the regions included in Fig. 1 and for each of the last 20 centuries, the invasions by European empires across metaethnic frontiers and the resultant appearance of new nationalities as resistive movements and states [122].

This mathematical model for empire expansion lacks “a well-developed theory that would connect micro-level individual actions”—like those deriving from structural cohesion—to macro-level dynamics of *asabiya* [121] al-

though the altruism of *asabiya* is seen to follow a conditional altruist model (like that of kin-selection [118]) that depends on cohesion with other altruists—discussed in Sect. “Cooperation, Connectivity-*k* and “Critical Mass” in Collective Action” below. A provisional model (later improved) is given, in its simplest form (pp. 64–66 in [121]), for a polity with a spatial scale h of power projection (imperial “reach”) over an area $A > 0$ and the resistant cohesion $0 < S < 1$ of *asabiya* with an everpresent/constant minimum geopolitical pressure a from the hinterland across a metaethnic frontier of size b . This is given as two dynamical equations with negative feedback that give an unstable equilibrium with a single boom/bust cycle (c_0 and r_0 in these equations are constants):

$$\dot{A} = c_0 A S \left(1 - \frac{A}{h} \right) - a$$

$$\dot{S} = r_0 \left(1 - \frac{A}{2b} \right) S (1 - S).$$

Here, change in area is a function of the polity area and of cohesion, limited by overextension, while the function for change in cohesion has an in-built oscillatory dynamic affected by the size of the metaethnic frontier. These dynamics, although intended only to characterize the problem, are informative as to how its parameters play into dynamical complexity. If “reach” h is not much greater than frontier width b , the empire can reach a stable equilibrium, while if $b < h/4$ the boom/bust cycle, but only one can occur, not more. Only when the model incorporates the discounting of expansionary power with distance in a spatial simulation is a second-order type of effect is obtained, that of oscillatory growth and decline.

Rather than having this model serve to study attacks and resistance, and the influence of *relative* cohesiveness in outcomes of politicomilitary contests (which is difficult to measure), Turchin’s frontier theory is tested instead with the time-lagged prediction for each of two millennia that *when the metaethnic frontier is intense in the first half of the millennium, for one of the cultural areas in Fig. 1, then large territorial polities (empires) will originate in the second* [121]. The evaluation is whether the expansive tendency *originated* in a contest of respectively cohesive entities rather than trying to predict the outcome of the battle, the more relevant outcome being that—having developed its cohesion through external conflict—the unit that is initially attacked may eventually enlarge to become an empire. This holds for 11 out of 15 cultural regions that were on the metaethnic frontier, while out of 34 regions that were not on the frontier, only 1 developed an empire in the first millennium AD; and it holds for 22 of 28 Fig. 1 frontiers in 500 AD–1500 and empires in 1000 AD–2000.

The four exceptions in the first case and the six in the second were regions incorporated into an empire centered in a neighboring region.

Human Behavior, Dynamics of, Table 1

Cross tabulations for politics that start on frontiers and end as empire a millenium later

0–1000CE	Starts as frontier	No frontier
Becomes empire	11	1
No Empire	4	34

1000–1900CE	Starts as frontier	No frontier
Becomes empire	22	3
No Empire	6	19

50 regions, $p < .0000004$

How many empires *were observed that lacked the temporal precondition of a metaethnic frontier* (with subsequent growth of resistant cohesion)? The exceptions are 1 and 3 for the two periods, respectively. The first, and two of the latter cases, occur where the existence of the frontier was of short duration. One polity (Savoy-Sardinia, founding Italy) remained as a true exception, in a population formed by Celts and Romans, but with no clear causal path from metaethnic frontiers to polity expansion. But the major result is that the empires of the later periods *did* (and not just may) result in almost all cases from cohesive resistance to the attacks of the previous period, at long time scales.

Thus, these results are fully consistent with the scalability of structural cohesion as a basis of sociopolitical support for military expansion of polities (potentially into empires) but more importantly are supportive of the theory that *k*-cohesive structural resistance, which grows slowly on the metaethnic frontiers of expanding empires in such a way as to facilitate the growth of resistive “nationalistic” ethnic solidarity and eventually of consolidation of resistive metaethnic frontier groups themselves into expanding polities and empires, with long time-lags in their development.

Lim, Metzler and Bar-Yam [72] analyze local conflicts between distinct ethnic or cultural groups within multi-ethnic states (India and former Yugoslavia), matching actual conflicts to spatial population structure in a simulation model of type separation, where cohesion emerges through movement to more homogeneous regions and through avoidance of conflict. Conflicts are predicted due to the structure of boundaries rather than between the groups themselves, consistent with Turchin’s [121] findings. The local ethnic patch serves as an “order parameter” to which aspects of behavior are coupled in the dynam-

ics of a universality class of collective behavior. Similarly, the multilevel evolutionary model of Garcia and van den Bergh [36] shows how parochialism, as altruistic behavior specifically targeted towards in-group members, can result from group selection operating on direct conflict between groups.

Cooperation, Connectivity- k and “Critical Mass” in Collective Action

If structural cohesion is scalable, what are the factors, aside from external conflict, that would prevent or facilitate the scale-up of cohesion? Or of group size generally, assuming some modicum of cooperation [73]? The major problem in explaining why cooperation should occur at all in human groups, in the absence of external conflict, is that of the benefits of selfishness to free-riders when others bear the cost of altruism. One component of “The Tragedy of the Commons” [46] is that collective goods [92] are nonexcludable: Once achieved (like peace, clean water or air, public transport, or wage contracts) they are available to everyone. Many if not most such goods have jointness of supply (available to all), i. e., their cost does not increase proportionally to group size. The initial problem is that if it takes only some initial investment and costs by those who bring such goods into existence, why should anyone else bear these costs when they can have them for free? This creates “the dilemma of cooperation” [92] and of collective action. And the larger the group the easier it is to ride free. Evolutionary game theory [89], with a replicator dynamic that favors those with lower cost for the same benefits, predicts that without some compensation for altruism, even starting from a small number of free-riders in a population, selfishness becomes the norm. The secondary problem of collective goods is who will bear the costs to maintain them?

Reputation may attach positively to altruism and negatively to free-riding. In this respect, two recent experimental papers are strongly supportive (although unaware) of connectivity- k in helping to explain cooperation in human groups [18]. In one study the judged veracity of gossip is shown to increase considerably if it came from more different sources [54], not if one source kept repeating the same gossip, while another relates gossip to reputation and cooperation in general [114]. James West in *Plainsville* (1945) [130] was the first to connect gossip and the maintenance of the unity of groups. According to Gluckman (p. 308 in [37]), however, West misinterpreted the extent to which “gossip does not have isolated roles in community life, but is part of the very blood tissue of that life.” Gluckman refers to Colson’s *Makah Indians* [23] ethnog-

raphy to illustrate the importance of gossip to the unity of groups.

While diffusion of reputation along the node-disjoint paths of k -components can provide benefits to altruism, its influence diminishes as groups grow larger and average network distance grows large, reducing the scalability of k -components with high levels of cooperation. Further, if a group has *too much* k -connectivity (as in completely connected cliques), the benefits of reputation diminish because of the “echoing” effects of conformity and diminution of independent sources of information [19,55,145]. In cliques or overly-connected groups, single dominant individuals have the potential to influence everyone and thus to distort the robust veracity of information. Further, studies of human friendships and other long-term relationships show that the success of reciprocal strategies (such as tit-for-tat) relies on a combination of medium-term accounting, forgiveness, and propensity to defect with strangers if they already have an adequate number of partners [55].

Benefits of punishment within a group have a similar profile of optimality to reputation. Like gossip and reputation, punishment can be effectively delivered through k -cohesive independent paths and thus diffuse coherently respecting the boundaries of k -cohesion [18] (although for a given k -cohesive group, the paths used to diffuse reputation need not be the same as those used to deliver punishments). This works best for groups with moderate average distances in the network and with (or defined by) moderate connectivity- k . Similar to reputation, “cohesion extends punishment even beyond the community network and protects insiders against trouble-making outsiders, [especially] when community members come to defend fellow community members against norm-violating outsiders” [18], while incidental defectors at the margins of the cohesive group may have little impact on behavior within the group. Henrich et al. [50,51] ethnographic-psychological study of 15 societies from five continents, representing the breadth of human production systems, found that willingness to use punishment in the dictator game covaries with altruism across populations in a manner consistent with coevolutionary theories. But while appropriate punishments diminish the relative rewards of free-riding, they also incur costs to the enforcers.

Punishment as third-party intervention tends to rely more on dominance and perceptions of the use of force, entailing higher risks in some cases, than on reputation in the modeling of advantages of cooperative behavior. After observing a group of macaques in captivity in which a small number of individuals fulfilled policing tasks, making interventions into dyadic conflicts, temporary ex-

perimental “network knockout” removals of the policing monkeys showed it was their presence that prevented the group from falling apart into small clusters [34,35]. Here, Jessica Flack and coworkers note [33], “the degree to which one individual perceives another as capable of using force is communicated using a special dominance signal. Group consensus about an individual’s capacity to use force arises from the network of signaling interactions.” Consistent with studies of k -cohesion, this research found that “coarse-grained information stored at the group level—behavioral macrostates – “was more useful than detailed information at the individual level”. Because “successful intervention relies on consensus among combatants about the intervener’s capacity to use force,” use of “a formalism to quantify consensus in the network,” and with consensus as a measure of power, showed that “the power distribution is fat tailed and power [here: consensus] is a strong predictor of social variables including request for support, intervention cost, and intensity.” This modeling of power distributions shows how dominance signaling strategies “promote robust power distributions despite individual signaling errors” [34].

Third-party interventions in conflicts resemble recognition of community membership in that such interventions rarely occur with respect to outsiders. Recognized community boundaries (as distinct from k -cohesion, which may extend beyond these boundaries) provide the most probable context for the dyadic construction of cooperativity through reciprocity [118], dominance in third-party intervention, and dyadic game theoretic strategies that achieve cooperativity (such as tit-for-tat or lose-shift in Prisoner’s dilemma) [113]. These, together with generalized reciprocity [71], i.e., altruism in the expectation of indirect return, are also among the most potent constructors of community-building strong ties in social networks [135], especially if they are navigable [1], as in many elite groups and non-Western [136] societies. Bowles and Gintis [15] summarize the game-theoretic work on cooperation showing that the critical condition for cooperative outcomes, which otherwise deteriorate with increases in group size, is the presence of *strong reciprocators*, who cooperate with one another and punish defectors, even if they sustain net costs, provided that they are more likely to interact with one another than at random. Thus, network structure and preferences (positive assortment) prove to be central to an evolutionary path to large-scale cooperativity. Pepper and Smuts [107] show how positive assortment through environmental feedback can play the same role. There are, then, evolutionary paths to the scale-up of k -cohesion for indefinitely large groups.

Putting together these principles of primate (reciprocity, policing) and human social networking, we can also see compatibility of k -connectivity with the theory of “critical mass” in collective action [77,90,91]. Group size *does* increase the probability of a critical mass of people who develop common goods through collective or cooperative action. This relates directly to the scalability of k -connectivity, wherein as the size *nof* such a structurally cohesive group expands, it is still only k links per person that are needed for k -connectivity. But there is always an expected excess of ties, upwards of k , for some members of such a group, and an increase in n increases the probability of formation for a group with a critical mass of connectivity $k + 1$ or higher ($k + l > 1$). This relates to the “paradox of group size” for collective action groups: “When groups are heterogeneous and a good has high jointness of supply [i.e., with cost that does not increase proportionally to size], a larger interest group [size n] can have a [relatively] smaller critical mass,” which could also be a critical mass with connectivity $k + l$. The problem of mobilizing collective action is whether there is a mechanism that connects enough people with appropriate interests and resources so that they can act to construct a collective good [77]. Structural cohesion provides just such a mechanism [17,83,97].

An extended feature of this model of critical mass, which has been investigated through simulation [77], is an *accelerative function* for what has been called network externality [4], where every new participant in creating a collective good makes it more attractive for the next participant to join. Different forms of collective action have some mix of this source of nonindependent decisions and/or a *decelerative function* wherein free-riding is more likely on the belief that others will do the job. Since success in collective action is partly a problem of coordination, there is some advantage to members of a critical mass in collective action having greater centrality. But again, if the collective action group at large has connectivity k and the leadership critical mass has connectivity $k + l$, the latter is achieved by the hierarchical embedding of higher orders of connectivity and not necessarily by greater centrality of a single leader.

So there are two aspects to consider for the dynamics of growth and decline in size of cooperative human groups: (1) reinforcement mechanisms of community, which tend to be self-limiting with respect of structural cohesion, and (2) critical mass in collective action or positive assortative reciprocators [15], which both tend to be self-enabling. While collective action to produce a collective good also requires a model of group process that cannot be deduced from simple models of individual behavior [77], the prob-

lem for understanding how large-scale societies and polities can achieve sustainability may be solved by assortative strong reciprocators [15]. The former problem—of sustainability of cooperation in a community—is different. All of the mechanisms there—reciprocity, third-party intervention, reputation, and punishment—depend on relative stability of community membership. Prior to the electronic age (which poses somewhat different problems of stability in virtual communities but makes for more independence of k -connectivity from local density and geographic distance), stable communities carried designations such as “settlement” or “nomadic group” as nominal indicators of relative stability in proximal spatial interactions among their members.

There is a herd cohesion solution to the stability problem (follow the *surest* neighbor!) [25,26], but also an advanced social cohesion solution (the formation of k -connectivity groups with a stable core) [18]. Empirical studies of neighborhoods [105,106] show that a certain threshold of residential stability is a crucial factor for the efficacy of mechanisms for community-level enforcement of the cooperative norms of third-party intervention, reputation, and punishment (all but strictly dyadic reciprocity unless it is strong and assortative). “For cooperation to be maintained at the community level, the network as a whole must be relatively more stable than patterns of individual actions” [18]. Combining community mechanisms and critical mass in collective action, we have the foundations for an evolutionary theory of cooperativity and cohesion in human groups. Many of these features (but not structural cohesion) have been brought under the Darwinian umbrella in a way that shows how the co-evolution of culture and genes jointly influence cultural transmission (dual inheritance theory) through the vehicles of human behavior and psychology [48,52]. This framework allows the integration of work on kinship, friendship, reciprocity, reputation, social norms, and ethnicity into a generally applicable mathematical characterization that may contribute to solving the problem of cooperation and extending on to the evolution of evolution and of economic systems [49,50].

Beyond adding k -connectivity into dual inheritance theory, there are also newer models of achieving minimum punishment and maximum crime reduction through policing concentration on an arbitrary push-down set of offenders [64]. The theory here, validated in simulation and case study, is that a fair and effective law enforcement strategy can only succeed if it approximates one with a stable target set of offenders at whom punishment is directed until recidivism ceases, individual by individual, replacing each nonrecidivist on the pushdown list by another known

offender chosen with a probability *proportional* to rate of current offenses but otherwise *arbitrarily*, i. e., *fairly*. Policing an arbitrarily stabilized set of offenders mirrors the requirement for stability in cooperative neighborhoods. *Stability seems to be a key ingredient for cooperativity*.

What are the implications of these findings for considerations of the scalability of human communities and of human polities? Although k -connected groups are scalable, the properties of third-party intervention, reputation, and punishment to maintain cooperativity are not scalable, nor is dyadic reciprocity except under very special conditions [135,136]. Scalability through conflict—resistance to threat—and through collective action to produce collective goods, organized by a “critical mass” or through assortative strong reciprocity [15] is, however, scalable.

So why do human groups not simply grow larger at all scales [73], as challenged by competing groups, or by possibilities such as establishing collective goods capable of sustaining growth? Prestate societies only rarely sustain continued growth in size, but rather split, and then remix through intermarriage and mating (fission and remixing), with transition to a higher-order political form occurring extremely sporadically. It might be thought that if politically independent groups fission but still remain linked, through intermarriage or k -connectivity, then fusion into larger political groups would be easy. Many anthropological theories assume a stage-wise progression such as band to tribe or tribe to chiefdom [110]. Comparative ethnographic, historical, and archaeological studies, such as those of Wright [143,144], however, make it clear that passages from band to tribe (concepts with serious conceptual problems) to chiefdom to state are extremely difficult and unlikely transitions. And as we have seen [121], growth in state societies and empires is followed by collapse and the rise of other polities instead. Models of political fission might provide necessary conditions for transitions in successions of forms of leadership as polities develop with different sets of roles. New role set configurations might also create founder effects in the emergence of economic or political forms.

Transition Models with Thresholds

Transitions such as chiefdom to state can be modeled in an evolutionary dynamics of human behavior framework that includes the interaction of ethological characteristics—general human behavioral tendencies—and forms of sociopolitical organization. Social anthropologist Christopher Boehm [13,14], whose field studies range from Montenegro [12] to wild chimpanzees at Gombe, called at-

tention to the *human tendency to resist domination* (consistent with Turchin's findings [121] in Sect. "Networks, and Cohesion in HB Dynamics"), which is not shared with other great apes (consistent with Henrich and Gil-White [47]). In a substantial cross-cultural survey of societies in a wide variety of social and ecological settings, Boehm selected those with egalitarian behavior, and found that their behavior was not shaped by these settings but rather was deliberately shaped by their members, guided by a nearly universal ethos in these societies "that disapproves of hierarchical behavior in general and of bossiness in leaders in particular." His survey reveals the wide variety of means by which "the political rank and file" evict leaders who evince excessive authoritarian tendencies. This "creates a reverse dominance hierarchy, a social arrangement that has important implications for cross-phylogenetic comparisons and for the theory of state formation" [13] that might be called a "law of human ethological resistance", consistent with [121]. One of these mechanisms of resistance is fission, the break-away from a group that is growing large or with too many settlements under a single leader.

Surveys of archaeological, historical, and ethnographic cases not only show transitions from chiefdoms to states to be very rare but also show that states are based on a radically different principle of a hierarchy of roles to which decision-makers are recruited. Henry Wright [143,144] shows that primary and secondary states have three or more levels of mobilization of resources upwards and passing of information both upwards and downwards through a hierarchy of divided offices and a division of political labor [115]. Chiefdoms, unlike states, are characterized by paramount leaders who delegate as little authority as possible, in contrast to states with their delegated division of labor for authority [144]. Paramount chiefs may govern subdivided territories with village chiefs and ritual specialists, but there are nearly always no more than *two* levels of chiefly resource mobilization conducting directly to the chief and *all political decisions are integrated into the chiefly persona*.

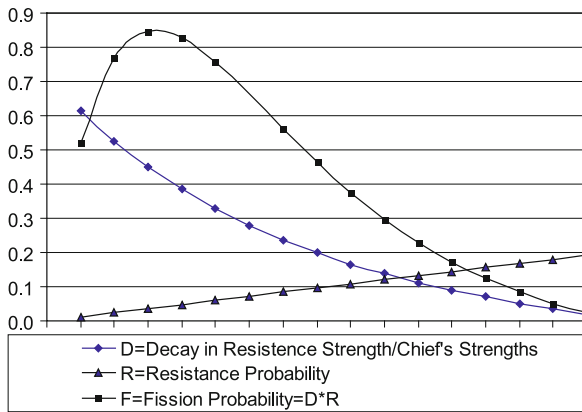
To assume a simple quantitative increase in network size and complexity as chiefdoms develop into states is therefore inappropriate. Chiefdoms are also characterized by a reverse ranking hierarchy [13,14], not an actual reversal of dominance but one of prestige ranking [47] in which leaders are expected to exhibit altruism to followers through redistribution of goods or forms of reciprocity and bestowal of favors or gifts to counterbalance the processes whereby resources were concentrated through interpersonal network ties, although the reciprocity is rarely balanced in any material sense [27,96,98]. In their dy-

namics of growth, chiefdoms—with their structural cohesion and cohesive hierarchies based on intermarriage, exchange, and cross-cutting ties—tend to increase in size through internal growth or annexation of settlements, then to give way to fission at times of crisis following growth, especially if these crises coincide with issues of political succession. There is no tendency in these dynamics for gradual cumulative evolution in complexity toward state organization. The mosaic of sub-chief territories mapped into the chiefly ranking are segments that recurrently separate and then re-form in successive periods of political change

Griffin and Stanish (p. 2,24 in [43]) provide evidence and a model for a tipping-point synchronicity threshold in the transition from chiefdoms to emergent pristine states in the Lake Titicaca case of Tiwanaku, c. 500 AD (outgrowing the territorially larger political formation at Pucara). The transition occurred archeologically and in a detailed simulation model after a long period of cycling in which multiple chiefdoms climb the population size gradient only to be fragmented by fission. There is strong empirical evidence for cycling in growth and fission. During the period of cycling, primary centers, population concentrations, and increase in both the overall productivity and population of the region occur sporadically without synchronization. Then, in one rapid burst, archeologically and in repeated probabilistic simulations, these previously unsynchronized features emerge synchronously, pushing past a probability threshold for fission. Figure 2, reflecting results from the simulation model, shows the variables affecting the fission of chiefdoms plotted against time for growth; then, as cycling occurs, setting the cycling time back to that of an earlier equal scale in size. The simulation data were also reaggregated for X as number of settlements under the chief, and could be estimated analytically for other variables such as X for communication time from center to furthest outlier. The first variable, Y_1 , is one of exponential decay in the ratio of resistance strength to the leader's strength, which can be expressed as an exponential probability density function supported on the interval $[0, \infty)$, where $\lambda > 0$ is the rate parameter of the distribution. Since this is a discrete exponential distribution with $X \geq 1$:

$$p(X; \lambda) \sim \lambda e^{-\lambda X}.$$

The second variable, Y_2 , increases with X on the assumption that the likelihood of resistance to the chief increases with time, or variables that cycle with time, such as number of settlements. The third variable, $Y_3 = Y_1 \bullet Y_2$, is the probability of fission due to resistance, which is humped because it is a product of distributions that have higher



Human Behavior, Dynamics of, Figure 2

The transition threshold from Chiefdom to State: $Y_1 \sim$ Exponential Decay in the ratio of Resistance strength/Leader's strength; $Y_2 \sim$ Increasing] probability of Resistance as number of settlements increases; $Y_3 = Y_1 \bullet Y_2 =$ the probability of fission

resistance probabilities at opposite magnitudes of X . The shape of the Y_3 distribution emerges as an average over many simulation runs, and opens further questions for investigation. What emerged in the Lake Titicaca region, consistent with the simulation, were two dominant polities, separated in time, one an incipient state, the other smaller in population but larger in territory, along with extensive trade networks including the smaller centers. This is a typical multisite trading configuration of early states [3].

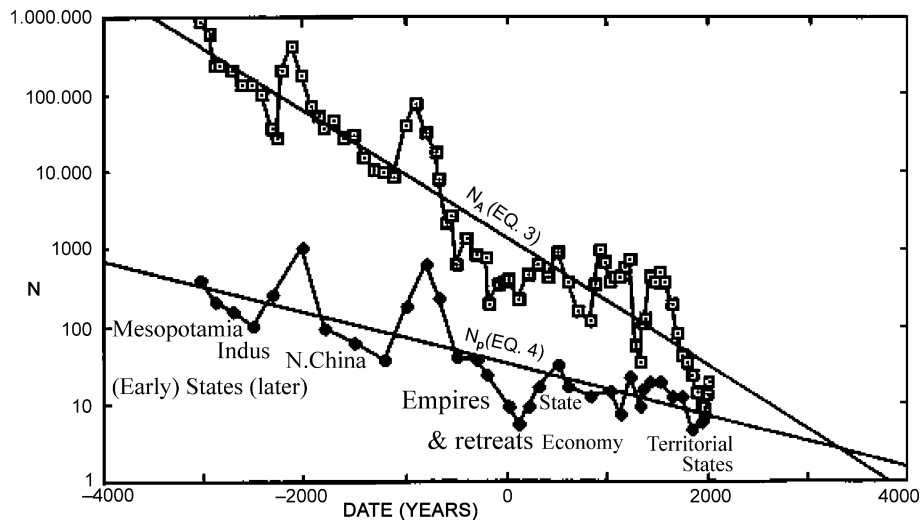
One can speculate as to whether resistive transition thresholds might occur also from band to tribe or big man (having occasional fission), to chiefdom and from chiefdoms (having occasional fission) to minimal states. Fissioning is by no means universal, and one study of the Titicaca region itself shows that village but not chiefly fissioning had ceased long before state formation with emergence of a regional religious tradition [7]. Are there resistive transition thresholds from minimal states (having occasional fission) to urbanized states, from urbanized states (having occasional fission through colonization) to dynastic states, or from dynastic states (with occasional fission with the death of a ruler and partition of domains under obligatory personal inheritance) to territorial agrarian states? Or do nonterritorial state expansions collapse, replaced by others? The territorial state, given institutional sovereignty over territory, is less likely to fission at a size threshold and its growth dynamics are shown in Sect. "Aggregate ("Sufficient Unit") Equation-based Modeling" to involve shrinkage following times of scarcity in population/resource ratios. This creates amplifications of in-

equality and internal conflict [126]. Modern mega-corporation growth is often arrested by national and international legal regulations mandating breakup of monopolies but there are no early barriers against corporate growth in size, although some corporations do fission for reasons other than size constraints.

The temporal scaling of long-term transitions in populations and sizes of the largest polities does show clear transitions, over 5000 years of world history, as shown on Fig. 3 [116]. The lower line in the figure is an exponential fitting of the effective number of polities (Laakso-Taagepera concentration index $1/\sum_i p_i$, where p_i is the effective proportion-weighting for each unit) weighted by their populations, and the upper line by their geographical areas. More even proportions for p_i , such as $\{.4 .3 .2 .1\}$, compared to higher concentrations like $\{.7 .1 .1 .1\}$, will have a higher effective numbers, 3.33 versus 1.92, while extremely concentrated proportions, e.g., $\{.97 .01 .01 .01\}$ with effective number 1.06, approach unity. The declining slopes in Fig. 3 show a decrease in effective polity numbers $1/\sum_i p_i$ with greater concentration of population than of area (slopes differ by 2, the fitted exponential population roughly the square root of area). Over these five millennia the fitted effective number of political entities weighted by area decreased from circa one million to circa 64, and from circa one thousand to circa 8 weighted by population. For Fig. 3:

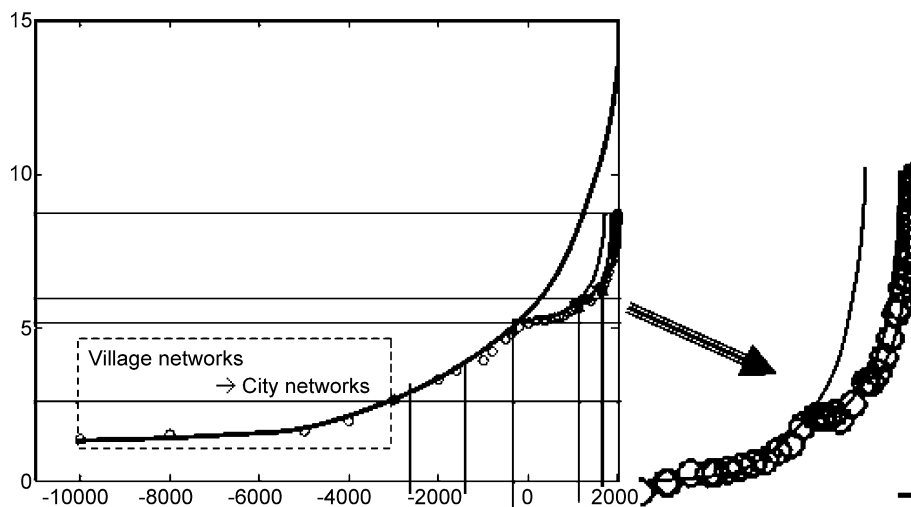
Three sudden increases in polity sizes occur: [fewer large polity concentrations] around 3000 BC [urban revolution in Mesopotamia], 600 BC, and AD 1600 [the seafaring trade revolution]. This study tests the exponential model against area and population data [for polities] over five millennia. It also gives tables and graphs of area versus time for all major polities since AD 600. The median duration of large polities at more than half the peak size has been 130 years, and it has not changed over 5000 years. (p. 475 in [116])

Two of the three solid lines superimposed on the original figure show how two of the three elbows of change in the lower of the empirical data lines (circa 2600 BC, 1200 BC, and 200 AD), from polity population concentrations to dispersals, are followed with short time-lags by two similar elbows of change in the upper line from polity area concentrations to dispersals over the next hundred years. Whether these transitions represent eras of crises in urban empires is unclear, as are most extrapolations from so few data points. The first case might reflect the short-lived breakup of early Bronze-age polities (Mesopotamian and Indus) and the second the breakup



Human Behavior, Dynamics of, Figure 3

Transition thresholds for States and Empires (effective number of polities), based on area and on population (Taagepera 1997 Fig. 5, courtesy of the author, who notes that individual polities that expand slower tend to last slightly longer (p. 475 in [116]); arrows mark his dates for large polity concentrations; others are marked by lighter lines)



Human Behavior, Dynamics of, Figure 4

World population power-law growth spurts and flattening as shown in a semilog plot of Kremer's (1993) [69] data with successive power-law fits

of North China states. At 50 AD the third elbow of transition from population concentration to dispersion (e.g., for classical empires such as those of the Romans, Svataphana and Han, which actively discouraged market developments) occurs in the dispersion phase commensurate with that of polity area. Population and area reconcentration in the next phase (ca. 500–850 AD) are also roughly commensurate. A downward spike of population concentration recurs circa 1050 AD when again it has a lagged

effect on polity area concentration. It might be surmised that changes in population-area interactions in the era of power-law city growth are increasingly subject to market-driven trade routes (e.g., Silk Roads in Eurasia from 100 BC–1300 AD). This is a context, from 900 AD forward, of new national market economies that diffuse from Sung China to the west, where the Abbasid, Carolingian and related polities encouraged widely articulated market systems. Such changes are studied and modeled by

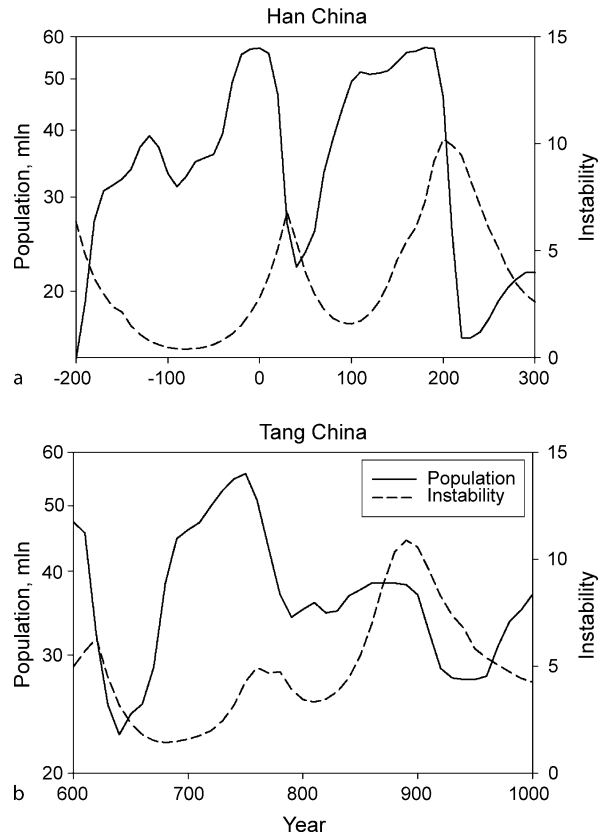
Modelski and Thompson [82]. Variability in the ways that markets change these political oscillations is particularly evident from 1800 and the industrial revolution, as nationalism and markets consolidate the effective number of polities geometrically weighted by size, up until 1990 with the breakups of the Soviet Union. There has been little dynamical modeling of the multiple causality in these coupled/decoupled oscillations. Taagapera (p. 488 in [116]) notes that while population concentration can continue to increase in the present era area concentration must stabilize because jumps to higher concentrations in earlier eras occurred with acquisition of control over large and sparsely populated areas (desert, steppe, deserts, tundra, respectively for Sargon, Mongol, British, Russian empires), and such areas are much less available now.

To express some of the consequences of the transition to networks of cities connected by trade routes, and eventually to market-driven trade, Fig. 4 sketches the suggestion that world population begins to grow not exponentially but in power-law growth spurts, correlated initially with the transitions noted in Fig. 3 [133].

Cities act as attractors for skilled, unskilled and intellectual labor as well as entrepreneurs and merchants, with a concomitant drain on settlements of smaller size. This enables power-law growth, at growth rates proportional to city size, i. e., cities as “attractors” as in the scale-free network model of Barabási 2000 [8] (but see [139]), while rural areas and smaller settlements do not diminish their population but with elevated birth rates can replenish their losses from outmigrants. This pattern allows world population to grow in power-law spurts, but power-law growth is self-limited by population crashes as it would otherwise grow to infinity in a finite time [129]. The places where the polity transition crises occur in Fig. 3, e. g., 2600–2400 BC, 1200–1100 BC, and 200–100 BC, and 1300 AD correspond to those crises in the larger states and cities where power-law growth in their (and world) population hits some sort of limit, growth flattens, and resets the starting parameters for a new upswing of power-law growth (a pattern first noticed but not explained in Korotayev et al. 2006 [66]). The largest world empires, of the Golden Horde Mongols and the British, appear as the result of two of these more recent upswings (a topic currently under investigation by Christopher Chase-Dunn in one of the National Science Foundation’s Human Social Dynamics research awards).

Aggregate (“Sufficient Unit”) Equation-Based Modeling

This approach aggregates to the unit size and boundaries at which to define causal variables and interactions and to



Human Behavior, Dynamics of, Figure 5

Population and Sociopolitical Instability for Han and Tang China (Turchin 2005a [124], courtesy of the author)

attempt to explain behavioral dynamics of these units by appropriate equations. This requires the “sufficient unit” condition that the aggregate units of study have the kinds of cohesive mass or entitvity for causal interactions to act on their aggregate characteristics. Time-series will have periods in which this condition is satisfied because of relative “endogeneity” of interactions where there are few external disturbances or exogenous shocks to the unit.

Using this approach, Peter Turchin [123,124] extended a realistic and empirical approach to historical processes—not caricatures of imperial collapse—for basic Malthusian models of population pressure on resources and time-lagged negative feedback effects with internal conflict (see also [125]). He simply uses “standard quantitative methods of natural sciences, such as time-series analysis, regression, and cross-validation. The statistical analysis reveals strong and repeatable patterns in the data on population numbers and the intensity of internal war. And history of science suggests that strong empirical regularities are usually associated with the action of funda-

mental laws” [22], some yet to be discovered for complex systems science. Examples from the Han and Tang China data [123,124] (Chap. 8 in [121]) are shown in Fig. 5 for population and sociopolitical instability (internecine wars), which are related by time-lagged feedback effects.

For the stationary X (population) and Y (internecine wars) variables in these figures, standard time-lagged regression is used to estimate regression constants $a_i = \{a_0, a_1, a_2\}$ where τ is the time lag of 30 years (approximating a human generation), t is time, and ε_t is an error term assumed to be normally distributed [16]:

$$X(t) = a_0 + a_1X(t - \tau) + a_2Y(t - \tau) + \varepsilon_t, \text{ (Model(1))}$$

(and an analogous model for $Y(t)$, reversing the definitions of X and Y). Further:

One possible objection to the procedure outlined above is that there is some positive autocorrelation between $X(t)$ and $X(t - \tau)$ due to the time-series nature of the data, and it is conceivable that the excellent correlations between the observed $X(t)$ and predicted $X(t)^*$ are entirely due to this “inertial” effect. To eliminate this possibility [the analyzes were redone] with a different dependent variable, $\Delta X(t) = X(t) - X(t - \tau)$. $\Delta X(t)$ is a measure of the rate of change, and by using it we break the autocorrelation arising from the time-series nature of the data. In fact, $\Delta X(t)$ is none other than the realized per capita rate of population change, which is the standard dependent variable in the analyzes of population data... There can still be some predictive relationship between $\Delta X(t)$ and $X(t)$, so we need to compare two alternative models:

$$\Delta X(t) = a_0 + a_1X(t - \tau) + \varepsilon_t, \text{ (Model(2))}$$

... the *inertial* model (with an analogous (2) for $Y(t)$), and

$$\Delta X(t) = a_0 + a_1X(t - \tau) + a_2Y(t - \tau) + \varepsilon_t, \text{ (Model(3))}$$

... the *interactive* model (with an analogous (3) for $Y(t)$). The interactive model has an extra parameter, but in a cross-validation setting this does not matter (if the extra independent variable does not have a systematic influence on the dependent variable, then adding it to the model actually *decreases* to the ability of the model to predict out-of-sample data) [124].

The comparisons of the inertial and interactive predictions in Table 2 show consistent effects of dynamical time-lagged interactions between population and sociopolitical instability (civil conflict) that cannot be attributed simply to the inertial dynamics of each of these variables separately. The interactive effects, documented in detailed case studies [126], are those of oscillations: rising population creating resource scarcity, which amplifies inequality, making the value of property rise while that of labor falls, which, if lasting longer than a generation, causes civil unrest and conflict, causing population in turn to decline, with a lag until the cycle recurs as civil conflict ceases, allowing population to rise again (see [38]). Replications of similar findings are obtained by Turchin for the English Tudor cycle (1485–1730) [124], the Medieval English Plantagenet Cycle (1150–1485), French Capetian cycle (1150–1450) and Valois (1450–1660) cycles, Roman republican (350–30 BCE) and principate (30 BCE–285 CE) cycles, Russian Muscovite and Romanov cycle [126], and the Pueblo cycle, where Kohler et al. [65] examine the Turchin model with data from Southwest Colorado between AD 600 and 1300. They find that “it fits well during those periods when this area is

Human Behavior, Dynamics of, Table 2

Comparing Out-of-Sample Predictions of the Inertial and Interactive Models (Turchin 2005a [124], courtesy of the author)

Source of data	Dependent variable	Correlation between predicted and observed			
		1st half → 2nd half		2nd half → 1st half	
		Inertial	Interactive	Inertial	Interactive
England	Population	−0.57	0.94	−0.07	0.44
England	Instability	−0.13	0.80	−0.53	0.89
Han China	Population	0.45	0.57	0.73	0.48
Han China	Instability	0.39	0.87	0.37	0.68
Tang China	Population	0.56	0.80	0.61	0.90
Tang China	Instability	0.57	0.78	0.66	0.92

a more or less closed system. It fits poorly during the time from about AD 1000–1200 when this area is heavily influenced first by the spread of the Chacoan system, and then, by its collapse and the local political reorganization that follows. The model is helpful in isolating periods in which the relationship between violence and population size is not as expected.”

Institutions, Network, Economic Models and Experiments: Testing Causality

Studies of historical HB dynamics often lead to different conclusions. In many cases these differences result from the aspects of social process that are focused upon. Contending views may have more general points of consensus when we look at these processes more abstractly. The concepts of structural (k -)cohesion and resistance may help to provide more points of consensus.

There are many views of the formative processes of a market economy based on impersonal exchange and its prior institutional bases. Conceptualized as a network, a market economy requires k -cohesiveness simply to attain $k > 3$ alternatives for buyers and sellers, the minimum “many” players in the market without which the advantages of competitive pricing cannot be obtained. Competition itself, however, is simultaneously a *resistive* as well as a cohesive process, a differentiation of the interests and identities of the competitors. The goods exchanged, for competitive markets, must be alienable, which entails a change of hands in property rights. Players at one time and place may be groups or corporations as property owners party to exchange. At other times they are individuals; or, parties to exchange may be a heterogeneous mix of individuals and groups. For parties to exchange they must have rights: rights to hold property and to alienate property, rights that can be agreed upon by contract, rights and *institutions* that can enforce the contract. Effective “coercion-constraining” institutions that prevent the abuse of others’ property rights “influence whether individuals will bring their goods to the market in the first place” (p. 727 in [41]). These give rise to agency, as the capacity for human beings to make choices within a social world and to enforce the rights that those choices impose on the world, whether agency is for the selfsame agent or on behalf of another. The social world is complexly layered at the level of rights, obligations, agents, agency—and institutions as cohesive and resistant social constructions exist for the enforcement of norms. Competing views and agendas are entailed.

These kinds of interlocking components of social worlds do not fall into place quickly, but are built up in-

crementally over time, just as social networks are built up incrementally and their structural configurations may change slowly even while specific individuals come and go. Market institutions, for example, “co-evolve through a dynamic inter-play between contract-enforcement and coercion constraining institutions” (p. 727 in [40]) along with resistive social movements, movements to create collective goods against the resistance of free-riders, and more episodic events.

The institution-building perspective is one that has received very detailed effort in modeling actual social networks and institutional change in their historical context, abstracting the ways that social players and agents have come to effectively optimize their interactions from their multiple interests and perspectives. One of the most formidable projects of this sort over the last decade, building on the earlier work of North [88], has been to trace social foundations and historical development of institutions in pre-modern Eurasia that facilitate impersonal exchange and lead to paths toward competitive markets, while other developmental paths lead in a variety of other directions [42]. In the words of one reviewer, this work of Avner Greif:

strips economic transactions down to their elements [and] focuses on the core question: who (or what) were the watchdogs that allowed the merchants to trust one another and to bear with the princes who could confiscate the fruits of all their efforts? And who (or what) were the watchdogs’ watchdogs? [The work] repeatedly and carefully relates these questions to economic theory [and] illustrates them with real transactions of medieval merchants. He takes the right approach to economic development, and thereby achieves an original and important new perspective on its causes [2].

In each of Greif’s case studies, dynamical game theory is used to test the fit between the observed data and the known historical development of institutions as well as the cultures and behaviors of the players and actors. One of the shortcomings of Greif’s work is that the early modern merchants did not face the same problems as those developing markets *de novo* in early Mesopotamia, India, China, and Mesoamerica. But further evaluation of the replicability of Greif’s model is carried on by network economists using experimental real-world simulations that engage participants in the knowledge, payoffs, and choices of the context that is modeled, testing the experimental models against the observed or recorded historical processes and outcomes [62]. To quote:

North (2005) argues that belief systems and the stock of local knowledge, the internal representations of the human experience, are intimately intertwined with the external institutions that humans build. We investigate this relationship by varying the degree to which property rights are enforced in yesterday's institutions before the opportunities for long-distance trade present themselves with perfectly enforced property rights. Specifically, in the new experiment we report here, three-fourths of the subjects in an economy are drawn from two different treatment histories in *Build8* sessions, one in which property rights in personal goods are perfectly enforced for all of the participants, *though they must rely on trust and repeat interactions to enforce exchange agreements*, and another in which no property rights of any kind are enforced. Hence, in both sets of history-inducing sessions, there is no external enforcement of exchange contracts and, as found [in an earlier experiment], no need for such [62].

The findings of the experimental study are “that a history of un-enforced property rights hinders our subjects’ ability to develop the requisite *personal* social arrangements necessary to support specialization and effectively exploit *impersonal* long-distance trade.” Thus we might understand through network economic experiments some replicable elements of the origin of impersonal market system. These, like cooperativity, require but go beyond structural cohesion to the social constructions of institutions that secure trust and the benefits of interpersonal trade, i. e., network elements that reinforce the scalability and benefits of structural *k*-cohesion as discussed above.

In Greif's analysis, while the institutional supports for impersonal long-distance trade only developed slowly in medieval Europe (and elsewhere) the full protections of “coercion-constraining” institutions “that prevent the abuse of others’ property rights” and “influence whether individuals will bring their goods to the market in the first place” were still not in place even in England after the “Glorious Revolution of 1688,” which did not secure such rights beyond “the landed, commercial, and financial elite” (p. 786 [41]). Rarely is linear progression of rights entailed in the ups and downs of the precursor elements of fully competitive markets that vary from one country to another. In England, after 1688, for example, although

parliament gained supremacy, it was not in the business of protecting property rights per se. Its policy reflected the interests of those who controlled it. . . . The subsequent history is thus marked by gross

abuses of property rights. . . . Yet, a state controlled by its landed, commercial, and financially elite and later empowered by the Industrial Revolution was a boon for the extensions of markets. The evolution of the modern markets reached its zenith. . . . Europeans shared a common heritage of individualism, self-governance, a broad distribution of coercive powers, and man-made laws. Reversing their institutional developments and enabling market extensions was relatively easy (pp. 775–776 in [41]).

Eurocentrism is not intended in the use of this example, as this project entailed equally detailed historical and modeling analyzes of China and the Muslim world.

Greif's analysis of land-based institutions and exchange example provides a contrastive comparison against Erikson and Bearman's 2006 [30]) network study of English maritime trade between 1600 and 1831. Here an entirely different account is given of the emergence of the competitive market system. The shared elements are the *k*-cohesive extensions of trade routes, extensive by sea as by land, and the institutional development of English rights in property, commercial exchange, protection, and agency. Here, however, the resistive element is paramount, and the “new economy” arises through malfeasance of the sea captains of the English East India Company. Their work is carried on preemptively, out of self-interest, exploiting the opportunity of delay. Instead of bringing English goods to the orient and returning with oriental goods in one single return cycle, in order to stay beyond the time when the ships could return by the monsoon winds, they traded from port to port on their own behalf with their own goods and retained the profits. Over time, the density of this network became so great that the sheer volume of overlapping circular routes, crisscrossing the net of visited ports, pushed the *k*-connectivity of the market exchanges beyond 0, 1, or 2 for different subregions often up to 7–8, a veritable revolution in creating new market opportunities and competitive market pricing through sheer volume of malfeasance behavior: malfeasance because this was all conducted against the policy of the home company, which was powerless to prevent it.

Future Directions

The topics covered here, of cohesion and resistance as measurable social forces in human behavior, and the multiple ways that these two social forces dynamically interact—and what enhances or limits scale-up and scale-down of both cohesion and conflict or resistance – leaves open many researchable questions. Lim, Metzler and Bar-Yam's (2007) study supports a more advanced even if

partial view of the dynamics of cohesion and resistance, group separations, and segregative conflicts along insufficiently demarcated boundaries. Other parts of the human cohesion/resistance dynamics covered in this review show some of the other ways in which cohesion and resistance interact. Human capacities for structural cohesion, for example, support cultural differentiation of groups. Transition thresholds characterize evolutionary bouts of scale-up in group size through central authority, oscillating against resistance from egalitarian preferences for autonomy. With scale-up in size, expansions of political units encounter boundaries of cultural and ethnic differentiation where resistance scales up as oppositional cohesion in positive feedback cycles, creating further expansion of polities that began only as resistive groups. These support growth of population sizes, which lead in turn to scarcity relative to resources within regions. With generational time lags there develop both greater differential inequality and conflictual resistances to inequality. Large polities develop institutional and economic frameworks that can provide benefits to internal differentiation, while the enhanced potential for cohesiveness and economic growth can find ways, as in the biotechnology industry illustration, to utilize the recruitment of diversity to create innovation [93] while stabilizing the costs of cohesive integration.

The problems of modern states and institutions may be seen to devolve on how to minimize the costs of the conflicts that are generated by the oscillations between oppositional cohesion and integrative cohesion. For HB dynamics more generally, solid causal analysis using the most advanced techniques is only possible with current and future data collected systematically on historically documented entities compared over different time scales, up to millennial time series. These data can be analyzed with processual models, network analytic models, institutional, cultural, and evolutionary game-theoretic and economic analyzes. Many of the algorithms needed at this level of complexity have developed in computer science, e. g., by Pearl [5,95] and, by including the crucial element of agency in a new econometrics framework [140,141,142] economics—the otherwise dismal science—can be investigated by causal modeling algorithms. In the modeling of causality that is relevant here to HB dynamics, the analytical power recently gained in econometric models may be neatly illustrated by a comparison of statistical results and conclusions reached by fractal economics, survivorship analysis of successful mutual fund managers, bootstrap models of the same problem, and market simulations of intelligent agents that place orders to trade at random. The first case involves the discovery of fractal pricing

in cotton markets [75] and the Dow-Jones [76], contradicting the standard assumption in economics hypothesized by Bachelier [6] that Brownian movement (Gaussian price deviations) is descriptive of market price dynamics. If volatility is predictable in markets, but not price and direction, the implication is that value might not be useful as a concept in economics [75] (consider market collapse when no trader wants to trade in an uncertain market (pp. 3–4 in [58])). Parallel evidence from experimental studies rejected the reference-independent framing of judgments for the “value” of expected utility theory as originally framed by Bernoulli [11], and questioning the assumption that utilities are stable [61]. Similarly, survivorship analysis of successful mutual fund managers showed no evidence that the top ranked funds were any better than random as they lacked measurable persistence [20]. Finely tuned bootstrap estimation models that are oriented toward testing causal models in econometrics, however, showed that while income fund managers did no better than random, growth fund managers showed persistence in their ability to pick stocks (Kosowski et al. [67]). And finally, a baseline market simulation model of for intelligent agents that place orders to trade at random, with only one free parameter, accounted for 96% of the best buying and selling prices (the spread), and 76% of the variance of the price diffusion rate [32], which “demonstrates the existence of simple laws relating prices to order flows, and in a broader context, because it suggests that there are circumstances where the strategic behavior of agents may be dominated by other considerations.” “One of the virtues of this model is that it provides a benchmark to separate properties that are driven by the statistical mechanics of the market institution from those that are driven by the strategic behavior of agents. It suggests that institutions strongly shape our behavior, so that some of the properties of markets may depend more on the structure of institutions than on the rationality of individuals.” These examples are all indicators of complex dynamics.

The challenge of HB dynamics is to assemble better data related to aspects of the problems modeled, including those of competing hypotheses: more complete data, data better grounded in diverse historical circumstances, and more contextual detail. A second challenge is to have better statistical estimators, identification and correction for sources of bias, attention to nonindependence, careful modeling of richly grounded historical data, attention to causal modeling, and multiple-level models. These efforts are facilitated by sharing of data, collaborative analysis of potential biases in data, sharing of documentation of software and source code, verification of source code,

and replication of results. Extensive effort has gone into the archaeological and geocoded data related to the evaluation of the model in Fig. 2 of transition probabilities. Ten years of effort went in locating and coding the data in Fig. 3, for example. Good data on population numbers at all levels, such as aggregated in Figs. 3 and 4, is extremely important for modeling, hard to come by, and demands careful analysis for bias detection, bias correction and data reconstruction. The data on internecine warfare in Fig. 5 were patiently transcribed episode by episode in two compendia of scholarly work over millennia. These are but a few of many thousands of databases, many of which have not been made sharable or are not conserved or not well documented. Also needed are data analytic routines able to make accurate estimates using probabilistic bootstrap methods with small samples and for different kinds of data, e. g., continuous or discrete, nominal or ordinal. A great deal of documented open source code is now available.

Causal modeling is the core of dynamical analysis, and future directions will include modeling of the types illustrated here, and many more, but with integrated datasets for different foci and levels of analysis. A host of intersecting and mutually enriching integratable time-coded longitudinal datasets—like Turchin’s data for the 50-region data in Fig. 1 [121], or the sufficient size data for Fig. 2 [124]—are needed from comparable local contexts and processes up to the global, e. g., google-earth-like sharable data structures, equipped with analytic routines for time-series causal modeling and testing. There are many separate projects on shared issues, but overall integration is needed. Geographic Information Systems (GIS), for example, need to be reintegrated around open source code (e. g., GRASS, written in open source R) that includes temporal and network modeling. Auto-correlation and other techniques and models for dealing with nonindependence of cases will figure heavily in causal modeling.

Among new network analytic methods that are becoming standard in many disciplines are the censuses of different types of cycles (“motifs”) in large networks that make up cohesive k -components, and accessible software to compute k -connectivity. Analysis of these sorts of data allow testing of where and how the internal micro and middle-range structures come from in k -components [44,81,132]. Which structures come from preferences and which from the marginals or limits on how data were collected or spatially distributed? This kind of work is now being done in biology but also in anthropological network studies, where new software packages have been developed that are specifically designed to deal with cer-

tain problems, such as kinship networks or the kinds of generative kinship computations that people actually use in their social cognition [70,99].

Entropy maximization “open system” models conditioned on biased random processes will increasingly become integrated with HB dynamics as we come to understand how to connect them to foundational problems in the human sciences, some of which are discussed in the ENTROPY entry in these volumes. Simple entropy models, for example, are currently being used to fill in missing data from what is known from an archaeological site [31]. Tsallis entropy [119], in contrast, would provide a one-parameter modification for least energy maximization channeled through networks, with diffusion gradients that have multiplicative effects. Generative network models for cohesive cycles, as studied to date [137] show consistent distributions of numbers of links that are all in the Tsallis entropy family, so there are promising avenues in this research area.

As seen in examples here, more integration of theory and data is needed, from macrohistorical models where large-unit aggregation relates to sufficient statistics for causality, through cascades of spatio-temporal processes to the micro level of interactions between individuals [94]. Kirman [63], for example, argues that “The emergence and evolution of the networks that govern the interaction in the economy plays a crucial role and it may well be the case that the standard notion of equilibrium is irrelevant in such as context.” Multilevel network analyzes will be aggregated structurally in new ways for which new modeling techniques are needed to analyze the composition of units and of processes [131].

The construction of theory and hypotheses in this article illustrate only a few potential causal links among major topics on evolutionary and historical dynamics, institutional and economic models, game theory and social networks, organized around a few core dimensions of ethological importance (structural cohesion and resistance). The point of these illustrations has been that there are truly major forces in history— k -cohesion and cohesive resistance for example—but these have very different properties than forces in physics, or the dynamics of chemistry and biology (although one LANL biologist, asked to pinpoint the major threats to survival, responded with “human behavior” [57]), and they require very different measurements and theory. But there is no closure on the topics of HB dynamics: rather, there is an abundance of theory and results in social, historical, and simulation modeling that lend themselves to the evaluation of causality. Taking causality and dynamics seriously rather than dismissively leads to very different theoretical and analytical perspec-

tives. Longitudinal data on human social, historical, and network phenomena are sufficient to support high-level theoretical and integrative research that can further benefit from the most advanced of methods in the complexity sciences. But there is a pressing work to be done in analytic methods and in constructing valid and reliable datasets and variables on appropriate and comparable units and processes under analysis.

Acknowledgment

As a Santa Fe Institute external faculty member, the author greatly appreciates SFI support and the benefits of interactions with scores of SFI researchers, visitors, and staff. Special thanks to Jeroen Bruggeman for sharing his pre-publication book manuscript and the cross-fertilization of ideas presented in Sect. “Cooperation, Connectivity- k and “Critical Mass” in Collective Action”, to Henry Wright for detailed commentary and suggestions, Eric Smith for comments on the coalescence of the argument, Peter Turchin for sharing historical data, and to Bruggeman, Turchin, Art Griffin, Charles Stanish, and Emily Erikson for suggestions on the discussions of their findings, to Lilyan Brudner-White for many editorial suggestions, and to D. Eric Smith and many other colleagues at SFI for discussions.

Bibliography

- Adamic L, Lukose RM, Huberman BA (2002) Local search in unstructured networks. In: Bornholdt S, Schuster HG (eds) *Handbook of graphs and networks: From the genome to the internet*. Wiley-VCH, Berlin
- Akerlof GA (2007) Book commentary on Avner Greif (2007a) *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Amazon.com editorial review
- Algaze G (2005) The Sumerian takeoff. *Struct Dyn* 1(1):2
- Arthur B (1994) *Increasing returns and path dependence in the economy*. University of Michigan Press, Ann Arbor
- Avin C, Shpitser I, Pearl J, Identifiability of path-specific effects. UCLA Cognitive Systems Laboratory, Technical Report (R-321), June 2005. In: *Proceedings of International Joint Conference on Artificial Intelligence*, Edinburgh, August 2005
- Bachelier L (1900) *Théorie de la Spéculation*, 1900, *Annales de l'Ecole normale supérieure* (trans. Random Character of Stock Market Prices)
- Bandy MS (2004) Fissioning, scalar stress, and social evolution in early village societies. *Am Anthropol* 106(2):322–333
- Barabási A-L (2002) *Linked: The New Science of networks*. Perseus Publishing, Cambridge. <http://www.nd.edu/~networks/linked>
- Barfield TJ (1989) *The perilous frontier: Nomadic empires and China*. Blackwell, Oxford
- Bearman PS, Burns L (1998) Adolescents, health and school: Early findings from the national longitudinal study of adolescent health. *NASSP Bull* 82:601–23
- Bernoulli D (1738) Exposition of a new theory on the Measurement of Risk (trans. 1954). *Econometrica* 22:123–136
- Boehm C (1983) *Montenegrin social organization and values: Political ethnography of a refuge area tribal adaption*. AMS Press, New York
- Boehm C (1993) Egalitarian behavior and reverse dominance hierarchy. *Curr Anthropol* 34:227–254
- Boehm C (1999) *Hierarchy in the forest: Egalitarian society and the evolution of democratic politics*. Harvard University Press, Cambridge
- Bowles S, Gintis H (2008) Cooperation. In: Blume LE, Durlauf SN (eds) *New palgrave encyclopedia of economics*, 2nd edn. Palgrave Macmillan, Basingstoke. http://www.dictionaryofeconomics.com/article?id=pde2008_C00059789=GintisAopicid=result_number=2
- Box GEP, Jenkins G (1976) *Time series analysis: Forecasting and control*. Holden-Day, San Francisco
- Brudner LA, White DR (1997) Class, property and structural endogamy: Visualizing networked histories. *Theory Soc* 26:161–208
- Bruggeman J (2008) *Social networks: An introduction*. Routledge, New York
- Burt RS (2001) Bandwidth and echo: Trust, information, and gossip in social networks, In: Casella A, Rauch JE (eds) *Networks and markets*. Russell Sage Foundation, New York
- Carhart MM, Carpenter JN, Lynch AW, Musto DK (2002) Mutual fund survivorship. *Rev Financ Stud* 15:1439–1463
- Carvalho R, Iori G (2007) Socioeconomic networks with long-range interactions. ECCS 2007, European Conference on Complex Systems 2007—Dresden. Paper #334. <http://arxiv.org/abs/0706.0024>
- Clodynamics (harvested 2006), Web site of Peter Turchin, <http://www.eeb.uconn.edu/people/turchin/Clio.htm>
- Colson E (1953) *The Makah indians*. Manchester University Press, Manchester, University of Minnesota Press, Minneapolis
- Coser L (1967) *Continuities in the study of social conflict*. The Free Press, New York
- Couzin I (2007) Collective minds. *Nature* 445:715
- Couzin I et al (2005) Effective leadership and decision making in animal groups on the move. *Nature* 433:513–516
- Earle TK (1977) A reappraisal of redistribution: Complex Hawaiian chiefdoms. In: Earle TK, Ericson J (eds) *Exchange Systems in Prehistory*. Academic Press, New York
- Eells E, Skyrms B (eds) (1994) *Probability and conditionals, Belief revision and rational decision*. In: *Cambridge studies in probability, induction, and decision theory*. Cambridge University Press, Cambridge
- Encyclopedia Britannica (2007) Ibn-Khaldun. Harvested December 8. <http://www.britannica.com/eb/article-225307/Ibn-Khaldun>
- Erikson E, Bearman PS (2006) Routes into networks: The structure of English East Indian trade, 1600–1831. *Am J Sociol* 112(1):195–230
- Evans T, Knappett C, Rivers R (2008) Using statistical physics to understand relational space: A case study from mediterranean prehistory. In: Lane D, Pumain D, van der Leeuw S, West G (eds) *Complexity perspectives on innovation and social change*. Springer Methodos series, in press. <http://www3.imperial.ac.uk/pls/portallive/docs/1/7292491.PDF>

32. Farmer JD, Patelli P, Zovko II (2005) The predictive power of zero intelligence in financial markets. *PNASUSA* 102(11):2254–2259, <http://www.santafe.edu/~jdf/papers/zero.pdf>
33. Flack JC, Krakauer DC (2006) Encoding power in communication networks. *Am Nat* 168:97–102
34. Flack JC, de Waal FBM, Krakauer DC (2005) Social structure, robustness, and policing cost in a cognitively sophisticated species. *Am Nat* 165:E126–E139
35. Flack JC et al (2006) Policing stabilizes construction of social niches in primates. *Nature* 439:426–429
36. Garcia J, van den Bergh J (2007) Evolution of parochialism requires group selection. *ECCS 2007, European Conference on Complex Systems 2006, Oxford. Paper #120*
37. Gluckman M (1963) Gossip and scandal. *Curr Anthropol* 4:307–315
38. Goldstone JA (1991) Revolution and rebellion in the early modern world. University of California Press, Berkeley
39. Granovetter M (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380
40. Greif A (1997) On the social foundations and historical development of institutions that facilitate impersonal exchange: From the community responsibility system to individual legal responsibility in pre-modern europe, Working Papers 97016, Stanford University, Department of Economics. <http://www-econ.stanford.edu/faculty/workp/swp97016.pdf>
41. Greif A (2007) Commitment, coercion and markets: The nature and dynamics of institutions supporting exchange. In: Ménard C, Shirley MM (eds) *Handbook of new institutional economics*. Springer, New York
42. Greif A (2007) Institutions and the path to the modern economy: Lessons from medieval trade. In: *Political economy of institutions and decisions*. Stanford University Press, Palo Alto
43. Griffin AF, Stanish C (2007) An agent-based model of prehistoric settlement patterns and political consolidation in the lake Titicaca basin of Peru and Bolivia. *Struct Dyn* 1(1):2
44. Hamberger K, Houseman M, Daillant I, White DR, Barry L (2004) Matrimonial ring structures. *Mathématiques et sciences humaines* 43(168):83–121
45. Hammond M (2002) Review of [15]. *Canad J Sociol Online* Nov–Dec 2002
46. Hardin G (1968) The tragedy of the commons. *Science* 162(3859):1243–1248
47. Henrich J, Gil-White F (2001) The Evolution of prestige: freely conferred status as a mechanism for enhancing the benefits of cultural transmission. *Evol Hum Behav* 22:1–32
48. Henrich J, Henrich N (2006) Culture, evolution and the puzzle of human cooperation. *Cogn Syst Res* 7:221–245
49. Henrich J, McElreath R (2003) The evolution of cultural evolution. *Evol Anthropol* 12:123–135
50. Henrich J et al (2005) Economic man in cross-cultural perspective: Ethnography and experiments from 15 small-scale societies. *Behav Brain Sci* 28:795–855
51. Henrich J et al (2006) Costly punishment across human societies. *Science* 312:1767–1770. <http://www.sciencemag.org/cgi/content/full/312/5781/1767>
52. Henrich N, Henrich J (2007) *Why humans cooperate*. Oxford University Press, New York
53. Henry A (1906, 1918) *The education of Henry Adams*. Houghton Mifflin, Boston
54. Hess NH, Hagen EH (2006) Psychological adaptations for assessing gossip veracity. *Hum Nat* 17:337–354
55. Hruschka D, Henrich J (2006) Friendship, cliquishness, and the emergence of cooperation. *J Theor Biol* 239:1–15
56. Jackson MO, Wolinsky A (1996) A strategic model of social and economic networks. *J Econ Theory* 71:44–74
57. Johnson N, LANL scientist, Personal communication
58. Jorion PJ (2007) Reasons vs. causes: Emergence as experienced by the human agent. *Struct Dyn* 2(1):1–6
59. Jost J (2005) Formal aspects of the emergence of institutions. *Struct Dyn* 1(2):2
60. Khaldun I (1958) [c.1379] *The Muqaddimah: An introduction to history*, (trans. from the Arabic by Rosenthal F). Pantheon books, New York
61. Kahneman D, Tversky A (2003) Maps of bounded rationality: A perspective on intuitive judgment and choice. In: Frangsmyr T (ed) *Les Prix Nobel 2002*. Almquist & Wiksell International, Stockholm
62. Kimbrough E, Smith VL, Wilson B (2006) Historical property rights, sociality, and the emergence of impersonal exchange in long-distance trade, *Am Econ Rev* (forthcoming) Interdisciplinary Center for Economic Science working paper at <http://www.ices-gmu.net/article.php/433.html>
63. Kirman A (1999) Aggregate activity and economic organisation. *Revue européenne des sciences sociales* 37(113):189–230
64. Kleiman MAR (2008) When brute force fails: Strategy for crime control (in progress)
65. Kohler TA, Cole S, Ciupe S (2008) Population and warfare: A test of the Turchin model in Pueblo societies. In: Shennan S (ed) *Pattern and process in cultural evolution*. University of California Press, Santa Fe, in press Institute Working Paper 06-06-018: <http://www.santafe.edu/research/publications/wpabstract/200606018>
66. Korotayev A, Malkov A, Khalitourina D (2006) Introduction to social macrodynamics: Secular cycles and millennial trends. URSS, Moscow
67. Kosowski R, Timmermann A, White H, Wermers R (2006) Can mutual fund “stars” really pick stocks? New evidence from a bootstrap analysis. *J Financ* 61(6):2551–2595. http://weber.ucsd.edu/~mbacci/white/pub_files/hwcv-101.pdf
68. Krakauer D (2007) Self-description of work in progress, SFI web site, Harvested 12/11
69. Kremer M (1993) Population growth and technological change: One million BC to (1990). *Q J Econ* 108(3):681–716
70. Leaf M (2007) Empirical formalism. *Structure and Dynamics* 2(1):2
71. Lévi-Strauss C (1949, 1969) *The elementary structures of kinship*. Beacon, Boston
72. Lim M, Metzler R, Bar-Yam Y (2007) Global pattern formation and ethnic/cultural violence. *Science* 317:1540–1544
73. Machalek R (1992) The evolution of macrosociety: Why are large societies rare? In: Freese L (ed) *Advances in human ecology*, vol 1. JAI Press, Greenwich, pp 33–64
74. Malinvaud E (1985) Econometric methodology at the Cowles commission: Rise and maturity. Abstracted from the Cowles fiftieth anniversary volume. <http://cowles.econ.yale.edu/archive/reprints/50th-malinvaud.htm>
75. Mandelbrot B (1963) The variation of certain speculative prices. *J Bus* 36:394–419

76. Mandelbrot B, Hunter RL (2004) The (mis)behavior of markets: A fractal view of risk, ruin, and reward. Basic Books, Hudson
77. Marwell G, Oliver P (1993) The critical mass in collective action: A micro-social theory. Cambridge University Press, Cambridge
78. Maryanski AR (1987) African ape social structure: Is there strength in weak ties? *Soc Netw* 9:191–215
79. McNeill WH (1963) The Rise of the west. New American Library, New York
80. Menger K (1927) Zur allgemeinen Kurventheorie. *Fundamenta Mathematicae* 10:96–115
81. Milo R et al (2002) Network motifs: Simple building blocks of complex networks. *Science* 298:824–827
82. Modelski G, Thompson WR (1996) Leading sectors and world power: The coevolution of global politics and economics. University of South Carolina, Columbia
83. Moody J, White DR (2003) Structural cohesion and embeddedness. *Am Sociol Rev* 69:103–127
84. Newman MEJ (2004) Detecting community structure in networks. *Eur Phys J B* (38):321–330
85. Newman MEJ (2006) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103:8577–8582
86. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 74:036104
87. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69:026113
88. North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
89. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314:1560–1563
90. Oliver P, Marwell G, Teixeira R (1985) A theory of the critical mass, I. interdependence, group heterogeneity, and the production of collective goods. *Am J Sociol* 91:522–556
91. Oliver PE, Marwell G (1988) The paradox of group size in collective action: A theory of the critical mass. II. *Am Sociol Rev* 53(February):1–8
92. Olson M (1965) The logic of collective action. Harvard University Press, Cambridge
93. Page SE (2007) The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton University Press, Princeton
94. Pande, Rohini and Christopher Udry. Institutions and development: A view from below. Economic Growth Center, Discussion Paper No. 928, Yale University. http://www.econ.yal.edu/growth_pdf/cdp928.pdf
95. Pearl J (2000) Causality: Models, reasoning and inference. Press MIT, Cambridge
96. Peebles C, Kus S (1977) Some archaeological correlates of ranked societies. *Am Antiq* 42:421–448
97. Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *Am J Sociol* 110(4):1132–1205
98. Pryor LP, Graburn NH (1980) The myth of reciprocity. In: Gergen KJ, Greenberg MS, Willis RH (eds) Social exchange: Advances in theory and research. Plenum Press, New York
99. Read D, Behrens C (1990) KAES: An expert system for the algebraic analysis of kinship terminologies. *J Quant Anthropol* 2:353–393
100. Reichardt J, White DR (2007) Role models for complex networks. *Eur Phys J B* 60:217–224
101. Richardson LF (1960) Statistics of deadly quarrels. Boxwood Press, Pacific Grove. <http://shakti.trincoll.edu/~pbrown/armsrace.html>
102. Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
103. Rubin D (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66:688–701
104. Ruths DA, Nakhleh L, Iyengar SM, Reddy SAG, Ram PT (2006) Hypothesis generation in signaling networks. *J Comput Biol* 13(9):1546–1557
105. Sampson RJ, Raudenbush SW, Earls F (1997) Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277:918–924
106. Sampson RJ, Morenoff JD, Earls F (1999) Beyond social capital: Spatial dynamics of collective efficacy for children. *Am Sociol Rev* 64:633–660
107. Scharf L (1991) Statistical signal processing: Detection, estimation, and time series analysis. Addison-Wesley, New York
108. Schelling T (1960) The strategy of conflict. http://en.wikipedia.org/wiki/Schelling_point
109. Schelling T (1978) Micromotives and macrobehavior. W.W. Norton, New York
110. Service ER (1962) Primitive social organization: An evolutionary perspective. Random House, New York
111. Simmel G (1890) The sociology of Georg Simmel (trans: 1950). Free Press, New York
112. Skyrms B (1988) Probability and causation. *J Economet* 39(1-2):53–68
113. Skyrms B (1996) Evolution of the social contract. Cambridge University Press, Cambridge
114. Sommerfeld RD et al (2007) Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences* 104:17435–17440
115. Spencer C (2004) Primary state formation in Mesoamerica. *Annu Rev Anthropol* 33:173–199
116. Taagepera R (1997) Expansion and contraction patterns of large polities: Context for Russia? *Int Stud Q* 41:482–504
117. Tambayong L (2007) Dynamics of network formation processes in the co-author model. *J Artif Soc Soc Simul* 10(3):2
118. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46:35–57
119. Tsallis C (2008) Entropy. *Encyclopedia of complexity and systems science*. Springer, Berlin
120. Turchin P (2003) Complex population dynamics. Princeton University Press, Princeton
121. Turchin P (2003) Historical dynamics. Princeton University Press, Princeton
122. Turchin P (2004) Dynamic maps of evolution of the state system and metaethnic frontiers in Europe during the two millennia. <http://www.eeb.uconn.edu/people/turchin/Clio.htm>
123. Turchin P (2005) A primer on statistical analysis of dynamical systems in historical social sciences (with a particular emphasis on secular cycles). *Struct Dyn* 1(1):4
124. Turchin P (2005) Dynamical feedbacks between population growth and sociopolitical instability in agrarian states. *Struct Dyn* 1(1):3

125. Turchin P, Korotayev A (2003) Population dynamics and internal warfare: a reconsideration. *Soc Evolut Hist* 5(2):112–147
126. Turchin P, Nefedov S (2008) *Secular cycles*. Princeton University Press, Princeton
127. Turner JH (1921) *The frontier in american history*. H. Holt and Company, New York
128. Turner JH (2002) *Face to face: toward a sociological theory of interpersonal behavior*. Stanford University Press, Palo Alto
129. von Foerster H, Mora PM, Amiot LW (1960) *Doomsday: Friday, 13 November, A.D., (2026)*. *Science* 132:1291–1295
130. West J (1945) *Plainsville*. Columbia University Press, New York
131. White DR (1974) Mathematical anthropology. In: Honigmann JJ (ed) *Handbook of social and cultural anthropology*. Rand McNally, Chicago, pp 369–446. <http://eclectic.ss.uci.edu/~drwhite/pub/MathAnth74-1.pdf>
132. White DR (2004) Ring cohesion theory in marriage and social networks. *Mathématiques et sciences humaines* 43(168): 5–28
133. White DR (2008) Innovation in the context of networks, hierarchies, and cohesion. In: Lane D, Pumain D, van der Leeuw S, West G (eds) *Complexity perspectives on innovation and social change*. Springer Methodos series, in press. <http://eclectic.ss.uci.edu/~drwhite/pub/ch5revMay-20.pdf>
134. White DR, Harary F (2001) The cohesiveness of blocks in social networks: Node connectivity and conditional density. *Sociolog Methodol* 31:305–359
135. White DR, Houseman M (2002) The navigability of strong ties: small worlds, tie strength and network topology. *Complexity* 8(1):72–81. <http://eclectic.ss.uci.edu/~drwhite/Complexity/K&C-a.pdf>
136. White DR, Johansen U (2005) *Network analysis and ethnographic problems: Process models of a Turkish nomad clan*. Lexington, Oxford. A summary is at http://en.wikipedia.org/wiki/Network_Analysis_and_Ethnographic_Problems
137. White DR, Kejžar N, Tsallis C, Farmer D, White S (2006) A generative model for feedback networks. *Phys Rev E* 11:016119
138. White DR, Owen-Smith J, Moody J, Powell WW (2004) *Comput Math Organ Theory* 10(1):95–117
139. White DR, Tambayong L, Kejžar N (2008) Oscillatory dynamics of city-size distributions in world historical systems. In: Moderski G, Devezas T, Thompson W (eds) *Globalization as evolutionary process: Modeling, simulating, and forecasting global change*. Routledge, London, pp 190–225
140. White H (2009) *Settable systems: An extension of Pearl's causal model with optimization, equilibrium, and learning*. Submitted to J Machine Learn Res. http://eclectic.ss.uci.edu/~drwhite/center/ppt_pdf/A_Comparison_of_Pearl_s_Causal_Models_and_Settable_Systems.pdf
141. White H, Chalak K (2006). A unified framework for defining and identifying causal effects. Department UCSD of Economics Discussion Paper http://www.economics.ucr.edu/seminars/spring06/econometrics/HalWhite_Lect_6-2-06.pdf
142. White H, Chalak K (2007) Independence and conditional independence in causal systems. http://eclectic.ss.uci.edu/~drwhite/center/ppt_pdf/Independence_and_Conditional_Independence_in_Causal_Systems.pdf
143. Wright HT (2000) Instability, collapse, and the transformation to state organization. *System shocks-system resilience*. Workshop Paper, Abisko, Sweden, May 20–23:2000
144. Wright HT (2006) *Atlas of Chiefdoms and Early States. Structure and Dynamics*: 1(4)1
145. Zuckerman EW (2003) *On Networks and Markets* by Rauch and Casella (eds). *J Econ Lit* 41:545–565

Human–Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development

LENORE LAURI NEWMAN
Royal Roads University, Victoria, Canada

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Placing Dynamic Sustainable Development in a Historical Context
 Feedback Loops and Reactive, Proactive, and Adaptive Management
 Path Dependence and Lock-in
 Future Directions
 Bibliography

Glossary

Emergence The appearance of complex structures or behaviors within a complex adaptive system that is unpredictable from the starting conditions or materials.

Lock-in The inability to change non-optimal behavior due to the engrained social, financial, or technical cost of changing the behavior.

Negative feedback Outcomes of an action within a complex adaptive system that tend to decrease the magnitude of the originating action.

Panarchy A model of behavior of complex adaptive systems proposing that such systems progress through cycles of growth, collapse, and renewal.

Path dependence The dependence on system behavior upon prior system behavior.

Positive feedback Outcomes of an action within a complex adaptive system that tend to increase the magnitude and impact of the originating action.

Resilience The ability of a complex adaptive system to maintain its form in the face of disruption.

Transformability The ability of a complex adaptive system to transform into a different state better suited to existing conditions.

Definition of the Subject

Dynamic sustainable development is a process-based environmental management theory based upon the recognition of human systems as complex adaptive systems that are, in turn, subsets of ecosystems that are also complex adaptive systems. Human systems and ecosystems exist in an evolving relationship, and as human systems grow in size adaptive management techniques must be employed to ensure that the containing ecosystems are not overwhelmed by human resource demands or by the waste products of human activity. Dynamic sustainable development is the latest step in an ongoing process of evolution within the larger sustainable development discourse away from goal-based, utopian models of sustainability to a process-based, iterative, adaptive approach.

Introduction

The concept of dynamic sustainable development arose as a means of reconciling the desire to create less environmentally damaging human enterprises with the complexity inherent in the field of resource management. Dynamic sustainable development suggests that there are two central pillars to the successful adaptive management of human–environment interactions; the need for resilience and the need for transformability. Resilience, defined as the magnitude of disturbance that can be absorbed before a structural change occurs within a system [30], has been identified as a desirable goal of adaptive management. However while a system should be able to maintain a degree of stability in the face of surrounding change, sometimes a system change is required for long term sustainability; the system must possess transformability, the ability to totally alter itself if needed [71]. This combination of stability and change represents a break with the goal-based models of steady-state human–environmental interactions more prevalent in earlier environmental discourses.

Some of the general principles of dynamic sustainable development are set out in [46]; these general principles draw on the continuing evolution of the general sustainable development concept. As such dynamic sustainable development is part of a larger trend to incorporate complex adaptive systems theory into environmental management. The first general principle is that *dynamic sustainable development is a process, not a goal*. This principle reflects a broader shift within the sustainable development discourse. In recent literature, sustainable development is described as a continuous process of change [36], and is described as a process that must be treated as an evolution of ideas [53]. This shift is well founded; complexity theory

has shown us that change is the norm in social and ecological systems. Any static or “climax community” will eventually fall prey to the “inevitable accident” [38]. Change is the process that allows evolving systems to thrive. Rammel notes, “... there cannot be any best state, or stable equilibrium, or optimal path of development” [51].

One of the key figures of the shift from goal to process within the environmental management and planning field is C.S. Holling, whose contributions include panarchy theory and many of the formative writings on adaptive management. The panarchy theory was developed to help clarify the role of change within adaptive systems. The term was chosen as a reference to the Greek God of nature, Pan, and as a reference to the combination of change and persistence found within complex systems. Panarchy theory focuses on the interplay between different levels of activity within a system and outlines an evolutionary cycle of growth, collapse, and rebirth that systems undergo. Different time scales within systems are also highlighted: fast scales of events experiment and test, while slower levels stabilize and conserve. The interplay of these scales and levels create dynamic structures. Holling argues that seeing sustainable development in the light of complex adaptive systems resolves the critique that sustainable development is an oxymoron:

Sustainability is the capacity to create, test, and maintain adaptive capability. Development is the process of creating, testing, and maintaining opportunity. The phrase that combines the two, ‘sustainable development’ thus refers to the goal of fostering adaptive capabilities and creating opportunities. It is, therefore, not an oxymoron but a term that describes a logical partnership [29].

Holling’s vision of sustainable development is an iterative one; it can emerge organically from unsustainable behavior in manageable steps. Norms cannot be imposed in advance [56], but will emerge as part of an adaptation process. Instead of being a final objective, sustainable development has to be understood as a continuous process of change [36], and a fruitful approach to this process is to treat it as an evolution [53]. Treating sustainable development as a process creates the need for an indefinite program of monitoring and adjustment. Every successful adaptation is only a temporary “solution” to changing selective conditions [53]. This principle of sustainable development means that it is always a moving target [60].

The second principle of dynamic sustainable development is that *dynamic sustainable development must cope with the inherent unpredictability of the systems it addresses*.

Uncertainty is inherent to complex adaptive systems; in particular the behavior of ecosystems and social systems are notoriously difficult to predict. These systems manifest emergent properties; human society in particular is highly heterogeneous, far from equilibrium, and complexity is our society's defining feature. Human society is also highly non-ergodic. Ergodicity is the tendency of a system to move towards equilibrium, maximizing entropy. Human societies do not settle down into stable patterns for long. They constantly innovate, grow and change, posing a challenge for those trying to adjust our interactions with the biosphere.

Though we might wish to design a perfect and stable society, history suggests such experiments end in failure. Sustainable development models must therefore be robust enough to mitigate the ecological effects of a non-ergodic society. Positive feedback loops, which will be discussed in detail later, allow accidents of history to get magnified in outcome [70]. This leads to many results of small actions being unintended [35] and unpredictable from the initial conditions. Our predictions of the future are at best temporary guides, leaving us in the need to iteratively monitor feedback loops and continually adjust our models and our actions accordingly. This inherent unpredictability represents a “strong uncertainty”; not only are we unable to predict the consequences of events we are unable to determine which events are the ones that will lead to future change [65]. To use the language of complex adaptive systems theory, when human societies interact with natural systems they show sensitive dependence on initial conditions. The changes that can arise in a complex system involving society and the environment can be particularly perplexing, as they can involve changes in human knowledge and awareness, changes in technology, and also changes in public perception [24]. These three aspects can all be present at the same time and interact.

The third guiding principle of dynamic sustainability is that *innovation processes greatly affect sustainable development initiatives*. Early models of environmental societies were almost always static, steady state societies that changed very little over time. Goal orientated utopian models of environmental action range from Skinner's “Walden Two” [65] to the steady state economics proposed by Herman Daly [14] to the models presented in the landmark document “Limits to Growth” [43]. These models, however, ignore innovation as a fundamental component of human society. We use technical ingenuity to create new technology, but social ingenuity reforms old institutions and social arrangements into new ones [31]. Managing our interaction with the Earth's ecosystems would be much easier without the complicating factor of innovation

constantly changing the nature of this interaction, but this process is constantly occurring on a number of scales. At the smaller scale we see incremental innovations, which are small refinements that occur relatively continuously. At a larger scale, there are radical innovations representing large shifts in technologies. These are not predictable, and may happen at any time. There are systematic innovations that can create entirely new fields [48]. Such sudden shifts can provide new technologies to protect ecosystems, can shift our resource use from one resource base to another, and can also increase our impact on ecosystems in new and unexpected ways.

The process of innovation should be rather familiar to those who work with complex systems; Complexity has been called the “science of surprise” precisely because unpredictable behaviors and structures are emergent within complex systems. Emergence is the creation of new behavior and properties that cannot be understood from an investigation of the system's parts [20]; emergent behavior underlies the need for transformability within systems as the need for sudden changes to preserve the integrity of ecological systems and social systems can arise at any time. Lewin calls emergence “the central feature of the new science of complexity”, p. 175 in [40]. It is certainly one of the more surprising features of complex adaptive systems. Emergence can be observed in the biosphere by even a casual investigator. As ecologist Chaia Heller says, a seed doesn't grow; it becomes something new p. 106 in [27]. An emergence process drives the agricultural processes that allow us to harness energy and survive. C.D Broad, a philosopher, coined the term emergence in the 1920s, to describe those things that appear even though there is no hint of them at lower levels, p. 28 in [9].

The problem is that while it is easy to understand effects that can be intuited from the behavior of the system, predicting emergent behavior is a different matter. This unpredictability deeply affects the quest for sustainable development. The existence of emergent behavior requires our definitions of sustainability to be dynamic emergent behavior can provide sudden unexpected innovation, and it can also spawn bursts of ecological destruction. For this reason, the existence of emergent structure raises the need for safety factors and resiliency.

The prominence of emergence and innovation do not sit well with many members of the environmental movement. There is a historical uneasiness with innovation that informs the development of sustainable development dialogs. Though technology can be seen as an “adaptive answer” to problems [53], there is a fundamental disconnect between the world of the information society and the groundings of sustainable development due to differing

values held by the actors involved [36]. This uneasiness is made more acute by the inherent uncertainty in the process of innovation [7].

Placing Dynamic Sustainable Development in a Historical Context

The application of complex systems theory to questions of resource management and environmental preservation was first attempted relatively early in the twentieth century, long before the concept of sustainable development itself was developed. What was then called general systems theory addressed environmental issues under the assumption that human organizations could not be successfully analyzed by only considering their parts. In effect, systems theory reverse engineered the study of complexity from the study of specific complex adaptive systems associated with human society.

General systems theory arose from the necessity of managing complex technologies in difficult situations. General systems theory sought common principles in the structure and the operation of systems of all shapes and sizes [63], a quest that continues in the field of adaptive resource management. In 1950, Ludwig von Bertalanffy, one of the founders of general systems theory, brought together several of his ideas to argue that systems have properties independent of discipline. Von Bertalanffy felt ever-increasing specialization was not useful to the study of such systems [69]. The study of cybernetics had already proven useful during the Second World War, and there was a growing movement among ecologists to consider systems in a transdisciplinary way.

In 1954, Anotol Rapoport, Ludwig von Bertalanffy, Kenneth Boulding, Ralph Gerrard, and James Miller founded the Society for General Systems Research [25]. Von Bertalanffy had already introduced the concept of general systems theory in 1937. He understood that social systems are far from equilibrium and coined the term “open system”.

General systems theory is devoted to understanding function of systems, including organizations [69]. It was meant to be a very practical study, and has been applied widely to the understanding of business organizations. Boulding, who coined the term “spaceship earth”, divided study into special systems theory, which focused on modeling, and general systems theory, which was a broader philosophical inquiry into the overall dynamics of social systems [25].

Several of the elements of complex adaptive systems studied by physicists have also been studied by systems theorists. Erich Jantsch felt dissipative structures were of

particular importance to the understanding of human systems [34], as they are far from equilibrium but resilient. He felt dissipative structures grow to limits governed by internal, not external constraints, a belief that reflects a general interest among systems theorists in autopoiesis, the process by which dissipative structures continually renew and regulate themselves in order to maintain themselves. The Society for General Systems Research contributed to the *Limits to Growth* Report, which is discussed below as a foundational document of sustainable development.

An unlikely contributor to the incorporation of complex systems theory within environmental discourse emerged from the mathematical study of discontinuous phase transitions. Mathematician Rene Thom developed a method for examining such discontinuous changes in mathematical and physical systems, a method he came to call catastrophe theory. In Thom’s theory, a catastrophe is the discontinuous transition between stable but divergent paths [73]. Thom based the theory on rules of topology, and the assumption of underlying structural stability. This assumption relies on the presence of unseen order; Thom assumed there was a hidden order beneath the discontinuity that he could not quantify. He also understood the uniqueness of each region of discontinuity. Even though it went against the grain of traditional model of mathematics, Thom insisted catastrophes were local, and that no global models of such systems exist, p. 7 in [66].

Catastrophe theory provided a new way of thinking about change, and gave a metaphor for the abrupt changes seen in natural and human systems, unlike the continuous, linear models used at the time. The theory thus represented a rather significant mathematical breakthrough, and enjoyed a spectacular rise in general popularity, perhaps due to its very striking name. Catastrophe theory was quickly trumpeted as describing the actual workings of natural and social systems. Controversy began to surround the theory as it was used as a blanket metaphor for a wide range of systems. The application of catastrophe theory became a craze that spread far beyond the mathematical sciences. Unfortunately, the application of catastrophe theory to social and environmental issues could not actually live up to the hype surrounding it. This was partly due to the difficulty of the mathematics underlying the theory, and partly due to the fact that many of the social systems it was applied to did not lead at the time to quantifiable models [20]. The bubble of popularity Thom and his theory enjoyed collapsed. However the brief flirtation with this theory planted a seed of interest in complex adaptive systems within practitioners of ecological management.

The evolution of the sustainable development concept itself really began in earnest with the publication of *Limits to Growth* by the Club of Rome, group which was founded by British scientist Alex King and Italian industrialist Aurelio Peccei. The first meeting of roughly forty international thinkers occurred on the seventh of August, 1967 in Rome, Italy. The *Limits to Growth* report features computer models of resource use created by Donella and Dennis Meadows. Their most advanced model, World Three, shows three possible futures: overshoot and collapse, overshoot and oscillation, and sigmoid growth, p. 123 in [44]. The futures predicted in World Three were highly pessimistic of future progress. Overshoot refers to a society that is using resources and creating wastes at a rate faster than can be supplied or absorbed by the biosphere. This crude beginning founded an ongoing movement to build complex models of human–ecological interactions.

Limits to Growth focused on what the authors called the “world problematique”. This group of linked problems included poverty amid plenty, environmental destruction, urban sprawl, and economic problems p. 10 in [43]. In the opinion of Meadows and her co-authors, we are far past the point at which overshoot has occurred. Her models commonly predict overshoot and collapse. In short, not only must our societies not grow any more, they must contract.

Response was immediate and was highly polarized. It was pointed out that World Three underestimates the ability of technology to postpone catastrophe [12]. The models rely on tables composed of extrapolated data generated through an iteration process, causing runaway errors, p. 32 in [12]. The *Limits to Growth* report was heavily critiqued for ignoring innovation [10]. In many ways what was missing from the models were non-linear effects such as feedback and emergence.

When the Club of Rome published *Limits to Growth* in 1970, it sparked intense international debate over its basic premise, which challenges the basic tenets of traditional economics. The prevalent economic vision sees the economy as an isolated system: a circular flow of exchange value between firms and households. The economy is the system of interest and natural systems are simply sources of resources and sinks for wastes. Nature may be finite, but these natural sources and sinks can be infinitely substituted for by human capital, without limiting overall growth in any important way.

The problem with exponential growth in a finite system is simple enough to describe; the economy strains nature with respect to sources and waste sinks [32]. In effect, the finite expanse of the natural world has a carrying capacity that can be expressed in different ways; economic

carrying capacity is the maximum global economic welfare derivable from the sustainable throughput flows of the ecosphere [72].

Following *Limits to Growth* it became popular to speak of a finite and firm upper limit to the scope of human activity within the biosphere, and to imagine static societies that would exist at an equilibrium point with the biosphere that respected clear limits. The first Earth Day was held in 1970, attracting 20 million people to peaceful demonstrations across the US. The deteriorating state of the Great Lakes prompted a number of clean-up efforts. Many governments began to allocate resources to establish ministries of the environment. Of particular importance was the founding of Greenpeace in 1971. Greenpeace mustered grass-roots support for campaigns against clear-cut ecological abuse. Spurred by public interest in the environment, several major commissions and reports established by various levels of government began to highlight the environmental damage caused by economic growth. The Stockholm Conference of 1972 focused attention on the growth of industrial pollution. The *Global 2000* report to President Carter of the United States warned that “if present trends continue, the world in 2000 will be more crowded, more polluted, less stable ecologically . . . despite greater material output the world’s people will be poorer in many ways than they are today”, p. 1 in [5].

The conservation movement helped to fuel the debate over humanity’s appropriation of natural resources and the physical limitations scarcity posed to economic growth. However, the early movement was often fixated with scaling back human activity to remove the complexity from our interactions with the environment, rather than addressing issues through a complex adaptive lens. In 1983, the United Nations assembled the World Commission on Environment and Development, and charged them with the task of gathering opinion on the state of the environment and its potential effect on development. In 1987, the group produced *Our Common Future*, which introduced the concept of sustainable development into common use. Known as the Brundtland Report in reference to Chairperson Gro Harlem Brundtland, this document drew on scientists, government and communities to pinpoint the resource issues most pressing to global development. The Brundtland report set out as an objective the achievement of a sustainable society by the year 2000, and quickly became a milestone in the discussion of Sustainable Development. *Our Common Future* also called for a conference to be called in five years time to set out an action plan for the achievement of these goals.

The lasting effects of the Brundtland report were mixed, and reflect the tensions between those still work-

ing within a goal-based mindset and those arguing for process-based management. *Our Common Future* created consensus between many different stakeholders, and formalized many of the assumptions now taken for granted in the discussion of sustainability. Two of these are particularly important; the Brundtland report made clear the link between economic and ecological health, condemning the treatment of the natural sphere and the human sphere as separate entities. The report defines development as what we do in an attempt to improve our lot within these conjoined spheres. *Our Common Future* begins by giving a very vague definition of “sustainable development”, saying simply that it should “meet the needs of the present without compromising the ability of future generations to meet their own needs” [8]. The Brundtland report puts forward many initiatives designed to reach a state of sustainability, which was still framed in a goal-oriented way. They call for action on all levels, including monitoring at the global level. They ask that “the ability to anticipate and prevent environmental damage requires that the ecological dimensions of policy be considered at the same time as the economic, trade, energy, agricultural and other dimensions”, p. 10 in [8]. In short, they wish for less reactive repair of environmental problems and more proactive monitoring to prevent environmental damage.

The lasting importance of the Brundtland report is due to several factors. The commission’s use of roundtables and participatory democracy allowed the widest possible input of voices to be heard, leading to a document that spoke to a variety of stakeholders. The Brundtland report made it clear that sustainability is a needed goal, but left the definitions vague enough to allow further discussion. The report also called for a follow up meeting to be held five years later at which a roadmap to sustainability would be set out.

The years following the publication of *Our Common Future* saw a peak in environmental concern, cumulating in the publication of *Agenda 21*, the follow-up report to *Our Common Future* [58]. *Agenda 21* is a broad action plan adopted at the 1992 Rio Conference to promote environmentally sound and sustainable development in all countries of the world. *Agenda 21* was signed on 13 June 1992 by over one hundred heads of state representing ninety eight percent of the world’s population. *Agenda 21* is not legally binding; it is a flexible guide for achieving a sustainable world.

Agenda 21 is divided into six themes composed of sub-areas with specified action plans. The first theme, quality of life, addresses areas such as limiting poverty, changing consumption patterns, controlling population growth, and ensuring the availability of adequate health care. The

second theme, efficient use of resources, focuses on land use planning, water conservation and management, energy resources, food production, forest management, and the protection of biodiversity. The third theme, protection of the global commons, discusses management of the atmosphere and the oceans. The fourth theme, management of human settlements, considers urban issues and the provision of adequate shelter. The fifth theme, waste management, focuses on the classification and disposal of chemical, solid, and radioactive wastes. The final theme, sustainable economic growth, discusses trade, development, and technology transfer.

Agenda 21 has been criticized for not including strong positions on transport, energy issues, and tourism. The action plan has also been criticized for being too focused on increasing trade. *Agenda 21* stresses that removing distortions in international trade is essential and environmental concerns should not restrain trade, positions that are critiqued by antiglobalisation groups.

The success of *Agenda 21* has been mixed. The action plan has been successful at linking environment and poverty, and many local working groups have been formed that include the multiple stakeholders such as youth, indigenous people, scientists, and farmers called for in the plan. The Commission on Sustainable Development, which is charged with monitoring the progress of *Agenda 21*’s implementation, has reported positive developments in the areas of controlling population growth, increasing food production, and improving local environments. However they also report an increase in inequality, increasing water scarcity, and extensive loss of agricultural land. Implementation in the European countries has been more successful than in other regions.

The action plan’s mixed success can be attributed in part to a lack of commitment of funding for the initiatives in the plan. Every action in *Agenda 21* included a projected cost; but no source of funding was secured at the time of signing. The Commission on Sustainable Development continues to monitor the implementation of *Agenda 21*, and follow-up meetings known as the Earth Summits and nicknamed Rio +5 and Rio +10, were held in 1997 in New York and in 2002 in Johannesburg. In 2002 the United Nations General Assembly called the progress of *Agenda 21*’s implementation extremely disappointing, and at Johannesburg a plan was developed to speed the implementation of *Agenda 21*.

Agenda 21 and *Our Common Future* set out as a goal the achievement of a sustainable world by the year 2000. This goal was not achieved, though much good did come out of these documents. This failed legacy will also shape any attempt to form a dynamically sustainable society. We

can learn a lot by considering this large attempt at building a sustainable world. The importance of securely funding initiatives is made clear, for example. Also, the importance of entrenching a culture of sustainability at the individual level is made clear as well. *Our Common Future* and *Agenda 21* were largely off the radar of popular culture, and so did not get the attention they needed to move forward.

Though *Our Common Future* propelled the concept of sustainable development into the public eye, the vague definition given within the report of sustainable development led to a long and ongoing debate over what sustainable development actually means. Since the introduction of sustainable development into common parlance, numerous variations have emerged, such as sustainability, sustainable growth, sustainable economic growth, and sustainable environmental or ecological development. Although disagreement exists among different sectors and communities about the usefulness of the concept of sustainable development, it is recognized internationally and it does avoid most of the traditional left-right polarization and discourse about growth versus no-growth, by bringing together the terms sustainable and development in a constructive ambiguity that has stimulated greater dialog between sectors. Despite its ambiguity, it has succeeded in uniting widely divergent theoretical and ideological perspectives into a single conceptual framework [19]. More fundamentally, it has brought a wide diversity of industrialists, environmentalists, public policy practitioners and politicians to round tables, in their attempts to define, deal with and actualize the concept. In order to provide some contextual appreciation, a brief examination of some of the earlier definitions of sustainable development follows. Human societies everywhere will place a different emphasis on the former and on the latter, according to their ecological, social and economic conditions.

Though the term was popularized by the Brundtland commission, it predates *Our Common Future*. In 1980, the World Conservation Strategy, IUCN, UNEP, WWF, and others argued that integration of conservation and development is particularly important, because unless patterns of development that also conserve living resources are widely adopted, it will become impossible to meet the needs of today without foreclosing the achievement of tomorrow's needs. Meadows et al. [44] defined a sustainable society as one that had in place informational, social and institutional mechanisms to keep in check the positive feedback loops that cause exponential population and capital growth. In other words, means that birth rates roughly equal death rates, and investment rates roughly equal depreciation rates, unless and until technical changes and so-

cial decisions justify a considered and controlled change in the levels of population or capital. In order to be socially sustainable the combination of population, capital and technology in the society would have to be configured so that the material living standard is adequate and secure for everyone. In order to be physically sustainable the society's material and energy throughput has to meet economist Herman Daly's [15] three conditions: its rates of use of renewable resources do not exceed their rates of regeneration; its rates of use of nonrenewable resources do not exceed the rate at which sustainable renewable substitutes are developed, and its rates of pollution emission do not exceed the assimilative capacity of the environment. Some scholars differentiate between sustainability and sustainable development, while others use them interchangeably. Sustainability derives from the Latin root *sus-tinere*, which literally means to under-hold or hold up from underneath. Sustainability describes a characteristic of relations (states or processes) that can be maintained for a very long time or indefinitely [37].

The outcome of this dialog produced a generic, rather than a specific definition for sustainable development. It is a process of reconciliation of three imperatives. These are the ecological imperative to live within global biophysical carrying capacity and to maintain biodiversity, the social imperative to ensure the development of democratic systems of governance that can effectively propagate and sustain the values that people wish to live by, and the economic imperative to ensure that basic needs are met worldwide [13,57]. This definition, however, remains general enough to allow for sustainable development to be interpreted differently in specific socio-geographic situations and to be sufficiently responsive in the face of unpredictable change and uncertainty. It also responds to the dynamic interplay between the imperatives, reflecting the complexity of modern human society.

The move from the above "three pillar model" to a process oriented version of the same model is largely a product of the last ten years. Various organizations have fed this shift by studying ecological systems with complex adaptive techniques. Perhaps the most well-known of these organizations is the Resilience Alliance, a multidisciplinary group based in the United States that focuses on the exploration of the dynamics of social-ecological systems. Their mission is to assemble knowledge that embraces resilience, adaptability and transformability and that influences sustainable development policy and practice. The Resilience Alliance publishes the journal "Ecology and Society", which was started by C.S. Holling under the name "Conservation Ecology". The Society for Human Ecology has a similar focus as the Resilience Alliance,

and holds one of the central annual conferences in the field. The journal is one of the central ones in the areas of sustainable development as process, along with the journals *Sustainable Development* and *International Journal of Sustainable Development*. On the modeling side of the field, the Santa Fe Institute is a definite leader. The Santa Fe Institute was founded as an independent multidisciplinary research institute to study complex adaptive systems theory and how it can be applied to address environmental, technological, biological, economic, and political challenges. These centers of knowledge creation continue to fuel the robust development of dynamic sustainable development.

Feedback Loops and Reactive, Proactive, and Adaptive Management

The next few sections look at the complex adaptive features that shape dynamic sustainable development; the first of these is the existence of feedback process within social-environmental systems. Complex adaptive systems are far more than a collection of elements; they are bound together by the flow of energy, matter and information. This flow is often two-way, forming feedback loops within the complex system. Feedback loops are a central feature of all complex adaptive systems, and an understanding of feedback loops is critical to the understanding of human complex adaptive systems. Achieving sustainability is fundamentally a question of observing and responding to feedback. Feedback allows control within complex adaptive systems, and also allows growth. Feedback loops form the nervous systems of complex adaptive systems, allowing the flow of information between elements and between the system and the environment.

Feedback is a process in which a change in an element alters other elements, which in turn affect the original element [35]. Feedback is an iterative process, and is a fundamental part of what makes a system both complex and adaptive.

There are two main types of feedback within complex systems: positive and negative feedback. Negative feedback loops moderate a system, but this process does not always lead to stability. Too much negative feedback can cause a system to become stagnant and unable to adapt to suddenly changing situations. A system composed only of negative feedbacks will become out of step with its surrounding environment and perish.

In Waldrop's opinion, it is the mixture of positive and negative feedback that creates complex systems [70]. Negative feedback provides stability [35] that holds systems away from run-away growth and collapse. Negative feed-

back is often associated with the concept of equilibrium. As an example, a thermostat uses negative feedback to maintain homeostatic equilibrium, keeping the surrounding area at a stable temperature. Most systems, thus, have built in processes that can bypass negative feedback in emergency situations. Writer and urban planner Jane Jacobs provides as an example of such an "override" mechanism: our reflexive ability to stop breathing upon being immersed suddenly in water, p. 115 in [33].

Positive feedback is the driving force behind sensitive dependence on initial conditions, which is also called the "butterfly effect" after an example given by Lorentz in a 1972 talk entitled, "Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?" [41]. The concept of positive feedback can be difficult to grasp, as we tend to believe in a "conservation of complexity" [11] in which simple causes lead to simple effects, and complex causes lead to complex effects. This is a restatement of Newton's second law: *for every action there is an equal and opposite reaction*. This correlation does not hold in complex adaptive systems. Positive feedback can reinforce a small event again and again until it becomes a system-wide phenomenon.

Positive feedback allows growth, and fuels expansion and diversity [33,35]. As Waldrop comments, positive feedback loops allow accidents of history to get magnified in outcome [70]. If negative feedback loops hold a system stable, positive feedback loops allow systems to explore their environment and follow new paths of development. As they magnify random small variations, positive feedback loops add an element of surprise to the system's behavior. This leads to many results of small actions being unintended and unpredictable from the initial conditions. A forest free from disturbances will evolve towards an equilibrium state that was once called "climax" ecology. Little changes in such an ecosystem until a positive feedback process occurs; such as a tiny blaze that builds upon itself to raze the forest and make room for new growth. By its nature, positive feedback has the ability to lead complex systems into precarious territory. Positive feedback loops are either dampened by negative feedback loops or they crash [33]. The ability to dampen positive feedback loops is necessary to the survival of complex adaptive systems.

Both positive and negative feedback take a non-zero amount of time to propagate through a complex system. If this time is overly large the feedback loop may be ineffective at regulating the system or might have an unintended effect on the system. Delayed feedback is quite common within natural ecosystems' responses to human stresses, as the two complex systems function on such different time scales. Delayed feedback poses a large challenge to the

achievement of dynamic sustainability. As it might take us some time for the biosphere to respond to our actions, we can suffer creeping environmental decay. Damage occurs slowly and by the time we notice, it has already become severe.

The existence of positive feedback and sensitive dependence on initial conditions within society has profound consequences for sustainable development. As we can never trust our predictions of the future entirely, there can be no perfect model of a sustainable society that will hold up for all time. Instead, we must monitor feedback loops carefully and continually adjust our models and our actions accordingly. In the larger context of dynamic sustainable development, negative feedback processes provide resilience and positive feedback processes provide transformability.

Within the environmental movement there is a great desire to move away from the reactive approach of dealing with environmental problems after they develop to a proactive approach in which feedback processes are anticipated before they begin.

One proposed method used to mitigate uncertainty is to use what is called a precautionary principle. The concept of a precautionary principle is often credited to the German principle of *Vorsorgeprinzip*, or foresight planning, which began to receive attention in the 1970s [45]. The concept has evolved over time; what began as a “measure” shifted to an “approach” and finally to a “principle” [1].

In the United Nation’s Rio Declaration, the use of a precautionary principle is urged. Principle 15 of the Rio Declaration states that where there are threats of serious or irreversible damage, lack of full scientific certainty shouldn’t be used as a reason for postponing cost-effective measures to prevent environmental degradation [45]. A stronger definition known as the “Wingspread definition” emerged from a conference of environmental thinkers in 1998. The Wingspread definition of the precautionary principle states that when an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically [49].

Intuitively, the precautionary principle is straightforward, but it is notoriously difficult to apply [1,49]. The general idea of the precautionary principle is to avoid serious and irreversible damage [64]. As Raffensperger states, the precautionary principle can be used to prevent, not just redress, harm [52]. What is simple to describe, however, is not necessarily simple to put into use. There must be some evidence a hazard exists if the precautionary principle isn’t

to lead to ruling out any action [61]. If the precautionary principle is not to stifle progress, it should be coherent, utilize known information and theories, have explanatory power, and possess simplicity [55]. Low complexity solutions should be preferred to high complexity solutions if the precautionary principle is to avoid simply creating further problems [64]. And given that not all results of an action within a complex system can be predicted before they occur, no purely proactive management regime will be sufficient.

A solution to this problem that is gaining in popularity is a combination of proactive and reactive management known as adaptive management. Adaptive co-management involves the combination of proactive and reactive management techniques through an iterative process of feedback monitoring. Adaptive management keeps careful track of uncertainties within the system, partially through local control of day to day ecosystem management, as it is assumed that undesirable effects will appear locally first. Adaptive co-management seeks to include past, present and future stakeholders. Adaptive co-management is an emergent and self-organizing.

Adaptive co-management attempts to correct for runaway feedback through a place-based approach. Local voices are incorporated into Adaptive co-management of ecosystems in which both local actors and larger level effects are considered is critical to the creation of resilience and transformability. Use of local ecological knowledge builds resilience [6]; local solutions tailored to local conditions are necessary for healthy ecosystem interactions, and if many variables are to be monitored changes in those variables will first be observed “on the ground” locally. Without local involvement, management tends to shift towards exploitation; as local resources are depleted new resources are substituted in other locations.

The need for local knowledge and observation is not, however, an argument in favor of moving to locally isolated small-scale enterprise. As even local action can have global consequences, resilience emerges from both cross-scale and within scale interactions [47]. Transformability in the face of external changes requires an outward focus to the larger scale. Connectivity allows resilience and movement, but the existence of local network structure buffers against cascades of disaster from the larger world [2]. That said, for adaptive co-management to work it must be collaborative; without a shared sense of purpose stakeholders at different scales are likely to have very different interests [26]. In some cases global stakeholders are likely to value short-term financial gain over local ecosystem integrity, but in other cases local actors might value short-term employment prospects over larger ecological

needs. Actors on all scales must work in concert, not at cross-purposes.

Complex systems are filled with uncertainty, and no amount of precaution will eliminate all risks. Management must, thus, build system resilience. There are several factors that influence the resilience of a system. These include the system's latitude, which is the maximum amount of stress that a system can absorb without changing to a new state, the system's ability to resist change, and its precariousness, which refers to how fragile it is [71]. The more resilient an ecosystem or society is, the better it will be at responding and adapting to unpredictable changes.

There are several ways system resilience can be increased. Primarily, we can increase resilience by ensuring that as we undertake a course of action we leave room for alternatives. Preventative measures should allow for more flexibility in the future [22]. Especially in cases of irreversibility, options should be kept open [4]. We can increase resilience through increasing the buffering capacity of the system, managing for processes at multiple scales, and nurturing sources of renewal [23].

Path Dependence and Lock-in

A critical feature of complex adaptive systems that has direct application on human–environment interactions is the occurrence of branching points during the evolution of complex systems. These points occur when the system can evolve in multiple and exclusive ways. Such systems are thus path dependent; the state of the system depends on its earlier history. Path dependence is important to the study of complex adaptive systems such as human societies and ecosystems because it limits the ability to act reactively when attempting to intervene in a complex system. The field of restoration ecology was one of the first to encounter this aspect of complex systems. To restore a damaged ecosystem, one cannot simply plant the species that were found in the undamaged ecosystem; one must plant the pioneer species for that ecosystem and then allow the ecosystem to follow a succession cycle to the desired state. To do this, one must know the ecosystem's *trajectory*; the history of its succession over time. Determining this can be very difficult as the pioneer species are often not found in the desired mature ecosystem. In short, the status of a system does not only depend on the state of a set of variables, but how they were reached [35]. Path dependence makes the job of understanding complex adaptive systems much more difficult.

Path dependence is central to the development of ecosystems and human development. Historical records are, thus, critically important. Knowledge gathered over time

locally can provide historical knowledge that is not accessible to more abstract methods of data collection [21]. Local knowledge allows for adaptive management of local resources [6] precisely because it captures the on-going path dependence of complex systems.

Local complex problems require general scientific knowledge, but local knowledge plays a complimentary role. That said, the needed historical data are not often easily found; most local knowledge is passed on orally and is based upon generations of observation; this knowledge as younger generations move to urban areas and older generations die. Practitioners have uncovered a wealth of unlikely environmental information locally, including records of planting times, bird logs, temperature records, and species lists.

Local knowledge can add another layer of uncertainty even as it provides crucial data. There is an assumption that locals have intimate knowledge of the local environment, but not all locals have the same knowledge [16]. Experts can usually be identified by asking locals who holds the greatest knowledge locally. Using local knowledge can be frustrating for all involved [21], and thus must be used appropriately. Local knowledge can determine changes in local conditions, outline ecological trajectories, and identify path dependent processes.

Path dependence can be described as “reactive sequences” in which each event precipitated by previous self-reinforcing sequences [42]. In short, history matters, and a random event can ensure a suboptimal technology or process becomes the norm. Rammell points out sometimes rather mediocre solutions dominate a natural selection process in the short term (2003); systems, particularly ones of great complexity, can prove very inflexible. Arthur calls this re-enforcement of certain historical paths nonergodic behavior; in his words, path dependence matters [3] deeply. Though this problem could be minimized by careful use of precautionary principles at the beginning of the development of a technological path, in many cases negatives of new technologies tend to appear after implementation [39]. The ozone-depleting properties of CFCs are a good example of a technology that had to be completely replaced due to an unforeseen danger that became apparent after lock-in had occurred. Identifying a problem and choosing a solution is difficult, but it is often, as Homer-Dixon notes, implementation that is the true problem, p. 23 in [31].

From a practical standpoint, the largest impact of path dependence is that it actively constrains the actions one can take to implement sustainable development programs. Even if one is carefully monitoring feedback, and a signal arrives within a system suggesting change is needed,

the needed change might not be possible. The problem of technologies' ideas and behavior patterns becoming entrenched is known in the literature as lock-in. Lock in occurs because within complex societies innovations do not stand alone, they co-develop as entire networks of supporting technologies, much as keystone species co-develop in nature. If we suddenly need to change one of these keystone technologies we can find there is significant social and economic reluctance to do so. Diamond calls this the "sunk cost effect" as we are reluctant to abandon what we have even if it doesn't work, p. 432 in [18]. Lock-in arises naturally out of two properties of complex systems; path dependence and increasing returns. One cannot simply change them without setting off cascading changes throughout the system.

Recognition of this problem is not new. In his work on the need to shift away from fossil fuels, Unruh calls lock-in a technological "cul de sac" that at its worst cumulates in an embedded techno-institutional complex [67]. There is a need to tackle path dependence and lock in [53], but as Unruh notes "the question of how to overcome lock-in has been little explored" [69]. Lock-in has been related to the issue of diversity discussed as condition three: the encouragement of niche markets is a possible way of breaking lock in [68]. However research has shown that in extremely locked in systems, research into options falls to near zero [54]. As Arthur notes, there is a minimum cost for a transition and changing by fiat sometimes necessary [3].

Future Directions

The role of complex systems theory in sustainable development continues to develop. At the qualitative level the greatest challenge is an educational one; factors such as feedback loops, sensitive dependence on initial conditions and path dependence are not widely understood by decision makers. One of the largest policy challenges will be understanding and controlling lock-in within complex markets. It is likely that the leaders in the field will continue to be found in the areas of ecosystem management led by members of groups such as the Resilience Alliance. Given the difficulty of the underlying mathematics and physics, knowledge of complex systems theory is likely to remain outside of the general knowledge employed by sustainable development practitioners aside from qualitative models and metaphorical applications of general concepts such as self organization and feedback. Fostering this understanding of general concepts should be encouraged.

If sustainable development initiatives are to create the kind of changes needed to ensure the survival of our soci-

eties and the ecosystems they reside in, the move to adaptive and process-based environmental management will need to proceed at a faster pace. It has been rightly remarked that the transition to sustainable development is alarmingly slow [52], partly due to the magnitude of the changes required.

The greatest factor in the continuing development of quantitative complex models in this field is the ongoing increase in available computational power. New computational techniques such as evolutionary computing [17] will also increase the range of techniques for modeling the sustainability of complex systems. These computational techniques could add needed rigor to popular sustainable development tools such as the Natural Step and the Ecological Footprint, both of which are widely used and are based on iterative processes that would benefit from complex systems approaches. It is likely the Santa Fee institute will continue to play a leading role in approaches to complex modeling. Areas that will benefit from these technological advances could include the management of electricity grids that include a large number of small intermittent renewable energy sources, traffic management to reduce emissions, and climate modeling.

The lasting influence of sustainable development lies in its ability to evolve as a concept. Given the advances in the understanding of complex adaptive systems and the application of this understanding to ecological and social systems, there are likely many fruitful avenues that combine sustainable development and complex systems theory. In the long run, the most successful sustainable development initiatives will likely look rather a lot like ecosystems; diverse, complex, and evolving.

Bibliography

Primary Literature

1. Adams M (2002) The precautionary principle and the rhetoric behind it. *J Risk Res* 5:301–316
2. Anderson E (2006) Urban landscapes and sustainable cities. *Ecol Soc* 11(1):34. <http://www.ecologyandsociety.org/vol11/iss1/art34/>
3. Arthur B (1994) Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbor
4. Arrow K, Fisher A (1974) Environmental preservation, uncertainty, and irreversibility. *Q J Econ* 88:312–319
5. Barney G (1980) The global 2000 report to the president. US Government Printing Office, Washington
6. Berkes F, Colding J, Folkes C (2000) Rediscovery of traditional ecological knowledge as adaptive management. *Ecol Appl* 10:1251–1262
7. Buenstorf G (2000) Self-organization and sustainability: Energetics of evolution and implications for ecological economics. *Ecol Econ* 33:119–134

8. Brundtland G (1987) *Our common future: World commission on environment and development*. Oxford University Press, New York
9. Capra F (1996) *The web of life*. Anchor Books, New York
10. Cleveland C, Ruth M (1997) When, where, and by how much do biophysical limits constrain the economic process? *Ecol Econ* 22:203–223
11. Cohen J, Steward I (1994) *The collapse of chaos: Discovering simplicity in a complex world*. Penguin Books, New York
12. Cole H (1973) Introduction. In: Cole H, Freeman C, Jahoda M, Pavittet K (eds) *Models of doom: A critique of the Limits to Growth*. Universe Books, New York
13. Dale A (2001) *At the edge: sustainable development in the 21st Century*. UBC Press, Vancouver
14. Daly H (1991) *Steady state economics*, 2nd edn. Island Press, Washington
15. Daly H, Cobb J (1994) *For the common good*, 2nd edn. Beacon Press, Boston
16. Davis A, Wagner J (2003) Who knows? On the importance of identifying experts when researching local ecological knowledge. *Human Ecol* 31:463–489
17. De Jong K (2006) *Evolutionary computation: a unified approach*. MIT Press, Cambridge
18. Diamond J (2005) *Collapse: How societies choose to fail or succeed*. Viking Press, New York
19. Estes R (1993) Toward sustainable development: from theory to praxis. *Soc Dev Issues* 15:1–29
20. Eve R, Horsefall S, Lee M (1997) *Chaos, complexity and sociology*. Sage Publications, London
21. Fischer F (2000) Citizens, experts, and the environment: the politics of local knowledge. Duke University Press, Durham
22. Gollier C, Jullien B, Treich N (2000) Scientific progress and irreversibility: An economic interpretation of the precautionary principle. *J Pub Econ* 75:229–253
23. Gunderson L (2000) Ecological resilience: In theory and application. *Annu Rev Ecol Syst* 31:425–439
24. Hadfield L (1999) A Co-evolutionary model of change in environmental management. *Futures* 31:577–592
25. Hammond D (2003) *The science of systems: Exploring the social implications of general systems theory*. University Press of Colorado, Boulder
26. Hein L, Van Kloppen K, De Groot R, Van Lerland E (2005) Spatial scales, stakeholders, and the valuation of Ecosystem services. *Ecol Econ* 57:209–228
27. Heller C (1999) *Ecology of everyday life*. Black Rose Books, Montreal
28. Holling C (1976) Resilience and stability of ecosystems. In: Jantsch E, Waddington C (eds) *Evolution and consciousness: human systems in transition*. Addison Wesley, Reading, pp 73–92
29. Holling C (2001) Understanding the complexity of economic, ecological, and social systems. *Ecosystems* 4:390–405
30. Holling C, Gunderson L (2002) Resilience and adaptive cycles. In: Gunderson L, Holling C (eds) *Panarchy: Understanding transformations in Human and Natural Systems*. Island Press, Washington
31. Homer-Dixon T (2000) *The ingenuity gap*. Alfred A Knopf, New York
32. Hudson R (2005) Towards sustainable economic practices, flows, and spaces: Or is the necessary impossible and the impossible necessary? *Sustain Dev* 13:239–252
33. Jacobs J (2001) *The nature of economies*. Vintage Canada, Toronto
34. Jantsch E (1975) *Design for evolution: Self-organization and planning in the life of human systems*. George Braziller, New York
35. Jervis R (1997) *System effects: Complexity in political and social life*. Princeton University Press, Princeton
36. Jokinen P, Malaska P, Kaivo-Oja J (1998) The environment in an information society: A transition stage towards more sustainable development. *Futures* 30:485–498
37. Jude U (2000) Towards a culture of sustainability. In: Filho W (ed) *Communicating sustainability*. Peter Land, Frankfurt
38. Kay J, Regier H, Boyle M, Francis G (1999) An ecosystem approach for sustainability: addressing the challenge of complexity. *Futures* 1999:721–742
39. Konnola T, Unruh G, Carrillo-Hermosilla J (2006) Prospective voluntary agreements for escaping techno-institutional lock-in. *Ecol Econ* 57:239–252
40. Lewin R (1992) *Complexity: life at the edge of chaos*. Macmillan Publishing Company, New York
41. Lorentz E (1993) *The essence of chaos*. University of Washington Press, Seattle
42. Mahoney J (2000) Path dependence in historical sociology. *Theor Soc* 29:507–548
43. Meadows D et al (1972) *The limits to growth: A report for the Club of Rome's project on the predicament of mankind*. Universe Books, New York
44. Meadows D et al (1992) *Beyond the limits: confronting global collapse, envisioning a sustainable future*. McClelland and Stewart, Toronto
45. Morris J (2000) Defining the precautionary principle. In: Morris J (ed) *Rethinking risk and the precautionary principle*. Butterworth, Oxford, pp 1–14
46. Newman L (2005) Uncertainty, innovation, and dynamic sustainable development. *Sustain: Sci Pract Policy* 1(2):25–31. <http://ejournal.nbii.org/archives/vol1iss2/TOC.html>
47. Peterson G (2000) Political ecology and ecological resilience: An integration of human and ecological dynamics. *Ecol Econ* 35:323–336
48. Pereira P (1994) New technologies: Opportunities and threats. In: Salomon J, Sagasti J, Sachs-Jeantet C (eds) *The uncertain quest: science, technology, and development*. United Nations University Press, Tokyo, pp 448–462
49. Raffensperger C (2002) Precaution and security: The labyrinthine challenge. *Whole Earth* 113. http://findarticles.com/p/articles/mi_m0GER/is_2002_Fall/ai_93135763
50. Raffensperger C (2003) Constitutional experiments: protecting the environment and future generations. *Conserv Biol* 17:1587–1488
51. Rammel C (2003) Sustainable development and innovations: Lessons from the red queen. *Int J Sustain Dev* 6:395–416
52. Rammel C, Kastenhofer K (2006) Obstacles to and potentials of the societal implementation of sustainable development. *Sustain: Sci Pract Policy* 1(2):5–13. <http://ejournal.nbii.org/archives/vol1iss2/TOC.html>
53. Rammel C, Van Den Berg J (2003) Evolutionary policies for sustainable development. *Ecol Econ* 47:121–133
54. Redding S (2002) Path dependence, endogenous innovation, and growth. *Int Econ Rev* 43:1215–1248
55. Resnik D (2003) Is the precautionary principle unscientific? *Stud Hist Philos Biol Biomed Sci* 34:329–344

56. Robinson J (2003) Future subjunctive: Backcasting as social learning. *Futures* 35:839–856
57. Robinson J, Tinker J (1997) Reconciling ecological, economic and social imperatives: a new conceptual framework. In: Schrecker T (ed) *Surviving globalism Social and environmental dimensions*. Macmillan, London
58. Robinson N (1993) *Agenda 21: Earth's action plan*. Oceana Publications, New York
59. Saltelli A, Funtowicz S (2005) The precautionary principle: Implications for risk management strategies. *Human Ecol Risk Manag* 11:69–83
60. Salwasser H (1993) Sustainability needs more than better science. *Ecol App* 3:587–589
61. Sandin P, Peterson M, Hansson S, Ruden C, Juthe A (2002) Five charges against the precautionary principle. *J Risk Res* 5:287–299
62. Skinner B (1976) *Waldon two*. Person Education Canada, Toronto
63. Skjottner L (2001) *General systems theory: Ideas and applications*. World Scientific, Singapore
64. Som C, Hilty L, Ruddy T (2004) The precautionary principle in the information society. *Human Ecol Risk Assess* 10: 787–799
65. Spash C (2002) *Greenhouse economics: Values and ethics*. Routledge Press, London
66. Thom R (1975) *Structural stability and morphogenesis*. W. Benjamin, Reading
67. Unruh G (2000) Understanding carbon lock-in. *Energy Policy* 28:817–830
68. Unruh G (2002) Escaping carbon lock-in. *Energy Policy* 30:317–325
69. Von Bertalanffy L (1968) *General systems theory: Foundations, development, applications*. George Braziller Press, New York
70. Waldrop M (1992) *Complexity: The emerging science at the edge of order and chaos*. Simon and Schuster, New York
71. Walker B, Holling C, Carpenter S, Kinzig A (2004) Resilience, adaptability, and transformability in Social-Ecological systems. *Ecol Soc* 9(2):5. <http://www.ecologyandsociety.org/vol9/iss2/art5>
72. Wetzel K, Wetzel J (1995) Sizing the earth: Recognition of economic carrying capacity. *Ecol Econ* 12:13–21
73. Woodcock A, Davis M (1978) *Catastrophe theory*. EP Dutton, New York

Books and Reviews

- Bak P (1996) *How nature works: The science of self-organized criticality*. Copernicus Press, New York
- Fisher M, Frohlich F (2001) *Knowledge, complexity and innovation systems*. Springer, Berlin
- Gladwell M (2000) *The tipping point: How little things can make a big difference*. Little, Brown and Company, Boston
- Holland J (1995) *Hidden order: How adaptation builds complexity*. Helix Books, Reading
- Kauffman S (1996) *At home in the universe: The search for laws of self-organization and complexity*. Oxford University Press, New York
- King I (2000) *Social science and complexity: The scientific foundations*. Nova Science Publishing, Huntington
- Ormerod P (1998) *Butterfly economics: A new general theory of social and economic behavior*. Faber and Faber, New York

Human Robot Interaction

DAVID FEIL-SEIFER¹, MAJA J. MATARIĆ²

¹ Computer Science Department, University of Southern California, Los Angeles, USA

² USC Center for Robotics and Embedded Systems, University of Southern California, Los Angeles, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Major HRI Influences in Popular Culture

Prominent Research Challenges

Benchmarks and Ethical Issues for HRI

Notable Conferences

Bibliography

Glossary

Anthropomorphic Resembling, or having the attributes, human form. Human qualities have been ascribed to inanimate objects, computer-animated characters, and mechanical objects, among others. When referring to a robot, anthropomorphism refers to how human-like that robot is.

AR Assistive Robotics refers to robot systems that are designed to give aid or support to a human user. Traditionally, the assistance has been physical. However, AR has more recently expanded to encompass other types of assistance, including social, motivational, and cognitive. AR thus includes socially assistive robotics (defined below), rehabilitation robots, wheelchair robots and other mobility aids, companion robots, and educational robots.

Autonomy The ability to exert independent control, to self-direct. When referring to robotics, autonomy is used to indicate how much control of a robot results from the robot itself (based on its sensory inputs and internal computation), and how much is exerted by a human operator through tele-operation (defined below).

Benchmarks A standard used to measure performance. In the robotics context, benchmarks can be practical, relating to the safety and task performance of a system, or more abstract, relating to the ethical and other aspects of the system.

Embodied Having physical form, a body. Robots are inherently embodied, having physical form and existing in the physical world. In contrast, computer characters

may or may not be embodied. Some animated characters have three-dimensional bodies with simulated physics, thereby satisfying the condition of being embodied, but not being in the physical world.

GSR Galvanic Skin Response. Measure of electrodermal activity (EDA), or skin conductance, a function of the eccrine gland. GSR has been shown to be related to emotional stimuli, making it a potential sensor for determining user state.

HCI Human–Computer Interaction. HCI is the study of interaction between humans and computers. HCI includes interface design, issues with usability, ethics, interaction, and hardware and software design.

HRI Human–Robot Interaction. HRI is the study of interaction dynamics between humans and robots. In contrast to HCI, which addresses human–computer interaction, HRI addresses the dynamics of interaction between humans and physical, embodied robots.

SAR Socially Assistive Robotics is the study of robots capable of providing assistance through social rather than physical interaction. SAR is the intersection of SIR (defined below) and AR (defined above). SAR work is focused on addressing societal needs, such as eldercare, education, and cognitive, physical, and social therapy.

SIR Socially Interactive Robotics describes robots that interact with humans through social interaction rather than through tele-operation. SIR is a subset of HRI that addresses the challenges of social interaction between humans and robots. SIR can also be referred to as social robotics.

Robot A mechanical system that takes inputs from sensors, processes them, and acts on its environment to perform tasks.

Tele-Operation The act of controlling a device (such as a robot) remotely. The use of tele-operation for a robot decreases the autonomy of that robot.

Definition of the Subject

Human–robot interaction (HRI) is the interdisciplinary study of interaction dynamics between humans and robots. Researchers and practitioners specializing in HRI come from a variety of fields, including engineering (electrical, mechanical, industrial, and design), computer science (human–computer interaction, artificial intelligence, robotics, natural language understanding, and computer vision), social sciences (psychology, cognitive science, communications, anthropology, and human factors), and humanities (ethics and philosophy).

Introduction

Robots are poised to fill a growing number of roles in today's society, from factory automation to service applications to medical care and entertainment. While robots were initially used in repetitive tasks where all human direction is given a priori, they are becoming involved in increasingly more complex and less structured tasks and activities, including interaction with people required to complete those tasks. This complexity has prompted the entirely new endeavor of Human–Robot Interaction (HRI), the study of how humans interact with robots, and how best to design and implement robot systems capable of accomplishing interactive tasks in human environments. The fundamental goal of HRI is to develop the principles and algorithms for robot systems that make them capable of direct, safe and effective interaction with humans. Many facets of HRI research relate to and draw from insights and principles from psychology, communication, anthropology, philosophy, and ethics, making HRI an inherently interdisciplinary endeavor.

Major HRI Influences in Popular Culture

Robots got their name in Čapek's play *R.U.R.* (Rossum's Universal Robots, 1921) [18]. In *R.U.R.*, robots were man-made beings created to work for people and, as in many fictional stories thereafter, they went on to rebel and destroy the human race. In the 1950s, Isaac Asimov coined the term "robotics" and first examined the fundamental concepts of HRI, most prominently in his book *I, Robot* [3].

HRI has continued to be a topic of academic and popular culture interest. In fact, real-world robots have come into existence long after plays, novels, and movies developed them as notions and began to ask questions regarding how humans and robots would interact, and what their respective roles in society could be. While not every one of those popular culture works has affected the field of robotics research, there have been instances where ideas in the research world had their genesis in popular culture. In this section, significant popular culture products relating to HRI are overviewed, and their impact discussed.

The original benchmarks for HRI were proposed by Isaac Asimov in his now famous three laws of robotics:

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey orders given it by human beings except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

In *I, Robot* [3], the three laws were examined relative to commands that humans give robots, methods for humans to diagnose malfunctions, and ways in which robots can participate in society. The theoretical implications of how the three laws are designed to work has impacted the way that robot and agent systems operate today [140], even though the type of autonomous reasoning needed for implementing a system that obeys the three laws does not exist yet.

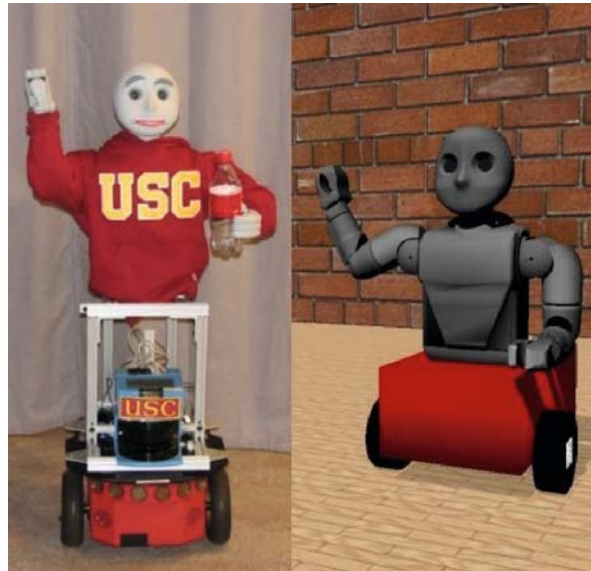
Philip K. Dick's novel *Do Androids Dream of Electric Sheep* [23] is set in a future world (originally in the late '90s) where robots (called replicants) mingle with humans. The replicants are humanoid robots that look and act like humans, and special tests are devised to determine if an individual is a human or a replicant. The test is related to the Turing Test [132], in that both involve asking probing questions that require human experiences and capacities in order to answer correctly. As is typical, the story also featured a battle between humans and replicants.

George Lucas' *Star Wars* movies (starting in 1977) feature two robot characters (C3PO and R2D2) as key characters, which are active, intuitive, even heroic. One of the most interesting features from a robot design point of view is that, while one of the robots is humanoid in form (C3PO) and the other (R2D2) is not, both interact effectively with humans through social, assistive, and service interactions. C3PO speaks, gestures, and acts as a less-than-courageous human. R2D2, on the other hand, interacts socially only through beeps and movement, but is understood and often preferred by the audience for its decisiveness and courage.

In the television show *Star Trek: The Next Generation* (1987–1994), an android named Data is a key team member with super-human intelligence but no emotions. Data's main dream was to become more human, finally mastering emotion. Data progressed to becoming an actor, a poet, a friend, and often a hero, presenting robots in a number of potentially positive roles.

The short story and movie *The Bicentennial Man* [4], features a robot who exhibits human-like creativity, carving sculptures from wood. Eventually, he strikes out on his own, on a quest to find like-minded robots. His quest turns to a desire to be recognized as a human. Through cooperation with a scientist, he develops artificial organs in order for him to bridge the divide between himself and other humans, benefiting both himself and humanity. Eventually, he is recognized as a human when he creates his own mortality.

These examples, among many others, serve to frame to scope of HRI research and exploration. They also provide some of the critical questions regarding robots and



Human Robot Interaction, Figure 1

An example of an HRI testbed: a humanoid torso on a mobile platform, and a simulation of the same system

society that have become benchmarks for real-world robot systems.

Prominent Research Challenges

The study of HRI contains a wide variety of challenges, some of them of basic research nature, exploring concepts general to HRI, and others of domain-specific nature, dealing with direct uses of robot systems that interact with humans in particular contexts. In this paper, we overview the following major research challenges within HRI: multi-modal sensing and perception; design and human factors; developmental and epigenetic robotics; social, service and assistive robotics; and robotics for education. Each is discussed in turn.

Multi-modal Perception

Real-time perception and dealing with uncertainty in sensing are some of the most enduring challenges of robotics. For HRI, the perceptual challenges are particularly complex, because of the need to perceive, understand, and react to human activity in real-time.

The range of sensor inputs for human interaction is far larger than for most other robotic domains in use today. HRI inputs include vision and speech, both major open challenges for real-time data processing. Computer vision methods that can process human-oriented data such as facial expression [10] and gestures [25] must be capa-

ble of handling a vast range of possible inputs and situations. Similarly, language understanding and dialog systems between human users and robots remain an open research challenge [49,141]. Tougher still is to obtain understanding of the connection between visual and linguistic data [106] and combining them toward improved sensing [110] and expression [14].

Even in the cases where the range of input for HRI-specific sensors is tractable, there is the added challenge of developing systems that can accomplish the sensory processing needed in a low-latency timeframe that is suitable for human interaction. For example, Kismet [13], an animated robotic head designed for infant-like interactions with a human, using object tracking for active vision, speech and prosody detection and imitation, and an actuated face for facial expressions, required several computers running in tandem to produce engaging if non-sensical facial and speech behavior. The humanoid ASIMO has been adapted to use a combination visual-auditory system for operation in indoor environments [109]. ASIMO's subsystems were used for perception, planning, and action with the goal of enabling human-robot interaction. Adding meaning to the facial and physical expressions and speech, and combining all of those capabilities in real time on a mobile, self-contained robot platform, is still an open research problem in HRI.

Even though most implemented HRI systems are necessarily domain-specific, as all physical systems, they still require the additional step of generalization to make them work beyond the research lab context. Computer vision solutions often depend on specific lighting conditions [51], ambient colors [116], and objects in the scene [15]. Beyond the lab, either the environment must be constrained to match the acceptable conditions for system operation [130], or the system capabilities must be extended in order to meet the range of conditions in the specific destination environment [83].

In addition to robot sensors that mimic the functionality of human perception (speech recognition, computer vision, etc.), sensors are being developed that cater to alternative sensing opportunities presented by an autonomous system such as a robot. These sensors enable a machine to observe people and the environment in ways that may be beyond human capability. Physiological signals, such as heart rate, blood pressure, galvanic skin response (GSR, the measure of skin conductance using a galvanometer), provide information about the user's emotional state [62,86,114] that may not otherwise be observable. Work by Mower et al. [90] used GSR as part of an HRI system to model and predict when a user is about to quit a rehabilitation-type task.

Body pose and movement are important sources of information for social interaction [106]. For example, social and expressive gestures are crucial components of human-human and human-robot interaction [121]. Computer vision can provide such information in limited contexts. In others, wearable sensors may be an effective means of obtaining human activity data in real time with high accuracy [85]. Such wearable systems have been used in HRI tasks applied to physical rehabilitation post-stroke [32], and for social interaction [124].

In addition to developing new and improving existing sensors toward particular needs of HRI, researchers are also developing algorithms for integrating multi-sensor multi-modal data inherent to HRI domains [29,44,91,93,109]. For example, Kapoor and Picard [57] implemented an affect recognition system that applies Gaussian models to fuse multiple sensors. Multi-modal sensing has also been used for a robot to detect the attention of human users in order to determine if a user is addressing the robot [71], integrating person tracking, face recognition [12], sound source localization [135], and leg detection [84].

Design and Human Factors

The design of the robot, particularly the human factor concerns, are a key aspect of HRI. Research in these areas draws from similar research in human-computer interaction (HCI) but features a number of significant differences related to the robot's physical real-world embodiment. The robot's physical embodiment, form and level of anthropomorphism, and simplicity or complexity of design, are some of the key research areas being explored.

Embodiment The most obvious and unique attribute of a robot is its physical embodiment. By studying the impact of physical embodiment on social interaction, HRI researchers hope to find measurable distinctions and trade-offs between robots and non-embodied systems (e. g., virtual companion agents, personal digital assistants, intelligent environments, etc.).

Little empirical work to date has compared robots to other social agents. Work by Bartneck et al. [9] claimed that robotic embodiment has no more effect on people's emotions than a virtual agent. Compelling recent work [60] used three characters, a human, a robot, and an animated character, to verbally instruct participants in a block stacking exercise. The study reported differences between the embodied and non-embodied agents: the robot was more engaging to the user than a simulated agent. Woods et al. [146] studied perception dif-

ferences between live and video recorded robot performances. They proposed using video recordings during system development as a complementary research tool for HRI.

Recent findings [138,139] suggest that there are several key differences between a robot and virtual agent in the context of human-machine interaction. The three conditions explored in that work (a physical robot body, a physical robot located elsewhere through a video link, and a simulation of a robot) were an attempt to control variables in order to isolate the effects of embodiment from realism. The researchers surveyed the participants regarding various properties related to the interaction. The results showed that the embodied robot was viewed by participants as more watchful, helpful, and appealing than either the realistic or non-realistic simulation.

Much work remains to be done in order to address the complex issues of physical embodiment in human-machine interaction. One confounding factor of this study involves the robot's form, discussed next.

Anthropomorphism The availability and sophistication of humanoid robots has recently soared. The humanoid form allows for exploring the use of robots for a vast variety of general tasks in human environments. This propels forward the various questions involved in studying the role of anthropomorphism in HRI. Evidence from communications research shows that people anthropomorphize computers and other objects, and that that anthropomorphism affects the nature of participant behavior during experiments [104].

HRI studies have verified that there are differences in interaction between anthropomorphic and non-anthropomorphic robots. For example, children with autism are known to respond to simple mobile car-like robots as well as to humanoid machines. However, pilot experiments have suggested that humanoid robots may be overwhelming and intimidating, while others have shown therapeutic benefit [107,111]. Biomimetic, and more specifically, anthropomorphic forms all for human-like gestures and direct imitation movements, while non-biomimetic form preserves the appeal of computers and mechanical objects.

Several examinations have been performed of the effects of anthropomorphic form on HRI [28]. These include studies of how people perceive humanoid robots compared to people and non-humanoid robots [99], possible benchmarks for evaluating the role of humanoid robots and their performance [54], and how the design of humanoid robots can be altered to affect user interacts with robots [24].

Simplicity/Complexity of Robot Design The simplicity/complexity of the robot's expressive behavior is related to the biomimetic/anthropomorphic property. Researchers are working to identify the effect that simple v. complex robot behavior has on people interacting with robots. For example, Parise et al. [100] examined the effects of life-like agents on task-oriented behavior. Powers and Kiesler [103] examined how two forms of agent embodiment and realism affect HRI for answering medical questions. Wainer et al. [138,139] used a similar experimental design to explore the effects of realism on task performance. In those studies, the more realistic or complex a robot was, the more watchful it seemed. However, it was also found that participants were less likely to share personal information with a realistic or complex robot.

Other Attributes In Reeves and Nass [104], several human factors concepts are explored in relation to human-computer interaction (HCI). As researchers work to better understand human-robot interaction, human factors insights from HCI can be valuable, but may not always be relevant. Lee and Nass [75] examined the relationship between a virtual agent's voice and its personality. The authors found that users experienced a stronger sense of social presence from the agent when the voice type and personality matched, than when they did not. In an HRI study, Tapus and Mataric [126] showed that when a robot's expressive personality matched the user's personality, task performance was better than when the personalities were mismatched. Robles et al. [108] used agents that gave feedback for a speed-dating application to examine users' feelings regarding monitoring (public and private), conformity, and self-consciousness. This study correlated users' actions with surveyed perceptions regarding feedback to determine how feedback can be most effectively given, and how it can be given in as effective a context as possible. Kidd and Breazeal [60] used a similar design to evaluate how a robot (compared to a computer agent or to a human) can give feedback for making decisions.

Ongoing research is also exploring how cultural norms and customs can affect the use of computer agent and robot systems. For example, Takeuchi et al. [125] designed an experiment to test the differences in behavior reciprocity between users of a virtual agent in the US and users in Japan. They discovered that users from both countries expressed attitudes consistent with behavior reciprocity, but only US users exhibited reciprocal behavior. However, they discovered that when recognizable brands from popular culture were used, then reciprocal behavior was exhibited in Japanese users as well.

Developmental/Epigenetic Robotics

Developmental robotics, sometimes referred to as epigenetic robotics, studies robot cognitive development. Developmental roboticists are focused on creating intelligent machines by endowing them with the ability to autonomously acquire skills and information [142]. Research into developmental/epigenetic robotics spans a broad range of approaches. One effort has studied teaching task behavior using shaping and joint attention [15], a primary means used by children in observing the behavior of others in learning tasks [92,94]. Developmental work includes the design of primitives for humanoid movements [26], gestures [69], and dialog [115].

While developmental/epigenetic robotics is not a direct subset of HRI research, there is significant overlap in the goals of the two areas. Developmental techniques for information acquisition share much in common with multi-modal perception. Epigenetic research into pronoun learning has overlap with social robotics [41]. Finally, techniques for automated teaching and learning of skills has direct applications for algorithm development for education robotics [66,95]. This work involves estimating behavior from human actions [67]. In the broader field of robot learning, a variety of methods are being developed for robot instruction from human demonstration [44,68,96,102], from reinforcement learning [134], and from genetic programming [97], among others.

Social, Service, and Assistive Robotics

Service and assistive robotics [31] include a very broad spectrum of application domains, such as office assistants [5,43], autonomous rehabilitation aids [79], and educational robots [128]. This broad area integrates basic HRI research with real-world domains that required some service or assistive function. The study of social robots (or socially interactive robots) focuses on social interaction [35], and so is a proper subset of problems studied under HRI.

Assistive robotics itself has not been formally defined or surveyed. An assistive robot is broadly defined as one that gives aid or support to a human user. Research into assistive robotics includes rehabilitation robots [16,27,47,53,78], wheelchair robots and other mobility aides [2,40,118,147], companion robots [8,101,136], manipulator arms for the physically disabled [39,42,59], and educational robots [55]. These robots are intended for use in a range of environments including schools, hospitals, and homes. In the past, assistive robotics (AR) has largely referred to robots developed to assist people through physical interaction. This definition has been significantly broadened in the last several years, in response to the

growing field of AR in which assistive robots provide help through non-contact, social interaction, defining the new field of socially assistive robotics (SAR).

Socially assistive robotics (SAR) is a growing area of research with potential benefits for elder care, education, people with social and cognitive disorders, and rehabilitation, among others [33]. SAR is the intersection of assistive robotics, which focuses on robots whose primary goal is assistance, and socially interactive robotics [35], which addresses robots whose primary feature is social interaction. SAR arose out of the large and growing body of problem domains suitable for robot assistance that involves social rather than physical interaction [77,129,144].

In rehabilitation robotics, an area that has typically developed physically-assistive robots, non-contact assistive robots are now being developed and evaluated. These robots fulfill a combined role of coach, nurse, and companion in order to motivate and monitor the user during the process of rehabilitation therapy. Observing the user's progress, the robots provide personalized encouragement and guidance. Applications for post-operative cardiac surgery recovery [56] and post-stroke rehabilitation [79] have been studied. Other rehabilitation projects have explored using a robot as a means of motivating rehabilitation through mutual storytelling [72,101]. In these experiments, a robot and a user constructs a story, which, when acted out, require the user to perform physical therapy exercises.

A variety of assistive robotics systems have been studied for use by the elderly. Such robots are meant to be used in the home, in assisted living facilities, and in hospital settings. They work to automate some physical tasks that an elderly person may not be able to do, including feeding [59], brushing teeth [131], getting in and out of bed, getting into and out of a wheelchair, and adjusting a bed for maximum comfort [52]. In some cases, the robots are envisioned as part of a ubiquitous computing system [52], which combines cameras and other sensors in the environment and computer controlled appliances (such as light switches, doors, and televisions) [8]. In others, the robots serve SAR roles such as promoting physical and cognitive exercise [127].

HRI systems have been used as companion robots in the public areas of nursing homes, aimed at increasing resident socialization. These robots are designed not to provide a specific therapeutic function, but to be a focus of resident attention. One such example is the huggable, a robot outfitted with several sensors to detect different types of touch [123]. Another such example is NurseBot, a robot used to guide users around a nursing home [87]. Paro [136,137], an actuated stuffed seal, behaves in re-



Human Robot Interaction, Figure 2

Examples of SAR research. Left: post-cardiac surgery convalescence. Middle: post-stroke rehabilitation. Right: cognitive and physical exercises

sponse to touch and sound. Its goal is to provide the benefits of pet-assisted therapy, which can affect resident quality of life [30], in nursing homes that cannot support pets. Initial studies have shown lowered stress levels in residents interacting with this robot, as well as an overall increase in the amount of socialization among residents in the common areas of the same facility.

Finally, HRI is being studied as a tool for diagnosis [111,112] and socialization [22,70,83,143] of children with autism spectrum disorders (ASD). When used for diagnosis, robots can observe children in ways that humans cannot. In particular, eye-tracking studies have shown remarkable promise when evaluating children for the purposes of diagnosing ASD. In terms of socialization, robots are a more comfortable social partner for children with ASD than people. These robots encourage social behavior, such as dancing, singing, and playing, with the robot and with other children or parents in the hope of making such behavior more natural.

Educational Robotics

Robotics has been shown to be a powerful tool for learning, not only as a topic of study, but also for other more general aspects of science, technology, engineering, and math (STEM) education. A central aspect of STEM education is problem-solving, and robots serve as excellent means for teaching problem-solving skills in group settings. Based on the mounting success of robotics courses world-wide, there is now an active movement to develop robot hardware and software in service of education, starting from the youngest elementary school ages and up [50, 80,81]. Robotics is becoming an important tool for teaching computer science and introductory college engineering [81].

Robot competition leagues such as Botball [122], RoboCup [120] and FIRST [98] have become vastly popular. The endeavors encourage focused hands-on problem solving, team work, and innovation, and range from

middle- and high-school-age children up to university teams. Educators are also using robots as tools for service learning, where projects are designed for assistive domains. Innovative teaching methods include competitions to develop robot toys for children with ASD [82] and other assistive environments [48].

In some specific domains, robots have been shown to be better for instruction than people [73]. While some automated systems are used for regular academic instruction [45], others are used for social skill instruction. In particular, robots can be used to teach social skills such as imitation [107], and self-initiation of behavior [65], in addition, they are being explored as potentially powerful tools for special education [58].

Benchmarks and Ethical Issues for HRI

As HRI systems are being developed, their impact on users and society at large are increasingly being considered [34]. Currently, it is difficult to compare robotic systems designed for different problem domains, yet it is important to do so in order to establish benchmarks for effective and ethical HRI design. Kahn et al. [54] argued for comparative methods and proposed benchmarks for HRI, with a particular focus on gaining a better understanding humanoid robots designed for HRI.

One of the most challenging aspects of establishing such benchmarks is that many aspects of HRI are difficult to measure. Establishing whether or not a robot can make eye contact with a person is comparatively simple (if not always easy to implement), but evaluating how the person reacts to and is affected by the robot's gaze and behavior is much more difficult. Does the user get bored or frustrated? Does the user consider the robot helpful and effective? Is the robot perceived as competent? Is it trusted to perform its intended tasks?

These and related questions lead to ethical considerations and legal guidelines that need to be addressed when developing HRI systems. Not only do roboticists

need to act ethically, the robots themselves must do so as well. Challenges to be considered include unintended uses of the robot, allowable tasks, and unintended situations that might be encountered. For example, if the user needs emergency attention, what is the robot's responsibility? Furthermore, the issue of control has important implications. While it is assumed the user is in control, in a variety of situations (dispensing medicine, dealing with cognitively incapacitated users) the control responsibility must rest with the machine. The issue of control and authority thus extends to all involved with the machine, including caretakers, and even designers and programmers. Well-studied ethical challenges are gradually making their way into HRI as the systems are growing in complexity and usefulness, and as their likelihood of entering human daily life increases.

General Benchmarks and Ethical Theory

While no specific ethical guidelines have yet been established, active discussions and task forces have taken up this challenging problem. Turkle [133] addressed the attachment that occurs between humans and robots when residents of a nursing home are asked to "care for" a baby-like robot. The users in the experiment ascribed human-like qualities to the robot, resulting in side-effects with ethical ramifications. What happens when the robot breaks down? What if the robot is taken away? Some benchmarks address the disparity between machines that exist only to serve a human "master" and those that exist in cooperation with their users and act with autonomy [54]. Is it acceptable for people to treat a social being like a slave?

The nature of morality for androids and other artificially intelligent entities has also been explored [140] and the difference between top-down and bottom-up morality defined. A top-down approach to morality is any approach that takes an ethical theory and guides the design and implementation of algorithms and subsystems capable of implementing that ethical theory. A bottom-up approach involves treating values as implicit to the design of the robot. In that work, morality (either implied or explicitly programmed) helps guide the behavior of robots to effectively work with humans in social situations.

Yanco [147] described the evaluation of an assistive robot, stating that such evaluation can be done through user tests and comparison to a human in the same assistive role. Long-term studies were recommended in order to evaluate effectiveness in real-world settings. Others advocated a human-centered approach to design, suggesting ecological studies of the use of the robots in the intended environment rather than long-term user studies [37].

Robot Evaluation

Any robot is a physical and technological platform that must be properly evaluated. In this section, two evaluation benchmarks of particular concern to HRI, safety and scalability, are discussed.

Safety Safety is an important benchmark for HRI: *How safe is the robot itself, and how safe can the robot make life for its user?*

A robot's safety in its given domain is the primary concern when evaluating an HRI system. If a robot is not designed with safety in mind, it could harm the very users it is designed to interact with. A key advantage of HRI over physically assistive robots is the minimization of the inherent safety risk associated with physical contact. When discussing safety pertaining to a mobile platform, we refer to the ability to maneuver about a scene without unwanted contact or collisions. Safety also refers to protection (as much as it is possible) of a robot's user and of the robot itself. This concept, as a benchmark, refers to safety in a bottom-up fashion, rather than Asimov's laws which refer to the concept in a top-down fashion [140].

Safety for assistive robots has been studied in depth in the contexts of obstacle avoidance for guide-canes and wheelchairs [7,105,147]. Robots have also been designed to help users navigate through a nursing home [40,89]. The need for safety assessment for HRI systems designed for vulnerable user populations is a topic of growing importance as HRI systems are increasingly being developed toward users from such populations.

Scalability The majority of current HRI work occurs in research laboratories, where systems are engineered for one environment and a pre-determined prototype user population. As HRI becomes more widespread in homes, schools, hospitals, and other daily environments, the question of scalability and adaptability arises: *How well will such HRI systems perform outside of the lab?* and *How well does a robot perform with users from the general population?*

The scalability benchmark does not imply that roboticians should design each robot for a large a variety of situations where assistance is required. Rather, it is important to stress that, even within a group that needs assistance, there is a great difference between a "prototypical" user or environment and the range of real-world users and environments.

Another key question to address is: *How many people can be helped by such a robot?* Consider, for example, a robot that uses speech recognition for understand-

ing a user's intentions. How does speech recognition perform when the speaker has recently suffered a stroke? Can the robot interact with someone who cannot speak? If the robot is meant to be a companion for a user, can the robot adapt its behavior to different users? How difficult is it for the robot to be modified for different needs?

In addition to user population scalability, the range of usable environments is an important benchmark. Most systems to date have been tested in research labs or controlled hospital and managed care settings. In the future, however, HRI systems will be used in homes and other more unpredictable environments. In such domains, the following benchmark becomes relevant: *Can the robot operate in the most relevant environments for the user?*

Social Interaction Evaluation

A critical benchmark of HRI is the evaluation of the robot as a social platform. Social interaction and engagement are both the primary means of interaction and the driving force behind the design. When assessing a robot in terms of social performance, we must also consider the larger goal of the robot in its application context.

Previously proposed benchmarks for humanoid robots [54] are directly relevant to HRI as well. In many respects, the same comparisons and evaluations that hold for humanoid robotics also hold for HRI. However, the goal of HRI is not to make as interesting or realistic a robot as possible, but to make a robot that can best carry out its task. It is important, therefore, to evaluate HRI not only from a perspective of modeling human characteristics, but also from a user-oriented perspective. The following sections describe how some of the previously identified humanoid benchmarks that relate to HRI.

Autonomy Autonomy is a complex property in the HRI context. It is favorable, when constructing a system that is designed to stand in for a human in a given situation, to have a degree of autonomy which allows it to perform well in its desired tasks. Autonomy can speed up applications for HRI by not requiring human input, and by providing rich and stimulating interactions. For example, HRI systems for proactive social interaction with children with ASD [22] and motivational robot tools [79,126,138] require such autonomy. However, autonomy can also lead to undesirable behavior. In situations such as medication dispensation and therapy monitoring [38], for example, autonomy is not desirable.

In general, HRI contexts require engaging and believable social interaction, but the user must clearly retain authority. For example, rehabilitation should terminate if the

user is in pain. Social interaction should only occur when it is tolerable for the user. Partial or adjustable autonomy on the part of the HRI system allows for an appropriate adjustment of both authority and autonomy.

Imitation Alan Turing proposed a test of artificial intelligence (AI), whereby a system is evaluated by whether it could fool a human user communicating with it through teletype [132]. This test was later elaborated to the Total Turing Test [46], where a system communicating in human-like ways (text, speech, facial expressions) tries to fool a human user into believing it is human. Since that time, one of the benchmarks for success in AI and HRI has been how well the system can imitate human behavior. However, when dealing with goal-oriented systems not primarily relating to human behavior but rather to assistance and treatment, imitating human behavior is necessarily desirable.

It has been shown that a robot's personality can effect a user's compliance with that robot [61]. When exhibiting a serious personality, the robot could provoke a greater degree of compliance than displaying a playful personality. It has also been shown that when the robot's extroversion/introversion personality traits matched the user's, task performance was improved [126]. Thus, the imitation benchmark proposed by Kahn could be revised for HRI: *How do imitation and reciprocity affect task performance?*

While no definitive evidence yet exists, there is a good deal of theory regarding a negative correlation between the robot's physical realism and its effectiveness in human-robot interaction. Realistic robotics introduces new complications to social robot design [28] and it has been implied that anthropomorphism has a negative influence on social interaction when the robot's behavior does not meet a user's expectations [117]. The Uncanny Valley theory suggests that as a robot becomes very similar in appearance to a human, that robot appears less, rather than more, familiar [88]. Physical similarity that attempts in imitation of human-like appearance and behavior could cause discord. This leads to two possible benchmark for imitation: *Does the interaction between the human and the robot reflect an accurate and effective impression of the robot's capabilities?* and *Does the interaction between the human and the robot allow for the expression of the human's capabilities?*

Privacy The presence of a robot inherently affects a user's sense of privacy [54]. In contrast to ubiquitous systems [11,63,76] where a user has no idea of when the system may be watching, robots are tangible and their perception limited and observable. A robot can be told to

leave when privacy is desired, and the user can observe when privacy is achieved. Because of its synthetic nature, a robot is perceived as less of a privacy invasion than a person, especially in potentially embarrassing situations. Since privacy is such a concern for designers of assistive systems [6]. Therefore, a possible benchmark from an HRI perspective asks: *Does the user's perceived sense of privacy relate to better robot performance as an assistive presence?*

Task-Oriented Benchmarks

The interactive, task-oriented nature of HRI suggests some additional benchmarks. Task performance is described as the ability of the robot to assist a user in a given task. The benchmarks then pertain to how the social aspects of the robot affect the overall task performance of the robot and its user. As with the other benchmarks, discussed above, these could apply to all social robots, but when put into an assistive context, the task-related effects highlight these features.

Social Success *Does the robot successfully achieve the desired social identity?* This is perhaps the most amorphous of benchmarks, but its evaluation is simple. When the robot is intended to be playful, do users find the robot playful? If the robot is supposed to be a social peer, do users act as if it were a social peer? How does the intended social identity compare to what occurs in practice? This benchmark is not meant to judge the ability of the robot system designer to generate a suitable robot personality. The social success of the robot is a fundamental component of HRI applications. As discussed above, the social identity of the robot (both the personality and the role of the robot) has an effect on the user's task performance.

Understanding of Domain Understanding of social dynamics is a critical component of HRI. Roboticists employ user and activity modeling as means of achieving such understanding. Efforts to understand a user of an HRI system include emotion recognition [19,20], and integration of vocalizations, speech, language, motor acts, and gestures [17,74] for effectively modeling user state.

Sensing social understanding and engagement can be assessed through a variety of means. Roboticists have also used radio frequency identification (RFID) tags and position tracking to observe children in school hallways to detect when users were in social range, and who they were interacting with over time [55], to help the robot determine appropriate social responses. Thus, social understanding in HRI can come from both human-oriented social perception (such as the interpretation of gestures,

speech, and facial expressions), and from an evaluation of user physiologic state (such as GSR, heart rate, temperature, etc.). How such data are used leads to the following benchmark: *Does a robot's social understanding of human behavior help task performance?*

Evaluation as an Assistive Tool

For the domains of HRI, impact on user's care, impact on caregivers, impact on the user's life, and the role of the robot are the key benchmarks for an assistive platform. An important way to view how an assistive robot performs when caring for people is by first observing how people care for other people in similar situations. The role of an assistive robot may be that of a stand-in for a human caregiver, a complement for a human caregiver, or an assistant to a human caregiver. Naturally, the benchmarks have different application in various scenarios. As with the other benchmarks, discussed above, this is not meant to be a comprehensive list, but a consideration of some of the most relevant benchmarks.

Success Relative to Human Caregiver A good place to start when evaluating the effect a robot has on a user's care is to compare the results of care with a robot caregiver to that of care with a human caregiver: *How does the robot perform relative to a human performing the same task?* When such evaluation is possible, existing metrics can be applied. For example, in rehabilitation tasks, functional improvement can be a metric [79]. For learning tasks, overall learning measures such as grades, tests, or evaluations can be used. In a spirometry task where a robot instructed a cardiac surgery patient to do breathing exercises [56], compliance with the robot compared to compliance with a human was a suitable metric. For companion robots, evaluating user satisfaction is most relevant.

A key role of assistive HRI is to provide care where human care is not available. In many cases, the type of interaction that is established in HRI is not directly comparable to human care, and in some instances, human care is not available for comparison. In all cases, the user satisfaction and motivation to engage in the relevant activities is a key metric of system effectiveness, on par with functional measures of task performance.

Cost/Benefit Analysis The robot can perform in several different capacities for any given task. For example, in a rehabilitation setting a robot could serve as therapist, giving advice on specific movements, a motivational coach, giving general encouragement and monitoring progress, a cognitive orthotic, reminding the users of important

items, a companion, a learning aid, or as a demonstration, showing a user how to do specific exercises. The role of the robot for a given task can inform the complexity and sophistication of the robot and its social and assistive capacities.

HRI is intended as a tool for creating robotic systems capable of providing cost-effective solutions to a variety of applications. Cost/benefit analysis can thus be a benchmark for success for such systems. In domains where no alternatives exist, and where HRI systems provide a novel and only solution, have the potential of creating major societal impact. Health care is one such domain. This suggests two benchmarks for HRI: *Does the use of the robot (a) change the cost/benefit ratio of providing such care or (b) make such care available where it was not previously possible?*

Impact on Caregivers In some cases, the goal of automation is not to increase the efficiency, productivity, or standard of care, but to make the user's or caregivers' job easier and more manageable. For example, the goal of the robot described above in Kang et al. [56] was to reduce the overall workload for cardiac nurses, given the overall nurse shortage in the US and world-wide. The robot visited cardiac patients post-surgery, approached each patient's bed, encouraged the patient to perform the breathing exercise, monitored the number and depth of the breaths taken, and collected performance data. By automating the prompting and monitoring of spirometry, which must be performed ten times per hour for the critical post-surgery period, the robot made it possible for caregivers to attend to other tasks and provide more individualized services. However, in this case, the robot did not provide any care not already provided by a human caregiver.

Caregiver impact is thus a useful benchmark: *Does the job condition of the caregiver improve as a result of the robot?* Additionally, it is important to observe cooperation: *How well does the caregiver work with the robot?* This arises out of a concern that trained and experienced caregivers are not used to working with robots, and may need to adjust their work habits [113].

Satisfaction with Care User satisfaction is an important aspect of assistive therapy success. Users' impression of a nurse robot's personality affects compliance with that robot, both positively and negatively [61]. Satisfaction, therefore can be a useful benchmark for success. Questionnaires are being explored [138,139] to measure satisfaction, although little work to date has directly related satisfaction with a robot system to task performance or user compliance. An important question when designing an assistive system is raised: *Does user satisfaction with*

a system affect the assistive task performance and/or user compliance?

Existing Quality of Life Measures Evaluating the effects of a particular therapy regimen must be done relative to the overall quality of life (QoL) of the user [145]. Some recommend using repeated measures with the same survey to capture changes over time. The SF-36 survey is designed for patient rating of health-related quality of life [1]. This survey assesses the comprehensive quality of life from the patient's point of view. The 15-D survey produces quality of life numbers along several dimensions [119]. In addition to such quantifiable measures, experiential measures, such as the Dementia Care Mapping (DCM), are also used broadly [64]. Such measures bring to the forefront the users of a particular type of service [148], as well as the notion that socially-sensitive care (involving eye-contact, favorable attention, etc.) is important to the overall outcome. This leads to a suitable HRI benchmark: *Does the robot result in a general increase in the quality of life as perceived by the user?*

Impact on the Role in Community/Society The introduction of automation and HRI-capable systems has an affect the user community. When fish tanks were introduced into a nursing home environment to test the effects on residents, observers found an overall increase in nutrition on the part of the participating residents [30]. A side-effect of the installation of the fish tanks was that residents gathered around those situated in common areas and engaged in more conversation than was previously observed. The introduction of new objects of social interest into an environment can thus change the dynamics of the community.

When roboticists introduced the robot seal Paro into the common areas of a nursing home [136,137], they found a reduction of stress proteins in the urine of the participants. Another positive effect of the experiment was that residents were in the common areas longer and socialized more. The Robovie project was able to use a robot to stimulate social interaction among a group of elementary school students [55]. By telling "secrets" about itself, the robot was able to elevate a student's status in the group by giving him/her special information [21].

An ethnographic study used readily-available low-cost robot vacuum cleaners to determine the role that the robots played in household [36]. The study used home tours and semi-structured interviews to create an ecological model of the home. The data provided insights into how a service robot might be treated, and how close the real users came to the design intention of the robot. Some treated the robot as if it were a member of the house-

hold, with status roughly between the vacuum cleaner and a pet. Others treated it strictly as a device with a purpose. An interesting observation is that men got more involved in cleaning tasks associated with the Roomba (pre-cleaning, activation, and emptying the unit when the task was completed).

A potential critique of assistive robotics is that social robots capable of HRI could reduce the amount of human contact for their users. Thus, when assessing a particular robot-assisted therapy, it is important to note not only the immediate effects on a single user, but also the effects that the robot has on the community as a whole: *Does the robot increase or decrease the amount of socialization in its user community?* and: *Are changes in community due to a robot positive or negative?*

Notable Conferences

HRI is an active and growing area of research. Progress in the field is discussed and showcased at a number of conferences, symposia, and workshops. Research results are published both in new and growing HRI conferences and journals, and the more established venues of the parent fields of HRI, namely robotics and AI.

Human–Robot Interaction-Specific Conferences

- **Conference on Human Robot Interaction (HRI)** This conference, created in 2006, is focused specifically on HRI research. Attendees and submissions to this conference are mostly from engineering (electrical engineering and computer science) with contributions from allied fields, such as psychology, anthropology, and ethics.
- **International Workshop on Robot and Human Interactive Communication (RO-MAN)** RO-MAN provides a forum for an interdisciplinary exchange for researchers dedicated to advancing knowledge in the field of human–robot interaction and communication. Importantly, RO-MAN has traditionally adopted a broad perspective encompassing research issues of human–machine interaction and communication in networked media as well as virtual and augmented tele-presence environments. RO-MAN is somewhat longer-standing than HRI.
- **International Conference on Development and Learning (ICDL)** This conference brings together the research community at the convergence of artificial intelligence, developmental psychology, cognitive science, neuroscience, and robotics, aimed at identifying common computational principles of development and learning in artificial and natural systems. The goal of the conference is to present state-of-the-art research on autonomous development in humans, animals and robots, and to continue to identify new interdisciplinary research directions for the future of the field.
- **Computer/Human Interaction (CHI) Conference** CHI is an established conference in Human–Computer Interaction (HCI). Every year, it is a venue for 2000 HCI professionals, academics, and students to discuss HCI issues and research and make lasting connections in the HCI community. HRI representation in this meeting is small, but the two fields (HRI and HCI) have much to learn and gain from each other.

General Robotics and AI Conferences

- **Association for the Advancement of Artificial Intelligence (AAAI)** AAAI's annual conference affords participants a setting where they can share ideas and learn from each other's artificial intelligence (AI) research. Topics for the symposia change each year, and the limited seating capacity and relaxed atmosphere allow for workshoplike interaction.
- **AAAI Spring and Fall Symposia** These annual symposia cover a broad range of focused topics. With the rapid growth of HRI, the topic and related areas (e. g., service robotics, socially assistive robotics, etc.) symposia are held in each session.
- **Epigenetic Robotics (EpiRob)** The Epigenetic Robotics annual workshop has established itself as an opportunity for original research combining developmental sciences, neuroscience, biology, and cognitive robotics and artificial intelligence is being presented.
- **International Conference on Robotics and Automation (ICRA)** This is one of two most major robotics conferences, covering all areas of robotics and automation. In recent years, the themes of the conference have included many areas of HRI research, such as "Humanitarian Robotics," "Ubiquitous Robotics," and "Human-Centered Robotics", reflecting the rapid growth in the field.
- **International Conference on Intelligent Robots and Systems (IROS)** This is the other major international robotics conference, featuring a very large number of papers, with a growing representation of HRI. Tutorials and workshops, as well as organized/special sessions in HRI are featured regularly.
- **International Symposium on Experimental Robotics (ISER)** ISER is a single-track symposium featuring around 50 presentations on experimental research in robotics. The goal of these symposia is to provide a forum dedicated to experimental robotics research with

principled foundations. HRI topics have become a regular part of this venue.

Bibliography

- Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, Bungay K, Fukuhara S, Gandek B, Keller S, Razavi D, Sanson-Fisher R, Sullivan M, Wood-Dauphinee S, Wagner A Jr JEW (2004) International quality of life assessment (iqola) project. *Qual Life Res* 1(5):349–351
- Aigner P, McCarragher B (1999) Shared control framework applied to a robotic aid for the blind. *Control Syst Mag IEEE* 19(2):40–46
- Asimov I (1950) *I, robot*. Doubleday, New York
- Asimov I (1976) *Bicentennial man*. Ballantine Books, New York
- Asoh H, Hayamizu S, Hara I, Motomura Y, Akaho S, Matsui T (1997) Socially embedded learning of the office-conversant mobile robot jijo-2. In: *International joint conference on artificial intelligence (IJCAI)*, Nagoya, Japan
- Baillie L, Pucher M, Képesi M (2004) A supportive multimodal mobile robot for the home. In: Stary C, Stephanidis C (eds) *User-centered interaction paradigms for universal access in the information society, Lecture notes in computer science*, vol 3196/2004. Springer, Berlin, pp 375–383
- Baker M, Yanco H (2005) Automated street crossing for assistive robots. In: *Proceedings of the international conference on rehabilitation robotics*, Chicago, IL, pp 187–192
- Baltus G, Fox D, Gemperle F, Goetz J, Hirsh T, Magaritis D, Montemerlo M, Pineau J, Roy N, Schulte J, Thrun S (2000) Towards personal service robots for the elderly. In: *Proceedings of the workshop on interactive robots and entertainment*, Pittsburgh, PA
- Bartneck C, Reichenbach J, v Breemen A (2004) In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In: *Proceedings of the design and emotion 2004 conference*, Ankara, Turkey
- Betke M, Mullally W, Magee J (2000) Active detection of eye scleras in real time. In: *Proceedings of the IEEE workshop on human modeling, analysis and synthesis*, Hilton Head, South Carolina
- Bien Z, Park K, Bang W, Stefanov D (2002) LARES: An Intelligent Sweet Home for Assisting the Elderly and the Handicapped. In: *Proceedings of the 1st Cambridge workshop on universal access and assistive technology (CWUAAT)*, Cambridge, UK, pp 43–46
- Bradski G et al (1998) Computer vision faxce tracking for use in a perceptual user interface. *Intel Technol J* 2(2):12–21
- Breazeal C (2000) Infant-like social interactions between a robot and a human caretaker. *Adapt Behav* 8(1):49–74
- Breazeal C, Edsinger A, Fitzpatrick P, Scassellati B (2001) Active vision for sociable robots. *IEEE Transactions on Man, Cybern Syst* 31(5)
- Breazeal C, Hoffman G, Lockerd A (2004) Teaching and working with robots as a collaboration. In: *Proceedings of the international joint conference on autonomous agents and multiagent systems*, vol 3. New York, pp 1030–1037
- Burgar C, Lum P, Shor P, van der Loos H (2002) Development of robots for rehabilitation therapy: The palo alto va/standford experience. *J Rehabil Res Dev* 37(6):663–673
- Busso C, Deng Z, Yildirim S, Bulut M, Lee C, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *Proceedings of the international conference on multimodal interfaces*, State Park, PA, pp 205–211
- Capek K (2001) *Rossum's universal robots*. Dover Publications, New York
- Cassell J, Sullivan J, Prevost S, Churchill E (2000) *Embodied conversational agents*. MIT Press, Cambridge
- Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. *IEEE Signal Process Mag* 18(1):32–80
- Cowley S, Kanda H (2005) Friendly machines: Interaction-oriented robots today and tomorrow. *Altern* 12(1a):79–106
- Dautenhahn K, Werry I (2002) A quantitative technique for analysing robot-human interactions. In: *Proceedings of the IEEE/RSJ, international conference on intelligent robots and systems*, Lausanne, Switzerland, pp 1132–1138
- Dick P (1968) *Do androids dream of electric sheep*. Doubleday, New York
- DiSalvo C, Gemperle F, Forlizzi J, Kiesler S (2002) All robots are not created equal: Design and the perception of humanoid robot heads. In: *Proceedings of the conference on designing interactive systems: processes, practices, methods, and techniques*, London, England, pp 321–326
- Drumwright E, Jenkins OC, Mataric MJ (2004) Exemplar-based primitives for humanoid movement classification and control. In: *IEEE International conference on robotics and automation*, pp 140–145
- Drumwright E, Ng-Thow-Hing V, Mataric MJ (2006) Toward a vocabulary of primitive task programs for humanoid robots. In: *International conference on development and learning*, Bloomington, IN
- Dubowsky S, Genot F, Godding S, Kozono H, Skwersky A, Yu H, Yu LS (2000) PAMM – a robotic aid to the elderly for mobility assistance and monitoring. In: *IEEE International conference on robotics and automation*, vol 1. San Francisco, CA, pp 570–576
- Duffy B (2003) Anthropomorphism and the social robot. *Robot Autonom Syst* 42(3):177–190
- Duquette A, Mercier H, Michaud F (2006) Investigating the use of a mobile robotic toy as an imitation agent for children with autism. In: *Proceedings of the international conference on epigenetic robotics: modeling cognitive development in robotic systems*, Paris, France
- Edwards N, Beck A (2002) Animal-assisted therapy and nutrition in Alzheimer's disease. *West J Nurs Res* 24(6):697–712
- Engelberger JF (1989) *Robotics in service*. MIT Press, Cambridge
- Eriksson J, Mataric MJ, Winstein C (2005) Hands-off assistive robotics for post-stroke arm rehabilitation. In: *Proceedings of the international conference on rehabilitation robotics*, Chicago, IL, pp 21–24
- Feil-Seifer D, Mataric MJ (2005) Defining socially assistive robotics. In: *Proceedings of the international conference on rehabilitation robotics*, Chicago, IL, pp 465–468
- Feil-Seifer DJ, Skinner KM, Mataric MJ (2007) Benchmarks for evaluating socially assistive robotics. *Interact Stud: Psychol Benchmarks Human-Robot Interact* 8(3):423–439

35. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robot Auton Syst* 42(3-4): 143–166
36. Forlizzi J, DiSalvo C (2006) Service robots in the domestic environment: A study of the Roomba vacuum in the home. In: *Proceeding of the 1st ACM SIGCHI/SIGART conference on human-robot interaction*. Academic, New York, pp 258–265
37. Forlizzi J, DiSalvo C, Gemperle F (2004) Assistive robotics and an ecology of elders living independently in their homes. *Human-Comp Interact* 19(1,2):25–59
38. Fortescue E, Kaushal R, Landrigan C, McKenna K, Clapp M, Federico F, Goldmann D, Bates D (2003) Prioritizing strategies for preventing medication errors and adverse drug events in pediatric inpatients. *Pediatr* 111(4):722–729
39. Gimenez A, Balaguer C, Sabatini SM, Genovese V (2003) The MATS robotic system to assist disabled people in their home environments. In: *Proceedings of the international conference on intelligent robots and systems*, vol 3. Las Vegas, Nevada, pp 2612–2617
40. Glover J, Holstius D, Manojlovich M, Montgomery K, Powers A, Wu J, Kiesler S, Matthews J, Thrun S (2003) A robotically-augmented walker for older adults. Tech. Rep. CMU-CS-03-170, Carnegie Mellon University, Computer Science Department, Pittsburgh, PA
41. Gold K, Scassellati B (2005) Learning about the self and others through contingency. In: *AAAI spring symposium on developmental robotics*, Stanford, CA
42. Graf B, Hans M, Kubacki J, Schraft R (2002) Robotic home assistant care-o-bot II. In: *Proceedings of the joint EMBS/BMES conference*, vol 3. Houston, TX, pp 2343–2344
43. Green A, Huttenrauch H, Norman M, Oestreicher L, Eklundh K (2000) User centered design for intelligent service robots. In: *Proceedings of the international workshop on robot and human interactive communication*, Osaka, Japan, pp 161–166
44. Grollman D, Jenkins O (2007) Learning elements of robot soccer from demonstration. In: *Proceedings of the international conference on development and learning (ICDL)*, London, England
45. Grynspan O, Martin J, Nadel J (2007) Exploring the influence of task assignment and output modalities on computerized training for autism. *Interact Stud* 8(2):241–266
46. Harnad S (1989) Minds, machines and searle. *J Exp Theor Artif Intell* 1:5–25
47. Harwin W, Ginige A, Jackson R (1988) A robot workstation for use in education of the physically handicapped. *IEEE Trans Biomed Eng* 35(2):127–131
48. Hobson RS (2000) The changing face of classroom instructional methods: servicelearning and design in a robotics course. In: *Frontiers in education conference*, vol 2. Kansas City, MO, pp F3C 20–25
49. Horvitz E, Paek T (2001) Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In: *Proceedings of the eighth international conference on user modeling*, pp 3–13
50. Hsiu T, Richards S, Bhav A, Perez-Bergquist A, Nourbakhsh I (2003) Designing a low-cost, expressive educational robot. In: *Proceedings of the conference on intelligent robots and systems*, vol 3, pp 2404–2409
51. Hunke M, Waibel A (1994) Face locating and tracking for human-computer interaction. In: *Conference record of the conference on signals, systems and computers*, Pacific Grove, CA, vol 2, pp 1277–1281
52. Jung J, Do J, Kim Y, Suh K, Kim D, Bien Z (2005) Advanced robotic residence for the elderly/the handicapped : Realization and user evaluation. In: *Proceedings of the international conference on rehabilitation robotics*, Chicago, IL, pp 492–495
53. Kahn L, Verbuch M, Rymer Z, Reinkensmeyer D (2001) Comparison of robot-assisted reaching to free reaching in promoting recovery from chronic stroke. In: *Proceedings of the international conference on rehabilitation robotics*. IOS Press, Evry, France, pp 39–44
54. Kahn PH, Ishiguro H, Friedman B, Kanda T (2006) What is a human? – Toward psychological benchmarks in the field of human-robot interaction. In: *IEEE Proceedings of the international workshop on robot and human interactive communication (RO-MAN)*, Hatfield, UK
55. Kanda T, Hirano T, Eaton D, Ishiguro H (2003) Person identification and interaction of social robots by using wireless tags. In: *IEEE/RSJ International conference on intelligent robots and systems (IROS2003)*, Las Vegas, NV, pp 1657–1664
56. Kang K, Freedman S, Matarić MJ, Cunningham M, Lopez B (2005) Hands-off physical therapy assistance robot for cardiac patients. In: *Proceedings of the international conference on rehabilitation robotics*, Chicago, IL, pp 337–340
57. Kapoor A, Picard RW (2005) Multimodal affect recognition in learning environments. In: *Proceedings of the 13th annual ACM international conference on Multimedia*, Singapore, pp 677–682
58. Karna-Lin E, Pihlainen-Bednarik K, Sutinen E, Virnes M (2006) Can robots teach? preliminary results on educational robotics in special education. In: *Proceedings of the sixth IEEE international conference on advanced learning technologies (ICALT)*, pp 319–321
59. Kawamura K, Bagchi S, Iskarous M, Bishay M (1995) Intelligent robotic systems in service of the disabled. *Proc IEEE Trans Rehabil Eng* 3(1):14–21
60. Kidd CD, Breazeal C (2004) Effect of a robot on user perceptions. In: *IEEE/RSJ International conference on intelligent robots and systems*, Sendai, Japan, pp 3559–3564
61. Kiesler S, Goetz J (2002) Mental models and cooperation with robotic assistants. In: *Proceedings of the conference on human factors in computing systems*. ACM Press, Minneapolis, Minnesota, USA, pp 576–577
62. Kim K, Bang S, Kim S (2004) Emotion recognition system using short-term monitoring of physiological signals. *Med Biol Eng Comput* 42(3):419–427
63. Kim Y, Park K, Seo K, Kim C, Lee W, Song W, Do J, Lee J, Kim B, Kim J et al (2003) A report on questionnaire for developing Intelligent Sweet Home for the disabled and the elderly in Korean living conditions. In: *Proceedings of the ICORR (the eighth international conference on rehabilitation robotics)*
64. Kitwood T, Bredin K (1992) A new approach to the evaluation of dementia care. *J Adv Health Nurs Care* 1(5):41–60
65. Koegel L, Carter C, Koegel R (2003) Teaching children with autism self-initiations as a pivotal response. *Top Lang Disord* 23:134–145
66. Koenig N, Matarić MJ (2006) Behavior-based segmentation of demonstrated task. In: *International conference on development and learning*, Bloomington, IN

67. Koenig N, Matarić MJ (2006) Behavior-based segmentation of demonstrated task. In: International conference on development and learning, Bloomington, IN
68. Koenig N, Matarić MJ (2006) Demonstration-based behavior and task learning. In: Working notes, AAAI spring symposium to boldly go where no human-robot team has gone before, Stanford, California
69. Kopp S, Wachsmuth I (2002) Model-based animation of coverbal gesture. In: Proceedings of the computer animation, IEEE computer society, Washington, DC, USA
70. Kozima H, Nakagawa C, Yasuda Y (2005) Interactive robots for communication-care: a case-study in autism therapy. In: IEEE International workshop on robot and human interactive communication (ROMAN), Nashville, TN, pp 341–346
71. Lang S, Kleinhagenbrock M, Hohenner S, Fritsch J, Fink GA, Sagerer G (2003) Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In: Proceedings of the international conference on multimodal interfaces. ACM, Vancouver, Canada, pp 28–35
72. Lathan C, Vice J, Tracey M, Plaisant C, Drui A, Edward K, Montemayor J (2001) Therapeutic play with a storytelling robot. In: Conference on human factors in computing systems. ACM Press, New York, NY, USA, pp 27–28
73. Lathan C, Boser K, Safos C, Frenz C, Powers K (2007) Using cosmo's learning system (CLS) with children with autism. In: Proceedings of the international conference on technology-based learning with disabilities, Dayton, OH, pp 37–47
74. Lee C, Narayanan S (2005) Towards detecting emotions in spoken dialogs. *IEEE Trans Speech Audio Process* 13(2): 293–302
75. Lee KM, Nass C (2003) Designing social presence of social actors in human computer interaction. In: Proceedings of the SIGCHI conference on human factors in computing systems, Ft. Lauderdale, FL, vol 5, pp 289–296, <http://portal.acm.org/citation.cfm?id=642662>
76. Lee N, Keating D (1994) Controllers for use by disabled people. *Comput Control Eng J* 5(3):121–124
77. Lord C, McGee J (eds) (2001) Educating children with autism. National Academy Press, Washington
78. Mahoney R, van der Loos H, Lum P, Burgar C (2003) Robotic stroke therapy assistant. *Robotica* 21:33–44
79. Matarić MJ, Eriksson J, Feil-Seifer D, Winstein C (2007) Socially assistive robotics for post-stroke rehabilitation. *J NeuroEng Rehabil* 4(5)
80. Matarić MJ, Koenig N, Feil-Seifer DJ (2007) Materials for enabling hands-on robotics and stem education. In: AAAI Spring symposium on robots and robot venues: resources for AI education, Stanford, CA
81. Matarić MJ, Fasola J, Feil-Seifer DJ (2008) Robotics as a tool for immersive, hands-on freshmen engineering instruction. In: American society for engineering education, Proceedings of the ASEE annual conference & exposition, Pittsburgh, PA
82. Michaud F, Clavet A (2001) Robotoy contest – designing mobile robotic toys for autistic children. In: Proceedings of the american society for engineering education (ASEE), Albuquerque, New Mexico, <http://citeseer.nj.nec.com/michaud01robotoy.html>
83. Michaud F, Laplante JF, Larouche H, Duquette A, Caron S, Le-tourneau D, Masson P (2005) Autonomous spherical mobile robot for child-development studies. *IEEE Trans Syst Man Cybern* 35(4):471–480
84. Mikolajczyk K, Schmid C, Zisserman A (2004) Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. *Computer Vision, ECCV 2004: Proceedings of the 8th European conference on computer vision, Prague, Czech Republic, 11–14 May 2004*
85. Miller N, Jenkins O, Kallman M, Matarić MJ (2004) Motion capture from inertial sensing for untethered humanoid teleoperation. In: Proceedings, IEEE-RAS International conference on humanoid robotics (Humanoids-2004), Santa Monica, CA
86. Mohan A, Picard R (2004) Health0: a new health and lifestyle management paradigm. *Stud Health Technol Inform* 108: 43–8
87. Montemerlo M, Prieau J, Thrun S, Varma V (2002) Experiences with a mobile robotics guide for the elderly. In: Proceedings of the AAAI national conference on artificial intelligence. Edmonton, Alberta, pp 587–592
88. Mori M (1970) Bukimi no tani (The uncanny valley). *Energy* 7:33–35
89. Morris A, Donamukkala R, Kapuria A, Steinfeld A, Matthews J, Dunbar-Jacob J, Thrun S (2003) A robotic walker that provides guidance. In: Proceedings of the 2003 IEEE international conference on robotics and automation. ICRA, Taipei, Taiwan, pp 25–30
90. Mower E, Feil-Seifer D, Matarić MJ, Narayanan S (2007) Investigating implicit cues for user state estimation in human robot interaction. In: Proceedings of the international conference on human-robot interaction (HRI)
91. MP Michalowski HK S, Sabanovic (2007) A dancing robot for rhythmic social interaction. In: Proceedings of the conference on human-robot interaction (HRI), Washington, DC
92. Mundy P, Card J, Fox N (2000) Fourteen-month cortical activity and different infant joint attention skills. *Dev Psychobiol* 36:325–338
93. Mutlu B, Krause A, Forlizzi J, Guestrin C, Hodgins J (2007) Robust, low-cost, non-intrusive recognition of seated postures. In: Proceedings of 20th ACM symposium on user interface software and technology, Newport, RI
94. Nagai Y, Hosoda K, Asada M (2003) How does an infant acquire the ability of joint attention?: A constructive approach. In: Proceedings of the third international workshop on epigenetic robotics: modeling cognitive development in robotic systems, Boston, MA, pp 91–98
95. Nicolescu M, Matarić MJ (2003) Linking perception and action in a control architecture for human-robot interaction. In: Hawaii international conference on system sciences, (HICSS-36), Hawaii, USA
96. Nicolescu M, Matarić MJ (2005) Task learning through imitation and human-robot interaction. In: Dautenhahn K, Nehaniv C (eds) Models and mechanisms of imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions. Cambridge University Press, New York
97. Nordin P (1997) An on-line method to evolve behavior and to control a miniature robot in real time with genetic programming. *Adapt Behav* 5(2):107–140
98. Oppliger D (2001) University-pre college interaction through FIRST robotics competition. Oslo, Norway, pp 11–16
99. Oztop E, Franklin DW, Chaminade T, Cheng G (2005) Human-humanoid interaction: Is a humanoid robot perceived as a human? *Int J Human Robot* 2(4):537–559

100. Parise S, Kiesler S, Sproull L, Waters K (1999) Cooperating with life-like interface agents. *Comp Human Behav* 15(2):123–142
101. Plaisant C, Druin A, Lathan C, Dakhane K, Edwards K, Vice J, Montemayor J (2000) A storytelling robot for pediatric rehabilitation. In: *Proceedings of the fourth international ACM conference on assistive technologies*, Arlington, VA, pp 50–55
102. Pomerleau D (1993) Knowledge-based training of artificial neural networks for autonomous robot driving. In: Connell JH, Mahadevan S (eds) *Robot Learning*. Kluwer, Boston, pp 19–43
103. Powers A, Kiesler S (2006) The advisor robot: Tracing people's mental model from a robot's physical attributes. In: *Proceedings of the 2006 ACM conference on human–robot interaction*. ACM Press, Salt Lake City, UT, pp 218–225
104. Reeves B, Nass C (1996) *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press, New York
105. Rentschler A, Cooper R, Blasch B, Boninger M (2003) Intelligent walkers for the elderly: Performance and safety testing of VA-PAMAID robotic walker. *J Rehabil Res Dev* 40(5): 423–431
106. Rizzolatti G, Arbib M (1998) Language within our grasp. *Trends Neurosci* 21(5):188–194
107. Robins B, Dautenhahn K, Boekhorst R, Billard A (2005) Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Univ Access Inf Soc* 4(2):105–120
108. Robles EA, Sukumaran A, Rickertsen K, Nass C (2006) Being watched or being special: how i learned to stop worrying and love being monitored, surveilled, and assessed. In: *Proceedings of the ACM SIGCHI conference on human factors in computing systems*, Montreal, Quebec, Canada, pp 831–839
109. Sakagami Y, Watanabe R, Aoyama C, Matsunaga S, Higaki N, Fujimura K, Ltd H, Saitama J (2002) The intelligent ASIMO: system overview and integration. In: *International conference on intelligent robots and system*, 2002, EPFL, Switzerland, vol 3, pp 2478–2483
110. Scassellati B (2003) Investigating models of social development using a humanoid robot. *Proc Int Joint Conf Neural Networks* 4:2704–2709
111. Scassellati B (2005) Quantitative metrics of social response for autism diagnosis. In: *IEEE International workshop on robots and human interactive communication (ROMAN)*, Nashville, TN, pp 585–590
112. Scassellati B (2005) Using social robots to study abnormal social development. In: *Proceedings of the fifth international workshop on epigenetic robotics: modeling cognitive development in robotic systems*, Nara, Japan, pp 11–14
113. Scholtz J (2002) Evaluation methods for human–system performance of intelligent systems. In: *Proceedings of the 2002 performance metrics for intelligent systems (PerMIS) workshop*, Gaithersburg, MD
114. Shalom DB, Mostofsky SH, Hazlett RL, Goldberg MC, Landa RJ, Faran Y, McLeod DR, Hoehn-Saric R (2006) Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism. *J Autism Dev Disord* 36(3):295–400
115. Shin J, Narayanan S, Gerber L, Kazemzadeh A, Byrd D (2002) Analysis of user behavior under error conditions in spoken dialogs. In: *Proceedings of ICSLP*, Denver, CO
116. Shin MC, Chang KI, Tsap LV (2002) Does colorspace transformation make any difference on skin detection? In: *Proceedings of sixth IEEE workshop on applications of computer vision*, 2002 (WACV 2002), pp 275–279, <http://citeseer.nj.nec.com/542214.html>
117. Shneiderman B (1989) A nonanthropomorphic style guide: Overcoming the humpty-dumpty syndrome. *Comput Teacher* 16(7):5
118. Simpson R, Levine S (1997) Development and evaluation of voice control for a smart wheelchair. In: *Proceedings of the rehabilitation engineering society of North America annual conference*, Pittsburgh, PA, pp 417–419
119. Sintonen H (1994) The 15-d measure of health related quality of life: reliability, validity and sensitivity of its health state descriptive system. Working Paper 41, Center for Health Program Evaluation, West Heidelberg, Victoria, Australia
120. Sklar E, Eguchi A, Johnson J (2003) RoboCupJunior: learning with educational robotics. *Robocup 2002: Robot Soccer World Cup VI*
121. Sowa T, Kopp S (2003) A cognitive model for the representation and processing of shape-related gestures. In: Schmalhofer F, Young R, Katz G (eds) *Proceedings of the European cognitive science conference (EuroCogSci03)*, Lawrence Erlbaum Assoc, New Jersey, p 441
122. Stein C (2002) Botball: Autonomous students engineering autonomous robots. In: *Proceedings of the ASEE conference*, Montreal, Quebec, Canada
123. Stiehl WD, Lieberman J, Breazeal C, Basel L, Lalla L, Wolf M (2006) The design of the huggable: A therapeutic robotic companion for relational, affective touch. In: *Proceedings of the AAAI fall symposium on caring machines: AI in eldercare*, Washington, DC
124. Stone M, DeCarlo D (2003) Crafting the illusion of meaning: Template-based specification of embodied conversational behavior. In: *Proceedings of the international conference on computer animation and social agents*, pp 11–16
125. Takeuchi Y, Katagiri Y, Nass CI, Fogg BJ (2000) Social response and cultural dependency in human–computer interaction. In: *Proceedings of the CHI 2000 conference*, Amsterdam, The Netherlands
126. Tapus A, Mataric MJ (2006) User personality matching with hands-off robot for post-stroke rehabilitation therapy. In: *Proceedings of the international symposium on experimental robotics (ISER)*, Rio de Janeiro, Brazil
127. Tapus A, Fasola J, Mataric MJ (2008) Socially assistive robots for individuals suffering from dementia. In: *ACM/IEEE 3rd human–robot interaction international conference, workshop on robotic helpers: user interaction, interfaces and companions in assistive and therapy robotics*, Amsterdam, The Netherlands
128. Tartaro A, Cassell J (2006) Authorable virtual peers for autism spectrum disorders. In: *Combined workshop on language enabled educational technology and development and evaluation of robust dialog system*, ECAI
129. Taub E, Uswatte G, King D, Morris D, Crago J, Chatterjee A (2006) A placebo-controlled trial of constraint-induced movement therapy for upper extremity after stroke. *Stroke* 37(4):1045–9
130. Thrun S, Bennewitz M, Burgard W, Cremers A, Dellaert F, Fox D, Hahnel D, Rosenberg C, Roy N, Schulte J, Schulz D (1999) MINERVA: A second-generation museum tour-guide

- robot. In: Proceedings of the IEEE international conference on robotics and automation (ICRA '99), Detroit, Michigan
131. Topping M, Smith J (1999) The development of handy, a robotic system to assist the severely disabled. In: Proceedings of the international conference on rehabilitation robotics, Stanford, CA, <http://rose.iinf.polsl.gliwice.pl/~kwadrat/www.csun.edu/cod/conf2001/proceedings/0211topping.html>
 132. Turing A (1950) Computing machinery and intelligence. *Mind* 49:433–460
 133. Turkle S (2005) Relational artifacts/children/elders: The complexities of cybercompanions. In: Toward social mechanisms of android science: A CogSci 2005 workshop, Stresa, Italy, p 62–33
 134. Uchibe E, Asada M, Hosoda K (1998) Cooperative behavior acquisition in multi-mobile robots environment by reinforcement learning based on state vector estimation. In: Proceedings of the international conference on robotics and automation, Leuven, Belgium, pp 1558–1563
 135. Valin J, Michaud F, Rouat J, Letourneau D (2003) Robust sound source localization using a microphone array on a mobile robot. Proceedings of the 2003 IEEE/RSJ international conference on intelligent robots and systems, 2003(IROS 2003), vol 2, pp 1228–1233
 136. Wada K, Shibata T, Saito T, Tanie K (2002) Analysis of factors that bring mental effects to elderly people in robot assisted activity. In: Proceedings of the international conference on intelligent robots and systems, Lausanne, Switzerland, vol 2, pp 1152–1157
 137. Wada K, Shibata T, Saito T, Sakamoto K, Tanie K (2005) Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), pp 2785–2790
 138. Wainer J, Feil-Seifer D, Shell D, Mataric MJ (2006) The role of physical embodiment in human–robot interaction. In: IEEE Proceedings of the international workshop on robot and human interactive communication, Hatfield, United Kingdom, pp 117–122
 139. Wainer J, Feil-Seifer D, Shell D, Mataric MJ (2007) Embodiment and human–robot interaction: A task-based perspective. In: Proceedings of the international conference on human–robot interaction
 140. Wallach W, Allen C (2005) Android ethics: Bottom-up and top-down approaches for modeling human moral faculties. In: Proceedings of the 2005 CogSci workshop: toward social mechanisms of android science, Stresa, Italy, pp 149–159
 141. Wang D, Narayanan S (2007) An acoustic measure for word prominence in spontaneous speech. *IEEE Trans Speech Audio Lang Process* 15(2):690–701
 142. Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E (2001) Autonomous mental development by robots and animals. *Science* 291(5504):599–600
 143. Werry I, Dautenhahn K, Ogden B, Harwin W (2001) Can social interaction skills be taught by a social agent? The role of a robotic mediator in autism therapy. Lecture notes in computer science, vol 2117. Springer, Heidelberg, pp 57–74
 144. Wolf S, Thompson P, Morris D, Rose D, Winstein C, Taub E, Giuliani C, Pearson S (2005) The EXCITE trial: Attributes of the wolf motor function test in patients with subacute stroke. *Neurorehabil Neural Repair* 19:194–205
 145. Wood-Dauphinee S (1999) Assessing quality of life in clinical research: From where have we come and where are we going? *J Clin Epidemiol* 52(4):355–363
 146. Woods S, Walters M, Koay KL, Dautenhahn K (2006) Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In: Proceedings the 9th international workshop on advanced motion control, Istanbul
 147. Yanco H (2002) Evaluating the performance of assistive robotic systems. In: Proceedings of the workshop on performance metrics for intelligent systems, Gaithersburg, MD
 148. Younger D, Martin G (2000) Dementia care mapping: an approach to quality audit of services for people with dementia in two health districts. *J Adv Nurs* 32(5):1206–1212

Human Sexual Networks

FREDRIK LILJEROS

Stockholm University, Stockholm, Sweden

Article Outline

Glossary

Definition of the Subject

Introduction

Non-Complex Models of Contagious Diseases

The Core Group Theory

Lessons Learned from the Early AIDS Epidemic

Clustering

The Effect of Geographical Space

The Long Tail

The Importance of Concurrent Relationship

Assortative Interaction

Data Sources

Future Directions

Bibliography

Glossary

Basic reproduction rate The most common way to calculate the epidemic threshold is to calculate the basic reproduction rate, R_0 , which is usually defined as the average number of secondary infections caused by one infectious individual that enters into a totally susceptible population. The basic reproduction rate may underestimate the risk of epidemic outbreaks if the variation in number of contacts is large, as is usually the case with sexual contacts.

Core group A subgroup of individuals in a population characterized by a high partner turnover rate and a high tendency for having sexual contacts within the group. The existence of a core group may push the population above the epidemic threshold.

Epidemic threshold The probability that an epidemic will occur is determined by the contagiousness of the disease, the duration of infectiousness, and the interaction structure in the population. Contagious diseases are nonlinear phenomena in the sense that small changes in any of these parameters may push the population from a state in which a large epidemic is not possible to a state in which an epidemic may easily occur if infection is introduced into the population. The specific point at which an epidemic is possible is referred to as the epidemic threshold.

Random homogeneous mixing When modeling outbreaks of contagious diseases in a population, the individuals are often assumed to have the same probability of interacting with everyone else in the population. This assumption has been shown to be less valid for sexually transmitted infections because they are characterized by a large variation in number of contacts.

Sexually transmitted infection Many contagious infections can be spread through sexual contact. Sexually transmitted infections are, however, generally defined as being spread through vaginal intercourse, anal intercourse, and oral sex. They include *Chlamydia trachomatis*, gonorrhea, and HIV. The reason why the expression “sexually transmitted infection” is used instead of “sexually transmitted disease” is that a state of infection and infectiousness do not necessarily result in disease.

Definition of the Subject

Human sexual networks are the network structures that emerge when individuals have sexual contact with each other. In general, use of the term “sexual contact” is restricted in this article to mean vaginal or anal intercourse or oral sex – contacts by which sexually transmitted infections (STIs) can be transmitted. Sexual networks are important because an understanding of their structure and how they facilitate the spread of infection can help us understand how the spread of this type of infection can best be prevented.

Introduction

Although the type of contact that spreads STIs occurs less frequently than is the case for most other types of contact that spread disease, the spread of STIs has turned out to be surprisingly hard to limit. The difficulties in getting STIs under control have led to an interest in sexual contact patterns [24,28,29]. In the present chapter we discuss a variety of explanations related to the structural properties of sexual networks that have been advanced for why STIs are so

widespread, and why such diseases are so hard to eradicate. We begin this chapter by presenting a family of models in which no explicit assumptions are made about the interaction structure other than that all individuals are assumed to have the same probability of interacting with everyone else in the population. These models are then used as a baseline when discussing more realistic assumptions about the sexual contact structure. We then move on to more realistic assumptions about the contact structure by introducing one of the first theories about sexual contact structure, the core-group theory.

The emerging AIDS epidemic in the early 1980s raised new questions about how sexually transmitted diseases are spread in human populations [19]. We discuss different structural explanations for why the number of HIV-infected not did grow as fast as models based on random homogeneous mixing predicted. We then look at two other structural properties of importance for the understanding of the spread of STIs, assortative interaction and concurrent relationship. Sexual networks are usually very difficult to study empirically for several different reasons. These difficulties and ways of handling them are introduced in the succeeding section. Some remarks on the future challenges for research on sexual networks conclude the chapter.

Non-Complex Models of Contagious Diseases

To help to understand the complexity of human sexual networks and how their structures may facilitate the spread of STIs, we will first introduce a simple family of models based on the assumption of random homogeneous mixing. Random homogeneous mixing means that every person in the population has the same probability of interacting with every other person in the population; hence no assumption is made about any structural properties of the contact network. Models based on this assumption often model outbreaks of highly contagious diseases surprisingly well, while they are usually less good for modeling less contagious diseases. These simple models and their characteristic behaviors will then be used as a baseline when we introduce and discuss different structural properties of human sexual networks that deviate from random homogeneous mixing.

In this type of standard model, individuals in a population are assumed to be in one of three states: susceptible (S), infected (I), or removed (R). The latter can, depending on the disease under study, result from immunity or death. It is conventional to distinguish between SI, SIS, and SIR models. Children’s diseases are best modeled by an SIR model because infection confers lifelong immunity, that is,

removal from the pool of those susceptible. For most sexually transmitted diseases, the *SIS* model makes most sense since few sexually transmitted diseases confer any immunity after infection, and people remain susceptible. An important exception is HIV, which is still appropriately described, at least in the Western world, using the *SI* model.

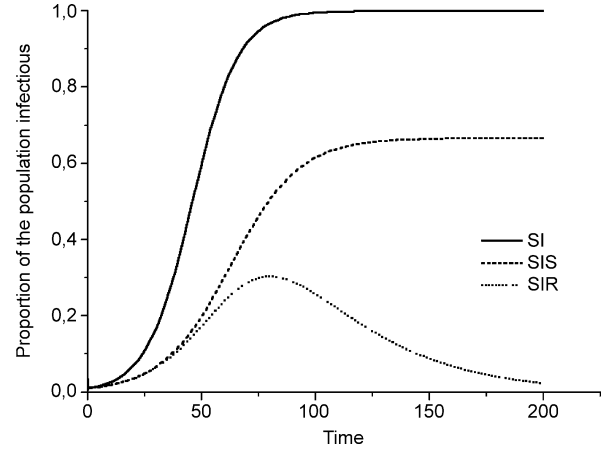
The random homogenous mixing assumption is explained by Eq. (1), which represents the *SI* model as a continuous model in its simplest form, consisting of a system of two differential equations:

$$\begin{aligned} \frac{dS}{dt} &= \frac{-c\beta S(t)I(t)}{N}, \\ \frac{dI}{dt} &= \frac{c\beta S(t)I(t)}{N}. \end{aligned} \quad (1)$$

There are two dependent variables in Eq. (1): The number of susceptible persons, S , and the number of infected persons, I . The number of susceptible and infectious persons are constrained by the size of the population N so that $S(t) + I(t) = N$. Note that this constraint means that the two variables, $S(t)$ and $I(t)$, are linearly dependent, so that the two differential equations actually are redundant. As is evident, this model is homogenous across the population as each person is assumed to have the same number of contacts per time unit, c and the same probability of infection per contact, B . This is what we refer to as the assumption of *random homogeneous mixing*.

For many diseases, such as measles or flu, that are spread by aerosolized droplets by all infected persons, random interaction is a reasonable assumption and probably a good approximation. It describes an abundance of everyday situations in which a person is exposed to such infections, for instance on public transportation, in the workplace, and in shops. A significant advantage of the random interaction assumption is that it can easily be modeled with differential equations, and these models can be studied analytically [1,7]. The equilibriums for the system, for example, can easily be found by first setting the left side of Eq. (1) to 0, and then solving the expression as a system of equations.

The solution to Eq. (1) yields an S-shaped trajectory as shown in Fig. 1. One important property of such models is that they predict that the number of infected persons will grow exponentially during the early stages of an outbreak. This growth cannot, however, continue to accelerate at this rate for long because the number of susceptible persons is low during later stages of the process. In later stages, most infected persons will interact only with other infectious persons. Consequently, the growth in the number of infected persons is largest at about the midpoint of the process.



Human Sexual Networks, Figure 1

The typical growth in the number of infectious persons during an outbreak for an *SI*, an *SIS* and an *SIR* model, when random homogeneous mixing is assumed for the population

The *SIS* model can be written as a system of two differential equations as follows:

$$\begin{aligned} \frac{dS}{dt} &= \frac{-c\beta S(t)I(t)}{N} + \frac{I(t)}{D}, \\ \frac{dI}{dt} &= \frac{c\beta S(t)I(t)}{N} - \frac{I(t)}{D}. \end{aligned} \quad (2)$$

The equation for the *SIS* model differs from that of the *SI* model in the sense that the term $\frac{I(t)}{D}$ that describes the rate of individuals who recover from the disease and become susceptible is added to both equations. The solution to the *SIS* equations also shows that we should expect an S-shaped trajectory with an exponential growth in the number of infected persons. The *SIS* trajectory, however, differs from the *SI* trajectories in the sense that the number of infected persons never reaches that of the entire population. The process equilibrates at a point where exactly as many infectious individuals become susceptible as susceptible ones become infected.

The last model we are going to discuss here is the *SIR* model, which can be formulated as a set of differential equations in its simplest form as:

$$\begin{aligned} \frac{dS}{dt} &= \frac{-c\beta S(t)I(t)}{N}, \\ \frac{dI}{dt} &= \frac{c\beta S(t)I(t)}{N} - \frac{I(t)}{D}, \\ \frac{dR}{dt} &= \frac{I(t)}{D}. \end{aligned} \quad (3)$$

The *SIR* model is portrayed here in the form of a system of three differential equations that describe the change in

the three different states (of which two are independent, since we now have $S(t) + I(t) + R(t) = N$). The solution to the *SIR* model also shows an S-shaped form during the early stages of the epidemic. The *SIR* model differs from the *SI* and the *SIS* models by its tendency to result in zero infectious individuals in the long run.

These three basic models can be adapted to the characteristics of specific diseases by, for example, letting persons be immune only for a certain time interval, making it possible for new individuals to enter into the population through birth and immigration, and allowing persons leave the population because of emigration or death. This makes it possible to generate more complex types of trajectories such as cyclic behavior [1].

A critical notion in disease epidemiology is the basic reproduction number, R_0 . In the homogenous deterministic *SIS*-model and *SIR*-model this number tells us how many uninfected persons an infectious individual will, on average, infect in a totally susceptible population [1]:

$$R_0 = c\beta D. \quad (4)$$

R_0 has received special attention because in the homogenous model it is quite an intuitive measure of the epidemic threshold. If R_0 is less than 1, then the disease will become extinct. With R_0 equaling exactly 1, we have an unstable equilibrium with no change in the number of infected or susceptible persons, that is, the disease is endemic. A value greater than 1 upsets the replacement conditions, which means that if $R_0 > 1$ the outcome is an epidemic. (Note that R_0 will always be larger than one in the *SI* model since an infectious individual is assumed to be infectious for an infinitely long time.)

The Core Group Theory

Core group theory is an early, and probably also the best known, explanation for why STIs can be endemic despite the fact that the average number of sexual contacts in most national populations is such that we should expect R_0 to be lower than 1, that is, below the epidemic threshold if the contacts were evenly distributed in the population [13]. According to core group theory, the reproduction of STIs can be explained by the existence of several distinct subgroups in the general population that are all characterized by high-risk sexual behavior (high partner turnover rate and unprotected sex) and extensive intergroup interaction. The existence of core groups, according to the theory, makes it possible for STIs to reproduce within the core group because R_0 is greater than 1 in these groups. The core groups thus constitute a reservoir allowing the sexually transmitted infection to remain endemic in a gen-

eral population in which the R_0 is lower than the critical value 1. The core groups also make it possible for the rest of the population to be infected through contacts with these groups.

Lessons Learned from the Early AIDS Epidemic

The discovery of a progressive outbreak of AIDS in the early 1980s gave rise to intensive research efforts to understand the path of contagion and the disease's dynamic course [8]. One of the most important findings was made by Anderson and May [1] who showed that the expression for R_0 in Eq. (4) is not suitable if the variance in number of potential infectious contacts is high, as is the case with STIs. They showed instead that R_0 would be more accurately estimated by using the following equation:

$$R_0 = \rho_0 \left(1 + \frac{\sigma^2}{\mu^2} \right), \quad (5)$$

where ρ_0 is the average number of infections produced by an infected person in an uninfected population, σ^2 is the variance in the number of contacts, and μ is the mean number of contacts in the population. From this equation, it is clear that the larger the variance in number of partners in the population for a given μ , the less infectious an infection needs to be to continue to reproduce itself, that is, to generate an epidemic.

It became clear relatively early that the AIDS epidemic did not behave the way traditional infection epidemiological models expected. The assumption of random homogeneous mixing makes these models predict an exponential increase in the number of contagious persons at the beginning of an epidemic (see Fig. 1). The AIDS epidemic did not seem to follow this pattern, however. Instead of increasing exponentially, it seemed to increase more slowly [5]. This was seen to occur in spite of the fact that the epidemic was very far from global saturation, so that the slowing down could not be due to a global saturation effect. Exponential growth is characterized by a constant time required for the number of infected persons to double. However, in the case of AIDS, that time became increasingly longer as the epidemic spread. At the end of the 1980s, the AIDS epidemic was demonstrated to exhibit a polynomial pattern of spread, and it was shown that epidemics have had this functional form before any changes in behavior patterns could have had an effect. Then Colgate et al. [5] argued that a polynomial pattern of spread could not be explained as a direct effect of interventions or by changed behavior patterns caused by a general understanding of how HIV was spread. They also showed that polynomial spread could be observed in groups of indi-

viduals of different ethnic backgrounds. There are at least three structural properties of contact structures that may slow down spread in the observed way: clustering, embedding in a low-dimensional space, and a large variation in partner turnover rate.

Clustering

One structural property of contact networks that has been shown to slow an epidemic's rate of spread is *clustering* (or *transitivity* as it is also called). A large amount of clustering is typical of many social networks. For example, if Charles is a good friend of Paul and Ben's, it is quite probable that Paul and Ben know each other. In an outbreak of a highly contagious disease, the contacts of an infected individual in a clustered network will often already have been infected by common contact or by a contact only a few steps away in the network. A common way of estimating clustering in a network is to estimate its relative number of triangles, or more exactly, to calculate the fraction C of all paths of length three in the network that form a triangle:

$$C = \frac{3n_{\text{triangle}}}{n_{\text{triple}}}, \quad (6)$$

where n_{triangle} is the number of triangles and n_{triple} is the number of triples of vertices connected by two or three contacts. The factor three is needed to normalize C to the interval $[0, 1]$. One useful property of C is that $1 - C$ gives the average proportion of outgoing links from all contacts directly connected to an individual that potentially can transmit the disease further to the rest of the network.

Clustering, that is, two individuals with a common sexual contact who also have sexual relations with each other by definition cannot exist in a heterosexual network unless bisexual relationships are allowed. It is possible, however, for two individuals of the same sex to have had two or more sexual partners in common. This phenomenon is called *mutuality*, M . For computational reasons, M is often defined in a way similar to $1 - C$ as

$$M = \frac{\text{mean number of nodes two steps away from a node}}{\text{mean number of paths of length two between those nodes}}, \quad (7)$$

where a node here would be a person, and a path would be a route along any number of links from one node to another. Mutuality can thus be thought of as the opposite of clustering (transitivity). For example, M for a given

node A reaches its maximum value Eq. (1) when A reaches as many two-hop neighbors as possible with the number of two-hop paths coming from A . Thus, high mutuality leads to faster epidemic spreading, while high transitivity tends to confine epidemic spreading.

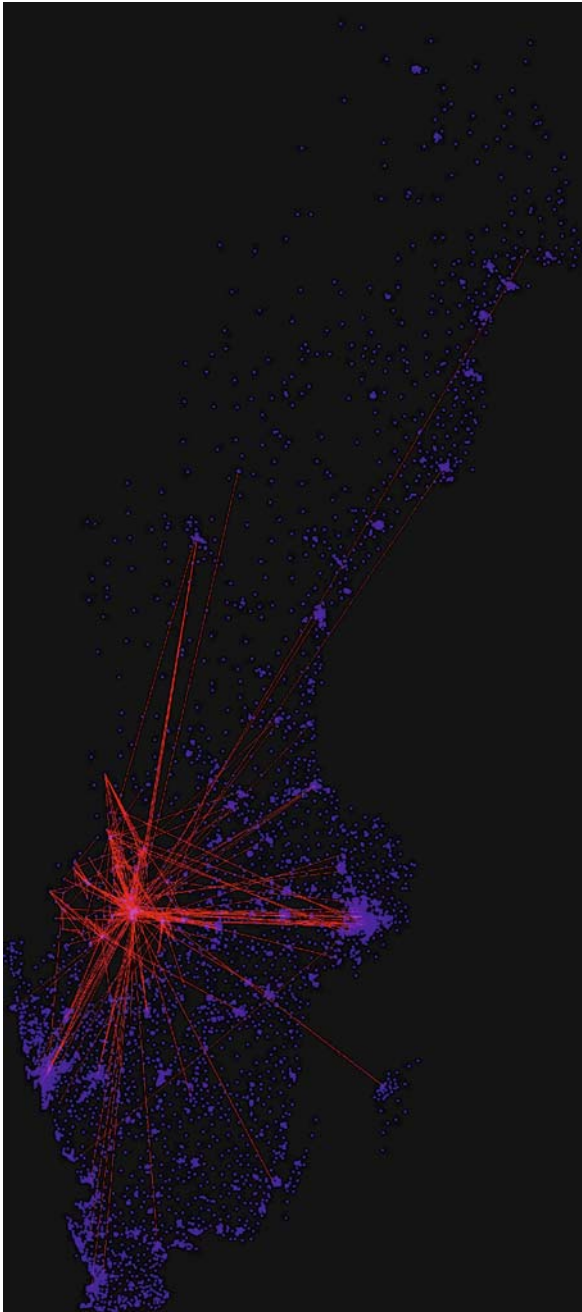
Recently, Balázs Szendroi and Gábor Csányi [45] have proposed that the polynomial pattern of spreading could be explained by the fact that sexual networks should show a high degree of transitivity. This may hold true for a population of gay men; it is not clear, however, that this thesis would be valid for heterosexual sexual contacts. In heterosexual networks, there are no tricycles per definition, so in these parts of the sexual network, an epidemic course must be slowed down by four cycles. One study of a romantic and sexual network presents a result that indicates the existence of a norm against changing each other's partner, which may decrease the effect of local clustering even more [3].

The Effect of Geographical Space

Contagious diseases that are spread by a contact network that is embedded in a two-dimensional space where only local interaction takes place will also have a polynomial pattern of spread. This property can easily be understood by the following thought experiment. Assume that we have a large population distributed on a square lattice so that there is one individual on each vertex of the lattice. If we then infect an individual in the center of the square lattice with a chronic disease that is spread to all neighbors every 24 h, it is easy to understand that the cumulative number of infected persons after t days and nights will be $(2 \cdot t + 1)^2$. That is to say, it will have a polynomial pattern of spread. It is not likely, however, that sexual contacts are sufficiently local for a polynomial pattern of spread to be observed. This is because strong evidence indicates that enough contacts extend over social and geographical distances, usually referred to as *social and spatial bridges* [47], for the sexual contact network to exhibit a so called *small world* quality [46]. That is to say, the average distance increases logarithmically with the network's size in the same way a random network does. Figure 2, for example, shows how different parts of the region Värmland in Sweden are sexually connected to each other and to the rest of Sweden by sexual contacts of individuals that tested positive for chlamydia in Värmland [36].

The Long Tail

According to Colgate et al. [5], the polynomial pattern of spread could be explained by a great variation in behavior that involves the risk of being infected by HIV.



Human Sexual Networks, Figure 2

The spatial distribution of the sexual contacts outside of Värmland County (contacts outside Sweden not shown) [36]

They demonstrated that if a behavior involving risk follows a power-law function, that is to say, the largest group indulges in a relatively safe kind of behavior, the next largest is a little riskier, etc., this would be enough, to-

gether with a tendency to have sex with individuals exhibiting similar risk behavior, to generate a polynomial pattern of spread. First, the small high-risk group will be quickly infected to the point of local saturation. The disease will then spread gradually to groups with lower and lower risk levels. In these groups, the infection will spread increasingly slowly as the epidemic continues. What is interesting is that Colgate et al. presented empirical evidence that at an STD clinic the distribution of the number of sexual contacts in the risk category of homosexual men followed a power-law with exponent -3 .

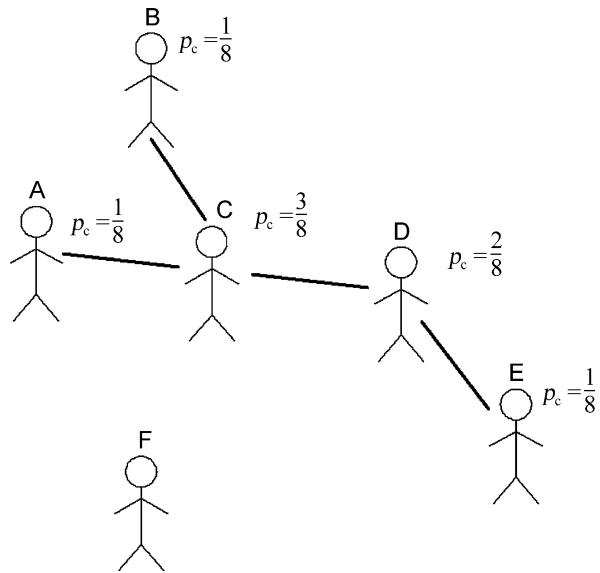
There are many examples that show the error of trying to generalize a certain behavior pattern from a group of individuals in a community to the whole community. To do so, it is necessary for these individuals to have been chosen at random, independently of each other. The first study indicating that the frequency distribution for the number of sexual contacts of an entire nation is very close to a power-law tail was reported by Liljeros et al. [23], using the results of a study of the sexual habits of Swedish citizens, "Sex in Sweden" [22]. It was shown that the number of sexual contacts for men and women did fit a power law for the upper tail of reported contacts during the past twelve months and reported contacts over the lifetime. The data for the Sex in Sweden study were gathered in 1996.

The method used to estimate tail inclination was criticized after the results of the Swedish study were published [11,17,18,24]. The study was conducted on a rather small dataset but had the additional strength of showing a power-law tail for both partners during the past 12 months and for the number of partners over a lifetime. Recently, however, Schneeberger et al. [43] reported a similar power-law-looking distributions on a dataset that can be described as the "Rolls Royce" of national sexual studies that has been carried out so far, the NATSAL 2000 [16]. The NATSAL 2000 was a study of 10 000 individuals in Great Britain who answered a survey at their home on a laptop that was brought to them by a research assistant. The research assistant did not sit in the same room while the respondent answered the questions, and the respondent was told that those answers were encrypted so that the research assistant would be unable to read them. Schneeberger et al. [43] reported the slope of the power law at 2.5 for heterosexual men, 3.1 for heterosexual women, 3.3 for homosexual women, and 1.6 for homosexual men. There will probably always be questions about human sexual behavior that cannot be answered due to lack of empirical data, such as the location of the upper cutoff of the distribution that by necessity must exist due to space and time constraints. Interestingly, a recent study of flirting on an Internet dating community reported a similar power-law-

like tail in the number of contacts that people had with each other on the site [15]. This data source is, however associated with large uncertainty when it comes to mapping the actual sexual network, because no information exists about actual sexual contact.

A power-law distribution in the number of sexual contacts has, under some specific circumstances, recently been shown to have serious consequences for the potential for eradicating sexually transmitted infections [37,38]. In an infinite population with homogeneous mixing, a slope smaller than three makes the second moment of the distribution infinitely large, and therefore also the variance of the distribution [25]. It can easily be seen from Eq. (5) that this will result in an infinitely large R_0 for the population. This has the bizarre consequence that all individuals must be tested and treated at the same time, otherwise it will be impossible to eradicate the disease. There are, however, several reasons to believe that this is not the case for sexual networks. We know, for example, that every human population must by definition have a finite size. Even though it has been shown that a power-law distribution in number of contacts will increase R_0 significantly in a finite population (as compared to a population with the same mean number of contacts but a low variance in number of contacts – see again Eq. (5), and [39], this does not imply that an epidemic outbreak of an STI cannot be stopped, or at least curtailed. Another factor that probably mitigates the effect of the skewed distribution is that there must also be an upper limit for how many sexual contacts a single individual can have per unit of time. A study of prostitutes in the United States shows, however, that this limit can be very large. The median value for number of partners was found to be as large as 103 during the previous six months [4]. One advantage of the skewed distribution, from an STI prevention perspective, is that it is predicted that R_0 can be drastically reduced and the epidemic eventually stopped if the individuals who change partners frequently can be tested and convinced to practice safe sex [6]. It may at first glance seem difficult to identify, for a specific targeted intervention, individuals who change partners frequently, except perhaps for specific groups such as prostitutes, or gay men who visit video clubs where anonymous sex takes place. Contact tracing, however (which is discussed in more detail in Sect. “Data Sources”), has the positive side effect that individuals who have many contacts have a higher probability of showing up – because they have a larger group of sources of infection – than do individuals with fewer contacts.

There are several ways to generate networks with a distribution of contacts similar to the one observed in sexual networks. The model so far given the most attention is



Human Sexual Networks, Figure 3

A snapshot at an early stage of a BA model for generating a network with a degree distribution power-law tail. The probability p_c that an already connected individual (i.e., one of A–E) will connect to the new individual F is proportional to the connected individual's number of contacts

one proposed by Albert-László Barabási and Réka Albert (the BA model) [2]. This model is based on an idea that can be traced back to the work of Herbert Simon [44] and J.D. Price [41] and works in the following way: start with a small number of vertices, and continuously add new vertices. Let the new vertices connect to one or several of the already existing vertices. Do this with a probability that is linear in proportion with the number of contacts these already existing vertices have.

Figure 3 shows a snapshot of the BA model, where the new vertex F is about to connect to the network of already connected vertices. The probability of F connecting to vertex D at this stage is twice the probability of F connecting to vertex E; and F's probability of connecting to C is three times the probability of F connecting to vertex E. The BA model generates a network, whose degree distribution could be described by a Yule distribution where $\alpha = 3$.

$$p(k|\alpha) = \frac{(\alpha - 1)\Gamma(k)\Gamma(\alpha)}{\Gamma(k + \alpha)}. \quad (8)$$

The BA model was originally proposed as a model for how the World Wide Web grows over time, that is to say, how a new homepage links to already existing homepages. As other types of networks have started to be analyzed, several modified BA models have been proposed that aim at finding special properties in these networks. The original

BA model assumes that new connections are made only between old and new vertices, which is not correct for sexual networks. In other words, it does not permit the updating of new links to already existing pages. Réka Albert and Albert-László Barabási [2] have shown that it is possible to generate power-law distributions leaning more steeply than 3 if the formation and resolving of links between already existing nodes is also permitted.

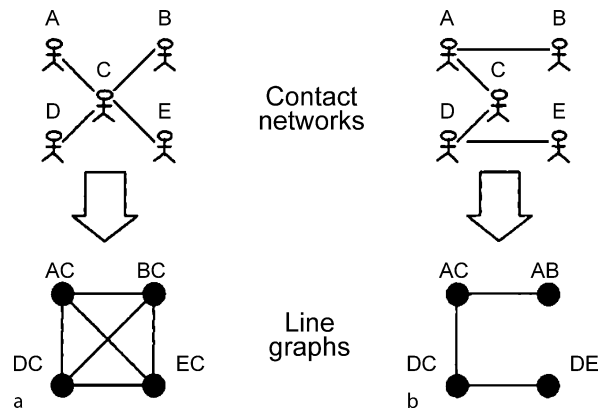
Surveys of sexual behavior are usually both too small and contain too little information about how the individual's number of partners grows over time to use standard methods to measure preferential attachment. One exception is an MLE-based expectation-maximization fitting technique developed especially for estimating preferential attachment in sexual survey data [10]. This method has demonstrated a significant effect of sublinear preferential attachment in partner growth over time on Norwegian survey data.

Theoretically, it is possible to argue for at least three different mechanisms that may cause preferential attachment. Like most other social behavior, it ought to be possible to get better at flirting and picking up through practice. An individual may also, under some circumstances, be considered to be more attractive the more partners he/she has had. Finally, we know that getting a new partner, like any other initially pleasurable behavior, can be psychologically addictive. To date we have no empirical results verifying the extent to which each of these different preferential attachment mechanisms are operative.

The Importance of Concurrent Relationship

The number of unprotected sexual contacts is clearly associated with an individual's risk for both being infected by a sexually transmitted infection, and for passing on the infection once infected. It is, however, possible that the number of sexual contacts is not the most important risk factor for getting an STI per se. Morris and Kretzschmar suggested in a series of articles [20,30,31] that it is the frequency of concurrent relations (partnerings that overlap over short time periods) in a population that is the most important factor in the transmission of sexually transmitted infections. Concurrent relations are important because potential contacts for the transmission of an STI come much closer in time if individuals have concurrent relations than if they practice serial monogamy.

This approach gives rise to a special type of graph called a line graph [12] in which the contacts between the persons are seen as nodes in a network. When we let a contact between two persons define a node on the graph, an edge is present whenever a person has more than one con-



Human Sexual Networks, Figure 4

Two contact networks and corresponding line graphs following Morris and Kretzschmar [30]. In the line graph, every edge in the contact network is translated into a node; for example, edge A–C in a becomes the node AC. Nodes with a degree > 1 in the contact network will contribute to new edges in the line graph. For example, in b there is an edge between D and C and D and E in the contact network (C has a degree of 3), thus there will be an edge between DC and DE in the line graph

tact. Two graphical examples of sexual contacts and their corresponding line graphs are shown in Fig. 4. Both contact networks displayed in Fig. 4a and b have the same average degree. Despite this, it is much easier for an STI to propagate in the left network than in the right one.

This is due to the fact that the level of concurrent sexual relations is much higher in network Fig. 4a than in network Fig. 4b, as can be seen by considering the corresponding line graphs. A measure for assessing the level of concurrency in a line graph has been suggested by Morris and Kretzschmar [20]. The concurrency, κ_2 , is given by the following equation:

$$\kappa_2 = L_2 \left(\frac{N_2(N_2 - 1)}{2} \right)^{-1} = \begin{cases} 1 & \text{all pairs adjacent} \\ 0 < \kappa_2 < 1 & \text{some pairs adjacent} \\ 0 & \text{no pairs adjacent (monogamy)} \end{cases} \quad (9)$$

Here L_2 is the number of links and N_2 is the number of nodes in the line graph (i.e., κ_2 is the density of the line graph). By further calculating the mean number of concurrent relationships per relationship, the index of concurrency κ_3 , it has been demonstrated that concurrency is a function of the mean and the standard deviation of the degree distribution [20]. Since these properties can be calculated solely on the basis of ego-network data (i.e., lo-

cal information), it is possible to estimate concurrency by using random samples of the population [20,30], which is very difficult for several other network measures such as density and component size [9].

The idea of using line graphs [20,30] has been developed even further by taking into consideration that an STI can propagate between two non-concurrent sexual relationships that occur relatively closely in time [42]. To handle this possibility in defining the line graph, it has been suggested that a sexual relation should be viewed as active for some period after the sexual partnership ended, depending on the type of disease.

Assortative Interaction

A tendency for individuals to prefer sexual contact with persons similar to themselves is usually referred to as *assortative interaction* in the STI literature. A tendency toward assortative interaction has been reported for social factors such as social class and ethnicity [24]. An important property of most sexually networks is that they are assortative by number of contacts [32,33,34]. This means that individuals who have many contacts tend to have contact with other individuals who also have many contacts. High assortativity decreases the epidemic threshold because a large interconnected component will emerge at a lower average density. The standard measure of assortativity is the assortative mixing coefficient r ; that is, Pearson's correlation coefficient between the individuals' degrees on each side of the edges [32,33]. As the edges are undirected, we need a coefficient that is invariant to edge reversal. This can be obtained for an edge (i, j) by including both (k_i, k_j) and (k_j, k_i) in the correlation coefficient, which can be expressed mathematically as:

$$r = \frac{4\langle k_1 k_2 \rangle - \langle k_1 + k_2 \rangle^2}{2\langle k_1^2 + k_2^2 \rangle - \langle k_1 + k_2 \rangle^2}, \quad (10)$$

where k_1 (k_2) is the degree of the first (and second) argument as it appears in the edge list.

Data Sources

The study of sexual networks can to some extent be compared to the study of planets in other solar systems in the sense that they can only be studied indirectly. Our knowledge comes from at least four different sources, each of which has its specific advantages and disadvantages. The first consists of national surveys of sexual behavior [16,21,22]. The advantage of such studies is that, given that they are based on random population samples, they are the only kind of study that can theoretically

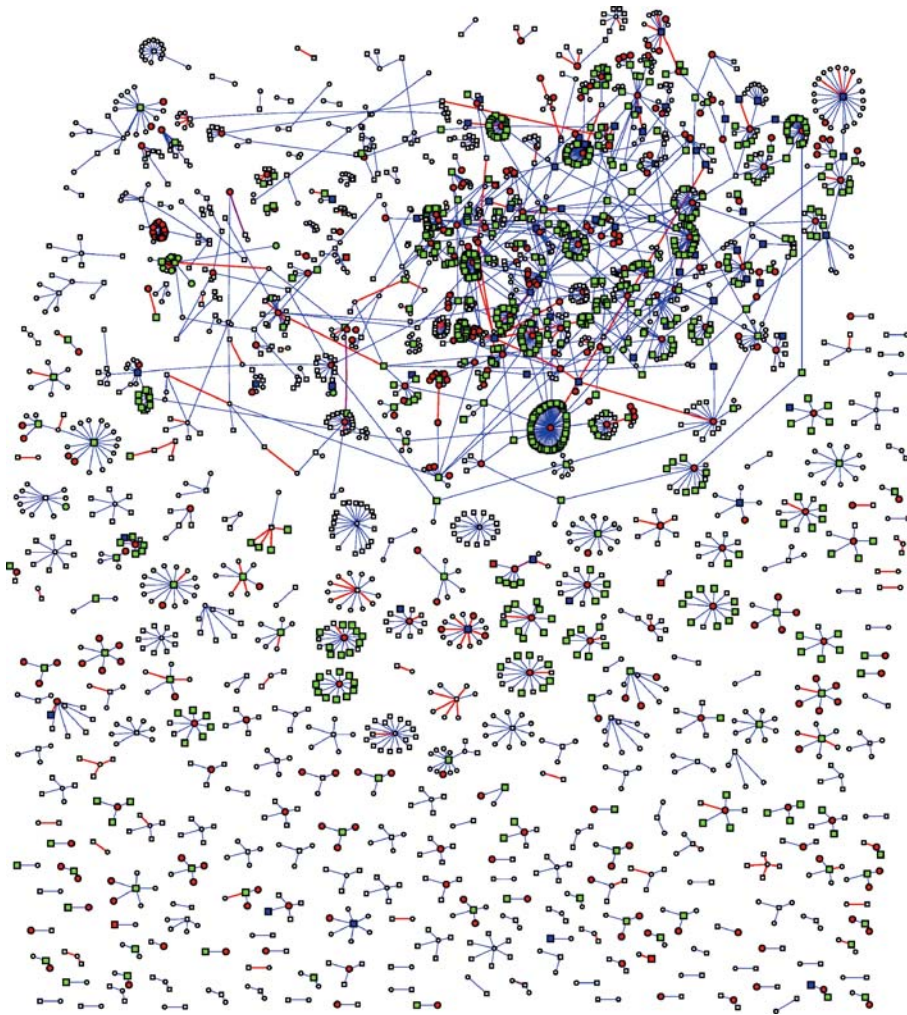
yield knowledge about sexual behavior for the whole population. Unfortunately, such studies have several drawbacks. For one thing, it is very hard to control for the accuracy of respondents' answers. Also, the response rate is also usually too low to guarantee reliable precision in the estimates.

One striking thing about national surveys, for example, is that men on average report a significantly higher number of sexual partners than do women. This discrepancy has been explained by a tendency for men as a group to over-report the number of sexual partners [27]. A recent study has also shown that this difference may be explained by the fact that prostitutes are usually not included in the samples [4]. National surveys have another disadvantage that is probably more important than the validity problem mentioned above, namely that they are only able to give us information about the behavior of the respondents, and not information about the behavior of their sexual partners (and *their* partners). National surveys cannot therefore give us information about the global properties of a network, such as level of clustering, size of the largest interconnected component, or average distance between the individuals.

A second source of information about sexual networks is the network data generated by *contact tracing* [47], which is the process whereby the contacts of an individual who has tested positive for an STI are traced and tested. If this procedure is also continued for the contacts that tested positive, and for their contacts in turn, it is eventually possible to generate a subgraph of a sexual network. This subnetwork can then be used to analyze global structural properties of the network that cannot be studied with national survey data. This kind of sampling is also associated with severe biases. It is not always possible for an individual, even if s/he is cooperative, to give enough information about any given sexual partner to be able to identify him or her. A more serious bias is that, by definition, contact tracing has a tendency to identify the subnetworks of the general sexual networks in which it is easiest for the disease to spread. The latter may not, however, always be a problem. If, for example, the purpose of a study is to find ways to mitigate the epidemic in the parts of the network in which the STIs are spread, this type of biased data really can be very useful.

A third type of data source is the mapping of the sexual network. This has so far only been carried out at the local level, probably for practical reasons, for example with a high risk group for HIV of drug users and prostitutes in a town in the United States (see Fig. 5) [40].

Another example is the study of a network of romantic relationships in a high school in the United States [3].



Human Sexual Networks, Figure 5

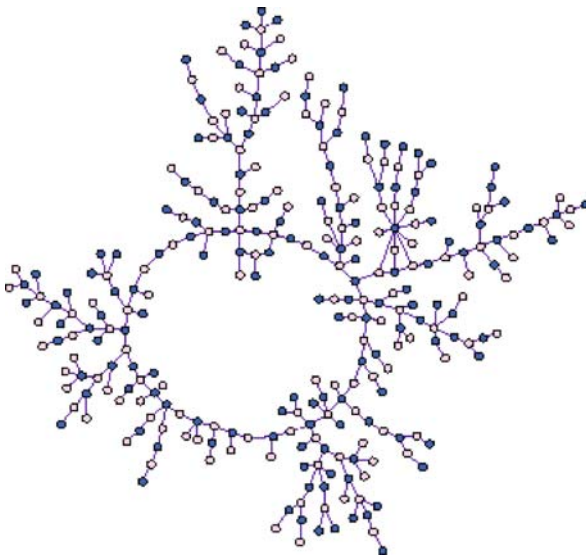
A sexual network of a high risk group for HIV infection in a town in the US. (Line colorings are *blue* for vaginal sexual contact, *red* for anal sexual contact. Node colorings are *red* for prostitutes, *blue* for pimps, and *green* for sex-buyers) [40]

These studies have the same problem as contact tracing studies when it comes to identifying and finding partners. It is also important to remember that these two studies cannot straightforwardly be generalized to the rest of the population, as studies based on a random sample can.

The fourth and last data source is data generated from Internet dating communities [15]. Because much of the activity that takes place in such communities is logged, it is possible to extract a network about which members interact. The advantage of this data source is that all contacts in the community can be traced, and that networks of a significant size can be generated. This data source is, however, probably the least valid network discussed here when

it comes to mapping the actual sexual network, because no information exists about actual sexual contact.

The classification of data sources into the four groups cited here should be seen as a preliminary one. There are also combinations of such as surveys based on non-random samples, for example, individuals who are seen at STI clinics [5] or convenient samples of university students. This short survey of data sources for sexual networks shows that all data sources are associated with different kinds of validity problems. This is quite a common situation in the social sciences. One way to mitigate, if not to solve, the problem is to try to confirm an empirical result with data from different sources, a procedure usually referred to as triangulation.



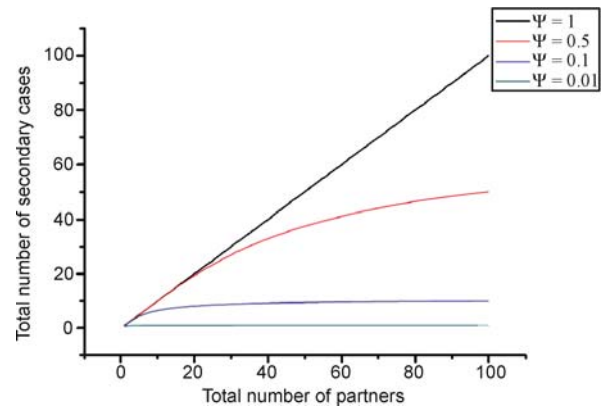
Human Sexual Networks, Figure 6

The largest component in network of romantic relationships in a high school in the United States (romantic relationships with non-students not included) [3]

Future Directions

To date, the risk of becoming infected with an STI and the risk of passing on the STI have been assumed to grow with partner turnover rate. A recent analysis of a small Swedish survey which contained unusually detailed information on sexual behavior has, to some extent, put this assumption into question [35]. The study suggested that the increased risk of high partner turnover rate may be compensated for by the fact that individuals with a high partner turnover rate seem to have a tendency to have fewer sexual intercourses with each partner, and therefore a lower risk for STI transmission per partner. The effect of this tendency seems to be especially important for STIs with a low risk of transmission per intercourse as shown in Fig. 7.

The high incidence of the most common STI, *Chlamydia trachomatis*, that has been reported in most Western societies may look like an anomaly in light of these results. Recent work based on a simulation model suggests, however, that a small set of individuals who report a large number of partners in sexual surveys are not necessarily as indispensable for transmitting STIs as generally thought [29]. The authors argue instead that multiconnected components, that is, network components such as network cycles in which each individual reaches each other individual in mutually exclusive ways, may be of greater importance for the spread of STIs. Interestingly, their preliminary simulation results show that large bicomponents,



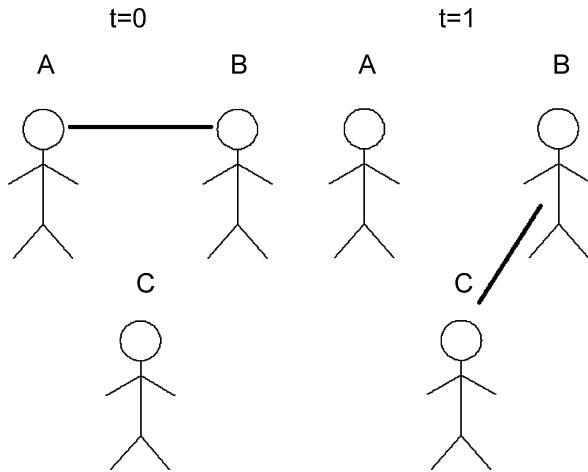
Human Sexual Networks, Figure 7

Total number of secondary cases per total number of partners, when each individual has 100 sexual intercourses evenly distributed among his or her partners (Ψ = probability of transmission per act of sexual intercourse) [35]

that is, components in which each individual can reach each other individual in two different, mutually exclusive ways, may emerge in networks that have a low variance and a relatively low mean number of contacts.

There is a tendency in most network research to simplify dynamic networks to their static equivalents, where the links that exist during a more or less long period are assumed to be concurrent. This also holds true of research about sexual networks even though the formation of and resolution of sexual relationship is to some extent taken into consideration in the notion of concurrent relations. Concurrence cannot, however, capture the fact that A can infect C, but not vice versa, in Fig. 8. One exception to this is Moody's pioneering "The importance of relationship timing for diffusion" [26] in which he presents a formalized framework for describing dynamic networks.

The results presented here indicate that many different structural properties must be taken into consideration to understand the spread of STIs. It is therefore not likely that a single solution to the problem will ever be found. The solution probably lies in a combination of broad and targeted interventions. It is also important to remember that efforts to control STIs in many Western societies have already drastically decreased their incidence, especially in the case of gonorrhea (even though the disease has not been totally eradicated). The fact that an increase in the incidence of STIs has been reported in some Western countries is a warning that must be taken seriously [14]. It is also important to remember that STIs are now a global problem, and as long as travel patterns between countries



Human Sexual Networks, Figure 8

An example of how the order of relationship may be of importance for the transmission of a sexually transmitted disease. A may indirectly infect C but not vice versa

and continents persist, STIs will never be eradicated in one part of the world as long as STIs are still endemic in the rest of the world.

Bibliography

- Anderson R, May RM (1991) Infectious diseases of humans. Oxford University Press, Oxford
- Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512
- Bearman PS, Moody J et al (2004) Chains of affection: The structure of adolescent romantic and sexual networks. *Am J Sociol* 110(1):44–91
- Brewer DD, Potterat JJ, Garrett SB, Muth SQ, Roberts JM, Kazprzyk D, Montano DE, Darrow WW (2000) Prostitution and the sex discrepancy in reported number of sexual partners. *Proc Natl Acad Sci USA* 97:12385–12388
- Colgate SA, Stanley EA et al (1989) Risk behavior-based model of the cubic growth of acquired immunodeficiency syndrome in the United States. *Proc Nat Acad Sci US* 86(12):4793–4797
- Dezso Z, Barabasi AL (2002) Halting viruses in scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 65(5 Pt 2): 055103
- Diekmann O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious disease. John Wiley and Son, Chichester
- Foulkes MA (1998) Advances in HIV/AIDS statistical methodology over the past decade. *Stat Med* 17(1):1–25
- Frank O (1971) Statistical inference in graphs. FOA, Stockholm
- Freiesleben de Blasio B, Svensson B et al (2007) Preferential attachment in sexual networks. *PNAS* 104(26):10762–10767
- Handcock MS, Jones JH (2004) Likelihood-based inference for stochastic models of sexual network formation. *Theor Popul Biol* 65(4):413–22
- Harary F (1969) Graph theory. Addison-Wesley, Reading
- Hethcote H, Yorke JA (1984) Gonorrhea transmission dynamics and control. Springer, New York
- Hiltunen-Back E, Haikala O et al (2003) Nationwide increase of Chlamydia trachomatis infection in Finland – Highest rise among adolescent women and men. *Sex Transm Dis* 30(10): 737–741
- Holme P, Edling CR et al (2004) Structure and time evolution of an Internet dating community. *Social Netw* 26(2):155–174
- Johnson AM, Mercer CH et al (2001) Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours. *Lancet* 358(9296):1835–1842
- Jones JH, Handcock MS (2003) An assessment of preferential attachment as a mechanism for human sexual network formation. *Proc Biol Sci* 270(1520):1123–1128
- Jones JH, Handcock MS (2003) Social networks: Sexual contacts and epidemic thresholds. *Nature* 423(6940):605–606; discussion 606
- Klov Dahl AS (1985) Social networks and the spread of infectious diseases: the AIDS example. *Soc Sci Med* 21(11):1203–1216
- Kretzschmar M, Morris M (1996) Measures of concurrency in networks and the spread of infectious disease. *Math Biosci* 133(2):165–195
- Laumann EO, Gagnon JH et al (1994) The social organization of sexuality. University of Chicago Press, Chicago
- Lewin B (ed) (2000) Sex in Sweden. The Swedish National Institute of Public Health, Stockholm
- Liljeros F, Edling CR et al (2001) The web of human sexual contacts. *Nature* 411(6840):907–908
- Liljeros F, Edling CR et al (2003) Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes Infect* 5(2):189–196
- Lloyd AL, May RM (2001) Epidemiology. How viruses spread among computers and people. *Science* 292(5520): 1316–1317
- Moody J (2002) The importance of relationship timing for diffusion. *Social Forces* 81(1):25–56
- Morris M (1993) Telling tails explain the discrepancy in sexual partner reports. *Nature* 365(6445):437–440
- Morris M (ed) (2004) Network epidemiology: A handbook for survey design and data collection. Oxford University Press Inc, New York
- Morris M, Goodreau S et al (2007) Sexual networks, concurrency, and STD/HIV. In: Holmes KK, Sparling PF, Stamm WE (eds) Sexually transmitted diseases. McGraw-Hill, New York
- Morris M, Kretzschmar M (1995) Concurrent partnerships and transmission dynamics in networks. *Social Netw* 17(3–4): 299–318
- Morris M, Kretzschmar M (1997) Concurrent partnerships and the spread of HIV. *Aids* 11(5):641–648
- Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89(20):1–4
- Newman MEJ (2003) Mixing patterns in networks. *Phys Rev E* 67(2):1–13
- Newman MEJ (2003) Properties of highly clustered networks. *Phys Rev E* 68(2):026121
- Nordvik MK, Liljeros F (2006) Number of sexual encounters involving intercourse and the transmission of sexually transmitted infections. *Sex Transm Dis* 33(6):342–349
- Nordvik MK, Liljeros F et al (2007) Spatial bridges and the spread of Chlamydia: the case of a county in Sweden. *Sex Transm Dis* 34(1):47–53
- Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and

endemic states in complex networks. *Phys Rev E* 63(066117): 1–8

38. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys Rev Lett* 86:3200–3203
39. Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 65(3 Pt 2A): 035108
40. Poterat JJ, Woodhouse DE et al (2004) Network dynamism: history and lessons of the Colorado Springs study. In: Morris M (ed) *Network epidemiology: A Handbook for survey design and data collection*. Oxford University Press Inc, New York, pp 87–114
41. Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. *J Am. Soc. Inform. Sci* 27: 292–306
42. Riolo CS, Koopman JS et al (2001) Methods and measures for the description of epidemiologic contact networks. *J Urban Health* 78(3):446–457
43. Schneeberger A, Mercer CH et al (2004) Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sex Transm Dis* 31(6):380–387
44. Simon HA (1955) On a class of skew distribution functions. *Biometrika* 42:425–440
45. Szendroi B, Csányi G (2004) Polynomial epidemics and clustering in contact networks. *Proc Biol Sci Aug* 7:271 Suppl 5:S364–6
46. Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393(6684):440–442
47. Wylie JL, Jolly A (2001) Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. *Sex Transm Dis* 28(1):14–24

Hybrid Control Systems

ANDREW R. TEEL¹, RICARDO G. SANFELICE²,
RAFAL GOEBEL³

¹ Electrical and Computer Engineering Department,
University of California, Santa Barbara, USA

² Department of Aerospace and Mechanical Engineering,
University of Arizona, Tucson, USA

³ Department of Mathematics and Statistics, Loyola
University, Chicago, USA

Article Outline

[Glossary](#)

[Notation](#)

[Definition of the Subject](#)

[Introduction](#)

[Well-posed Hybrid Dynamical Systems](#)

[Modeling Hybrid Control Systems](#)

[Stability Theory](#)

[Design Tools](#)

[Applications](#)

[Discussion and Final Remarks](#)

Future Directions

Bibliography

Glossary

Global asymptotic stability The typical closed-loop objective of a hybrid controller. Often, the hybrid controller achieves global asymptotic stability of a compact set rather than of a point. This is the property that solutions starting near the set remain near the set for all time and all solutions tend toward the set asymptotically. This property is robust, in a practical sense, for well-posed hybrid dynamical systems.

(Well-posed) Hybrid dynamical system

System that combines behaviors typical of continuous-time and discrete-time dynamical systems, that is, combines both flows and jumps. The system is said to be well-posed if the data used to describe the evolution (consisting of a flow map, flow set, jump map, and jump set) satisfy mild regularity conditions; see conditions (C1)–(C3) in Subsect. “[Conditions for Well-posedness](#)”.

Hybrid controller Algorithm that takes, as inputs, measurements from a system to be controlled (called the plant) and combines behaviors of continuous-time and discrete-time controllers (i.e. flows and jumps) to produce, as outputs, signals that are to control the plant.

Hybrid closed-loop system The hybrid system resulting from the interconnection of a plant and a controller, at least one of which is a hybrid dynamical system.

Invariance principle A tool for studying asymptotic properties of bounded solutions to (hybrid) dynamical systems, applicable when asymptotic stability is absent. It characterizes the sets to which such solutions must converge, by relying in part on invariance properties of such sets.

Lyapunov stability theory A tool for establishing global asymptotic stability of a compact set without solving for the solutions to the hybrid dynamical system. A Lyapunov function is one that takes its minimum, which is zero, on the compact set, that grows unbounded as its argument grows unbounded, and that decreases in the direction of the flow map on the flow set and via the jump map on the jump set.

Supervisor of hybrid controllers A hybrid controller that coordinates the actions of a family of hybrid controllers in order to achieve a certain stabilization objective. Patchy control Lyapunov functions provide a means of constructing supervisors.

Temporal regularization A modification to a hybrid controller to enforce a positive lower bound on the amount of time between jumps triggered by the hybrid control algorithm.

Zeno (and discrete) solutions A solution (to a hybrid dynamical system) that has an infinite number of jumps in a finite amount of time. It is discrete if, moreover, the solution never flows, i. e., never changes continuously.

Notation

- \mathbb{R}^n denotes n -dimensional Euclidean space. \mathbb{R} denotes the real numbers. \mathbb{R}_{\geq} denotes the nonnegative real numbers, i. e., $\mathbb{R}_{\geq} = [0, \infty)$. \mathbb{Z} denotes the integers. \mathbb{N} denotes the natural numbers including 0, i. e., $\mathbb{N} = \{0, 1, \dots\}$.
- Given a set S , \bar{S} denotes its closure.
- Given a vector $x \in \mathbb{R}^n$, $|x|$ denotes its Euclidean vector norm.
- \mathbb{B} is the closed unit ball in the norm $|\cdot|$.
- Given a set $S \subset \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$, $|x|_S := \inf_{y \in S} |x - y|$.
- Given sets S_1, S_2 subsets of \mathbb{R}^n , $S_1 + S_2 := \{x_1 + x_2 \mid x_1 \in S_1, x_2 \in S_2\}$.
- A function is said to be *positive definite with respect to a given compact set* in its domain if it is zero on that compact set and positive elsewhere. When the compact set is the origin, the function will be called *positive definite*.
- Given a function $h: \mathbb{R}^n \rightarrow \mathbb{R}$, $h^{-1}(c)$ denotes its c -level set, i. e. $h^{-1}(c) := \{z \in \mathbb{R}^n \mid h(z) = c\}$.
- The double-arrow notation e. g., $g: D \rightrightarrows \mathbb{R}^n$, indicates a set-valued mapping, in contrast to a single arrow used for functions.

Definition of the Subject

Control systems are ubiquitous in nature and engineering. They regulate physical systems to desirable conditions. The mathematical theory behind engineering control systems developed over the last century. It started with the elegant stability theory of linear dynamical systems and continued with the more formidable theory of nonlinear dynamical systems, rooted in knowledge of stability theory for attractors in nonlinear differential or difference equations. Most recently, researchers have recognized the limited capabilities of control systems modeled only by differential or difference equations. Thus, they have started to explore the capabilities of hybrid control systems. Hybrid control systems contain dynamical states that some-

times change continuously and other times change discontinuously. These states, which can flow and jump, together with the output of the system being regulated, are used to produce a (hybrid) feedback control signal. Hybrid control systems can be applied to classical systems, where their added flexibility permits solving certain challenging control problems that are not solvable with other methods. Moreover, a firm understanding of hybrid dynamical systems allows applying hybrid control theory to systems that are, themselves, hybrid in nature, that is, having states that can change continuously and also change discontinuously. The development of hybrid control theory is in its infancy, with progress being marked by a transition from ad-hoc methods to systematic design tools.

Introduction

This article will present a general framework for modeling hybrid control systems and analyzing their dynamical properties. It will put forth basic tools for studying asymptotic stability properties of hybrid systems. Then, particular aspects of hybrid control will be described, as well as approaches to successfully achieving control objectives via hybrid control even if they are not solvable with classical methods. First, some examples of hybrid control systems are given.

Hybrid dynamical systems combine behaviors typical of continuous-time dynamical systems (i. e., flows) and behaviors typical of discrete-time dynamical systems (i. e., jumps). Hybrid control systems exploit state variables that may flow as well as jump to achieve control objectives that are difficult or impossible to achieve with controllers that are not hybrid. Perhaps the simplest example of a hybrid control system is one that uses a relay-type hysteresis element to avoid cycling the system's actuators between "on" and "off" too frequently.

Consider controlling the temperature of a room by turning a heater on and off. As a good approximation, the room's temperature T is governed by the differential equation

$$\dot{T} = -T + T_0 + T_{\Delta}u, \quad (1)$$

where T_0 represents the natural temperature of the room, T_{Δ} represents the capacity of the heater to raise the temperature in the room by being always on, and the variable u represents the state of the heater, which can be either 1 ("on") or 0 ("off").

A typical temperature control task is to keep the temperature between two specified values T_{\min} and T_{\max} where

$$T_0 < T_{\min} < T_{\max} < T_0 + T_{\Delta}.$$

For purposes of illustration, consider the case when $T_{\min} = 70^\circ\text{F}$, $T_{\max} = 80^\circ\text{F}$. An algorithm that accomplishes this control task is

```

input T, u
if u=1 and T >= 80 then
    u=0
elseif u = 0 and T <= 70 then
    u=1
end

```

In words, if the heater is “on” and the temperature is larger than 80°F , then the heater is turned off, while if the heater is “off” and the temperature is smaller than 70°F , then the heater is turned on. By programming the controller with the algorithm above and closing the loop, the temperature of the system will evolve as shown in Fig. 1 (the values $T_0 = 60^\circ\text{F}$ and $T_\Delta = 30^\circ\text{F}$ were used in these simulations).

Following the logic in the algorithm above, the controller can be expressed by the following conditional difference equation

$$\begin{aligned}
 u^+ &= 0 & \text{if } u = 1, T \geq 80 \\
 u^+ &= 1 & \text{if } u = 0, T \leq 70,
 \end{aligned}$$

where u^+ is the value of u after a jump. Combining this difference equation with the differential equation (1) leads to the closed-loop system

$$\left. \begin{aligned} \dot{T} &= -T + T_0 + T_\Delta u \\ \dot{u} &= 0 \end{aligned} \right\} \begin{aligned} &u = 0, T \geq 70 \quad \text{or} \\ &u = 1, T \leq 80 \end{aligned} \quad (2)$$

$$\left. \begin{aligned} T^+ &= T \\ u^+ &= 1 - q \end{aligned} \right\} \begin{aligned} &u = 1, T \geq 80 \quad \text{or} \\ &u = 0, T \leq 70. \end{aligned} \quad (3)$$

This closed-loop system is a hybrid dynamical system. The state variables are T and u ; the continuous dynamics

or *flows* as well as the constraints on the continuous evolution are given by (2); the discrete dynamics or *jumps* as well as the constraints on the discrete evolution are given by (3).

As in the temperature control problem above, a hybrid control system can arise from controlling a classical system with a hybrid controller. In a more general setting, hybrid control systems emerge when controlling hybrid systems with nonlinear and/or hybrid controllers. Consider controlling the vertical motion of a ball through collisions with an actuated robot. Figure 2 depicts such a scenario. The impacts between the ball and the robot generate a jump in their velocity. When the control task is to stabilize the ball to a periodic motion, like the height pattern in Fig. 2b, this system is referred to as the *one degree-of-freedom juggler*.

The dynamics of the ball in between the collisions are given by classical physics laws and can be written in terms of the ball's height and vertical velocity, denoted by x_{11} and x_{12} , respectively, as follows:

$$\begin{aligned} \dot{x}_{11} &= x_{12} \\ \dot{x}_{12} &= -\gamma, \end{aligned} \quad (4)$$

where γ is the gravity constant. We denote the mass of the ball by m_1 .

We consider a robot with a vertical velocity control input u and denote the robot's height and velocity by x_{21} and x_{22} , respectively. Then, its dynamics are given by

$$\begin{aligned} \dot{x}_{21} &= x_{22} \\ \dot{x}_{22} &= u. \end{aligned} \quad (5)$$

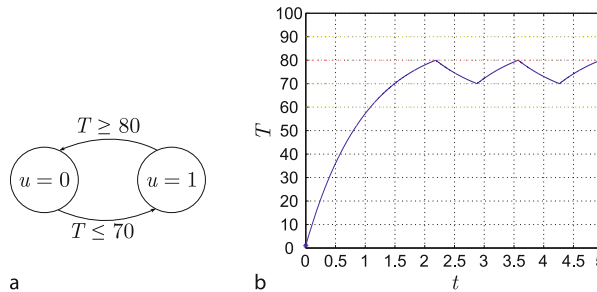
The mass of the actuated robot is denoted by m_2 .

Following [10,58], impacts between the ball and the robot are assumed to conserve momentum, i. e.,

$$m_1 x_{12}^+ + m_2 x_{22}^+ = m_1 x_{12} + m_2 x_{22}, \quad (6)$$

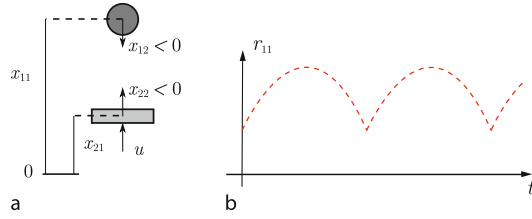
and, for some restitution coefficient $e \in (0, 1)$, satisfy

$$x_{12}^+ - x_{22}^+ = -e(x_{12} - x_{22}). \quad (7)$$



Hybrid Control Systems, Figure 1

Temperature control. **a** Flow diagram of control algorithm. **b** Evolution of temperature with control algorithm



Hybrid Control Systems, Figure 2

One degree-of-freedom juggler and a juggling task. Positions are denoted by x_{11} , x_{21} and velocities by x_{12} , x_{22} , respectively. The control input is denoted by u . A desired height pattern is denoted by r_{11} . **a** Ball and actuated robot. **b** Desired ball's height pattern

Defining $\lambda := \frac{m_1}{m_1 + m_2}$, Eqs. (6) and (7) can be combined to obtain the update law for velocities

$$\begin{bmatrix} x_{12}^+ \\ x_{22}^+ \end{bmatrix} = \begin{bmatrix} -e + \lambda(1+e) & (1-\lambda)(1+e) \\ \lambda(1+e) & 1 - \lambda(1+e) \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} \\ =: \Gamma(\lambda, e) \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix}.$$

The update law for positions is given by

$$x_{11}^+ = x_{11}, \quad x_{21}^+ = x_{21}.$$

Impacts between the ball and the actuated robot occur when their heights are the same, that is, $x_{11} = x_{21}$, and when their velocities indicate that they are not moving away from each other, that is, $x_{12} \leq x_{22}$.

The derivation above leads to the following model for the one degree-of-freedom juggler system in Fig. 2:

Flows:

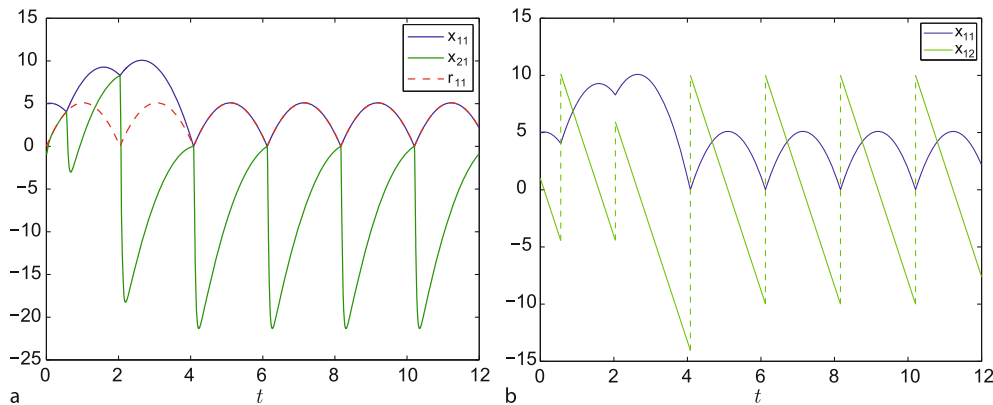
$$\left. \begin{aligned} \dot{x}_{11} &= x_{12}, & \dot{x}_{12} &= -\gamma \\ \dot{x}_{21} &= x_{22}, & \dot{x}_{22} &= u \end{aligned} \right\} \quad x_{11} - x_{21} \geq 0.$$

Jumps:

$$\left. \begin{aligned} x_{11}^+ &= x_{11} \\ x_{12}^+ &= [1 \quad 0] \Gamma(\lambda, e) \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} \\ x_{21}^+ &= x_{21} \\ x_{22}^+ &= [1 \quad 0] \Gamma(\lambda, e) \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} \end{aligned} \right\} \quad \begin{aligned} & x_{11} - x_{21} = 0 \\ & \text{and } x_{12} - x_{22} \leq 0. \end{aligned}$$

A controller designed to accomplish stabilization of the ball to a periodic pattern will only be able to measure the ball's state at impacts. During flows, it will be able to control the robot's velocity through u . Regardless of the nature of the controller, the closed-loop system will be a hybrid system by virtue of the dynamics of the one degree-of-freedom system. Figure 3 shows a trajectory to the closed-loop system with a controller that stabilizes the ball state to the periodic pattern in Fig. 2b (note the discontinuity in the velocity of the ball at impacts); see the control strategy in [55].

Following the modeling techniques illustrated by the examples above, the next section introduces a general modeling framework for hybrid dynamical systems. The framework makes possible the development of a robust stability theory for hybrid dynamical systems and prepares the way for insights into the design of robust hybrid control systems.



Hybrid Control Systems, Figure 3

Trajectories to the one degree-of-freedom juggler. Their positions are denoted by x_{11} , x_{21} and their velocities by x_{12} , x_{22} , respectively. The desired height pattern is denoted by r_{11} . **a** Juggling on ball. **b** Ball's position and velocity

Well-posed Hybrid Dynamical Systems

For numerous mathematical problems, well-posedness refers to the uniqueness of a solution and its continuous dependence on parameters, for example on initial conditions. Here, well-posedness will refer to some mild regularity properties of the data of a hybrid system that enable the development of a robust stability theory.

Hybrid Behavior and Model

Hybrid dynamical systems combine continuous and discrete dynamics. Such a combination may emerge when controlling a continuous-time system with a control algorithm that incorporates discrete dynamics, like in the temperature control problem in Sect. “Introduction”, when controlling a system that features hybrid phenomena, like in the juggling problem in Sect. “Introduction”, or as a modeling abstraction of complex dynamical systems.

Solutions to hybrid systems (sometimes referred to as *trajectories*, *executions*, *runs*, or *motions*) can evolve both continuously, i.e. *flow*, and discontinuously, i.e. *jump*. Figure 4 depicts a representative behavior of a solution to a hybrid system.

For a purely continuous-time system, flows are usually modeled by differential equations, and sometimes by differential inclusions. For a purely discrete-time system, jumps are usually modeled by difference equations, and sometimes by difference inclusions. Set-valued dynamics naturally arise as regularizations of discontinuous difference and differential equations and represent the effect of state perturbations on such equations, in particular, the effect of state measurement errors when the equations represent a system in closed loop with a (discontinuous) feedback controller. For the continuous-time case, see the work by Filippov [22] and Krasovskii [34], as well as [26,27,]; for the discrete-time case, see [33].

When working with hybrid control systems, it is appropriate to allow for set-valued discrete dynamics in order to capture decision making capabilities, which are typical in hybrid feedback. Difference inclusions, rather than

equations, also arise naturally in modeling of hybrid automata, for example when the discrete dynamics are generated by multiple so-called “guards” and “reset maps”; see, e.g. [8,11,40] or [56] for details on modeling guards and resets in the current framework.

Naturally, differential equations and difference inclusions will be featured in the model of a hybrid system. In most hybrid systems, or even in some purely continuous-time systems, the flow modeled by a differential equation is allowed to occur only on a certain subset of the state space \mathbb{R}^n . Similarly, the jumps modeled by a difference equation may only be allowed from a certain subset of the state space. Hence, the model of the hybrid system stated above will also feature a *flow set*, restricting the flows, and a *jump set*, restricting the jumps.

More formally, a hybrid system will be modeled with the following data:

- The flow set $C \subset \mathbb{R}^n$;
- The flow map $f: C \rightarrow \mathbb{R}^n$;
- The jump set $D \subset \mathbb{R}^n$;
- The (set-valued) jump map $G: D \rightrightarrows \mathbb{R}^n$.

A shorthand notation for a hybrid system with this data will be $\mathcal{H} = (f, C, G, D)$. Such systems can be written in the suggestive form

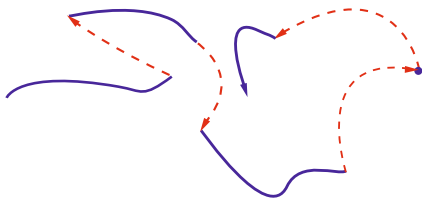
$$\mathcal{H}: \quad x \in \mathbb{R}^n \quad \begin{cases} \dot{x} = f(x), & x \in C \\ x^+ \in G(x), & x \in D, \end{cases} \quad (8)$$

where $x \in \mathbb{R}^n$ denotes the state of the system, \dot{x} denotes its derivative with respect to time, and x^+ denotes its value after jumps. In several control applications, the state x of the hybrid system can contain logic states that take value in discrete sets (representing, for example, “on” or “off” states, like in the temperature control problem in Sect. “Introduction”).

Two parameters will be used to specify “time” in solutions to hybrid systems: t , taking values in $\mathbb{R}_{\geq 0}$, and representing the elapsed “real” time; and j , taking values in \mathbb{N} , and representing the number of jumps that have occurred. For each solution, the combined parameters (t, j) will be restricted to belong to a hybrid time domain, a particular subset of $\mathbb{R}_{\geq 0} \times \mathbb{N}$. Hybrid time domains corresponding to different solutions may differ.

Note that with such a parametrization, both purely continuous-time and purely discrete-time dynamical systems can be captured. Furthermore, for truly hybrid solutions, both flows and jumps are parametrized “symmetrically” (cf. [8,40]).

A subset E of $\mathbb{R}_{\geq 0} \times \mathbb{N}$ is a *hybrid time domain* if it is the union of infinitely many intervals of the



Hybrid Control Systems, Figure 4

Evolution of a hybrid system: continuous motion during flows (solid), discontinuous motion at jumps (dashed)

form $[t_j, t_{j+1}] \times \{j\}$, where $0 = t_0 \leq t_1 \leq t_2 \leq \dots$, or of finitely many such intervals, with the last one possibly of the form $[t_j, t_{j+1}] \times \{j\}$, $[t_j, t_{j+1}) \times \{j\}$, or $[t_j, \infty) \times \{j\}$. On each hybrid time domain there is a natural ordering of points: we write $(t, j) \preceq (t', j')$ for $(t, j), (t', j') \in E$ if $t \leq t'$ and $j \leq j'$.

Solutions to hybrid systems are given by functions, which are called *hybrid arcs*, defined on hybrid time domains and satisfying the dynamics and the constraints given by the data of the hybrid system. A hybrid arc is a function $x: \text{dom } x \rightarrow \mathbb{R}^n$, where $\text{dom } x$ is a hybrid time domain and $t \mapsto x(t, j)$ is a locally absolutely continuous function for each fixed j . A hybrid arc x is a solution to $\mathcal{H} = (f, C, G, D)$ if $x(0, 0) \in C \cup D$ and it satisfies

Flow condition:

$$\dot{x}(t, j) = f(x(t, j)) \quad \text{and} \quad x(t, j) \in C \quad (9)$$

for all $j \in \mathbb{N}$ and almost all t such that $(t, j) \in \text{dom } x$;

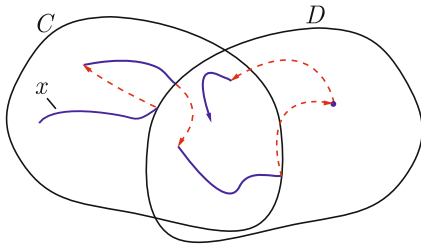
Jump condition:

$$x(t, j+1) \in G(x(t, j)) \quad \text{and} \quad x(t, j) \in D \quad (10)$$

for all $(t, j) \in \text{dom } x$ such that $(t, j+1) \in \text{dom } x$.

Figure 5 shows a solution to a hybrid system $\mathcal{H} = (f, C, G, D)$ flowing (as solutions to continuous-time systems do) while in the flow set C and jumping (as solutions to discrete-time systems do) from points in the jump set D .

A hybrid arc x is said to be *nontrivial* if $\text{dom } x$ contains at least one point different from $(0, 0)$ and *complete* if $\text{dom } x$ is unbounded (in either the t or j direction, or both). It is said to be *Zeno* if it has an infinite number of jumps in a finite amount of time, and *discrete* if it has an infinite number of jumps and never flows. A solution x to a hybrid system is *maximal* if it cannot be extended, i.e.,



Hybrid Control Systems, Figure 5

Evolution of a solution to a hybrid system. Flows and jumps of the solution x are allowed only on the flow set C and on the jump set D , respectively

there is no solution x' such that $\text{dom } x$ is a proper subset of $\text{dom } x'$ and x agrees with x' on $\text{dom } x$. Obviously, complete solutions are maximal.

Conditions for Well-Posedness

Many desired results in stability theory for dynamical systems, like invariance principles, converse Lyapunov theorems, or statements about generic robustness of stability, hinge upon some fundamental properties of the space of solutions to the system. These properties may involve continuous dependence of solutions on initial conditions, completeness and sequential compactness of the space of solutions, etc.

To begin addressing these or similar properties for hybrid systems, one should establish a concept of distance between solutions. In contrast to purely continuous-time systems or purely discrete-time systems, the uniform metric is not a suitable indicator of distance: two solutions experiencing jumps at close but not the same times will not be close in the uniform metric, even if (intuitively) they represent very similar behaviors. For example, Fig. 6 shows two solutions to the juggling problem in Sect. “Introduction” starting at nearby initial conditions for which the velocities are not close in the uniform metric.

A more appropriate distance notion should take possibly different jump times into account. We use the following notion: given $T, J, \varepsilon > 0$, two hybrid arcs $x: \text{dom } x \rightarrow \mathbb{R}^n$ and $y: \text{dom } y \rightarrow \mathbb{R}^n$ are said to be (T, J, ε) -close if:

- (a) for all $(t, j) \in \text{dom } x$ with $t \leq T, j \leq J$ there exists s such that $(s, j) \in \text{dom } y, |t - s| < \varepsilon$, and

$$|x(t, j) - y(s, j)| < \varepsilon,$$

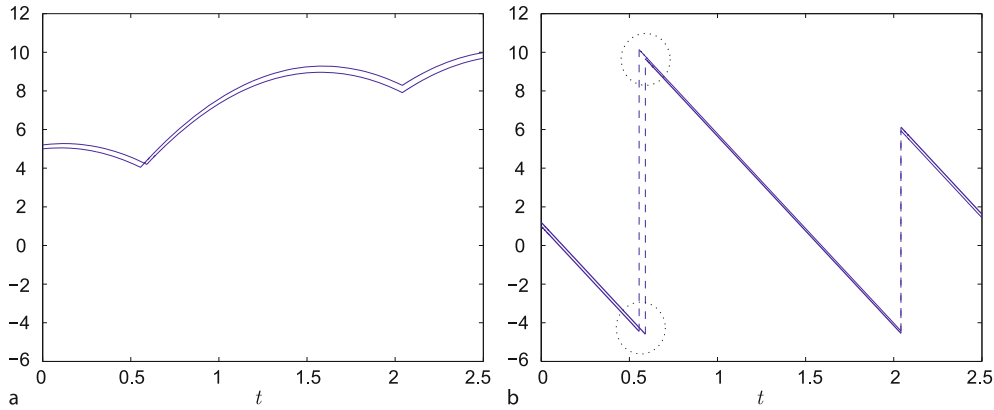
- (b) for all $(t, j) \in \text{dom } y$ with $t \leq T, j \leq J$ there exists s such that $(s, j) \in \text{dom } x, |t - s| < \varepsilon$, and

$$|y(t, j) - x(s, j)| < \varepsilon.$$

An appealing geometric interpretation of (T, J, ε) -closeness of x and y can be given. The graph of a hybrid arc $x: \text{dom } x \rightarrow \mathbb{R}^n$ is the subset of \mathbb{R}^{n+2} given by

$$\text{gph } x := \{(t, j, z) \mid (t, j) \in \text{dom } x, z = x(t, j)\}.$$

Hybrid arcs x and y are (T, J, ε) -close if the restriction of the graph of x to $t \leq T, j \leq J$, i.e., the set $\{(t, j, z) \mid (t, j) \in \text{dom } x, t \leq T, j \leq J, z = x(t, j)\}$, is in the ε -neighborhood of $\text{gph } y$, and vice versa: the restriction of the graph of y is in the ε -neighborhood of $\text{gph } x$. (The neighborhoods of $\text{gph } x$ and $\text{gph } y$ should be understood in the norm for which the unit ball is $[-1, 1] \times [-1, 1] \times \mathbb{B}$.)



Hybrid Control Systems, Figure 6

Two solutions to the juggling system in Sect. “Introduction” starting at nearby initial conditions. Velocities are not close in the uniform metric near the jump times. **a** Ball’s heights. **b** Ball’s velocities

The (T, J, ε) -closeness can be used to quantify the concept of graphical convergence of a sequence of hybrid arcs; for details, see [23]. Here, it is only noted that graphical convergence of a sequence of mappings is understood as convergence of the sequence of graphs of these mappings. Such a convergence concept does have solid intuitive motivation when the mappings considered are associated with solutions to hybrid systems.

It turns out that when (T, J, ε) -closeness and graphical convergence are used to study the properties of the space of solutions to a hybrid system $\mathcal{H} = (f, C, G, D)$, only mild and easy to verify conditions on the data of \mathcal{H} are needed to ensure that \mathcal{H} is “well-posed”. These conditions are:

- (C1) the flow set C and jump set D are closed;
- (C2) the flow map $f: C \rightarrow \mathbb{R}^n$ is continuous;
- (C3) the jump map $G: D \rightrightarrows \mathbb{R}^n$ is outer semicontinuous and locally bounded.

Only (C3) requires further comment: $G: D \rightrightarrows \mathbb{R}^n$ is outer semicontinuous if for every convergent sequence $x_i \in D$ with $x_i \rightarrow x$ and every convergent sequence $y_i \in G(x_i)$ with $y_i \rightarrow y$, one has $y \in G(x)$; G is locally bounded if for each compact set $K \subset \mathbb{R}^n$ there exists a compact set $K' \subset \mathbb{R}^n$ such that $G(x) \subset K'$ for all $x \in K$. Any system $\mathcal{H} = (f, C, G, D)$ meeting (C1), (C2), and (C3) will be referred to as well-posed.

One important consequence of a hybrid system \mathcal{H} being well-posed is the following:

(\star) Every sequence of solutions to \mathcal{H} has a subsequence that graphically converges to a solution to \mathcal{H} ,

which holds under very mild boundedness assumptions about the sequence of solutions in question. The assumptions hold, for example, if the sequence is uniformly bounded, i.e., there exists a compact set $K \subset \mathbb{R}^n$ such that, for each i , $x_i(t, j) \in K$ for all $(t, j) \in \text{dom } x_i$, where $\{x_i\}_{i=1}^\infty$ is a sequence of solutions.

Another important consequence of well-posedness is the following *outer-semicontinuity* (or *upper-semicontinuity*) property:

($\star\star$) For every $x^0 \in \mathbb{R}^n$, every desired level of closeness of solutions $\varepsilon > 0$, and every (T, J) , there exists a level of closeness for initial conditions $\delta > 0$ so that for every solution x_δ to \mathcal{H} with $|x_\delta(0, 0) - x^0| < \delta$, there exist a solution x to \mathcal{H} with $x(0, 0) = x^0$ such that x_δ and x are (T, J, ε) -close.

This property holds at each $x^0 \in \mathbb{R}^n$ from which all maximal solutions to \mathcal{H} are either complete or bounded.

Properties (\star) and ($\star\star$), while being far weaker than any kind of continuous dependence of solutions on initial conditions, are sufficient to develop basic stability characterizations. Continuous dependence of solutions on initial conditions is rare in hybrid systems, as in its more classical meaning it entails uniqueness of solutions from each initial point. If understood in a set-valued sense, in order for inner-semicontinuity (or lower-semicontinuity) to be present, it still requires many further assumptions on the data.

Property (\star) is essentially all that is needed to establish invariance principles, which will be presented in Subsect. “Invariance Principles”. Property ($\star\star$) is useful in describing, for example, uniformity of convergence and

of overshoots in an asymptotically stable hybrid system. For the analysis of robustness properties of well-posed hybrid systems, strengthened versions of (\star) and $(\star\star)$ are available. They take into account the effect of small perturbations; for example a stronger version of (\star) makes the same conclusion not about a sequence of solutions to \mathcal{H} , but about a sequence of solutions to \mathcal{H} generated with vanishing perturbations. (More information can be found in [23].) It is the stronger versions of the two properties that make converse Lyapunov results possible; see Subsect. “Converse Lyapunov Theorems and Robustness”, where more precise meaning to perturbations is also given.

In the rest of this article, the analysis results will assume that (C1)–(C3) hold, i. e., the hybrid system under analysis is well-posed. The control algorithms will be constructed so that the corresponding closed-loop systems are well-posed. Such algorithms will be called well-posed controllers.

Modeling Hybrid Control Systems

Hybrid Controllers for Classical Systems

Given a nonlinear control system of the form

$$\mathcal{P}: \begin{cases} \dot{x} = f(x, u), & x \in C_P \\ y = h(x), \end{cases} \quad (11)$$

where C_P is a subset of the state space where the system is allowed to evolve, a general output-feedback hybrid controller $\mathcal{K} = (\kappa, \phi, C_K, \psi, D_K)$ takes the form

$$\mathcal{K}: \begin{cases} u = \kappa(y, \eta) \\ \dot{\eta} = \phi(y, \eta), & (y, \eta) \in C_K \\ \eta^+ \in \psi(y, \eta), & (y, \eta) \in D_K, \end{cases}$$

where the output of the plant $y \in \mathbb{R}^p$ is the input to the controller, the input to the plant $u \in \mathbb{R}^m$ is the output of the controller, and $\eta \in \mathbb{R}^k$ is the controller state. When system (11) is controlled by \mathcal{K} , their interconnection results in a hybrid closed-loop system given by

$$\left. \begin{aligned} \dot{x} &= f(x, \kappa(h(x), \eta)) \\ \dot{\eta} &= \phi(h(x), \eta) \end{aligned} \right\} (x, \eta) \in C \quad (12)$$

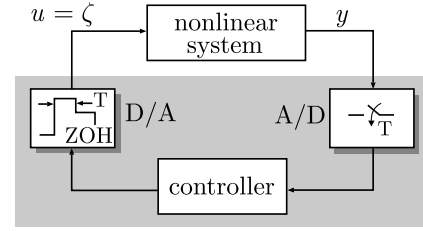
$$\left. \begin{aligned} x^+ &= x \\ \eta^+ &\in \psi(h(x), \eta) \end{aligned} \right\} (x, \eta) \in D,$$

where

$$C := \{(x, \eta) \mid x \in C_P, (h(x), \eta) \in C_K\}$$

$$D := \{(x, \eta) \mid (h(x), \eta) \in D_K\}.$$

We now cast some specific situations into this framework.



Hybrid Control Systems, Figure 7

Sample-and-hold control of a nonlinear system

Sample-and-hold Control

Perhaps the simplest example of a hybrid system that arises in control system design is when a continuous-time plant is controlled via a digital computer connected to the plant through a sample-and-hold device. This situation is ubiquitous in feedback control applications. A hybrid system emerges by considering the nonlinear control system (11), with $C_P = \mathbb{R}^n$, where the measurements y are sampled every $T > 0$ seconds, producing a sampled signal y_s that is processed through a discrete-time algorithm

$$\begin{bmatrix} z^+ \\ u_s \end{bmatrix} = \chi(z, y_s)$$

to generate a sequence of input values u_s each of which is held for T seconds to generate the input signal u .

When combined with the continuous-time dynamics, this algorithm will do the following:

- At the beginning of the sample period, it will update the value of u and the value of the controller's internal state z , based on the values of z and y (denoted y_s) at the beginning of the sampling period.
- During the rest of the sampling period, it will hold the values of u and z constant.

A complete model of this behavior is captured by defining the controller state to be

$$\eta := \begin{bmatrix} z \\ \zeta \\ \tau \end{bmatrix},$$

where ζ keeps track of the input value to hold during a sampling period and τ is a timer state that determines when the state variables z and ζ should be updated. The hybrid controller is specified in the form of the previous

section as

$$\left. \begin{array}{l} u = \zeta \\ \dot{z} = 0 \\ \dot{\zeta} = 0 \\ \dot{\tau} = 1 \end{array} \right\} (y, z, \zeta, \tau) \in C_K$$

$$\left. \begin{array}{l} \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ = \chi(z, y) \\ \tau^+ = 0 \end{array} \right\} (y, z, \zeta, \tau) \in D_K,$$

where

$$C_K := \{(y, z, \zeta, \tau) \mid \tau \in [0, T]\},$$

$$D_K := \{(y, z, \zeta, \tau) \mid \tau = T\}.$$

The overall closed-loop hybrid system is given by

$$\left. \begin{array}{l} \dot{x} = f(x, \zeta) \\ \dot{z} = 0 \\ \dot{\zeta} = 0 \\ \dot{\tau} = 1 \end{array} \right\} (x, z, \zeta, \tau) \in C$$

$$\left. \begin{array}{l} x^+ = x \\ \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ = \chi(z, h(x)) \\ \tau^+ = 0 \end{array} \right\} (x, z, \zeta, \tau) \in D,$$

where

$$C := \{(x, z, \zeta, \tau) \mid \tau \in [0, T]\},$$

$$D := \{(x, z, \zeta, \tau) \mid \tau = T\}.$$

Notice that if the function χ is discontinuous then this may fail to be a well-posed hybrid system. In such a case, it becomes well-posed by replacing the function χ by its set-valued regularization

$$\bar{\chi}(z, y) := \bigcap_{\delta > 0} \overline{\chi((z, y) + \delta \mathbb{B})}.$$

This corresponds to allowing, at points of discontinuity, values that can be obtained with arbitrarily small perturbations of z and y . Allowing these values is reasonable in light of inevitable control systems perturbations, like measurement noise and computer round-off error.

Networked Control Systems

Certain classes of networked control systems can be viewed as generalizations of systems with a sample-and-hold device. The networked control systems generalization allows for multiple sample-and-hold devices operating simultaneously and asynchronously, and with a variable sampling period. Compared to the sample-and-hold

closed-loop model in Subsect. “Sample-and-hold Control”, one can think of u as a large vector of inputs to a collection of i plants, collectively modeled by $\dot{x} = f(x, u)$. The update rule for u may only update a certain part of u at a given jump time. This update rule may depend not only on z and y but perhaps also on u and a logic variable, which we denote by ℓ , that may be cycling through the list of i indices corresponding to connections to different plants. Several common update protocols use algorithms that are discontinuous functions, so this will be modeled explicitly by allowing a set-valued update rule. Finally, due to time variability in the behavior of the network connecting the plants, the updates may occur at any time in an interval $[T_{\min}, T_{\max}]$ where $T_{\min} > 0$ represents the minimum amount of time between transmissions in the network and $T_{\max} > T_{\min}$ represents the maximum amount of time between transmissions. The overall control system for the network of plants is given by

$$\left. \begin{array}{l} u = \zeta \\ \dot{z} = 0 \\ \dot{\zeta} = 0 \\ \dot{\ell} = 0 \\ \dot{\tau} = 1 \end{array} \right\} (y, z, \zeta, \ell, \tau) \in C_K$$

$$\left. \begin{array}{l} \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ \in \chi(z, y, \zeta, \ell) \\ \ell^+ = (\ell \bmod i) + 1 \\ \tau^+ = 0 \end{array} \right\} (y, z, \zeta, \ell, \tau) \in D_K,$$

where

$$C_K := \{(y, z, \zeta, \ell, \tau) \mid \tau \in [0, T_{\max}]\},$$

$$D_K := \{(y, z, \zeta, \ell, \tau) \mid \tau \in [T_{\min}, T_{\max}]\}.$$

The closed-loop networked control system has the form

$$\left. \begin{array}{l} \dot{x} = f(x, \zeta) \\ \dot{z} = 0 \\ \dot{\zeta} = 0 \\ \dot{\ell} = 0 \\ \dot{\tau} = 1 \end{array} \right\} (x, z, \zeta, \ell, \tau) \in C$$

$$\left. \begin{array}{l} x^+ = x \\ \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ \in \chi(z, h(x), \zeta, \ell) \\ \ell^+ = (\ell \bmod i) + 1 \\ \tau^+ = 0 \end{array} \right\} (x, z, \zeta, \ell, \tau) \in D,$$

where

$$C := \{(x, z, \zeta, \ell, \tau) \mid \tau \in [0, T_{\max}]\} ,$$

$$D := \{(x, z, \zeta, \ell, \tau) \mid \tau \in [T_{\min}, T_{\max}]\} .$$

Reset Control Systems

The first documented reset controller was created by Clegg [19]. Consisting of an operational amplifier, resistors, and diodes, Clegg's controller produced an output that was the integral of its input subject to the constraint that the sign of the output and input agreed. This was achieved by forcing the state of the circuit to jump to zero, a good approximation of the behavior induced by the diodes, when the circuit's input changed sign with respect to its output. Consider such a circuit in a negative feedback loop with a linear control system

$$\dot{x} = Ax + Bu , \quad x \in \mathbb{R}^n , u \in \mathbb{R}$$

$$y = Cx , \quad y \in \mathbb{R} .$$

Use η to denote the state of the integrator. Then, the hybrid model of the Clegg controller is given by

$$u = \eta$$

$$\dot{\eta} = -y , \quad (y, \eta) \in C_K$$

$$\eta^+ = 0 , \quad (y, \eta) \in D_K ,$$

where

$$C_K := \{(y, \eta) \mid \eta y \leq 0\} ,$$

$$D_K := \{(y, \eta) \mid \eta y \geq 0\} .$$

One problem with this model is that it exhibits discrete solutions, as defined in Subsect. “Hybrid Behavior and Model”. Indeed, notice that the jump map takes points with $\eta = 0$, which are in the jump set D_K , back to points with $\eta = 0$. Thus, there are complete solutions that start with $\eta = 0$ and never flow, which corresponds to the definition of a discrete solution.

There are several ways to address this issue. When a reset controller like the Clegg integrator is implemented through software, a temporal regularization, as discussed next in Subsect. “Zeno Solutions and Temporal Regularization”, can be used to force a small amount of flow time between jumps. Alternatively, and also for the case of an analog implementation, one may consider a more detailed model of the reset mechanism, as the model proposed above is not very accurate for the case where η and y are small. This modeling issue is analogous to hybrid model-

ing issues for a ball bouncing on a floor where the simplest model is not very accurate for small velocities.

Zeno Solutions and Temporal Regularization

Like for reset control systems, the closed-loop system (12) may exhibit Zeno solutions, i.e., solutions with an infinite number of jumps in a finite amount of time. These are relatively easy to detect in systems with bounded solutions, as they exist if and only if there exist discrete solutions, i.e., solutions with an infinite number of jumps and no flowing time. Discrete solutions in a hybrid control system are problematic, especially from an implementation point of view, but they can be removed by means of temporal regularization. The temporal regularization of a hybrid controller $\mathcal{K} = (\kappa, \phi, C_K, \psi, D_K)$ is generated by introducing a timer variable τ that resets to zero at jumps and that must pass a threshold defined by a parameter $\delta \in (0, 1)$ before another jump is allowed. The regularization produces a well-posed hybrid controller $\mathcal{K}_\delta = (\tilde{\kappa}, \tilde{\phi}, C_{K,\delta}, \tilde{\psi}, D_{K,\delta})$ with state $\tilde{\eta} := (\eta, \tau) \in \mathbb{R}^{k+1}$, where

$$\tilde{\kappa}(\tilde{\eta}) := \kappa(\eta)$$

$$\tilde{\phi}(\tilde{\eta}) := \phi(\eta) \times \{1 - \tau\} ,$$

$$\tilde{\psi}(\tilde{\eta}) := \psi(\eta) \times \{0\} ,$$

$$C_{K,\delta} := (C_K \times \mathbb{R}_{\geq 0}) \cup (\mathbb{R}^k \times [0, \delta]) ,$$

$$D_{K,\delta} := D_K \times [\delta, 1] .$$

This regularization is related to one type of temporal regularization introduced in [32]. The variable τ is initialized in the interval $[0, 1]$ and remains there for all time. When $\delta = 0$, the controller accepts flowing only if $(y, \eta) \in C_K$, since $\dot{\tau} = 1 - \tau$ and the flow condition for τ when $(y, \eta) \notin C_K$ is $\tau = 0$. Moreover, jumping is possible for $\delta = 0$ if and only if $(y, \eta) \in D_K$. Thus, the controller with $\delta = 0$ has the same effect on the closed loop as the original controller \mathcal{K} . When $\delta > 0$, the controller forces at least δ seconds between jumps since $\dot{\tau} \leq 1$ for all $\tau \in [0, \delta]$. In particular, Zeno solutions, if there were any, are eliminated. Based on the remarks at the end of Subsect. “Conditions for Well-Posedness”, we expect the effect of the controller for small $\delta > 0$ to be close to the effect of the controller with $\delta = 0$. In particular, we expect that this temporal regularization for small $\delta > 0$ will not destroy the stability properties of the closed-loop hybrid system, at least in a practical sense. This aspect is discussed in more detail in Subsect. “Zeno Solutions, Temporal Regularization, and Robustness”.

Hybrid Controllers for Hybrid Systems

Another interesting scenario in hybrid control is when a hybrid controller

$$\mathcal{K} \begin{cases} u = \kappa(y, \eta) \\ \dot{\eta} = \phi(y, \eta), & (y, \eta) \in C_K \\ \eta^+ \in \psi(y, \eta), & (y, \eta) \in D_K \end{cases}$$

is used to control a plant that is also hybrid, perhaps modeled as

$$\begin{aligned} \dot{x} &= f(x, u), & x &\in C_P \\ x^+ &= g(x), & x &\in D_P \\ y &= h(x). \end{aligned}$$

This is the situation for the juggling example presented in Sect. “Introduction”. In this scenario, the hybrid closed-loop system is modeled as

$$\begin{cases} \dot{x} = f(x, \kappa(h(x), \eta)) \\ \dot{\eta} = \phi(h(x), \eta) \end{cases} \quad (x, \eta) \in C$$

$$\begin{bmatrix} x \\ \eta \end{bmatrix}^+ \in G(x, \eta), \quad (x, \eta) \in D$$

where

$$\begin{aligned} C &:= \{(x, \eta) \mid x \in C_P, (h(x), \eta) \in C_K\} \\ D &:= \{(x, \eta) \mid x \in D_P \text{ or } (h(x), \eta) \in D_K\} \end{aligned}$$

and

$$\begin{aligned} G_P(x, \eta) &:= \begin{bmatrix} g(x) \\ \eta \end{bmatrix} \\ G_K(x, \eta) &:= \begin{bmatrix} \{x\} \\ \psi(h(x), \eta) \end{bmatrix} \\ G(x, \eta) &:= \begin{cases} G_P(x, \eta), & x \in D_P, (h(x), \eta) \notin D_K \\ G_K(x, \eta), & x \notin D_P, (h(x), \eta) \in D_K \\ G_P(x, \eta) \cup G_K(x, \eta), & x \in D_P, (h(x), \eta) \in D_K. \end{cases} \end{aligned}$$

As long as the data of the controller and plant are well-posed, the closed-loop system is a well-posed hybrid system. Also, it can be verified that the only way this model can exhibit discrete solutions is if either the plant exhibits discrete solutions or the controller, with constant y , exhibits discrete solutions. Indeed, if the plant does not exhibit discrete solutions then a discrete solution for the closed-loop system would eventually have to have x , and thus y , constant. Then, if there are no discrete solutions

to the controller with constant y , there can be no discrete solutions to the closed-loop system.

Stability Theory

Lyapunov stability theory for dynamical systems typically states that asymptotically stable behaviors can be characterized by the existence of energy-like functions, which are called Lyapunov functions. This theory has served as a powerful tool for stability analysis of nonlinear dynamical systems and has enabled systematic design of robust control systems. In this section, we review some recent advances on Lyapunov-based stability analysis tools for hybrid dynamical systems.

Global (Pre-)Asymptotic Stability

In a classical setting, say of differential equations with Lipschitz continuous right-hand sides, existence of solutions and completeness of maximal ones can be taken for granted. These properties, together with a Lyapunov inequality ensuring that the Lyapunov function decreases along each solution, lead to a classical concept of asymptotic stability. On its own, the Lyapunov inequality does not say anything about the existence of solutions. It is hence natural to talk about a concept of asymptotic stability that is related only to the Lyapunov inequality. This appears particularly natural for the case of hybrid systems, where existence and completeness of solutions can be problematic.

The compact set $\mathcal{A} \subset \mathbb{R}^n$ is *stable* for \mathcal{H} if for each $\varepsilon > 0$ there exists $\delta > 0$ such that any solution x to \mathcal{H} with $|x(0, 0)|_{\mathcal{A}} \leq \delta$ satisfies $|x(t, j)|_{\mathcal{A}} \leq \varepsilon$ for all $(t, j) \in \text{dom } x$; it is *globally pre-attractive* for \mathcal{H} if any solution x to \mathcal{H} is bounded and if it is complete then $x(t, j) \rightarrow \mathcal{A}$ as $t + j \rightarrow \infty$; it is *globally pre-asymptotically stable* if it is both stable and globally pre-attractive. When every maximal solution to \mathcal{H} is complete, the prefix “pre” can be dropped and the “classical” notions of stability and asymptotic stability are recovered.

Globally Pre-Asymptotically Stable Ω -Limit Sets Suppose all solutions to the hybrid system are bounded, there exists a compact set S such that all solutions eventually reach and remain in S , and there exists a neighborhood of S from which this convergence to S is uniform. (We are not assuming that the set S is forward invariant and thus it may not be stable.) Moreover, assume there is at least one complete solution starting in S . In this case, the hybrid system admits a nonempty compact set $\mathcal{A} \subset S$ that is globally pre-asymptotically stable. Indeed, one such set is

the so-called Ω -limit set of S , defined as

$$\Omega_{\mathcal{H}}(S) := \left\{ y \in \mathbb{R}^n \left| \begin{array}{l} y = \lim_{i \rightarrow \infty} x_i(t_i, j_i), \\ t_i + j_i \rightarrow \infty, (t_i, j_i) \in \text{dom } x_i \\ x_i \text{ is a solution to } \mathcal{H} \\ \text{with } x_i(0, 0) \in S \end{array} \right. \right\}.$$

In fact, this Ω -limit set is the smallest compact set in S that is globally pre-asymptotically stable.

To illustrate this concept, consider the temperature control system in Sect. “Introduction” with additional heater dynamics given by

$$\dot{h} = -3h + (2h_{\Delta} + h)u,$$

where h is the heater temperature and h_{Δ} is a constant that determines how hot the heater can get due to being on. That is, when the heater is “on” ($u = 1$), its temperature rises asymptotically towards h_{Δ} . There is a maximum temperature $h_{\bar{\sigma}} < h_{\Delta}$ for which the heater can operate safely, and another temperature $h_{\underline{\sigma}} < h_{\bar{\sigma}}$ corresponding to a temperature far enough below $h_{\bar{\sigma}}$ that it is considered safe to turn the heater back on. For the desired range of temperatures for T given by $T_{\min} = 70^{\circ}\text{F}$, $T_{\max} = 80^{\circ}\text{F}$ and $T_{\Delta} = 30^{\circ}\text{F}$, let the overheating constant be $h_{\Delta} = 200^{\circ}\text{F}$, the maximum safe temperature be $h_{\bar{\sigma}} = 150^{\circ}\text{F}$, and the lower temperature $h_{\underline{\sigma}} = 50^{\circ}\text{F}$. Then, to keep the temperature T in the desired range and prevent overheating, the following algorithm is used:

- When the heater is “on” ($u = 1$) and either $T \geq 80$ or $h \geq 150$, then turn the heater off ($u^{+} = 0$).
- When the heater is “off” ($u = 0$) and $T \leq 70$ and $h \leq 50$, then turn the heater on ($u^{+} = 1$).

These rules define the jump map for u , which is given by $u^{+} = 1 - u$, and the jump set of the hybrid control system. The resulting hybrid closed-loop system, denoted by \mathcal{H}_T , is given by

$$\left. \begin{array}{l} \dot{T} = -T + T_0 + T_{\Delta}u \\ \dot{h} = -3h + (2h_{\Delta} + h)u \\ \dot{u} = 0 \end{array} \right\} \left\{ \begin{array}{l} u = 1, \\ (T \leq 80 \text{ and } h \leq 150) \\ \text{or} \\ u = 0, \\ (T \geq 70 \text{ or } h \geq 50) \end{array} \right.$$

$$\left. \begin{array}{l} T^{+} = T \\ h^{+} = h \\ u^{+} = 1 - u \end{array} \right\} \left\{ \begin{array}{l} u = 1, (T \geq 80 \text{ or } h \geq 150) \\ \text{or} \\ u = 0, (T \leq 70 \text{ and } h \leq 50) \end{array} \right.$$

For this system, the set

$$S := [70, 80] \times [0, 150] \times \{0, 1\} \quad (13)$$

is not forward invariant. Indeed, consider the initial condition $(T, h, u) = (70, 150, 0)$ which is not in the jump set, so there will be some time where the heater remains off, cooling the room to a value below 70°F . Nevertheless, all trajectories converge to the set S and a neighborhood of initial conditions around S produce solutions that reach the set S in a uniform amount of time. Thus, the set $\Omega_{\mathcal{H}_T}(S) \subset S$ is a compact globally pre-asymptotically stable set for the system \mathcal{H}_T .

Converse Lyapunov Theorems and Robustness

For purely continuous-time and discrete-time systems satisfying some regularity conditions, global asymptotic stability of a compact set implies the existence of a smooth Lyapunov function. Such results, known as converse Lyapunov theorems, establish a necessary condition for global asymptotic stability. For hybrid systems $\mathcal{H} = (f, C, G, D)$, the conditions for well-posedness also lead to a converse Lyapunov theorem.

If a compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable for the hybrid system $\mathcal{H} = (f, C, G, D)$, then there exists a smooth Lyapunov function; that is, there exists a smooth function $V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ that is positive definite with respect to \mathcal{A} , radially unbounded, and satisfies

$$\langle \nabla V(x), f(x) \rangle \leq -V(x) \quad \forall x \in C,$$

$$\max_{g \in G(x)} V(g) \leq \frac{V(x)}{2} \quad \forall x \in D.$$

Converse Lyapunov theorems are not of mere theoretical interest as they can be used to characterize robustness of asymptotic stability. Suppose that $\mathcal{H} = (f, C, G, D)$ is well-posed and that $V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ is a smooth Lyapunov function for some compact set \mathcal{A} . The smoothness of V and the regularity of the data of \mathcal{H} imply that V decreases along solutions when the data is perturbed. More precisely:

Global pre-asymptotic stability of the compact set $\mathcal{A} \subset \mathbb{R}^n$ for \mathcal{H} is equivalent to semiglobal practical pre-asymptotic stability of \mathcal{A} in the size of perturbations to \mathcal{H} , i.e., to the following: there exists a continuous, nondecreasing in the first argument, nonincreasing in second argument function $\beta: \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ with the property that $\lim_{s \searrow 0} \beta(s, t) = \lim_{t \rightarrow \infty} \beta(s, t) = 0$ and, for each $\varepsilon > 0$ and each compact set $K \subset \mathbb{R}^n$, there exists $\rho^ > 0$, such that for each perturbation level*

$\rho \in (0, \rho^*]$ each solution $x, x(0, 0) \in K$, to the ρ -perturbed hybrid system

$$\mathcal{H}_\rho: \quad x \in \mathbb{R}^n \quad \begin{cases} \dot{x} \in F_\rho(x), & x \in C_\rho \\ x^+ \in G_\rho(x), & x \in D_\rho, \end{cases}$$

where, for each $x \in \mathbb{R}^n$,

$$F_\rho(x) := \overline{\text{co}}f((x + \rho\mathbb{B}) \cap C) + \rho\mathbb{B},$$

$$G_\rho(x) := \{v \in \mathbb{R}^n \mid v \in z + \rho\mathbb{B}, z \in G((x + \rho\mathbb{B}) \cap D)\},$$

and

$$C_\rho := \{z \in \mathbb{R}^n \mid (z + \rho\mathbb{B}) \cap C \neq \emptyset\},$$

$$D_\rho := \{z \in \mathbb{R}^n \mid (z + \rho\mathbb{B}) \cap D \neq \emptyset\},$$

satisfies

$$|x(t, j)|_{\mathcal{A}} \leq \max\{\beta(|x(0, 0)|_{\mathcal{A}}, t + j), \varepsilon\} \quad \forall (t, j) \in \text{dom } x.$$

The above result can be readily used to derive robustness of (pre-)asymptotic stability to various types of perturbations, such as slowly-varying and weakly-jumping parameters, “average dwell-time” perturbations (see [16] for details), and temporal regularizations, as introduced in Subsect. “[Zeno Solutions and Temporal Regularization](#)”. We clarify the latter robustness now.

Zeno Solutions, Temporal Regularization, and Robustness Some of the control systems we will design later will have discrete solutions that evolve in the set we are trying to asymptotically stabilize. So, these solutions do not affect asymptotic stability adversely, but they are somewhat problematic from an implementation point of view. We indicated in Subsect. “[Zeno Solutions and Temporal Regularization](#)” how these solutions arising in hybrid control systems can be removed via temporal regularization. Here we indicate how doing so does not destroy the asymptotic stability achieved, at least in a semi-global practical sense.

The assumption is that stabilization via hybrid control is achieved as a preliminary step. In particular, assume that there is a well-posed closed-loop hybrid system $\mathcal{H} := (f, G, C, D)$ with state $\xi \in \mathbb{R}^n$, and suppose that the compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable. Following the prescription for a temporal regularization of a hybrid controller in Subsect. “[Zeno Solutions and Temporal Regularization](#)”, we consider the hybrid system $\mathcal{H}_\delta := (\tilde{f}, C_\delta, \tilde{G}, D_\delta)$ with the state $x := (\xi, \tau) \in \mathbb{R}^{n+1}$, where $\delta \in (0, 1)$ and

$$\tilde{f}(x) := f(\xi) \times \{1 - \tau\},$$

$$\tilde{G}(x) := G(\xi) \times \{0\},$$

$$C_\delta := (C \times \mathbb{R}_{\geq 0}) \cup (\mathbb{R}^n \times [0, \delta]),$$

$$D_\delta := D \times [\delta, 1].$$

As observed before, the system $\mathcal{H}_0 = (\tilde{f}, C_0, \tilde{G}, D_0)$ has the compact set $\tilde{\mathcal{A}} := \mathcal{A} \times [0, 1]$ globally pre-asymptotically stable. When $\delta > 0$, in each hybrid time domain of each solution, each time interval is at least δ seconds long, since $\tau \leq 1$ for all $\tau \in [0, \delta]$. In particular, Zeno solutions, if there were any, have been eliminated. Regarding pre-asymptotic stability, note that

$$C_\delta \subset \{z \in \mathbb{R}^{n+1} \mid (z + \delta\mathbb{B}) \cap C_0 \neq \emptyset\}$$

(with $\mathbb{B} \subset \mathbb{R}^{n+1}$), while $D_\delta \subset D_0$. Hence, following the discussion above, one can conclude that for \mathcal{H}_δ , the set $\tilde{\mathcal{A}}$ is semi-globally practically asymptotically stable in the size of the temporal regularization parameter δ . Broadly speaking, temporal regularization does not destroy (practical) pre-asymptotic stability of \mathcal{A} .

Lyapunov Stability Theorem

A Lyapunov function is not only necessary for asymptotic stability but also sufficient. It is a convenient tool for establishing asymptotic stability because it eliminates the need to solve explicitly for solutions to the system. In its sufficiency form, the requirements of a Lyapunov function can be relaxed somewhat compared to the conditions of the previous subsection.

For a hybrid system $\mathcal{H} = (f, C, G, D)$ the compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable if there exists a continuously differentiable function $V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq}$ that is positive definite with respect to \mathcal{A} , radially unbounded, and, with the definitions,

$$u_c(x) := \begin{cases} \langle \nabla V(x), f(x) \rangle & x \in C \\ -\infty & \text{otherwise} \end{cases} \quad (14)$$

$$u_d(x) := \begin{cases} \max_{g \in G(x)} V(g) - V(x) & x \in D \\ -\infty & \text{otherwise} \end{cases}, \quad (15)$$

satisfies

$$u_d(x) \leq 0 \quad \forall x \in \mathbb{R}^n \quad (16)$$

$$u_c(x) < 0, \quad u_d(x) < 0 \quad \forall x \in \mathbb{R}^n \setminus \mathcal{A}. \quad (17)$$

In light of the converse theorem of the previous subsection, this sufficient condition for global asymptotic stability is reasonable. Nevertheless, finding a Lyapunov function is often difficult to do. Thus, there is motivation for stability analysis tools that relax the Lyapunov conditions. There are several directions in which to go. One is in the direction of invariance principles, which are presented next.

Invariance Principles

An important tool to study the convergence of solutions to dynamical systems is LaSalle's invariance principle. LaSalle's invariance principle [35,36] states that bounded and complete solutions converge to the largest invariant subset of the set where the derivative or the difference (depending whether the system is continuous-time or discrete-time, respectively) of a suitable energy function is zero. In situations where the condition (17) holds with nonstrict inequalities, the invariance principle provides a tool to extract information about convergence of solutions.

By relying on the sequential compactness property of solutions in Subsect. "Conditions for Well-Posedness", several versions of LaSalle-like invariance principles can be stated for hybrid systems. Like for continuous-time and discrete-time systems, to make statements about the convergence of a solution one typically assumes that the solution is bounded, that its hybrid time domain is unbounded (i. e., the solution is complete), and that a Lyapunov function does not increase along it. To obtain information about the set to which the solution converges, an invariant set is to be computed. For hybrid systems, since solutions may not be unique, the standard concept of invariance needs to be adjusted appropriately.

Following [35], but in the setting of hybrid systems, we will insist that a (weakly) invariant set be both *weakly forward invariant* and *weakly backward invariant*. The word "weakly" indicates that only one solution, rather than all, needs to meet some invariance conditions. By requiring both forward and backward invariance we refine the sets to which solutions converge. For a given set \mathcal{M} and a hybrid system \mathcal{H} , these notions were defined, in [54], as follows:

- *Forward invariance*: if for each point $x^0 \in \mathcal{M}$ there exists at least one complete solution x to \mathcal{H} that starts at x^0 and stays in the set \mathcal{M} for all $(t, j) \in \text{dom } x$.
- *Backward invariance*: if for each point $q \in \mathcal{M}$ and every positive number N there exists a point x^0 from which there exists at least one solution x to \mathcal{H} and $(t^*, j^*) \in \text{dom } x$ such that $x(t^*, j^*) = q$ and $x(t, j) \in \mathcal{M}$ for all $(t, j) \in \text{dom } x, (t, j) \leq (t^*, j^*)$.

Then, the following invariance principle can be stated:

Let $V: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and suppose that $\mathcal{U} \subset \mathbb{R}^n$ is nonempty. Let x be a bounded and complete solution to a hybrid system $\mathcal{H} := (f, C, G, D)$. If x satisfies $x(t, j) \in \mathcal{U}$ for each $(t, j) \in \text{dom } x$ and

$$u_c(z) \leq 0, \quad u_d(z) \leq 0 \quad \text{for all } z \in \mathcal{U},$$

then, for some constant $r \in V(\mathcal{U})$, the solution x approaches the largest weakly invariant set contained in

$$[u_c^{-1}(0) \cup (u_d^{-1}(0) \cap G(u_d^{-1}(0)))] \cap V^{-1}(r) \cap \mathcal{U}. \quad (18)$$

Note that the statement and the conclusion of this invariance principle resemble the ones by LaSalle for differential/difference equations. In particular, the definition of the set (18) involves both the zero-level set of u_c and u_d as the continuous and discrete-time counterparts of the principle. For more details and other invariance principles for hybrid systems, see [54].

The invariance principle above leads to the following corollary on global pre-asymptotic stability:

For a hybrid system $\mathcal{H} = (f, C, G, D)$ the compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable if there exists a continuously differentiable function $V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq}$ that is positive definite with respect to \mathcal{A} , radially unbounded, such that, with the definitions (14)–(15),

$$u_c(x) \leq 0, \quad u_d(x) \leq 0 \quad \forall x \in \mathbb{R}^n,$$

and, for every $r > 0$, the largest weakly invariant subset in (18) is empty.

This corollary will be used to establish global asymptotic stability in the control application in Subsect. "Source Localization".

Design Tools

Supervisors of Hybrid Controllers

In this section, we discuss how to construct a single, globally asymptotically stabilizing, well-posed hybrid controller from several individual well-posed hybrid controllers that behave well on particular regions of the state-space but that are not defined globally.

Suppose we are trying to control a nonlinear system

$$\dot{x} = f(x, u), \quad x \in C_0 \quad (19)$$

and suppose that we have constructed a finite family of well-posed hybrid controllers \mathcal{K}_q , that work well individually, on a particular region of the state space. We will make this more precise below.

For simplicity, the controllers will share the same state. This can be accomplished by embedding the states of the individual controllers into a common state space. Each

controller \mathcal{K}_q , with $q \in \mathcal{Q}$ and \mathcal{Q} being a finite index set not containing 0, is given by

$$\mathcal{K}_q \begin{cases} u = \kappa_q(x, \eta) \\ \dot{\eta} = \phi_q(x, \eta), & (x, \eta) \in C_q \\ \eta^+ \in \psi_q(x, \eta), & (x, \eta) \in D_q. \end{cases} \quad (20)$$

The sets C_q and D_q are such that $(x, \eta) \in C_q \cup D_q$ implies $x \in C_0$.

The union of the regions over which these controllers operate is the region over which we want to obtain robust, asymptotic stability. We define this set as $\Theta := \bigcup_{q \in \mathcal{Q}} (C_q \cup D_q)$. To achieve our goal, we will construct a hybrid supervisor that makes decisions about which of the hybrid controllers should be used based on the state's location relative to a collection of closed sets $\Psi_q \subset C_q \cup D_q$ that cover Θ . The hybrid supervisor will have its own state $q \in \mathcal{Q}$. The composite controller, denoted \mathcal{K} with flow set C and jump set D , should be such that 1) $C \cup D = \Theta \times \mathcal{Q}$, 2) all maximal solutions of the interconnection of \mathcal{K} with (19), denoted \mathcal{H} , starting in $\Theta \times \mathcal{Q}$ are complete, 3) and the compact set $\mathcal{A} \times \mathcal{Q}$, a subset of $\Theta \times \mathcal{Q}$, is globally asymptotically stable for the system \mathcal{H} .

We now clarify what we mean by the family of hybrid controllers working well individually. Let \mathcal{H}_q denote the closed-loop interconnection of the system (19) with the hybrid controller (20). For each $q \in \mathcal{Q}$, the solutions to the system \mathcal{H}_q satisfy:

- I. The set \mathcal{A} is globally pre-asymptotically stable.
- II. Each maximal solution is either complete or ends in

$$\left(\bigcup_{i \in \mathcal{Q}, i > q} \Psi_i \right) \cup \overline{\Theta \setminus (C_q \cup D_q)}.$$

- III. No maximal solution starting in Ψ_q reaches

$$\overline{\Theta \setminus \left[C_q \cup D_q \cup \left(\bigcup_{i \in \mathcal{Q}, i > q} \Psi_i \right) \right]} \setminus \mathcal{A}.$$

Item III holds for free for the minimum index q_{\min} since $\Psi_{q_{\min}} \subset C_{q_{\min}} \cup D_{q_{\min}}$ and $\bigcup_{i \in \mathcal{Q}} \Psi_i = \Theta$. The combination of the three items for the maximum index q_{\max} implies that the solutions to $\mathcal{H}_{q_{\max}}$ that start in $\Psi_{q_{\max}}$ converge to \mathcal{A} .

Intuitively, the hybrid supervisor will attempt to reach its goal by guaranteeing completeness of solutions and making the evolution of q eventually monotonic while (x, η) does not belong to the set \mathcal{A} . In this way, the (x, η)

component of the solutions eventually converges to \mathcal{A} since q is eventually constant and because of the first assumption above. The hybrid supervisor can be content with sticking with controller q as long as $(x, \eta) \in C_q \cup D_q$, it can increment q if $(x, \eta) \in \bigcup_{i \in \mathcal{Q}, i > q} \Psi_i$, and it can do anything it wants if $(x, \eta) \in \mathcal{A}$. These are the only three situations that should come up when starting from Ψ_q . Otherwise, the hybrid controller would be forced to decrease the value of q , taking away any guarantee of convergence. This provides the motivation for item III above. Due to a disturbance or unfortunate initialization of q , it may be that the state reaches a point that would not otherwise be reached from Ψ_q . From such conditions, the important thing is that the solution is either complete (and thus converges to \mathcal{A}) or else reaches a point where either q can be incremented or where q is allowed to be decremented. This is the motivation for item II above.

The individual hybrid controllers are combined into a single, well-posed hybrid controller \mathcal{K} as follows: Define $\Phi_q := \bigcup_{i \in \mathcal{Q}, i > q} \Psi_i$ and then

$$\mathcal{K} \begin{cases} u = \kappa_q(x, \eta) \\ \dot{\eta} = \phi_q(x, \eta), & (x, \eta) \in \tilde{C}_q \\ \begin{bmatrix} \eta \\ q \end{bmatrix}^+ \in G_q(x, \eta), & (x, \eta) \in \tilde{D}_q, \end{cases} \quad (21)$$

where

$$\tilde{D}_q := D_q \cup \Phi_q \cup \overline{\Theta \setminus (C_q \cup D_q)},$$

\tilde{C}_q is closed and satisfies

$$\overline{C_q \setminus \Phi_q} \subset \tilde{C}_q \subset C_q,$$

and the set-valued mapping G_q is constructed via the following definitions:

$$\begin{aligned} D_{q,a} &:= \Phi_q \\ \overline{D_q \setminus \Phi_q} &\subset D_{q,b} \subset D_q \\ D_{q,c} &:= \overline{\Theta \setminus (C_q \cup D_q \cup \Phi_q)} \end{aligned}$$

and

$$\begin{aligned} G_{q,a}(x, \eta) &:= \left[\{i \in \mathcal{Q} \mid i > q, (x, \eta) \in \Psi_i\} \right] \\ G_{q,b}(x, \eta) &:= \left[\begin{matrix} \psi_q(x, \eta) \\ \{q\} \end{matrix} \right] \\ G_{q,c}(x, \eta) &:= \left[\{i \in \mathcal{Q} \mid (x, \eta) \in \Psi_i\} \right] \\ G_q(x, \eta) &:= \bigcup_{\{j \in \{a,b,c\}, (x, \eta) \in D_{q,j}\}} G_{q,j}(x, \eta). \end{aligned} \quad (22)$$

This hybrid controller is well-posed and induces complete solutions from $\Theta \times \mathcal{Q}$ and global asymptotic stability of the compact set $\mathcal{A} \times \mathcal{Q}$.

Uniting Local and Global Controllers As a simple illustration, consider a nonlinear control system $\dot{x} = f(x, u)$, $x \in \mathbb{R}^n$, and the task of globally asymptotically stabilizing the origin using state feedback while insisting on using a particular state feedback κ_2 in a neighborhood of the origin. In order to solve this problem, one can find a state feedback κ_1 that globally asymptotically stabilizes the origin and then combine it with κ_2 using a hybrid supervisor. Suppose that the feedback κ_2 is defined on a closed neighborhood of the origin, denoted C_2 , and that if the state x starts in the closed neighborhood $\Psi_2 \subset C_2$ of the origin then the closed-loop solutions when using κ_2 do not reach the boundary of C_2 . Then, using the notation of this section, we can take $\Psi_1 = C_1 = \mathbb{R}^n$ and $D_1 = D_2 = \emptyset$. With these definitions, the assumptions above are satisfied and a hybrid controller can be constructed to solve the posed problem. The controller need not use the additional variable η . Its data is defined as $G_q(x) := 3 - q$, $\tilde{D}_1 := \Psi_2 = \Phi_1$, $\tilde{D}_2 := \mathbb{R}^n \setminus C_2$, $\tilde{C}_1 := \overline{C_1 \setminus \Psi_2}$ and $\tilde{C}_2 := C_2$.

Additional examples of supervisors will appear in the applications section later.

Patchy Control Lyapunov Functions

A key feature of (smooth) control Lyapunov functions (CLFs) is that their decrease along solutions to a given control system can be guaranteed by an appropriate choice of the control value, for each state value. It is known that, under mild assumptions on the control system, the existence of a CLF yields the existence of a robust (non hybrid) stabilizing feedback. It is also known that many nonlinear control systems do not admit a CLF. This can be illustrated by considering the question of robust stabilization of a single point on a circle, which faces a similar obstacle as the question of robust stabilization of the set $\mathcal{A} = \{0, 1\}$ for the control system on \mathbb{R} given by $\dot{x} = f(x, u) := u$. Any differentiable function on \mathbb{R} that is positive definite with respect to \mathcal{A} must have a maximum in the interval $(0, 1)$. At such a maximum, say \tilde{x} , one has $\nabla V(\tilde{x}) = 0$ and no choice of u can lead to $\langle \nabla V(\tilde{x}), f(\tilde{x}, u) \rangle < 0$.

(Smooth) patchy control Lyapunov functions (PCLFs) are, broadly speaking, objects consisting of several local CLFs the domains of which cover \mathbb{R}^n and have certain weak invariance properties. PCLFs turn out to exist for far broader classes of nonlinear systems than CLFs, especially if an infinite number of patches (i. e., of local CLFs)

is allowed. They also lead to robust hybrid stabilizing feedbacks. This will be outlined below. A brief illustration of the concept, for the control system on \mathbb{R} mentioned above, would be to consider functions $V_1(x) = x^2$ on $(-\infty, 2/3)$ and $V_2(x) = (x - 1)^2$ on $(1/3, \infty)$. These functions are local CLFs for the points, respectively, 0 and 1; their domains cover \mathbb{R} ; and for each function, an appropriate choice of control will not only lead to the function's decrease, but will also ensure that solutions starting in the function's domain will remain there.

While the example just mentioned outlines a general idea of a PCLF, the definition is slightly more technical. For the purposes of this article, a *smooth patchy control Lyapunov function* for a nonlinear system

$$\dot{x} = f(x, u) \quad x \in \mathbb{R}^n, \quad u \in U \subset \mathbb{R}^m \quad (23)$$

with respect to the compact set \mathcal{A} consists of a finite set $Q \subset \mathbb{Z}$ and a collection of functions V_q and sets Ω_q, Ω'_q for each $q \in Q$, such that:

- (i) $\{\Omega_q\}_{q \in Q}$ and $\{\Omega'_q\}_{q \in Q}$ are families of nonempty open subsets of \mathbb{R}^n such that

$$\mathbb{R}^n = \bigcup_{q \in Q} \Omega_q = \bigcup_{q \in Q} \Omega'_q,$$

and for all $q \in Q$, the unit (outward) normal vector to Ω_q is continuous on $\partial\Omega_q \setminus \bigcup_{i>q} \Omega'_i$, and

$$\overline{\Omega'_q} \subset \Omega_q;$$

- (ii) for each q , V_q is a smooth function defined on a neighborhood of $\overline{\Omega_q \setminus \bigcup_{i>q} \Omega'_i}$;

and the following conditions are met: There exist a continuous, positive definite function $\alpha: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$, and positive definite, radially unbounded functions $\underline{\gamma}, \bar{\gamma}$ such that

- (iii) for all $q \in Q$, all $x \in \Omega_q \setminus \bigcup_{i>q} \Omega'_i$,

$$\underline{\gamma}(|x|_{\mathcal{A}}) \leq V_q(x) \leq \bar{\gamma}(|x|_{\mathcal{A}});$$

- (iv) for all $q \in Q$, all $x \in \Omega_q \setminus \bigcup_{i>q} \Omega'_i$, there exists $u_{q,x} \in U$ such that

$$\langle \nabla V_q(x), f(x, u_{q,x}) \rangle \leq -\alpha(|x|_{\mathcal{A}});$$

- (v) for all $q \in Q$, all $x \in \partial\Omega_q \setminus \bigcup_{i>q} \Omega'_i$, the $u_{q,x}$ of (iii) can be chosen such that

$$\langle n_q(x), f(x, u_{q,x}) \rangle \leq -\alpha(|x|_{\mathcal{A}}),$$

where $n_q(x)$ is the unit (outward) normal vector to Ω_q at x .

Suppose that, for each $x, v \in \mathbb{R}^n$ and $c \in \mathbb{R}$, the set $\{u \in U \mid \langle v, f(x, u) \rangle \leq c\}$ is convex, as always holds if $f(x, u)$ is affine in u and U is convex. For each $q \in Q$ let

$$C_q = \overline{\Omega_q \setminus \bigcup_{i>q} \Omega'_i}$$

and

$$\Psi_q = \overline{\Omega'_q \setminus \bigcup_{i \in Q, i>q} \Omega'_i}.$$

It can be shown, in part via arguments similar to those one would use when constructing a feedback from a CLF, that for each $q \in Q$ there exists a continuous mapping

$$k_q: C_q \rightarrow U$$

such that, for all $x \in C_q$,

$$\langle \nabla V_q(x), f(x, k_q(x)) \rangle \leq -\frac{\alpha(|x|_{\mathcal{A}})}{2};$$

all maximal solutions to

$$\dot{x} = f(x, k_q(x))$$

are either complete or end in

$$\left(\bigcup_{i>q, i \in Q} \Psi_i \right) \cup \overline{\mathbb{R}^n \setminus C_q};$$

and no maximal solution starting in Ψ_q reaches

$$\overline{\mathbb{R}^n \setminus \left(C_q \cup \bigcup_{i \in Q, i>q} \Psi_i \right)}.$$

The feedbacks k_q can now be combined in a hybrid feedback, by taking $D_q = \emptyset$ for each $q \in Q$, and following the construction of Subsect. “[Supervisors of Hybrid Controllers](#)”. Indeed, the properties of maximal solutions just mentioned ensure conditions I, II and III of that section; the choice of C_q and Ψ_q also ensures that $\Psi_q \subset C_q$ and the union of Ψ_q ’s covers \mathbb{R}^n .

Among other things, this construction illustrates that the idea of hybrid supervision of hybrid controllers applies also to combining standard, non hybrid controllers.

Applications

In this section, we make use of the following:

- For vectors in \mathbb{R}^2 , we will use the following multiplication rule, conjugate rule, and identity element:

$$z \otimes x := \begin{bmatrix} z_1 x_1 - z_2 x_2 \\ z_2 x_1 + z_1 x_2 \end{bmatrix}, x^c := \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}, \quad \mathbf{1} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The multiplication rule is commutative, associative, and distributive. Note that $x = \mathbf{1} \otimes x = x \otimes \mathbf{1}$ and note that $x^c \otimes x = x \otimes x^c = |x|^2 \mathbf{1}$. Also, $(z \otimes x)^c = x^c \otimes z^c$.

- For vectors in \mathbb{R}^4 , we will use the following multiplication rule, conjugate rule, and identity element (vectors are partitioned as $x = [x_1 \ x_2^T]^T$ where $x_2 \in \mathbb{R}^3$):

$$z \otimes x = \begin{bmatrix} z_1 x_1 - z_2^T x_2 \\ z_2 x_1 + z_1 x_2 + z_2 \times x_2 \end{bmatrix},$$

$$x^c = \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix},$$

$$\mathbf{1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The multiplication rule is associative and distributive but not necessarily commutative. Note that $x = \mathbf{1} \otimes x = x \otimes \mathbf{1}$ and $x^c \otimes x = x \otimes x^c = |x|^2 \mathbf{1}$. Also, $(z \otimes x)^c = x^c \otimes z^c$.

Overcoming Stabilization Obstructions

Global Stabilization and Tracking on the Unit Circle

In this section, we consider stabilization and tracking control of the constrained system

$$\dot{\xi} = \xi \otimes v(\omega), \quad v(\omega) := \begin{bmatrix} 0 \\ \omega \end{bmatrix} \quad \xi \in S^1, \quad (24)$$

where S^1 denotes the unit circle and $\omega \in \mathbb{R}$ is the control variable. Notice that S^1 is invariant regardless of the choice of ω since $\langle \xi, \xi \otimes v(\omega) \rangle = 0$ for all $\xi \in S^1$ and all $\omega \in \mathbb{R}$. This model describes the evolution of orientation angle of a rigid body in the plane as a function of the angular velocity ω , which is the control variable. We discuss robust, global asymptotic stabilization and tracking problems which cannot be solved with classical feedback control, even when discontinuous feedback laws are allowed, but can be solved with hybrid feedback control.

Stabilization First, we consider the problem of stabilizing the point $\xi = \mathbf{1}$. We note that the (classical) feedback control $\omega = -\xi_2$ would almost solve this problem. We would have $\dot{\xi}_1 = \xi_2^2 = 1 - \xi_1^2$ and the derivative of the energy function $V(\xi) := 1 - \xi_1$ would satisfy

$$\langle \nabla V(\xi), \xi \otimes v(\omega) \rangle = -(1 - \xi_1^2).$$

We note that the energy will remain constant if ξ starts at $\pm \mathbf{1}$. Thus, since the goal was (robust) global asymptotic stability, this feedback does not achieve the desired goal. One could also consider the discontinuous feedback $\omega = -\text{sgn}(\xi_2)$ where the function “sgn” is defined arbitrarily in the set $\{-1, 1\}$ when its argument is zero. This feedback is not robust to arbitrarily small measurement noise which can keep the trajectories of the system arbitrarily close to the point $\xi = -\mathbf{1}$ for all time. To visualize this, note that from points on the circle with $\xi_2 < 0$ and close to $\xi = -\mathbf{1}$, this control law steers the trajectories towards $\xi = \mathbf{1}$ counterclockwise, while from points on the circle with $\xi_2 > 0$ and close to $\xi = -\mathbf{1}$, it steers the trajectories towards $\mathbf{1}$ clockwise. Then, from points on the circle arbitrarily close to $\xi = -\mathbf{1}$, one can generate an arbitrarily small measurement noise signal e that changes sign appropriately so that $-\text{sgn}(\xi_2 + e)$ is always pushing trajectories towards $-\mathbf{1}$.

In order to achieve a robust, global asymptotic stability result, we consider a hybrid controller that uses the controller $\omega = -\xi_2$ when the state is not near $-\mathbf{1}$ and uses a controller that drives the system away from $-\mathbf{1}$ when it is near that point. One way to accomplish the second task is to build an almost global asymptotic stabilizer for a point different from $-\mathbf{1}$ such that the basin of attraction contains all points in a neighborhood of $-\mathbf{1}$. For example, consider the feedback controller $\omega = -\xi_1 =: \kappa_1(\xi)$ which would almost globally asymptotically stabilize the point $\zeta := (0, -1)$ with the only point not in the basin of attraction being the point ζ^c .

Strangely enough, each of the two controllers can be thought of as globally asymptotically stabilizing the point $\mathbf{1}$ if their domains are limited. In particular, let the domain of applicability for the controller $\omega = -\xi_1$ be

$$C_1 := S^1 \cap \{\xi \mid \xi_1 \leq -1/3\},$$

and let the domain of applicability for the controller $\omega = -\xi_2$ be

$$C_2 := S^1 \cap \{\xi \mid \xi_1 \geq -2/3\}.$$

Notice that $C_1 \cup C_2 = S^1 =: \emptyset$. Thus, we are in a situation where a hybrid supervisor, as discussed in Subsect. “Supervisors of Hybrid Controllers”, may be able to give us a hybrid, global asymptotic stabilizer. (There is no state η in the controllers we are working with here.) Let us take

$$\Psi_1 := C_1, \quad \Psi_2 := \overline{S^1 \setminus C_1}.$$

We have $\Psi_1 \cup \Psi_2 = S^1$. Next, we check the assumptions of Subsect. “Supervisors of Hybrid Controllers”. For each

$q \in \{1, 2\}$, the solutions of \mathcal{H}_q (the system we get by using $\omega = \kappa_q(\xi)$ and restricting the flow to C_q), are such that the point $\mathbf{1}$ is globally pre-asymptotically stable. For $q = 1$, this is because there are no complete solutions and $\mathbf{1}$ does not belong to C_1 . For $q = 2$, this is because C_2 is a subset of the basin of attraction for $\mathbf{1}$. We note that every maximal solution to \mathcal{H}_1 ends in Ψ_2 . Every maximal solution to \mathcal{H}_2 is complete and every maximal solution to \mathcal{H}_2 starting in Ψ_2 does not reach $S^1 \setminus C_2$. Thus, the assumptions for a hybrid supervisor are in place.

We follow the construction in Subsect. “Supervisors of Hybrid Controllers” to define the hybrid supervisor that combines the feedback laws κ_1 and κ_2 . We take

$$\omega := \kappa_q(\xi)$$

$$\tilde{C}_q := C_q$$

$$\tilde{D}_1 := \Psi_2 \cup (\overline{S^1 \setminus C_1}) = \overline{S^1 \setminus C_1}$$

$$\tilde{D}_2 := \overline{S^1 \setminus C_2}.$$

For this particular problem, jumps toggle the mode q in the set $\{1, 2\}$. Thus, the jump map G_q can be simplified to $G_q := 3 - q$.

Tracking Let $\zeta: \mathbb{R}_{\geq 0} \rightarrow S^1$ be continuously differentiable. Suppose we want to find a hybrid feedback controller so that the state of (24) tracks the signal ζ . This problem can be reduced to the stabilization problem of the previous section. Indeed, first note that $\zeta^c \otimes \zeta = \mathbf{1}$ and thus the following properties hold:

$$\begin{aligned} \dot{\zeta}^c \otimes \zeta &= -\dot{\zeta}^c \otimes \dot{\zeta} \\ \zeta^c \otimes \dot{\zeta} &= \begin{bmatrix} 0 \\ \zeta_1 \dot{\zeta}_2 - \zeta_2 \dot{\zeta}_1 \end{bmatrix}. \end{aligned} \quad (25)$$

Then, with the coordinate transformation $\xi = z \otimes \zeta$, we have:

- I. By multiplying the coordinate transformation on the right by ζ^c we get $z = \xi \otimes \zeta^c$ and $z \otimes z^c = \xi \otimes \xi^c = \mathbf{1}$, so that $z \in S^1$.
- II. $\xi = \zeta \iff z = \mathbf{1}$.
- III. The derivative of z satisfies

$$\begin{aligned} \dot{z} &= \dot{\xi} \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \\ &= \xi \otimes v(\omega) \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \otimes \zeta \otimes \zeta^c \\ &= \xi \otimes \left[v(\omega) + \dot{\zeta}^c \otimes \zeta \right] \otimes \zeta^c \\ &= z \otimes \zeta \otimes \left[v(\omega) - \zeta^c \otimes \dot{\zeta} \right] \otimes \zeta^c. \end{aligned}$$

Our desire is to pick ω so that we have

$$\dot{z} = z \otimes v(\Omega)$$

and then to choose Ω to globally asymptotically stabilize the point $z = \mathbf{1}$. Due to (25) and the properties of multiplication, the vectors $\zeta^c \otimes \dot{\zeta}$ and $\zeta^c \otimes v(\Omega) \otimes \dot{\zeta}$ are in the range of $v(\omega)$. So, we can pick $v(\omega) = \zeta^c \otimes \dot{\zeta} + \zeta^c \otimes v(\Omega) \otimes \dot{\zeta}$ to achieve the robust, global tracking goal. In fact, since the multiplication operation is commutative in \mathbb{R}^2 , this is equivalent to the feedback $v(\omega) = \zeta^c \otimes \dot{\zeta} + v(\Omega)$.

Global Stabilization and Tracking for Unit Quaternions

In this section, we consider stabilization and tracking control of the constrained system

$$\dot{\xi} = \xi \otimes v(\omega), \quad v(\omega) := \begin{bmatrix} 0 \\ \omega \end{bmatrix} \quad \xi \in S^3, \quad (26)$$

where S^3 denotes the hypersphere in \mathbb{R}^4 and $\omega \in \mathbb{R}^3$ is the control variable. Notice that S^3 is invariant regardless of the choice of ω since $\langle \xi, \xi \otimes v(\omega) \rangle = 0$ for all $\xi \in S^3$ and all $\omega \in \mathbb{R}^3$. This model describes the evolution of orientation angle of a rigid body in space as a function of angular velocities ω , which are the control variables. The state ξ corresponds to a unit quaternion that can be used to characterize orientation. We discuss robust, global asymptotic stabilization and tracking problems which cannot be solved with classical feedback control, even when discontinuous feedback laws are allowed, but can be solved with hybrid feedback control.

Stabilization First, we consider the problem of stabilizing the point $\xi = \mathbf{1}$. We note that the (classical) feedback control

$$\omega := \begin{bmatrix} 0 & I \end{bmatrix} v(-\xi_2) =: \kappa_2(\xi)$$

(i.e., $\omega = -\xi_2$ where ξ_2 refers to the last three components of the vector ξ) would almost solve this problem. We would have $\dot{\xi}_1 = \xi_2^T \xi_2 = 1 - \xi_1^2$ and the derivative of the energy function $V(\xi) := 1 - \xi_1^2$ would satisfy

$$\langle \nabla V(\xi), \xi \otimes v(\omega) \rangle = -(1 - \xi_1^2).$$

We note that the energy will remain constant if ξ starts at $\pm \mathbf{1}$. Thus, since the goal is (robust) global asymptotic stabilization, this feedback does not achieve the desired goal. One could also consider the discontinuous feedback $\omega = -\text{sgn}(\xi_2)$ where the function “sgn” is the component-wise sign and each component is defined arbitrarily in the set $\{-1, 1\}$ when its argument is zero. This feedback is not robust to arbitrarily small measurement noise which can keep the trajectories of the system arbitrarily close to the point $-\mathbf{1}$.

In order to achieve a robust, global asymptotic stabilization result, we consider a hybrid controller that uses the

controller above when the state is not near $-\mathbf{1}$ and uses a controller that drives the system away from $-\mathbf{1}$ when it is near that point. One way to accomplish the second task is to build an almost global asymptotic stabilizer for a point different from $-\mathbf{1}$ and so that the basin of attraction contains all points in a neighborhood of $-\mathbf{1}$. For example, consider stabilizing the point $\zeta := (0, -1, 0, 0)^T$ using the feedback controller (for more details see the next subsection)

$$z = \xi \otimes \zeta^c, \quad \omega = \begin{bmatrix} 0 & I \end{bmatrix} (\zeta^c \otimes v(-z_2) \otimes \zeta) =: \kappa_1(\xi)$$

(i.e. $\omega = (-\xi_1, \xi_4, -\xi_3)^T$ where now the subscripts refer to the individual components of ξ). This feedback would almost globally asymptotically stabilize the point ζ with the only point not in the basin of attraction being the point ζ^c .

The two feedback laws, $\omega = \kappa_2(\xi)$ and $\omega = \kappa_1(\xi)$, are combined into a single hybrid feedback law via the hybrid supervisor approach given in Subsect. “Supervisors of Hybrid Controllers”. In fact, the construction is just like the construction for the case of stabilization on a circle with the only difference being that S^1 is replaced everywhere by S^3 .

Tracking Let $\zeta: \mathbb{R}_{\geq 0} \rightarrow S^3$ be continuously differentiable. Suppose we want to find a hybrid feedback controller so that the state of (26) tracks the signal ζ . This problem can be reduced to the stabilization problem of the previous section. Indeed, first note that $\zeta^c \otimes \zeta = \mathbf{1}$ and thus the following properties hold:

$$\begin{aligned} \dot{\zeta}^c \otimes \zeta &= -\zeta^c \otimes \dot{\zeta} \\ \zeta^c \otimes \dot{\zeta} &= \begin{bmatrix} 0 \\ \zeta_1 \dot{\zeta}_2 - \dot{\zeta}_1 \zeta_2 - \zeta_2 \times \dot{\zeta}_2 \end{bmatrix}. \end{aligned} \quad (27)$$

Then, with the coordinate transformation $\xi = z \otimes \zeta$, we have:

- I. By multiplying the coordinate transformation on the right by ζ^c we get $z = \xi \otimes \zeta^c$ and $z^c \otimes z = \xi^c \otimes \xi = \mathbf{1}$ so that $z \in S^3$.
- II. $\xi = \zeta \iff z = \mathbf{1}$.
- III. The derivative of z satisfies

$$\begin{aligned} \dot{z} &= \dot{\xi} \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \\ &= \xi \otimes \widehat{\omega} \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \otimes \zeta \otimes \zeta^c \\ &= \xi \otimes \left[\widehat{\omega} + \dot{\zeta}^c \otimes \zeta \right] \otimes \zeta^c \\ &= z \otimes \zeta \otimes \left[\widehat{\omega} - \zeta^c \otimes \dot{\zeta} \right] \otimes \zeta^c. \end{aligned}$$

Our desire is to pick ω so that we have

$$\dot{z} = z \otimes v(\Omega)$$

and then to choose Ω to globally asymptotically stabilize the point $z = \mathbf{1}$. Due to (27) and the properties of multiplication, the vectors $\zeta^c \otimes \dot{\zeta}$ and $\zeta^c \otimes v(\Omega) \otimes \dot{\zeta}$ are in the range of $v(\omega)$. So, we can pick $v(\omega) = \zeta^c \otimes \dot{\zeta} + \zeta^c \otimes v(\Omega) \otimes \dot{\zeta}$ to achieve the robust, global tracking goal.

Stabilization of a Mobile Robot

Consider the global stabilization problem for a model of a unicycle or mobile robot, given as

$$\begin{aligned}\dot{x} &= \xi \vartheta \\ \dot{\xi} &= \xi \otimes v(\omega)\end{aligned}\tag{28}$$

where $x \in \mathbb{R}^2$ denotes planar position from a reference point (in meters), $\xi \in S^1$ denotes orientation, $\vartheta \in \mathcal{V} := [-3, 30]$ denotes velocity (in meters per second), and $\omega \in [-4, 4]$ denotes angular velocity (in radians per second). Both ϑ and ω are control inputs. Due to the specification of the set \mathcal{V} , the vehicle is able to move more rapidly in the forward direction than in the backward direction. We define \mathcal{A}_0 to be the point $(x, \xi) = (0, 1)$. The controllers below will all use a discrete state $p \in \mathcal{P} := \{-1, 1\}$. We take $\mathcal{A} := \mathcal{A}_0 \times \mathcal{P}$ and $\Theta := \mathbb{R}^2 \times S^1 \times \mathcal{P}$.

This system also can be modeled as

$$\begin{aligned}\dot{x} &= \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \vartheta \\ \dot{\theta} &= \omega\end{aligned}$$

where $\theta = 0$ corresponds to $\xi = \mathbf{1}$ and $\theta > 0$ is in the counterclockwise direction. The set \mathcal{A}_0 in these coordinates is given by the set $\{0\} \times \{\theta \mid \theta = 2k\pi, k \in \mathbb{Z}\}$. Even for the point $(x, \theta) = (0, 0)$, this control system fails Brockett's well-known condition for robust local asymptotic stabilization by classical (even discontinuous) time-invariant feedback [9,26,52]. Nevertheless, for the control system (28), the point $(0, 1)$ can be robustly, globally asymptotically stabilized by hybrid feedback. This is done by building three separate hybrid controllers and combining them with a supervisor. The three controllers are the following:

- The first hybrid controller, \mathcal{K}_1 , uses $\vartheta = \text{Proj}_{\mathcal{V}}(k_1 \xi^T x)$, where $k_1 < 0$ and $\text{Proj}_{\mathcal{V}}$ denotes the projection onto \mathcal{V} , while the feedback for ω is given by the hybrid controller in Subsect. "Global Stabilization and Tracking on the Unit Circle" for tracking on the

unit circle with reference signal for ξ given by $-x/|x|$. The two different values for q in that controller should be associated with the two values in \mathcal{P} . The particular association does not matter. Note that the action of the tracking controller causes the vehicle eventually to use positive velocity to move x toward zero. The controller's flow and jump sets are such that

$$C_1 \cup D_1 = \{x \in \mathbb{R}^2 \mid |x| \geq \varepsilon_{11}\} \times S^1 \times \mathcal{P}$$

where $\varepsilon_{11} > 0$, and C_1, D_1 are constructed from the hybrid controller in Subsect. "Global Stabilization and Tracking on the Unit Circle" for tracking on the unit circle.

- The second hybrid controller, \mathcal{K}_2 , uses $\vartheta = \text{Proj}_{\mathcal{V}}(k_2 \xi^T x)$, $k_2 \leq 0$, while the feedback for ω is given as in Subsect. "Global Stabilization and Tracking on the Unit Circle" for stabilization of the point $\mathbf{1}$ on the unit circle. Again, the q values of that controller should be associated with the values in \mathcal{P} and the particular association does not matter. The controller's flow and jump sets are such that

$$\begin{aligned}C_2 \cup D_2 &= (\{x \in \mathbb{R}^2 \mid |x| \leq \varepsilon_{21}\} \times S^1) \\ &\cap \{(x, \xi) \mid 1 - \xi_1 \geq \varepsilon_{22}|x|^2\} \times \mathcal{P},\end{aligned}$$

where $\varepsilon_{21} > \varepsilon_{11}$, $\varepsilon_{22} > 0$, and C_2, D_2 are constructed from the hybrid controller in Subsect. "Global Stabilization and Tracking on the Unit Circle" for stabilization of the point $\mathbf{1}$ on the unit circle.

- The third hybrid controller, \mathcal{K}_3 , uses $\vartheta = \text{Proj}_{\mathcal{V}}(k_3 \xi^T x)$, $k_3 < 0$, while the feedback for ω is hybrid as defined below. The controller's flow and jump sets are designed so that

$$\begin{aligned}C_3 \cup D_3 &= (\{x \mid |x| \leq \varepsilon_{31}\} \times S^1) \\ &\cap \{(x, \xi) \mid 1 - \xi_1 \leq \varepsilon_{32}|x|^2\} \times \mathcal{P} =: \Lambda_3,\end{aligned}$$

where $\varepsilon_{31} > \varepsilon_{21}$ and $\varepsilon_{32} > \varepsilon_{22}$. The control law for ω is given by $\omega = pk$, where $k > 0$ and the discrete state p has dynamics given by

$$\dot{p} = 0, \quad p^+ = -p.$$

The flow and jump sets are given by

$$\begin{aligned}C_3 &:= \Lambda_3 \cap \{\sigma(p)\xi_2 \leq 0\} \\ &\cup \{\sigma(p)\xi_2 \geq 0, 1 - \xi_1 \leq \varepsilon_{22}|x|^2\} \\ D_3 &:= \overline{\Lambda_3 \setminus C_3}.\end{aligned}$$

This design accomplishes the following: controller \mathcal{K}_1 makes ξ track $-x/|x|$ as long as $|x|$ is not too small, and

thus the vehicle is driven towards $x = 0$ eventually using only positive velocity; controller \mathcal{K}_2 drives ξ towards $\mathbf{1}$ to get the orientation of the vehicle correct; and controller \mathcal{K}_3 stabilizes ξ to $\mathbf{1}$ in a persistently exciting manner so that ϑ can be used to drive the vehicle to the origin.

This control strategy is coordinated through a supervisor by defining

$$\begin{aligned}\Psi_1 &:= C_1 \cup D_1 \\ \Psi_2 &:= (\overline{\Theta \setminus \Psi_1}) \cap (C_2 \cup D_2) \\ \Psi_3 &:= ((\{x \mid |x| \leq \varepsilon_{21}\} \times S^1) \\ &\quad \cap \{(x, \xi) \mid 1 - \xi_1 \leq \varepsilon_{22}|x|^2\}) \times \mathcal{P}.\end{aligned}$$

It can be verified that $\bigcup_{q \in \mathcal{Q}} \Psi_q = \Theta$ and that the conditions in Subject. “[Supervisors of Hybrid Controllers](#)” for a successful supervisor are satisfied.

Figure 8 depicts simulation results of the mobile robot with the hybrid controller proposed above for global asymptotic stabilization of $\mathcal{A} \times \mathcal{Q}$. From the initial condition $x(0, 0) = (10, 10)$ (in meters), $\xi(0, 0)$ corresponding to an angle of $\frac{\pi}{4}$ radians, the mobile robot backs up using controller \mathcal{K}_1 until its orientation ξ corresponds to about $\frac{3\pi}{4}$ radians, at which x is approximately $(10, 9.5)$. The green \star denotes a jump of the hybrid controller \mathcal{K}_1 . From this configuration, the mobile robot is steered towards a neighborhood of the origin with orientation given by $-x/|x|$. About a fifth of a meter away from it, a jump of the hybrid supervisor connects controller \mathcal{K}_3 to the vehicle input (the location at which the jump occurs is denoted by the red \star). Note that the trajectory is such that controller \mathcal{K}_2 is bypassed since at the jump of the hybrid

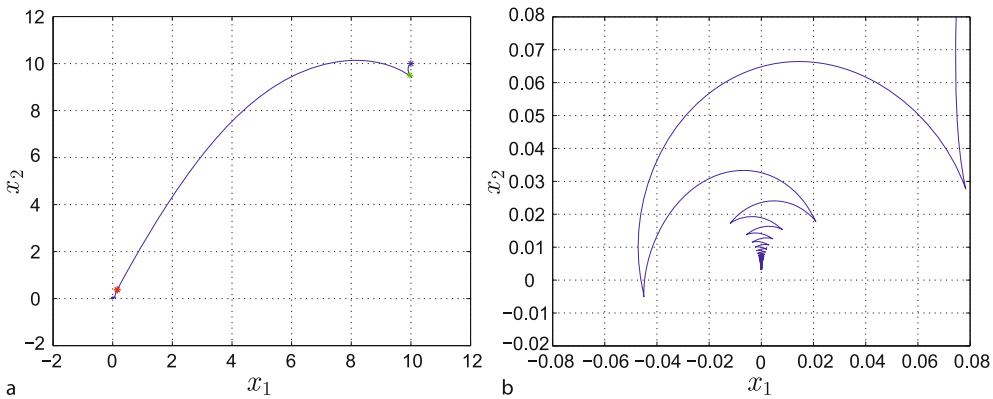
supervisor, while $|x|$ is small the orientation ξ is such that the system state does not belong to Ψ_2 but to Ψ_3 ; that is, the orientation is already close enough to $\mathbf{1}$ at that jump. Figure 8a shows a zoomed version of the x trajectory in Fig. 8b. During this phase, controller \mathcal{K}_3 is in closed-loop with the mobile robot. The vehicle is steered to the origin with orientation close to $\mathbf{1}$ by a sequence of “parking” maneuvers. Note that after about seven of those maneuvers, the vehicle position is close to $(0, 0.01)$ with almost the desired orientation.

Source Localization

Core of the Algorithm Consider the problem of programming an autonomous vehicle to find the location of a maximum for a continuously differentiable function by noting how the function values change as the vehicle moves. Like before, the vehicle dynamics are given by

$$\begin{aligned}\dot{x} &= \xi \vartheta \\ \dot{\xi} &= \xi \otimes v(\omega), \quad \xi \in S^1.\end{aligned}$$

The vehicle is to search for the maximum of the function $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$. The function is assumed to be such that its maximum is unique, denoted x^* , that $\nabla \varphi(x) = 0$ if and only if $x = x^*$, and the union of its level sets over any interval of the form $[c, \infty)$, $c \in \mathbb{R}$, yields a compact set. For simplicity, we will assume that the sign of the derivative of the function φ in the direction $\dot{x} = \xi \vartheta$ is available as a measurement. We will discuss later how to approximate this quantity by considering the changes in the value of the function φ along solutions. We also assume that the vehi-



Hybrid Control Systems, Figure 8

Global stabilization of a mobile robot to the origin with orientation $(1, 0)$. Vehicle starts at $x(0, 0) = (10, 10)$ (in meters) and $\xi(0, 0)$ corresponding to an angle of $\frac{\pi}{4}$ radians. **a** The vehicle is initially steered to a neighborhood of the origin with orientation $-x/|x|$. At about 1/5 meters away from it, controller \mathcal{K}_3 is enabled to accomplish the stabilisation task. **b** Zoomed version of trajectory in a around the origin. Controller \mathcal{K}_3 steers the vehicle to $x = (0, 0)$ and $\xi = \mathbf{1}$ by a sequence of “parking” maneuvers

cle's angle, ξ , can make jumps, according to

$$\xi^+ = \xi \otimes \zeta, \quad (\xi, \zeta) \in S^1 \times S^1.$$

We will discuss later how to solve the problem when the angle ξ cannot change discontinuously.

We propose the dynamic controller

$$\left. \begin{aligned} \dot{\vartheta} &= \bar{\vartheta} \\ \begin{bmatrix} \dot{z} \\ \omega \end{bmatrix} &= \phi(z) \end{aligned} \right\} & (x, \xi, z) \in C \\ \left. \begin{aligned} \begin{bmatrix} z^+ \\ \zeta \end{bmatrix} &\in \psi(z) \end{aligned} \right\} & (x, \xi, z) \in D$$

where $\bar{\vartheta}$ is a positive constant,

$$C := \{(x, \xi, z) \mid \langle \nabla \varphi(x), \xi \bar{\vartheta} \rangle \geq 0, \xi \in S^1, z \in \mathcal{T}\}$$

$$D := \{(x, \xi, z) \mid \langle \nabla \varphi(x), \xi \bar{\vartheta} \rangle \leq 0, \xi \in S^1, z \in \mathcal{T}\},$$

(note that C and D use information about the sign of the derivative of φ in the direction of the flow of x) and the required properties for the set \mathcal{T} , the function ϕ , and the set-valued mapping ψ are as follows:

- I. (a) The set \mathcal{T} is compact.
- (b) The maximal solutions of the continuous-time system

$$\begin{bmatrix} \dot{z} \\ \omega \end{bmatrix} = \phi(z) \quad z \in \mathcal{T}$$

and the maximal solutions to the discrete-time system

$$\begin{bmatrix} z^+ \\ \zeta \end{bmatrix} \in \psi(z) \quad z \in \mathcal{T}$$

are complete.

- II. There are no non-trivial solutions to the system

$$\left. \begin{aligned} \dot{x} &= \xi \bar{\vartheta} \\ \dot{\xi} &= \xi \otimes v(\omega) \\ \begin{bmatrix} \dot{z} \\ \omega \end{bmatrix} &= \phi(z) \end{aligned} \right\} & (x, \xi, z) \in C_0$$

where

$$C_0 := \{(x, \xi, z) \mid \langle \nabla \varphi(x), \xi \bar{\vartheta} \rangle = 0, \xi \in S^1, z \in \mathcal{T}\}.$$

- III. The only complete solutions to the system

$$\left. \begin{aligned} x^+ &= x \\ \xi^+ &= \xi \otimes \zeta \\ \begin{bmatrix} z^+ \\ \zeta \end{bmatrix} &\in \psi(z) \end{aligned} \right\} & (x, \xi, z) \in D$$

start from $x_0 = x^*$.

The first assumption on $(\mathcal{T}, \phi, \psi)$ above guarantees that the control algorithm can generate commands by either flowing exclusively or jumping exclusively. This permits arbitrary combinations of flows and jumps. Thus, since $C \cup D = \mathbb{R}^2 \times S^1 \times \mathcal{T}$, all solutions to the closed-loop system are complete. Moreover, the assumption that \mathcal{T} is compact guarantees that the only way solutions can grow unbounded is if x grows unbounded.

The second assumption on $(\mathcal{T}, \phi, \psi)$ guarantees that closed-loop flows lead to an increase in the function φ . One situation where the assumption is easy to check is when $\omega = 0$ for all $z \in \mathcal{T}$ and the maxima of φ along each search direction are isolated. In other words, the maxima of the function $\varphi_{\xi, x}: \mathbb{R} \rightarrow \mathbb{R}$ given by $\lambda \mapsto \varphi(x + \lambda \xi)$ are isolated for each $(\xi, x) \in S^1 \times \mathbb{R}^2$. In this case it is not possible to flow while keeping φ constant.

The last assumption on $(\mathcal{T}, \phi, \psi)$ guarantees that the discrete update algorithm is rich enough to be able to find eventually a direction of decrease for φ for every point x . (Clearly the only way this assumption can be satisfied is if $\nabla \varphi(x) = 0$ only if $x = x^*$, which is what we are assuming.) The assumption prevents the existence of discrete solutions at points where $x \neq x^*$.

One example of data $(\mathcal{T}, \phi, \psi)$ satisfying the three conditions above is

$$\mathcal{T} = \{0\}, \quad \phi(z) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \psi(z) = \begin{bmatrix} z \\ 0 \\ 1 \end{bmatrix}.$$

For this system, the state z does not change, the generated angular velocity ω is always zero, and the commanded rotation at each jump is $\pi/2$ radians. Other more complicated orientation-generating algorithms that make use of the dynamic state z are also possible. For example, the algorithm in [42] uses the state variable z to generate conjugate directions at the update times.

With the assumptions in place, the invariance principle of Subsect. "Invariance Principles" can be applied with the function $-\varphi$ to conclude that the closed-loop system has the compact set $\mathcal{A} := \{x^*\} \times S^1 \times \mathcal{T}$ globally asymptotically stable. Moreover, because of the robustness of global asymptotic stability to small perturbations, the results that we obtain are robust, in a practical sense, to slow variations in the characterization of the function φ , including the point where it obtains its maximum.

Practical Modifications The assumptions of the previous section that we would like to relax are that ξ can change discontinuously and that the derivative of φ is available as a measurement.

The first issue can be addressed by inserting a mode after every jump where the forward velocity is set to zero and a constant angular velocity is applied for the correct amount of time to drive ξ to the value $\xi \otimes \zeta$. If it is not possible to set the velocity to zero then some other open-loop maneuver can be executed so that, after some time, the orientation has changed by the correct amount while the position has not changed.

The second issue can be addressed by making sure that, after the direction is updated, values of the function φ along the solution are stored and compared to the current value of φ to determine the sign of the derivative. The comparison should not take place until after a sufficient amount of time has elapsed, to enhance robustness to measurement noise. The robustness of the nominal algorithm to temporal regularization and other perturbations permits such a practical implementation.

Discussion and Final Remarks

We have presented one viewpoint on hybrid control systems, but the field is still developing and other authors will give a different emphasis.

For starters, we have stressed a dynamical systems view of hybrid systems, but authors with a computer science background typically will emphasize a hybrid automaton point of view that separates discrete-valued variables from continuous-valued variables. This decomposition can be found in the early work [61] and [59], and in the more recent work [2,8,58]. An introduction to this modeling approach is given in [7]. Impulsive systems, as described in [4], are closely related to hybrid systems but rely on ordinary time domains and usually do not consider the case where the flow set and jump set overlap. The work in [18] on “left-continuous systems” is closely linked to such systems. Passing to hybrid time domains (under different names) can be seen in [20,24,39,40]; other generalized concepts of time domains can be found in [43] and in the literature on dynamical systems on time scales, see [41] for an introduction.

For simplicity, we have taken the flow map to be a function rather than a set-valued mapping. A motivation for set-valued mappings satisfying basic conditions is given in [56] where the notion of generalized solutions is developed and shown to be equivalent to solutions in the presence of vanishing perturbations or noise. Set-valued dynamics, with some consideration of the regularity of the mappings defining them, can be found in [1,2]. Implications of data regularity on the basic structural properties of the set of solutions to a system were outlined concurrently in [20,24]. The work in [24] emphasized the impli-

cations for robustness of stability theory. The latter work preceded the rigorous derivations in [23] where the proofs of statements in the Subsect. “Conditions for Well-Posedness” can be found. To derive stronger results on continuous dependence of solutions to initial conditions, extra assumptions must be added. An early result in this direction appeared in [59]; see also [11,20,40]. The work in [11] exhibited a continuous selection of solutions, continuous dependence in the set-valued sense was addressed in [17].

Sufficient Lyapunov stability conditions, for hybrid or switching systems, and relying on various concepts of a solution, appeared in [6,21,40,62]. The results in Subsect. “Lyapunov Stability Theorem” are contained in [54] which develops a general invariance principle for hybrid systems. A part of the latter result is quoted in Subsect. “Invariance Principles”. Other invariance results for hybrid or switching systems have appeared in [3,18,28,30,40]. Some early converse results for hybrid systems, relying on nonsmooth and possibly discontinuous Lyapunov functions, were given in [62]. The results quoted in Subsect. “Converse Lyapunov Theorems and Robustness” come from [15].

The development of hybrid control theory is still in its formative stages. This article has focused on the development of supervisors of hybrid controllers and related topics, as well as to applications where hybrid control gives solutions that overcome obstacles faced by classical control. For hybrid supervisors used in the context of adaptive control, see [63] and the references therein. Other results related to supervisors include [47], which considers the problem discussed at the end of Subsect. “Supervisors of Hybrid Controllers”, and [57]. The field of hybrid control systems is moving in the direction of systematic design tools, but the capabilities of hybrid control have been recognized for some time. Many of the early observations were in the context of nonholonomic systems, like the result for mobile robots we have presented as an application of supervisors. For example, see [29,31,37,44,49]. More recently, in [48] and [50] it has been established that every asymptotically controllable nonlinear system can be robustly asymptotically stabilized using logic-based hybrid feedback. Other recent results include the work in [25] and its predecessor [12], and the related work on linear reset control systems as considered in [5,45] and the references therein.

This article did not have much to do with the control of hybrid systems, other than the discussion of the juggling problem in the introduction and the structure of hybrid controllers for hybrid systems in Subsect. “Hybrid Controllers for Hybrid Systems”. A significant amount of work on the control of hybrid systems has been done, although

typically not in the framework proposed here. Notable references include [10,46,51] and the references therein.

Other interesting topics and open questions in the area of hybrid control systems are developed in [38].

Future Directions

What does the future hold for hybrid control systems? With a framework in place that mimics the framework of ordinary differential and difference equations, it appears that many new results will become available in directions that parallel results available for nonlinear control systems. These include certain types of separation principles, control algorithms based on zero dynamics (results in this direction can be found in [60] and [13]), and results based on interconnections and time-scale separation. Surely there will be unexpected results that are enabled by the fundamentally different nature of hybrid control as well.

The theory behind the control of systems with impacts will continue to develop and lead to interesting applications. It is also reasonable to anticipate further developments related to the construction of robust, embedded hybrid control systems and robust networked control systems. Likely the research community will also be inspired by hybrid control systems discovered in nature.

The future is bright for hybrid control systems design and it will be exciting to see the progress that is made over the next decade and beyond.

Bibliography

Primary Literature

1. Aubin JP, Haddad G (2001) Cadenced runs of impulse and hybrid control systems. *Internat J Robust Nonlinear Control* 11(5):401–415
2. Aubin JP, Lygeros J, Quincampoix M, Sastry SS, Seube N (2002) Impulse differential inclusions: a viability approach to hybrid systems. *IEEE Trans Automat Control* 47(1):2–20
3. Bacciotti A, Mazzi L (2005) An invariance principle for nonlinear switched systems. *Syst Control Lett* 54:1109–1119
4. Bainov DD, Simeonov P (1989) Systems with impulse effect: stability, theory, and applications. Ellis Horwood, Chichester; Halsted Press, New York
5. Beker O, Hollot C, Chait Y, Han H (2004) Fundamental properties of reset control systems. *Automatica* 40(6):905–915
6. Branicky M (1998) Multiple Lyapunov functions and other analysis tools for switched hybrid systems. *IEEE Trans Automat Control* 43(4):475–482
7. Branicky M (2005) Introduction to hybrid systems. In: Levine WS, Hristu-Varsakelis D (eds) *Handbook of networked and embedded control systems*. Birkhäuser, Boston, pp 91–116
8. Branicky M, Borkar VS, Mitter SK (1998) A unified framework for hybrid control: Model and optimal control theory. *IEEE Trans Automat Control* 43(1):31–45
9. Brockett RW (1983.) Asymptotic stability and feedback stabilization. In: Brockett RW, Millman RS, Sussmann HJ (eds) *Differential Geometric Control Theory*. Birkhauser, Boston, MA, pp 181–191
10. Brogliato B (1996) *Nonsmooth mechanics models, dynamics and control*. Springer, London
11. Broucke M, Arapostathis A (2002) Continuous selections of trajectories of hybrid systems. *Syst Control Lett* 47:149–157
12. Bupp RT, Bernstein DS, Chellaboina VS, Haddad WM (2000) Resetting virtual absorbers for vibration control. *J Vib Control* 6:61
13. Cai C, Goebel R, Sanfelice R, Teel AR (2008) Hybrid systems: limit sets and zero dynamics with a view toward output regulation. In: Astolfi A, Marconi L (eds) *Analysis and design of nonlinear control systems – In Honor of Alberto Isidori*. Springer, pp 241–261 <http://www.springer.com/west/home/generic/search/results?SGWID=4-40109-22-173754110-0>
14. Cai C, Teel AR, Goebel R (2007) Results on existence of smooth Lyapunov functions for asymptotically stable hybrid systems with nonopen basin of attraction. In: *Proc. 26th American Control Conference*, pp 3456–3461, <http://www.ccec.ece.ucsb.edu/~cai/>
15. Cai C, Teel AR, Goebel R (2007) Smooth Lyapunov functions for hybrid systems - Part I: Existence is equivalent to robustness. *IEEE Trans Automat Control* 52(7):1264–1277
16. Cai C, Teel AR, Goebel R (2008) Smooth Lyapunov functions for hybrid systems - Part II: (Pre-)asymptotically stable compact sets. *IEEE Trans Automat Control* 53(3):734–748. See also [14]
17. Cai C, Goebel R, Teel A (2008) Relaxation results for hybrid inclusions. *Set-Valued Analysis* (To appear)
18. Chellaboina V, Bhat S, Haddad W (2003) An invariance principle for nonlinear hybrid and impulsive dynamical systems. *Nonlin Anal* 53:527–550
19. Clegg JC (1958) A nonlinear integrator for servomechanisms. *Transactions AIEE* 77(Part II):41–42
20. Collins P (2004) A trajectory-space approach to hybrid systems. In: *16th International Symposium on Mathematical Theory of Networks and Systems*, CD-ROM
21. DeCarlo R, Branicky M, Pettersson S, Lennartson B (2000) Perspectives and results on the stability and stabilizability of hybrid systems. *Proc of IEEE* 88(7):1069–1082
22. Filippov A (1988) *Differential equations with discontinuous right-hand sides*. Kluwer, Dordrecht
23. Goebel R, Teel A (2006) Solutions to hybrid inclusions via set and graphical convergence with stability theory applications. *Automatica* 42(4):573–587
24. Goebel R, Hespanha J, Teel A, Cai C, Sanfelice R (2004) Hybrid systems: Generalized solutions and robust stability. In: *Proc. 6th IFAC Symposium in Nonlinear Control Systems*, pp 1–12, http://www-ccec.ece.ucsb.edu/7Ersanfelice/Preprints/final_nolcos.pdf
25. Haddad WM, Chellaboina V, Hui Q, Nersisov SG (2007) Energy- and entropy-based stabilization for lossless dynamical systems via hybrid controllers. *IEEE Trans Automat Control* 52(9):1604–1614, <http://ieeexplore.ieee.org/iel5/9/4303218/04303228.pdf>
26. Hájek O (1979) Discontinuous differential equations, I. *J Diff Eq* 32:149–170
27. Hermes H (1967) Discontinuous vector fields and feedback control. In: *Differential Equations and Dynamical Systems*, Academic Press, New York, pp 155–165

28. Hespanha J (2004) Uniform stability of switched linear systems: Extensions of LaSalle's invariance principle. *IEEE Trans Automat Control* 49(4):470–482
29. Hespanha J, Morse A (1999) Stabilization of nonholonomic integrators via logic-based switching. *Automatica* 35(3): 385–393
30. Hespanha J, Liberzon D, Angeli D, Sontag E (2005) Nonlinear norm-observability notions and stability of switched systems. *IEEE Trans Automat Control* 50(2):154–168
31. Hespanha JP, Liberzon D, Morse AS (1999) Logic-based switching control of a nonholonomic system with parametric modeling uncertainty. *Syst Control Lett* 38:167–177
32. Johansson K, Egerstedt M, Lygeros J, Sastry S (1999) On the regularization of zeno hybrid automata. *Syst Control Lett* 38(3):141–150
33. Kellet CM, Teel AR (2004) Smooth Lyapunov functions and robustness of stability for differential inclusions. *Syst Control Lett* 52:395–405
34. Krasovskii N (1970) *Game-Theoretic Problems of capture*. Nauka, Moscow
35. LaSalle JP (1967) An invariance principle in the theory of stability. In: Hale JK, LaSalle JP (eds) *Differential equations and dynamical systems*. Academic Press, New York
36. LaSalle J (1976) *The stability of dynamical systems*. SIAM's Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia
37. Lucibello P, Oriolo G (1995) Stabilization via iterative state steering with application to chained-form systems. In: *Proc. 35th IEEE Conference on Decision and Control*, pp 2614–2619
38. Lygeros J (2005) An overview of hybrid systems control. In: Levine WS, Hristu-Varsakelis D (eds) *Handbook of Networked and Embedded Control Systems*. Birkhäuser, Boston, pp 519–538
39. Lygeros J, Johansson K, Sastry S, Egerstedt M (1999) On the existence of executions of hybrid automata. In: *Proc. 38th IEEE Conference on Decision and Control*, pp 2249–2254
40. Lygeros J, Johansson K, Simić S, Zhang J, Sastry SS (2003) Dynamical properties of hybrid automata. *IEEE Trans Automat Control* 48(1):2–17
41. M Böhner M, Peterson A (2001) *Dynamic equations on time scales. An introduction with applications*. Birkhäuser, Boston
42. Mayhew CG, Sanfelice RG, Teel AR (2007) Robust source seeking hybrid controllers for autonomous vehicles. In: *Proc. 26th American Control Conference*, pp 1185–1190
43. Michel A (1999) Recent trends in the stability analysis of hybrid dynamical systems. *IEEE Trans Circuits Syst – I Fund Theory Appl* 45(1):120–134
44. Morin P, Samson C (2000) Robust stabilization of driftless systems with hybrid open-loop/feedback control. In: *Proc. 19th American Control Conference*, pp 3929–3933
45. Nesic D, Zaccarian L, Teel A (2008) Stability properties of reset systems. *Automatica* 44(8):2019–2026
46. Plestan F, Grizzle J, Westervelt E, Abba G (2003) Stable walking of a 7-dof biped robot. *IEEE Trans Robot Automat* 19(4): 653–668
47. Prieur C (2001) Uniting local and global controllers with robustness to vanishing noise. *Math Contr, Sig Syst* 14(2): 143–172
48. Prieur C (2005) Asymptotic controllability and robust asymptotic stabilizability. *SIAM J Control Opt* 43:1888–1912
49. Prieur C, Astolfi A (2003) Robust stabilization of chained systems via hybrid control. *IEEE Trans Automat Control* 48(10):1768–1772
50. Prieur C, Goebel R, Teel A (2007) Hybrid feedback control and robust stabilization of nonlinear systems. *IEEE Trans Automat Control* 52(11):2103–2117
51. Ronsse R, Lefèvre P, Sepulchre R (2007) Rhythmic feedback control of a blind planar juggler. *IEEE Transactions on Robotics* 23(4):790–802, <http://www.montefiore.ulg.ac.be/services/stochastic/pubs/2007/RLS07>
52. Ryan E (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. *SIAM J Control Optim* 32(6):1597–1604
53. Sanfelice R, Goebel R, Teel A (2006) A feedback control motivation for generalized solutions to hybrid systems. In: Hespanha JP, Tiwari A (eds) *Hybrid Systems: Computation and Control: 9th International Workshop*, vol LNCS, vol 3927. Springer, Berlin, pp 522–536
54. Sanfelice R, Goebel R, Teel A (2007) Invariance principles for hybrid systems with connections to detectability and asymptotic stability. *IEEE Trans Automat Control* 52(12):2282–2297
55. Sanfelice R, Teel AR, Sepulchre R (2007) A hybrid systems approach to trajectory tracking control for juggling systems. In: *Proc. 46th IEEE Conference on Decision and Control*, pp 5282–5287
56. Sanfelice R, Goebel R, Teel A (2008) Generalized solutions to hybrid dynamical systems. *ESAIM: Control, Optimisation and Calculus of Variations* 14(4):699–724
57. Sanfelice RG, Teel AR (2007) A “throw-and-catch” hybrid control strategy for robust global stabilization of nonlinear systems. In: *Proc. 26th American Control Conference*, pp 3470–3475
58. van der Schaft A, Schumacher H (2000) *An introduction to hybrid dynamical systems*. Lecture notes in control and information sciences. Springer, London
59. Tavernini L (1987) Differential automata and their discrete simulators. *Nonlinear Anal* 11(6):665–683
60. Westervelt E, Grizzle J, Koditschek D (2003) Hybrid zero dynamics of planar biped walkers. *IEEE Trans Automat Control* 48(1):42–56
61. Witsenhausen HS (1966) A class of hybrid state continuous-time dynamic systems. *IEEE Trans Automat Control* 11(2):161–167
62. Ye H, Mitchel A, Hou L (1998) Stability theory for hybrid dynamical systems. *IEEE Trans Automat Control* 43(4):461–474
63. Yoon TW, Kim JS, Morse A (2007) Supervisory control using a new control-relevant switching. *Automatica* 43(10): 1791–1798

Books and Reviews

- Aubin JP, Cellina A (1984) *Differential inclusions*. Springer, Berlin
- Haddad WM, Chellaboina V, Nersisov SG (2006) *Impulsive and hybrid dynamical systems: stability, dissipativity, and control*. Princeton University Press, Princeton
- Levine WS, Hristu-Varsakelis D (2005) *Handbook of networked and embedded control systems*. Birkhäuser, Boston
- Liberzon D (2003) *Switching in systems and control. Systems and control: Foundations and applications*. Birkhäuser, Boston
- Matveev AS, Savkin AV (2000) *Qualitative theory of hybrid dynamical systems*. Birkhäuser, Boston

Michel AN, Wang L, Hu B (2001) Qualitative theory of dynamical systems. Dekker
 Rockafellar RT, Wets RJ-B (1998) Variational analysis. Springer, Berlin

Hybrid Soft Computing Models for Systems Modeling and Control

OSCAR CASTILLO, PATRICIA MELIN
 Division of Graduate Studies and Research,
 Tijuana Institute of Technology, Tijuana, Mexico

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Genetic Algorithm for Optimization
 Evolution of Fuzzy Systems
 Application to Anesthesia Control
 Application to the Control of the Bar and Ball System
 Hierarchical Genetic Algorithms for Neural Networks
 Experimental Results for Time Series Prediction
 Conclusions
 Future Directions
 Bibliography

Glossary

Hybrid intelligent systems Intelligent Systems that are build using a combination of soft computing techniques. In particular, Soft Computing includes fuzzy logic, neural networks, genetic algorithms or hybrid approaches.

Intelligent control The application of intelligent techniques for achieving the control of non-linear plants. In particular, the use of fuzzy logic, neural networks, genetic algorithms or hybrid approaches for designing intelligent controllers.

Soft computing Soft Computing is a new area of Computer Science that deals with new intelligent methodologies that combine symbolic and numerical calculations. In particular, Soft Computing includes, at the moment, methodologies like fuzzy logic, neural networks, genetic algorithms or hybrid approaches.

Fuzzy systems Intelligent systems that are developed based on the theory of fuzzy logic, fuzzy inference and membership functions, and fuzzy rules. Fuzzy systems are able to manage the uncertainty of the decision process of humans, and for this reason are able to

mimic the expert decision process in automation applications.

Genetic algorithms Genetic algorithms are search optimization techniques that mimic natural evolution for finding solution to complex problems. In particular, genetic algorithms use operators to generate new candidate solutions based on the selection of previous good solutions.

Evolution of fuzzy systems Application of evolutionary algorithms to the optimization of number of fuzzy rules and membership functions, as well as the parameter values of the fuzzy system.

Definition of the Subject

The evolutionary design of hybrid intelligent systems using hierarchical genetic algorithms will be described in this paper. The evolutionary approach can be used for fuzzy system optimization in intelligent control. In particular, we consider the problem of optimizing the number of rules and membership functions using an evolutionary approach. The hierarchical genetic algorithm enables the optimization of the fuzzy system design for a particular application. We illustrate the approach with two cases of intelligent control. Simulation results for both applications show that we are able to find an optimal set of rules and membership functions for the fuzzy control system. We also describe the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction.

Introduction

The application of a Hierarchical Genetic Algorithm (HGA) for fuzzy system optimization [14] will be described in this paper. In particular, we consider the problem of finding the optimal set of rules and membership functions for a specific application [24]. The HGA is used to search for this optimal set of rules and membership functions, according to the data about the problem. We consider, as an illustration, the case of a fuzzy system for intelligent control [4].

Fuzzy systems are capable of handling complex, non-linear and sometimes mathematically intangible dynamic systems using simple solutions [10]. Very often, fuzzy systems may provide a better performance than conventional non-fuzzy approaches with less development cost [17]. However, to obtain an optimal set of fuzzy membership functions and rules is not an easy task [12]. It requires

time, experience and skills of the designer for the tedious fuzzy tuning exercise [22]. In principle, there is no general rule or method for the fuzzy logic set-up, although a heuristic and iterative procedure for modifying the membership functions to improve performance has been proposed [19]. Recently, many researchers have considered a number of intelligent schemes for the task of tuning the fuzzy system [1]. The noticeable Neural Network (NN) approach [9] and the Genetic Algorithm (GA) approach [8] to optimize either the membership functions or rules, have become a trend for fuzzy logic system development.

The HGA approach differs from the other techniques [5] in that it has the ability to reach an optimal set of membership functions and rules without a known fuzzy system topology [20]. During the optimization phase, the membership functions need not be fixed. Throughout the genetic operations [7], a reduced fuzzy system including the number of membership functions and fuzzy rules will be generated [25]. The HGA approach has a number of advantages:

1. An optimal and the least number of membership functions and rules are obtained,
2. no pre-fixed fuzzy structure is necessary, and
3. simpler implementing procedures and less cost are involved.

We consider in this paper the case of automatic anesthesia control in human patients for testing the optimized fuzzy controller. We did have, as a reference, the best fuzzy controller that was developed for the automatic anesthesia control [13], and we consider the optimization of this controller using the HGA approach. After applying the genetic algorithm the number of fuzzy rules was reduced from 12 to 9 with a similar performance of the fuzzy controller. Of course, the parameters of the membership functions were also tuned by the genetic algorithm. We did compare the simulation results of the optimized fuzzy controllers obtained with the HGA against the best fuzzy controller that was obtained previously with expert knowledge, and control is achieved in a similar fashion. Since simulation results are similar, and the number of fuzzy rules was reduced, we can conclude that the HGA approach is a good alternative for designing fuzzy systems.

We also describe the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction.

Genetic Algorithm for Optimization

In this paper, we used a floating-point genetic algorithm [3] to adjust the parameter vector θ , specifically we used the Breeder Genetic Algorithm (BGA). The genetic algorithm is used to optimize the fuzzy system for control that will be described later. A BGA can be described by the following equation:

$$\text{BGA} = (P_g^0, N, T, \Gamma, \Delta, \text{HC}, F, \text{term}) \quad (1)$$

where: P_g^0 = initial population, N = the size of the population, T = the truncation threshold, Γ = the recombination operator, Δ = the mutation operator, HC = the hill climbing method, F = the fitness function, term = the termination criterion.

The BGA uses a selection scheme called truncation selection. The $\%T$ best individuals are selected and mated randomly until the number of offspring is equal the size of the population. The offspring generation is equal to the size of the population. The offspring generation replaces the parent population. The best individual found so far will remain in the population. Self-mating is prohibited. As a recombination operator we used "extended intermediate recombination", defined as: If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ are the parents, then the successor $z = (z_1, \dots, z_n)$ is calculated by:

$$z_i = x_i + \alpha_i(y_i - x_i) \quad i = 1, \dots, n. \quad (2)$$

The mutation operator is defined as follows: A variable x_i is selected with probability p_m for mutation. The BGA normally uses $p_m = 1/n$. At least one variable will be mutated. A value out of the interval $[-\text{range}_i, \text{range}_i]$ is added to the variable. range_i defines the mutation range. It is normally set to $(0.1 \times \text{searchinterval}_i)$. searchinterval_i is the domain of definition for variable x_i . The new value z_i is computed according to

$$z_i = x_i \pm \text{range}_i \cdot \delta. \quad (3)$$

The $+$ or $-$ sign is chosen with probability 0.5. δ is computed from a distribution which prefers small values. This is realized as follows

$$\delta = \sum_{i=0}^{15} \alpha_i 2^i \quad \alpha_i \in 0, 1. \quad (4)$$

Before mutation we set $\alpha_i = 0$. Then each α_i is mutated to 1 with probability $p_\delta = 1/16$. Only $\alpha_i = 1$ contributes to the sum. On the average there will be just one α_i with value 1, say α_j . Then δ is given by

$$\delta = 2^{-j}. \quad (5)$$

The standard BGA mutation operator is able to generate any point in the hypercube with center x defined by $x_i \pm \text{range}_i$. But it generates values much more often in the neighborhood of x . In the above standard setting, the mutation operator is able to locate the optimal x_i up to a precision of $\text{range}_i \cdot 2^{-150}$.

We also solved the problem with a LMS algorithm, with the purpose of have a good reference mark. To monitor the convergence rate of the LMS algorithm, we computed a short term average of the squared error $e^2(n)$ using

$$\text{ASE}(m) = \frac{1}{K} \sum_{k=n+1}^{n+K} e^2(k) \quad (6)$$

where $m = n/K = 1, 2, \dots$. The averaging interval K may be selected to be (approximately) $K = 10N$. The effect of the choice of the step size parameter Δ on the convergence rate of LMS algorithm may be observed by monitoring the $\text{ASE}(m)$.

Genetic Algorithm for Optimization

The proposed genetic algorithm is as follows:

1. We use real numbers as a genetic representation of the problem.
2. We initialize variable i with zero ($i = 0$).
3. We create an initial random population P_i , in this case (P_0). Each individual of the population has n dimensions and, each coefficient of the fuzzy system corresponds to one dimension. Since we are using a fuzzy system of 25 coefficients, then our search space is of $n = 25$. The generated individuals have their coefficients in $[-3, 3]$.
4. We calculate the normalized fitness of each individual of the population using linear scaling with displacement, in the following form:

$$f'_i = f_i + \frac{1}{N} \sum |f_i| + \left| \min_i(f_i) \right| \quad \forall i.$$

5. We normalize the fitness of each individual using:

$$F_i = \frac{f'_i}{\sum_{i=1}^N f'_i} \quad \forall i.$$

6. We sort the individuals from greater to lower fitness.
7. We use the truncated selection method, selecting the %T best individuals, for example if there are 500 individuals and, then we select $0.30 \cdot 500 = 150$ individuals.

8. We apply random crossover, to the individuals in the population (the 150 best ones) with the goal of creating a new population (of 500 individuals). Crossover with it self is not allowed, and all the individuals have to participate. To perform this operation we apply the genetic operator of extended intermediate recombination as follows:

If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ are the parents, then the successors $z = (z_1, \dots, z_n)$ are calculated by, $z_i = x_i + \alpha_i(y_i - x_i)$ for $i = 1, \dots, n$ where α is a scaling factor selected randomly in the interval $[-d, 1 + d]$. In intermediate recombination $d = 0$, and for extended $d > 0$, a good choice is $d = 0.25$, which is the one that we used.

9. We apply the mutation genetic operator of BGA. In this case, we select an individual with probability $p_m = 1/n$ (where n represents the working dimension, in this case $n = 25$, which is the number of coefficients in the membership functions). The mutation operator calculates the new individuals z_i of the population in the following form: $z_i = x_i \pm \text{range}_i \delta$ we can note from this equation that we are actually adding to the original individual a value in the interval: $[-\text{range}_i, \text{range}_i]$ the range is defined as the search interval, which in this case is the domain of variable x_i , the sign \pm is selected randomly with probability of 0.5, and is calculated using the following formula,

$$\delta = \sum_{i=0}^{m-1} \alpha_i 2^{-i} \quad \alpha_i \in 0, 1.$$

Common used values in this equation are $m = 16$ and $m = 20$. Before mutation we initiate with $\alpha_i = 0$, then for each α_i we mutate to 1 with probability $p_\delta = 1/m$.

10. Let $i = i + 1$, and continue with step 4.

Evolution of Fuzzy Systems

Ever since the very first introduction of the fundamental concept of fuzzy logic [26], its use in engineering disciplines has been widely studied. Its main attraction undoubtedly lies in the unique characteristics that fuzzy logic systems possess [27]. They are capable of handling complex, non-linear dynamic systems using simple solutions. Very often, fuzzy systems provide a better performance than conventional non-fuzzy approaches with less development cost [24].

However, to obtain an optimal set of fuzzy membership functions and rules is not an easy task [11]. It requires time, experience, and skills of the operator for the tedious

fuzzy tuning exercise [2]. In principle, there is no general rule or method for the fuzzy logic set-up [15]. Recently, many researchers have considered a number of intelligent techniques for the task of tuning the fuzzy set.

Here, another innovative scheme is described [23]. This approach has the ability to reach an optimal set of membership functions and rules without a known overall fuzzy set topology. The conceptual idea of this approach is to have an automatic and intelligent scheme to tune the membership functions and rules, in which the conventional closed loop fuzzy control strategy remains unchanged, as indicated in Fig. 1.

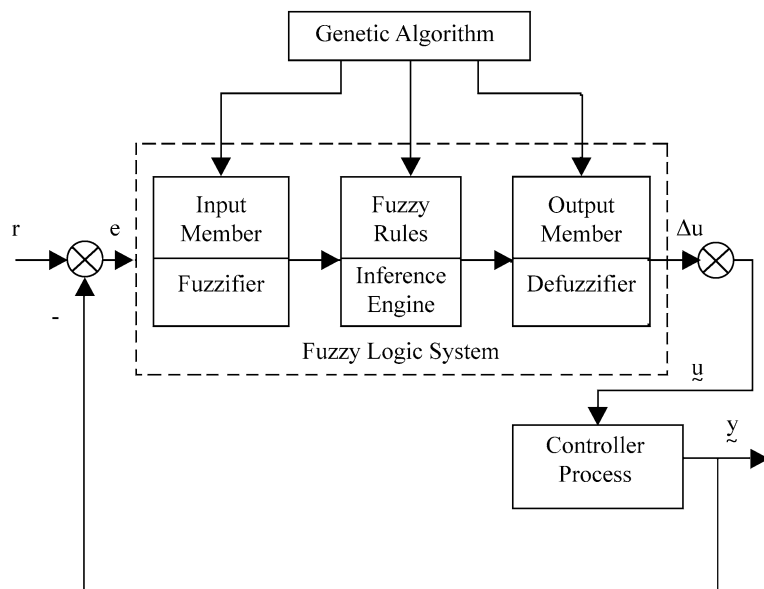
In this case, the chromosome [6] of a particular system is shown in Fig. 2. The chromosome consists of two types of genes, the control genes and parameter genes. The control genes, in the form of bits, determine the membership function activation, whereas the parameter genes are in the form of real numbers to represent the membership functions.

To obtain a complete design for the fuzzy control system, an appropriate set of fuzzy rules is required to en-

sure system performance [25]. At this point it should be stressed that the introduction of the control genes is done to govern the number of fuzzy subsets in the system.

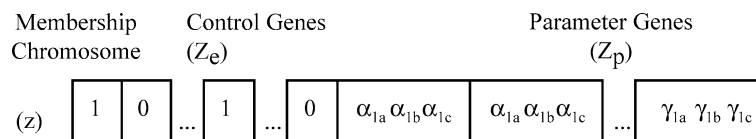
Once the formulation of the chromosome has been set for the fuzzy membership functions and rules, the genetic operation cycle can be performed. This cycle of operation for the fuzzy control system optimization using a genetic algorithm is illustrated in Fig. 3.

There are two population pools, one for storing the membership chromosomes and the other for storing the fuzzy rule chromosomes. We can see this in Fig. 3 as the membership population and fuzzy rule population, respectively. Considering that there are various types of gene structure, a number of different genetic operations can be used. For the crossover operation, a one-point crossover is applied separately for both the control and parameter genes of the membership chromosomes within certain operation rates. There is no crossover operation for fuzzy rule chromosomes since only one suitable rule set can be assisted.



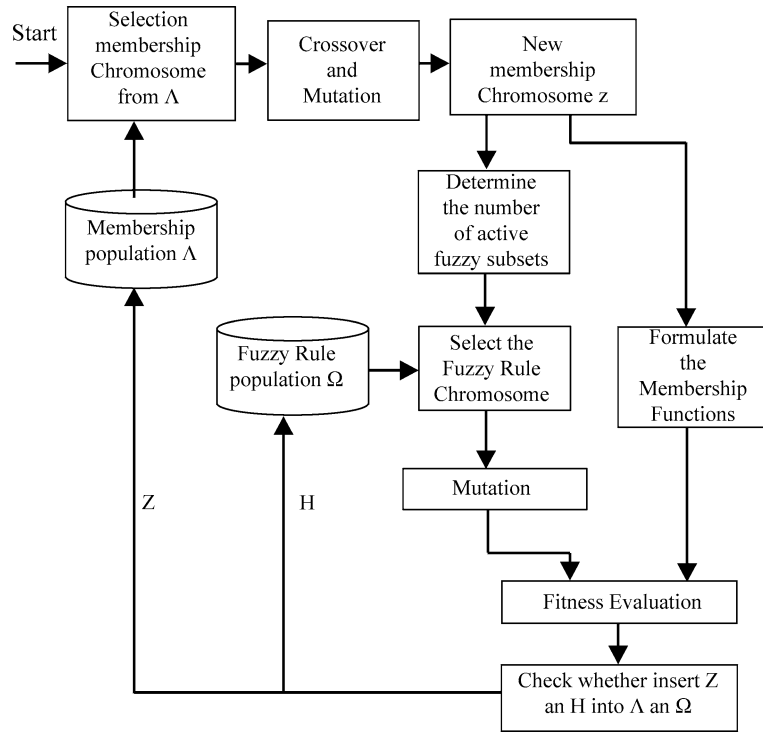
Hybrid Soft Computing Models for Systems Modeling and Control, Figure 1

Genetic algorithm for a fuzzy control system



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 2

Chromosome structure for the fuzzy system



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 3
Genetic cycle for fuzzy system optimization

Bit mutation is applied for the control genes of the membership chromosome. Each bit of the control gene is flipped if a probability test is satisfied (a randomly generated number is smaller than a predefined rate). As for the parameter genes, which are real number represented, random mutation is applied.

The complete genetic cycle continues until some termination criteria, for example, meeting the design specification or number of generation reaching a predefined value are fulfilled.

The fitness function can be defined in this case as follows:

$$f_i = \sum |y(k) - r(k)| \quad (7)$$

where \sum indicates the sum for all the data points in the training set, and $y(k)$ represents the real output of the fuzzy system and $r(k)$ is the reference output. This fitness value measures how well the fuzzy system is approximating the real data of the problem.

Application to Anesthesia Control

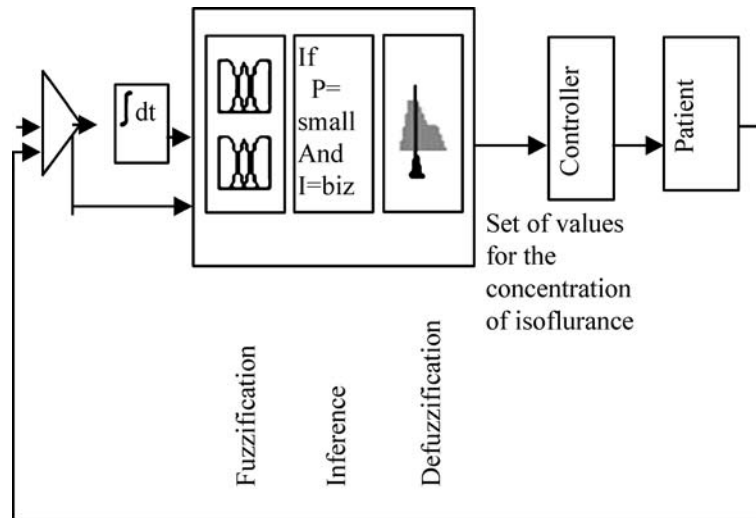
We consider the case of controlling the anesthesia given to a patient as the problem for finding the optimal fuzzy system for control [13]. The complete implementation was

done in the MATLAB programming language. The fuzzy systems were built automatically by using the Fuzzy Logic Toolbox, and the genetic algorithm was coded directly in the MATLAB language. The fuzzy systems for control are the individuals used in the genetic algorithm, and these are evaluated by comparing them to the ideal control given by the experts. In other words, we compare the performance of the fuzzy systems that are generated by the genetic algorithm, against the ideal control system given by the experts in this application. We give more details below.

Anesthesia Control Using Fuzzy Logic

The main task of the anesthetist, during and operation, is to control anesthesia concentration. In any case, anesthesia concentration can't be measured directly. For this reason, the anesthetist uses indirect information, like the heartbeat, pressure, and motor activity. The anesthesia concentration is controlled using a medicine, which can be given by a shot or by a mix of gases. We consider here the use of isoflurane, which is usually given in a concentration of 0 to 2% with oxygen. In Fig. 4 we show a block diagram of the controller.

The air that is exhaled by the patient contains a specific concentration of isoflurane, and it is recirculated to the



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 4
Architecture of the fuzzy control system

patient. As consequence, we can measure isoflurance concentration on the inhaled and exhaled air by the patient, to estimate isoflurance concentration on the patient's blood. From the control engineering point of view, the task by the anesthesiologist is to maintain anesthesia concentration between the high level W (threshold to wake up) and the low level E (threshold to success). These levels are difficult to be determined in a changing environment and also are dependent on the patient's condition. For this reason, it is important to automate this anesthesia control, to perform this task more efficiently and accurately, and also to free the anesthesiologist from this time consuming job. The anesthesiologist can then concentrate in doing other task during operation of a patient.

The first automated system for anesthesia control was developed using a PID controller in the 60's. However, this system was not very successful due to the non-linear nature of the problem of anesthesia control. After this first attempt, adaptive control was proposed to automate anesthesia control, but robustness was the problem in this case. For these reasons, fuzzy logic was proposed for solving this problem. An additional advantage of fuzzy control is that we can use in the rules the same vocabulary as the medical doctors use. The fuzzy control system can also be easily interpreted by the anesthesiologists.

Characteristics of the Fuzzy Controller

In this section we describe the main characteristics of the fuzzy controller for anesthesia control. We will define input and output variable of the fuzzy system. Also, the fuzzy

rules of fuzzy controller previously designed will be described.

The fuzzy system is defined as follows:

1. Input variables: Blood pressure and Error.
2. Output variable: Isoflurance concentration.
3. Nine fuzzy if-then rules of the optimized system, which is the base for comparison.
4. 12 fuzzy if-then rules of an initial system to begin the optimization cycle of the genetic algorithm.

The linguistic values used in the fuzzy rules are the following:

- PB = Positive Big
- PS = Positive Small
- ZERO = zero
- NB = Negative Big
- NS = Negative Small

We show below a sample set of fuzzy rules that are used in the fuzzy inference system that is represented in the genetic algorithm for optimization:

- if Blood pressure is NB and error is NB then conc_isoflurance is PS
- if Blood pressures is PS then conc_isoflurance is NS
- if Blood pressure is NB then conc_isoflurance is PB
- if Blood pressure is PB then conc_isoflurance is NB
- if Blood pressure is ZERO and error is ZERO then conc_isoflurance is ZERO
- if Blood pressure is ZERO and error is PS then conc_isoflurance is NS

- if Blood pressure is ZERO and error is NS then conc_isoflurance is PS
- if error is NB then conc_isoflurance is PB
- if error is PB then conc_isoflurance is NB
- if error is PS then conc_isoflurance is NS
- if Blood pressure is NS and error is ZERO then conc_isoflurance is NB
- if Blood pressure is PS and error is ZERO then conc_isoflurance is PS.

Genetic Algorithm Specification

The general characteristics of the genetic algorithm that was used are the following:

- **NIND** = 40; % Number of individuals in each subpopulation.
- **MAXGEN** = 100; % Maximum number of generations allowed.
- **GGAP** = .6; % “Generational gap”, which is the percentage from the complete population of new individuals generated in each generation.
- **PRECI** = 120; % Precision of binary representations.
- **SelCh** = select('rws', Chrom, FitnV, GGAP); % Roulette wheel method for selecting the individuals participating in the genetic operations.
- **SelCh** = recomb('xovmp', SelCh, 0.7); % Multi-point crossover as recombination method for the selected individuals.
- **ObjV** = FuncionObjDifuso120_555(Chrom, sdifuso); Objective function is given by the error between the performance of the ideal control system given by the experts and the fuzzy control system given by the genetic algorithm.

Representation of the Chromosome

In Table 1 we show the chromosome representation, which has 120 binary positions. These positions are divided in two parts, the first one indicates the number of

Hybrid Soft Computing Models for Systems Modeling and Control, Table 1

Binary chromosome representation

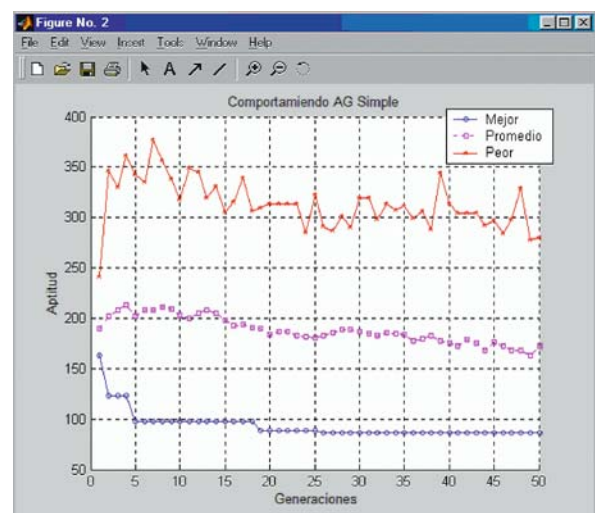
Bit assigned	Representation
1 to 12	Which rule is active or inactive
13 to 21	Membership functions active or inactive of rule 1
22 to 30	Membership functions active or inactive of rule 2
...	Membership functions active or inactive of rule ...
112 to 120	Membership functions active or inactive of rule 12

rules of the fuzzy inference system, and the second one is divided again into fuzzy rules to indicate which membership functions are active or inactive for the corresponding rule.

Simulation Results for the Case of Anesthesia Control

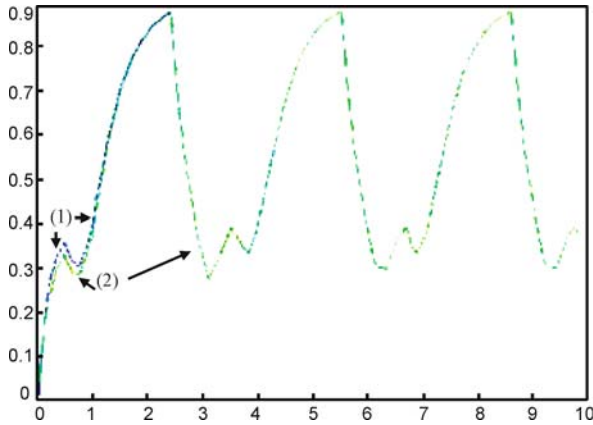
We describe in this section the simulation results that were achieved using the hierarchical genetic algorithm for the optimization of the fuzzy control system, for the case of anesthesia control. The genetic algorithm is able to evolve the topology of the fuzzy system for the particular application. We used 50 generations of 40 individuals each to achieve the minimum error. We show in Fig. 5 the final results of the genetic algorithm, where the error has been minimized. This is the case in which only nine fuzzy rules are needed for the fuzzy controller. The value of the minimum error achieved with this particular fuzzy logic controller was of 0.0064064, which is considered a small number in this application.

In Fig. 6 we show the simulation results of the fuzzy logic controller produced by the genetic algorithm after evolution. We used a sinusoidal input signal with unit amplitude and a frequency of 2 radians/second, with a transfer function of $[1/(0.5s + 1)]$. In this figure we can appreciate the comparison of the outputs of both the ideal controller (1) and the fuzzy controller optimized by the genetic algorithm (2). From this figure it is clear that both controllers are very similar and as a consequence we can conclude that the genetic algorithm was able to optimize



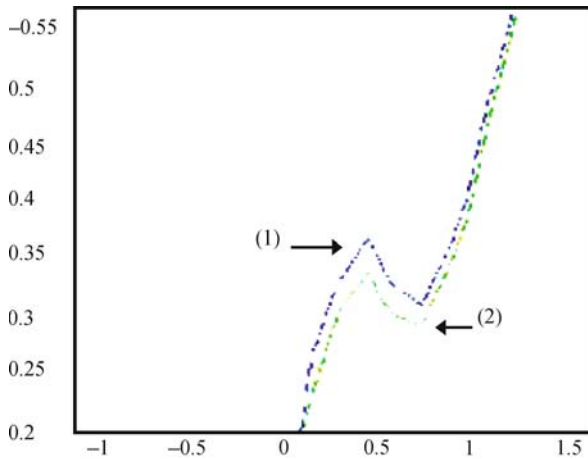
Hybrid Soft Computing Models for Systems Modeling and Control, Figure 5

Plot of the error after 50 generations of the HGA



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 6

Comparison between outputs of the ideal controller (1) and the fuzzy controller with the HGA (2)



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 7

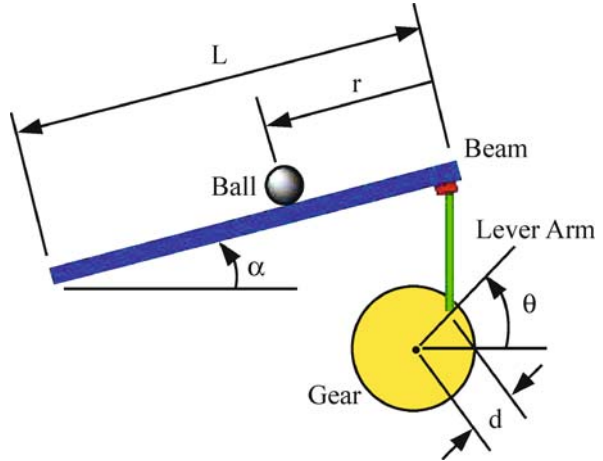
Zoom in of Fig. 6 to view in more detail the difference between the controllers

the performance of the fuzzy logic controller. We can also appreciate this fact more clearly in Fig. 7, where we have amplified the simulation results from Fig. 6 for a better view.

Application to the Control of the Bar and Ball System

In this section, we describe the ball and beam experiment [23], which was also used as a basis for testing the genetic approach of fuzzy controller optimization.

A ball is placed on a beam, see Fig. 8, where it is allowed to roll with one degree of freedom along the length of the beam. A lever arm is attached to the beam at one



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 8

Diagram of the ball and beam system

end and a servo gear at the other. As the servo gear turns by an angle θ , the lever changes the angle of the beam by a magnitude "alpha". When the angle is changed from the horizontal position, gravity causes the ball to roll along the beam. A controller will be designed for this system so that the ball's position can be manipulated.

For this problem, we will assume that the ball rolls without slipping and friction between the beam and ball is negligible. The constants for this example are defined as follows:

- M mass of the ball 0.11 kg
- R radius of the ball 0.015 m
- d lever arm offset 0.03 m
- g gravitational acceleration 9.8 m/s²
- L length of the beam 1.0 m
- J moment of inertia 9.99×10^{-6} kg m²
- r ball position coordinate
- α beam angle coordinate
- θ servo gear angle

The Lagrangian equation of motion for the ball is then given by the following equation:

$$(J/R^2 + m)r'' + mg \sin \alpha - mr(\alpha')^2 = 0. \quad (8)$$

Simulation Results with the Complete HGA

- NIND = 50; % Number of individuals in the population
- MAXGEN = 80; % Maximum number of generations
- GGAP = 0.8; % Generation gap
- PRECI = 120; % Length of the Chromosome

In this case, we use the complete HGA with the following parameters:

Hybrid Soft Computing Models for Systems Modeling and Control, Table 2

Fuzzy rules in indexed and linguistic form

Indexed Rules	Linguistic Rules
1 1, 1 (1): 1	If error is N and derror is N then Angle is NG
1 2, 2 (1): 1	If error is N and derror is Z then Angle is N
2 1, 2 (1): 1	If error is Z and derror is N then Angle is N
2 2, 3 (1): 1	If error is Z and derror is Z then Angle is Z
2 3, 4 (1): 1	If error is Z and derror is P then Angle is P
3 2, 4 (1): 1	If error is P and derror is Z then Angle is P
3 3, 5 (1): 1	If error is P and derror is P then Angle is PG

Hybrid Soft Computing Models for Systems Modeling and Control, Table 3

Fuzzy rules in indexed and linguistic form

Indexed Rules	Linguistic Rules
1 1, 1 (1): 1	If error is N and derror is N then Angle is NG
1 2, 2 (1): 1	If error is N and derror is Z then Angle is N
2 1, 2 (1): 1	If error is Z and derror is N then Angle is N
2 2, 3 (1): 1	If error is Z and derror is Z then Angle is Z
2 3, 4 (1): 1	If error is Z and derror is P then Angle is P
3 2, 4 (1): 1	If error is P and derror is Z then Angle is P

At the end of the genetic evolution seven fuzzy rules were obtained, which are shown in Table 2.

Simulation Results with a Method Dased Only on Mutation

In this section, we show results for the HGA method based only on mutation (no crossover was used). The parameters of the genetic algorithm are:

- NIND = 50; % Number of individuals in the population
- MAXGEN = 250; % Maximum number of generations
- GGAP = 0.8; % Generation gap
- PRECI = 120; % Length of the chromosome

Table 3 shows the fuzzy rules obtained at the end with this type of genetic algorithm.

Simulation Results with a Method Based Only on Crossover

In this section, we show results for the HGA method based only on crossover (no mutation was used). The parameters of the genetic algorithm are the same. The fuzzy rules obtained at the end of the genetic evolution are shown in Table 4.

Hybrid Soft Computing Models for Systems Modeling and Control, Table 4

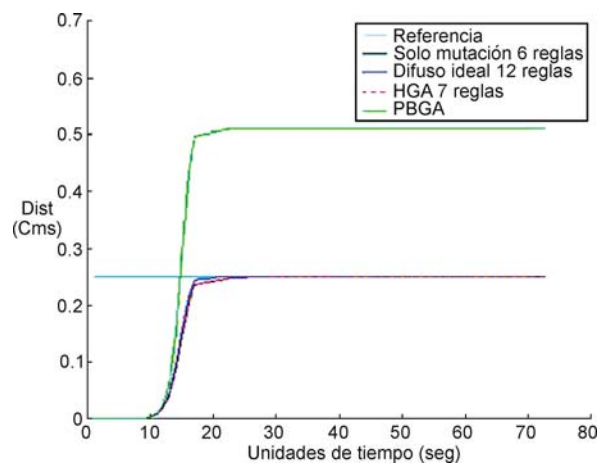
Fuzzy rules in indexed and linguistic form

Indexed rules	Linguistic rules
1 1, 1 (1): 1	If error is N and derror is N then Angle is NG
1 3, 3 (1): 1	If error is N and derror is P then Angle is Z
2 1, 2 (1): 1	If error is Z and derror is N then Angle is N
2 3, 4 (1): 1	If error is Z and derror is P then Angle is P
3 2, 4 (1): 1	If error is P and derror is Z then Angle is P
3 3, 5 (1): 1	If error is P and derror is P then Angle is PG

Comparison of the Simulation Results

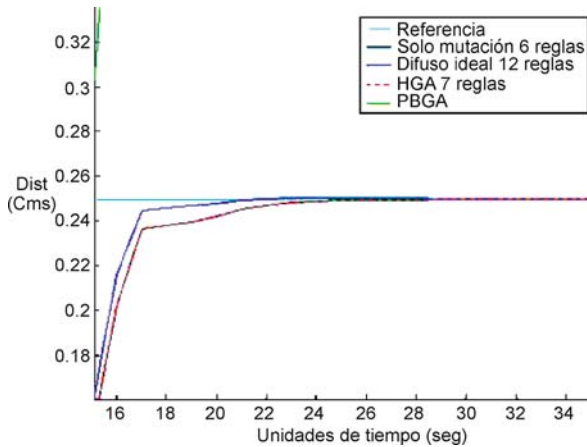
In this section we show the comparison of the fuzzy controllers that were obtained with the different types of genetic algorithms that were considered. We show in Figs. 9 and 10 the comparison of the responses of the different fuzzy controllers. From these figures we can appreciate that the different types of HGA work. However, a pseudobacterial genetic algorithm (PBGA) does not give a valid fuzzy controller. The PBGA is not described in this paper, but can be seen with detail in [16] and [18].

Table 5 shows the comparison between the different methods used. This table summarizes response times of the controllers and the number of fuzzy rules of the corresponding controllers. This table also contains the information of a PID controller. We also show in Fig. 11 the behavior of the PID controller to have a basis for comparison with the other controllers.



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 9

Responses of the fuzzy controllers obtained (green = PBGA, red = HGA complete, blue = ideal fuzzy, black = HGA with only mutation)



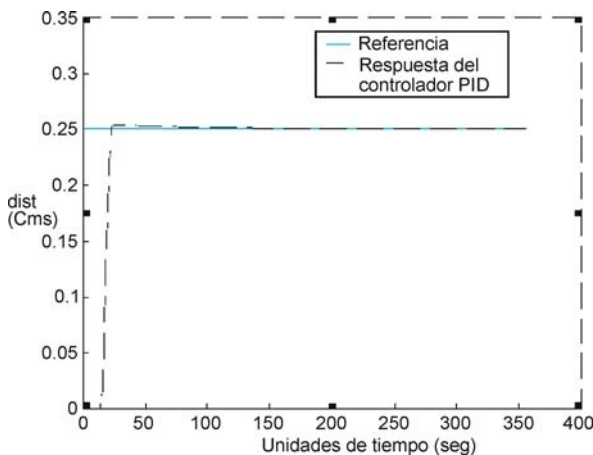
Hybrid Soft Computing Models for Systems Modeling and Control, Figure 10

Magnification of Fig. 9

Hybrid Soft Computing Models for Systems Modeling and Control, Table 5

Comparison of the different methods

Methods	Time	Reference	Number of rules
Only mutation	12.000	0.25	6
Ideal Fuzzy System (Not optimized)	13.8000	0.25	12
Complete HGA	12.000	0.25	7
PBGA	Did not control	0.25	9
Only crossover	Did not control	0.25	7
Traditional (PID)	400	0.25	Transfer Function



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 11

Response of the PID controller (blue = reference, black = response of the controller)

We can appreciate from Table 5 that the best result corresponds to the method using only mutation, because it achieved a lower number of fuzzy rules without losing the efficiency of the control goal. The complete HGA method also achieves control but with one more fuzzy rule. The ideal fuzzy controller (from the experts) also achieves control but with 12 fuzzy rules. The other methods do not achieve the control goal.

Hierarchical Genetic Algorithms for Neural Networks

The bottleneck problem for NN application lies within the optimization procedures that are used to obtain an optimal NN topology. Hence, the formulation of the Hierarchical Genetic Algorithm (HGA) is applied for this purpose. The HGA differs from the standard GA with a hierarchy structure in that each chromosome consists of multilevel genes. Each chromosome consists of two types of genes, i. e. control genes and connection genes. The control genes in the form of bits, are the genes for layers and neurons for activation. The connection genes, a real value representation, are the genes for connection weightings and neuron bias.

With such a specific treatment, a structural chromosome incorporates both active and inactive genes. It should be noted that the inactive genes remain in the chromosome structure and can be carried forward for further generations. Such an inherent genetic variation in the chromosome avoids any trapping at local optima, which has the potential to cause premature convergence. Thus it maintains a balance between exploiting its accumulated knowledge and exploring the new areas of the search space. This structure also allows larger genetic variations in chromosome while maintaining high viability by permitting multiple simultaneous genetic changes. As a result, a single change in high level genes will cause multiple changes (activation or deactivation in the whole level) in lower level genes. In the case of the traditional GA, this is only possible when a sequence of many random changes takes place. Hence the computational power is greatly improved.

The fitness function used in this work combines the information the error objective and also the information about the number of nodes as a second objective. This is shown in the following equation.

$$f(z) = \left(\frac{1}{\alpha * \text{Ranking}(\text{ObjV1}) + \beta * \text{ObjV2}} \right) * 10. \quad (9)$$

The first objective is basically the average sum of squared of errors as calculated by the predicted outputs of the MNN compared with real values of the function. This is

given by the following equation.

$$f_1 = \frac{1}{N} \sum_{i=1}^N (Y_i - y_i). \quad (10)$$

The parameters of the genetic algorithm for this case are as follows:

- Type of crossover operator: Two-point crossover
- Crossover rate: 0.8
- Type of mutation operator: Binary mutation
- Mutation rate: 0.05
- Population size per generation: 10
- Total number of generations: 100

The evolution of neural networks is used in the following section to optimize hybrid intelligent systems for time series prediction. In particular, in the neuro-genetic approach the evolution is used to optimize the neural network for prediction, and in the neuro-fuzzy-genetic approach the evolution is used to optimize the neuro-fuzzy system.

Experimental Results for Time Series Prediction

In this section, we illustrate the application of interval type-2 fuzzy logic to the problem of time series prediction.

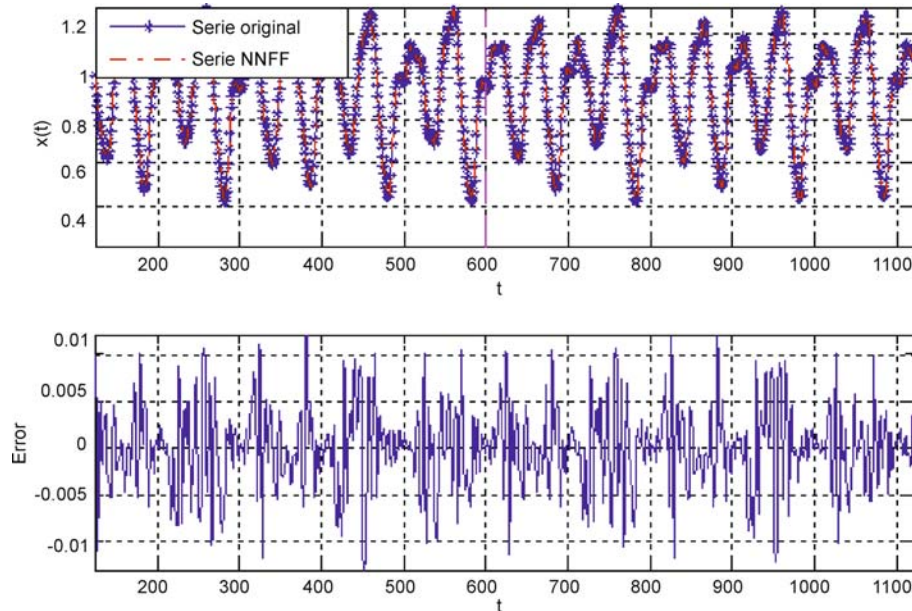
The Mackey-Glass time series is used to compare the results of interval type-2 fuzzy logic with the results of other intelligent methods. In particular, a comparison is made with type-1 fuzzy systems, neural networks, neuro-fuzzy systems and neuro-genetic and fuzzy-genetic approaches.

Prediction with a Neural Network

In this case, we did find a neural network model for this time series using 5 time delays and 6 periods. The architecture of the neural network used was 4–12–1 with 150 epochs, and 500 data points for training/500 data points for validation. The mean square error for prediction is 0.0043 (see Fig. 12).

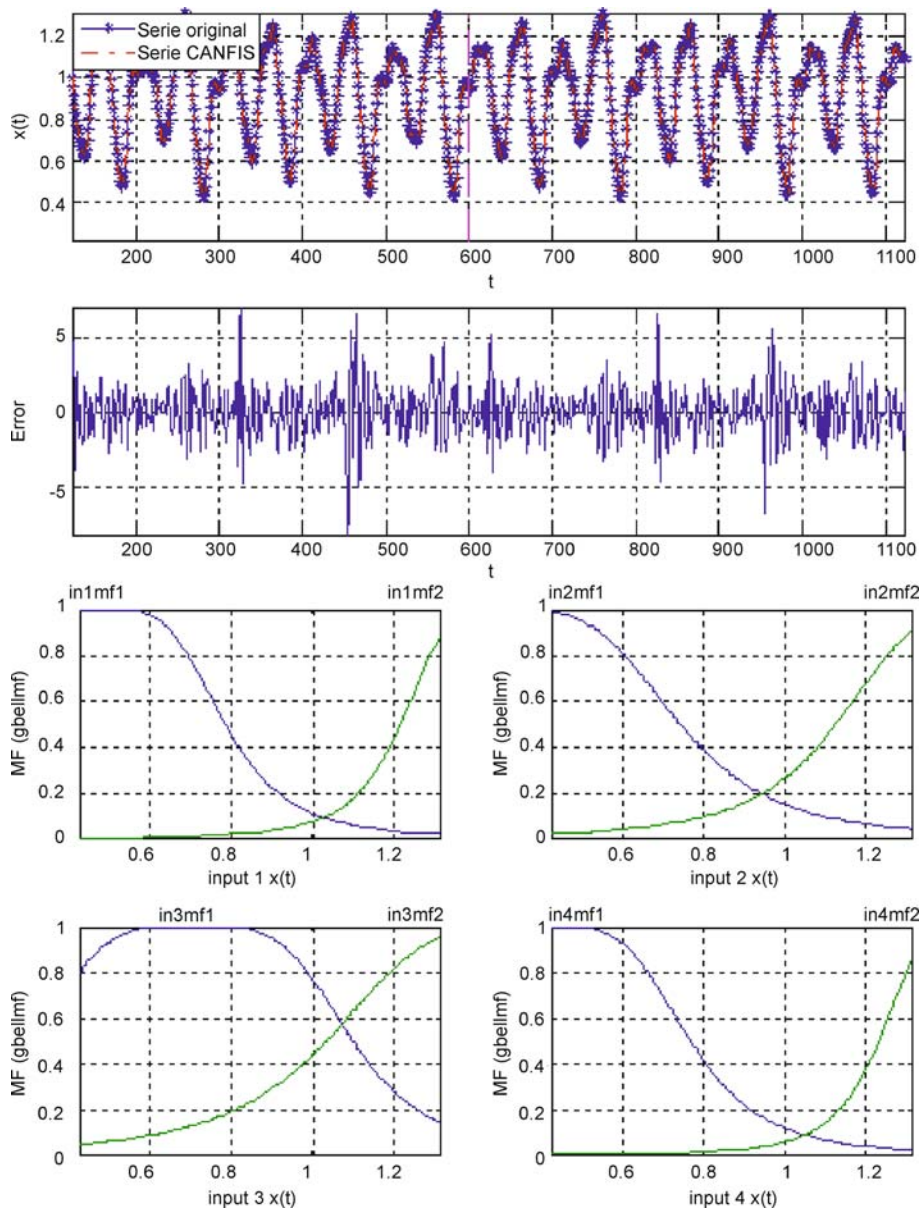
Prediction with an Adaptive Neuro-Fuzzy Inference System (ANFIS)

To find the ANFIS model an analysis was made of the data, and the decision was to use a 5 time delay with 6 periods. The time series was divided into 500 data points for training and 500 data points for validation. The ANFIS model was designed with 4 inputs (2 membership functions each) and 1 output with 16 linear functions, giving a total of 16 fuzzy rules. The training was of 50 epochs and the mean squared error of prediction is of 0.0016 (see Fig 13).



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 12

Forecasting the Mackey-Glass time series with a neural network. On top it is shown the comparison of the prediction with the neural network (red) and the original time series (blue). Below it is shown the forecasting error



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 13

Forecasting the Mackey-Glass time series with ANFIS. The figure on top shows the comparison of the predicted values and the original time series. The following figures show the error plot and the membership functions obtained with ANFIS

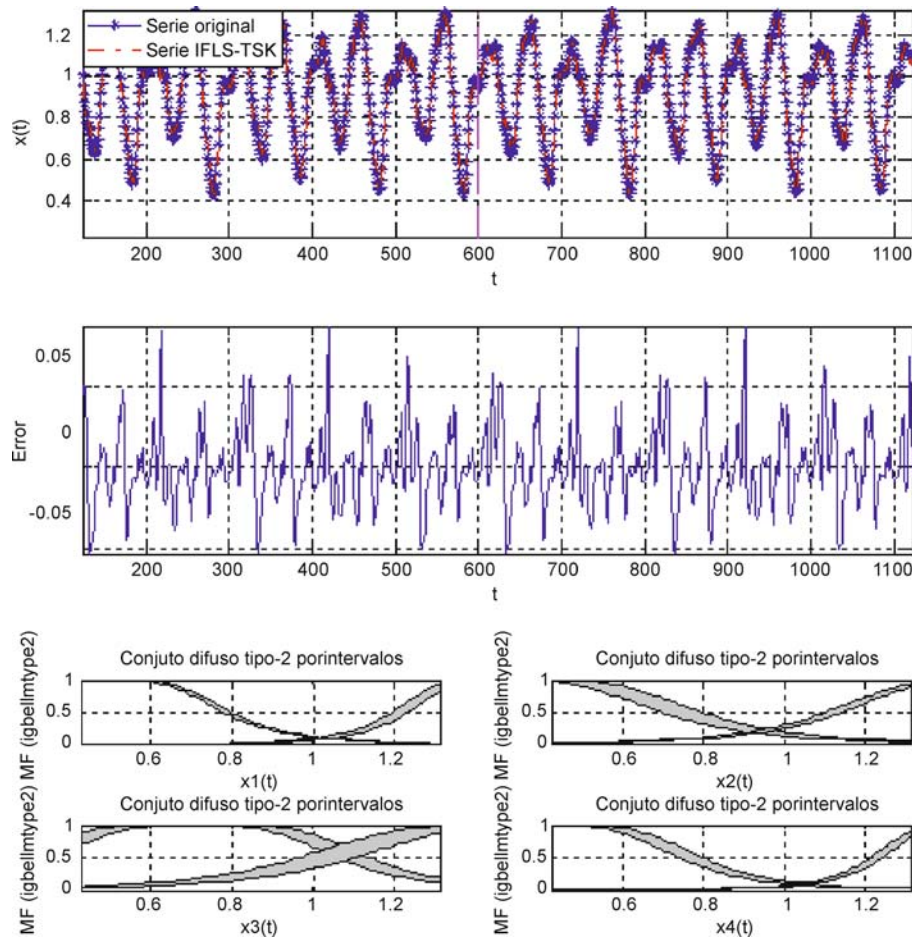
Prediction with an Interval Type-2 TSK Fuzzy Model

To find the type-2 fuzzy model an analysis was made of the data, and the decision was to use a 5 time delay with 6 periods. The time series was divided into 500 data points for training and 500 data points for validation. The model was designed with 4 inputs (with two interval type-2 (igbellmtype2) membership functions each) and one output

with 16 interval linear functions, giving a total of 16 fuzzy rules. The training was of 50 epochs and the mean squared error of prediction is of 0.00023 (see Fig. 14).

Prediction with a Neuro-Genetic Model

In this case, we use a genetic algorithm for training a neural network model for this time series using 5 time de-



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 14

Forecasting the Mackey-Glass time series with an interval type-2 TSK fuzzy model. The figure on *top* shows the comparison of the predicted values and the original time series. The *following* figures show the error plot and the interval type-2 membership functions obtained

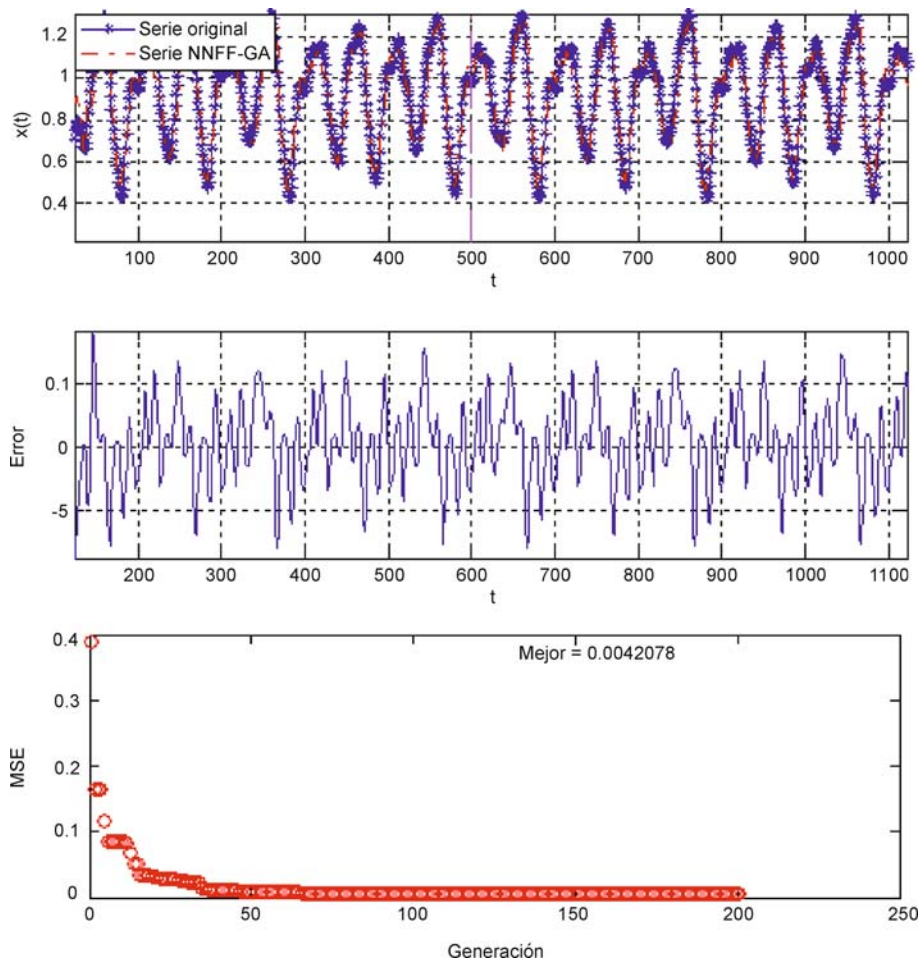
lays and 6 periods. The architecture of the neural network used was 4–12–1 with 200 generations, and 500 data points for training/500 data points for validation. The mean square error for prediction is 0.00064 (see Fig. 15). The genetic parameters are: population size = 30, crossover ratio = 0.75, and mutation ratio = 0.01.

Prediction with Type-1 Fuzzy Models Optimized with Genetic Algorithms

In this section, we show prediction results of type-1 fuzzy models that were optimized using genetic algorithms. In all cases, we have 4 inputs (with 2 membership functions) and 1 output with 16 functions. The parameters of the ge-

netic algorithms are the same as in the previous section. We show in Fig. 16 the results of a TSK model (mean squared error of 0.00647) and in Fig. 17 the results of a Mamdani model (mean squared error of 0.00692).

The Mackey-Glass time series shows chaotic behavior and for this reason has been chosen many times as a benchmark problem for prediction methods. We show in Table 6 a summary of the results using the methods mentioned previously. Based on mean squared error (RMSE) of forecasting, we can conclude that the interval type-2 fuzzy model (IT2 FLS on Table 6) is the best one to predict future values of this time series. Also, based on the training required, we can conclude that the interval type-2 fuzzy model is the best one because it only requires one epoch.



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 15

Forecasting the Mackey-Glass time series with a neuro-genetic model. The figure on *top* shows the comparison of the predicted values and the original time series. The figure on the *middle* shows the error plot. The figure on the *bottom* shows the evolution of the mean squared error as the generations increase up to 250

Hybrid Soft Computing Models for Systems Modeling and Control, Table 6

Summary of results for the mackey-glass time series

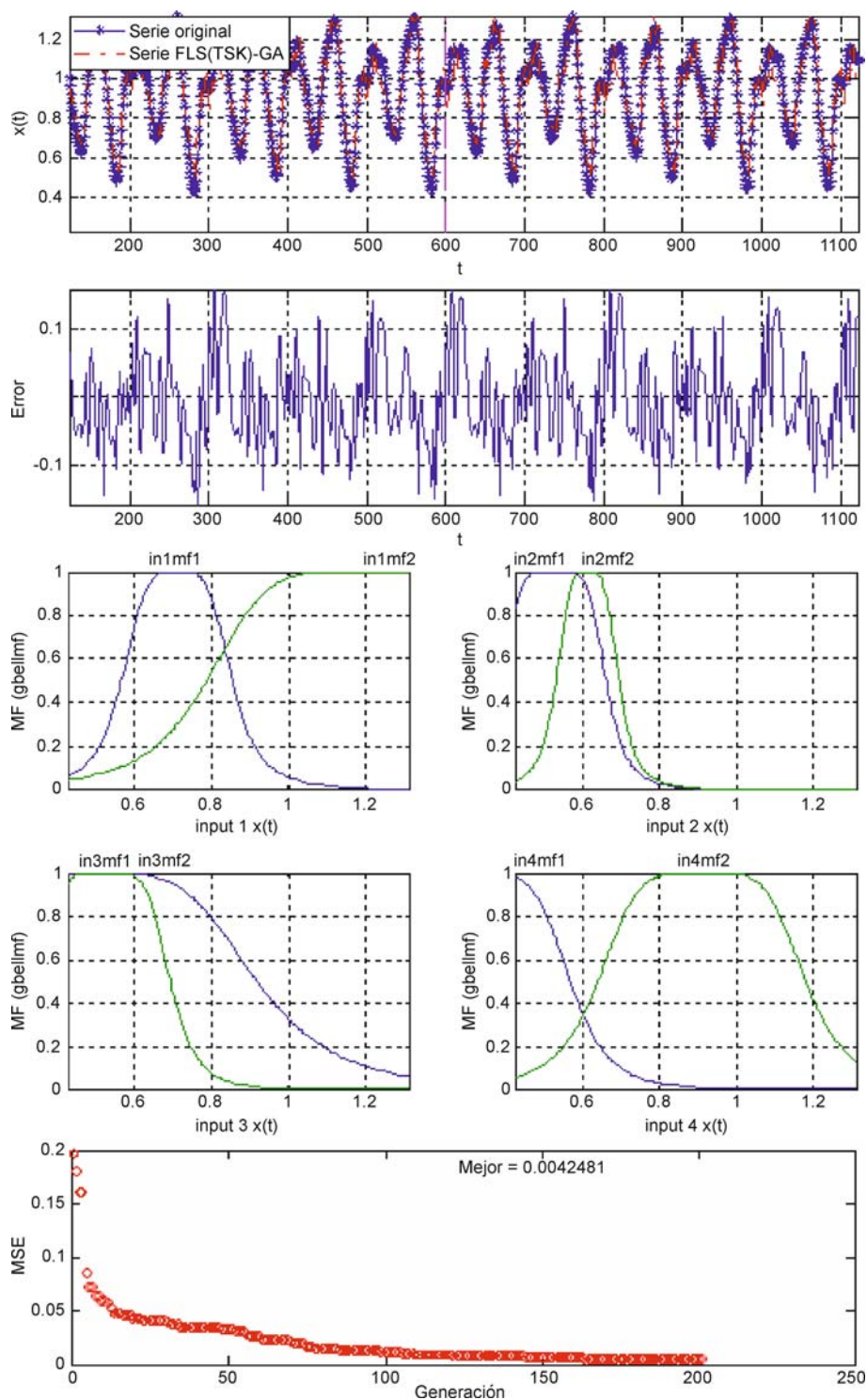
Method	RMSE	Data Training/Checking	Epochs or Generations
NNFF (Fig. 12)	0.00430	500/500	150
CANFIS (Fig. 13)	0.00160	500/500	50
IT2FLS(TSK) (Fig. 14)	0.00023	500/500	1
NNFF-GA (Fig. 15)	0.00064	500/500	150
FLS(TSK)-GA (Fig. 16)	0.00647	500/500	200
FLS(MAM)-GA (Fig. 17)	0.00693	500/500	200

Finally, from Table 6 we can say that the use of evolution helps in the design of hybrid intelligent systems. The

results of the hybrid intelligent systems shown in the last three rows (of Table 6) are very good, and this is a consequence of using genetic algorithm for optimizing the neural networks or the fuzzy systems.

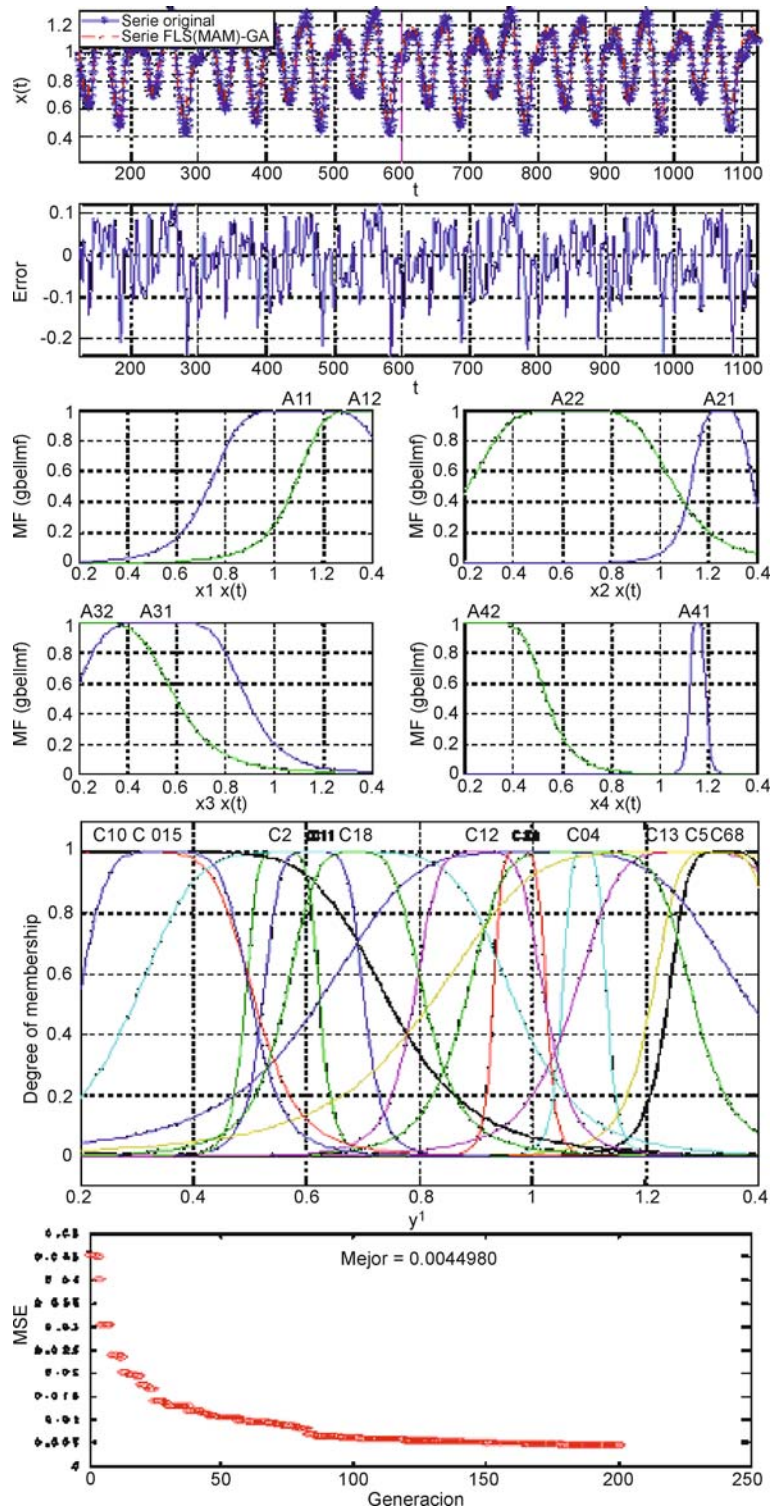
Conclusions

We consider in this paper the case of automatic anesthesia control in human patients for testing the optimized fuzzy controller. We did have, as a reference, the best fuzzy controller that was developed for the automatic anesthesia control, and we consider the optimization of this controller using the HGA approach. After applying the genetic algorithm the number of fuzzy rules was reduced from 12 to 9 with a similar performance of the fuzzy controller.



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 16

Forecasting the Mackey-Glass time series with a TSK fuzzy model optimized using a genetic algorithm



Hybrid Soft Computing Models for Systems Modeling and Control, Figure 17

Forecasting the Mackey-Glass time series with a Mamdani fuzzy model optimized using a genetic algorithm

Of course, the parameters of the membership functions were also tuned by the genetic algorithm. We did compare the simulation results of the optimized fuzzy controllers obtained with the HGA against the best fuzzy controller that was obtained previously with expert knowledge, and control is achieved in a similar fashion. Since simulation results are similar, and the number of fuzzy rules was reduced, we can conclude that the HGA approach is a good alternative for designing fuzzy systems. We also consider the case of controlling the bar and ball system, and the genetic approach was able to optimize the number of rules from 12 to 6. In conclusion, the HGA approach is a good alternative in optimizing fuzzy controllers. Future work will include testing the proposed approach with the optimization of other fuzzy controllers. We also described the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction.

Future Directions

The evolutionary approach is a good alternative in optimizing fuzzy controllers. Future work will include testing the proposed approach with the optimization of other fuzzy controllers and comparison with results of existing approaches. We also described the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction. In this case, future work will include applying the hybrid approach to other problems of time series prediction, and compare the results with existing approaches.

Bibliography

- Baruch S, Garrido R (2005) A direct adaptive neural control scheme with integral terms. *Int J Intell Syst* 20(2):213-224
- Castillo O, Melin P (2001) Soft computing for control of non-linear dynamical systems. Springer, Heidelberg
- Castillo O, Melin P (2003) Soft computing and fractal theory for intelligent manufacturing. Springer, Heidelberg
- Castillo O, Huesca G, Valdez F (2004) Evolutionary computing for fuzzy system optimization in intelligent control. *Proceedings of IC-AI'04, Las Vegas, vol 1*. CSREA Press, Las Vegas, pp 98-104
- Davis L (1991) Handbook of genetic algorithms. Van Nostrand Reinhold, New York
- Goldberg D (ed) (1989) Genetic algorithms in search, optimization and machine learning. Addison Wesley, Reading
- Holland J (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor
- Homaifar, McCormick E (1995) Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. *IEEE Trans Fuzzy Syst* 3:129-139
- Jang JSR, Sun CT (1995) Neurofuzzy fuzzy modeling and control. *Proc IEEE* 83:378-406
- Jang JSR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing, a computational approach to learning and machine intelligence. Prentice Hall, Upper Saddle River
- Karr CL, Gentry EJ (1993) Fuzzy control of pH using genetic algorithms. *IEEE Trans Fuzzy Systems* 1:46-53
- Langari R (1990) A framework for analysis and synthesis of fuzzy linguistic control systems. Ph D thesis, University of California, Berkeley
- Lozano (2004) Optimización de un sistema de control difuso por medio de algoritmos genéticos jerárquicos. Thesis, Dept of Computer Science, Tijuana Institute of Technology
- Man KF, Tang KS, Kwong S (1999) Genetic algorithms: Concepts and designs. Springer, London
- Melin P, Castillo O (2002) Modelling, simulation and control of non-linear dynamical systems. Taylor and Francis, London
- Nawa NE, Furuhashi T (1999) Fuzzy system parameters discovery by bacterial evolutionary algorithm. *IEEE Trans Fuzzy Syst* 7:608-616
- Procyk TJ, Mamdani EM (1979) A linguistic self-organizing process controller. *Automatica* 15(1):15-30
- Salmeri M, Re M, Petrongari E, Cardarilli GC (1999) A novel bacterial algorithm to extract the rule base from a training set. Technical Report, Dept. of Electronic Engineering, University of Rome
- Sepulveda R, Castillo O, Melin P, Montiel O, Rodriguez-Diaz A (2005) Handling uncertainty in controllers using type-2 fuzzy logic. *J Intell Syst* 14:237-262
- Tang KS, Man KF, Liu ZF, Kwong S (1998) Minimal fuzzy memberships and rules using hierarchical genetic algorithms. *IEEE Trans Ind Electron* 45(1):142-150
- Vachtsevanos G, Farinwata S (1996) Fuzzy logic control. In: Patyra MJ, Mlynek DM (eds) A systematic design and performance assessment methodology. Wiley, New York
- Valdes M, Gomez-Skarmeta AF, Botia JA (2005) Toward a framework for the specification of hybrid fuzzy modeling. *Int J Intell Syst* 20(2):225-252
- Valdez F, Castillo O (2004) Comparative study of evolutionary computing methods for fuzzy system optimization in intelligent control. *Proc IS-IC'04*. Tijuana, México, pp 1-5
- Yen J, Langari R (1999) Fuzzy logic: intelligence, control and information. Prentice Hall, Upper Saddle River
- Yoshikawa T, Furuhashi T, Uchikawa Y (1996) Emergence of effective fuzzy rules for controlling mobile robots using DNA coding method. *Proc ICEC'96*. Nagoya, Japan, pp 581-586
- Zadeh L (1965) Fuzzy sets. *J Inf Control* 8:338-353
- Zadeh L (1987) Fuzzy sets and applications. In: Yager RR, Ovchinnikov S, Tong RM, Nguyen HT (eds) Selected papers. Wiley, New York

Hyperbolic Conservation Laws

ALBERTO BRESSAN

Department of Mathematics, Penn State University,
University Park, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Examples of Conservation Laws

Shocks and Weak Solutions

Hyperbolic Systems in One Space Dimension

Entropy Admissibility Conditions

The Riemann Problem

Global Solutions

Hyperbolic Systems in Several Space Dimensions

Numerical Methods

Future Directions

Bibliography

Glossary

Conservation law Several physical laws state that certain basic quantities such as mass, energy, or electric charge, are globally conserved. A conservation law is a mathematical equation describing how the density of a conserved quantity varies in time. It is formulated as a partial differential equation having divergence form.

Flux function The flux of a conserved quantity is a vector field, describing how much of the given quantity moves across any surface, at a given time.

Shock Solutions to conservation laws often develop shocks, i.e. surfaces across which the basic physical fields are discontinuous. Knowing the two limiting values of a field on opposite sides of a shock, one can determine the speed of propagation of a shock in terms of the Rankine–Hugoniot equations.

Entropy An entropy is an additional quantity which is globally conserved for every smooth solution to a system of conservation laws. In general, however, entropies are not conserved by solutions containing shocks. Imposing that certain entropies increase (or decrease) in correspondence to a shock, one can determine a unique physically admissible solution to the mathematical equations.

Definition of the Subject

According to some fundamental laws of continuum physics, certain basic quantities such as mass, momen-

tum, energy, electric charge..., are globally conserved. As time progresses, the evolution of these quantities can be described by a particular type of mathematical equations, called conservation laws.

Gas dynamics, magneto-hydrodynamics, electromagnetism, motion of elastic materials, car traffic on a highway, flow in oil reservoirs, can all be modeled in terms of conservation laws. Understanding, predicting and controlling these various phenomena is the eventual goal of the mathematical theory of hyperbolic conservation laws.

Introduction

Let $u = u(x, t)$ denote the density of a physical quantity, say, the density of mass. Here t denotes time, while $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ is a three-dimensional space variable. A *conservation law* is a partial differential equation of the form

$$\frac{\partial}{\partial t} u + \operatorname{div} \mathbf{f} = 0. \quad (1)$$

which describes how the density u changes in time. The vector field $\mathbf{f} = (f_1, f_2, f_3)$ is called the *flux* of the conserved quantity. We recall that the divergence of \mathbf{f} is

$$\operatorname{div} \mathbf{f} = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \frac{\partial f_3}{\partial x_3}.$$

To appreciate the meaning of the above Eq. (1), consider a fixed region $\Omega \subset \mathbb{R}^3$ of the space. The total amount of mass contained inside Ω at time t is computed as

$$\int_{\Omega} u(x, t) dx.$$

This integral may well change in time. Using the conservation law (1) and then the divergence theorem, one obtains

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} u(x, t) dx &= \int_{\Omega} \frac{\partial}{\partial t} u(x, t) dx = - \int_{\Omega} \operatorname{div} \mathbf{f} dx \\ &= - \int_{\Sigma} \mathbf{f} \cdot \mathbf{n} d\Sigma. \end{aligned}$$

Here Σ denotes the boundary of Ω , while the integrand $\mathbf{f} \cdot \mathbf{n}$ denotes the inner product of the vector \mathbf{f} with the unit outer normal \mathbf{n} to the surface Σ . According to the above identities, no mass is created or destroyed. The total amount of mass contained inside the region Ω changes in time only because some of the mass flows in or out across the boundary Σ .

Assuming that the flux \mathbf{f} can be expressed as a function of the density u alone, one obtains a closed equation. If the initial density \tilde{u} at time $t = 0$ is known, then the values

of the function $u = u(x, t)$ at all future times $t > 0$ can be found by solving the initial-value problem

$$u_t + \operatorname{div} \mathbf{f}(u) = 0 \quad u(0, x) = \tilde{u}(x).$$

More generally, a *system of balance laws* is a set of partial differential equations of the form

$$\begin{cases} \frac{\partial}{\partial t} u_1 + \operatorname{div} \mathbf{f}_1(u_1, \dots, u_n) = \phi_1, \\ \dots \\ \frac{\partial}{\partial t} u_n + \operatorname{div} \mathbf{f}_n(u_1, \dots, u_n) = \phi_n. \end{cases} \quad (2)$$

Here u_1, \dots, u_n are the conserved quantities, $\mathbf{f}_1, \dots, \mathbf{f}_n$ are the corresponding fluxes, while the functions $\phi_i = \phi_i(t, x, u_1, \dots, u_n)$ represent the source terms. In the case where all ϕ_i vanish identically, we refer to (2) as a *system of conservation laws*.

Systems of this type express the fundamental balance equations of continuum physics, when small dissipation effects are neglected. A basic example is provided by the equations of non-viscous gases, accounting for the conservation of mass, momentum and energy. This subject is thus very classical, having a long tradition which can be traced back to Euler [21] and includes contributions by Stokes, Riemann, Weyl and von Neumann, among several others.

In spite of continuing efforts, the mathematical theory of conservation laws is still largely incomplete. Most of the literature has been concerned with two main cases: (i) a single conservation law in several space dimensions, and (ii) systems of conservation laws in one space dimension. For systems of conservation laws in several space dimensions, not even the global-in-time existence of solutions is presently known, in any significant degree of generality. Several mathematical studies are focused on particular solutions, such as traveling waves, multi-dimensional Riemann problems, shock reflection past a wedge, etc. . .

Toward a rigorous mathematical analysis of solutions, the main difficulty that one encounters is the lack of regularity. Due to the strong nonlinearity of the equations and the absence of dissipation terms with regularizing effect, solutions which are initially smooth may become discontinuous within finite time. In the presence of discontinuities, most of the classical tools of differential calculus do not apply. Moreover, the Eqs. (2) must be suitably reinterpreted, since a discontinuous function does not admit derivatives in a classical sense.

Topics which have been more extensively investigated in the mathematical literature are the following:

- Existence and uniqueness of solutions to the initial-value problem. Continuous dependence of the solutions on the initial data [8,10,24,29,36,49,57].
- Admissibility conditions for solutions with shocks, characterizing the physically relevant ones [23,38,39,46].
- Stability of special solutions, such as traveling waves, w.r.t. small perturbations [34,47,49,50,62].
- Relations between the solutions of a hyperbolic system of conservation laws and the solutions of various approximating systems, modeling more complex physical phenomena. In particular: vanishing viscosity approximations [6,20,28], relaxations [5,33,48], kinetic models [44,53].
- Numerical algorithms for the efficient computation of solutions [32,41,41,56,58].

Some of these aspects of conservation laws will be outlined in the following sections.

Examples of Conservation Laws

We review here some of the most common examples of conservation laws. Throughout the sequel, subscripts such as u_t, f_x will denote partial derivatives.

Example 1 (Traffic flow) Let $u(x, t)$ be the density of cars on a highway, at the point x at time t . This can be measured as the number of cars per kilometer (see Fig. 1). In first approximation, following [43] we shall assume that u is continuous and that the speed s of the cars depends only on their density, say $s = s(u)$. Given any two points a, b on the highway, the number of cars between a and b therefore varies according to

$$\begin{aligned} \int_a^b u_t(x, t) dx &= \frac{d}{dt} \int_a^b u(x, t) dx \\ &= [\text{inflow at } x = a] - [\text{outflow at } x = b] \\ &= s(u(a, t)) \cdot u(a, t) - s(u(b, t)) \cdot u(b, t) \\ &= - \int_a^b [s(u)u]_x dx. \end{aligned}$$

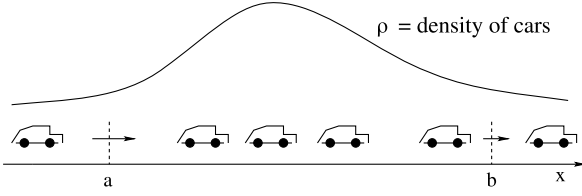
Since the above equalities hold for all a, b , one obtains the conservation law in one space dimension

$$u_t + [s(u)u]_x = 0, \quad (3)$$

where u is the conserved quantity and $f(u) = s(u)u$ is the flux function. Based on experimental data, an appropriate flux function has the form

$$f(u) = a_1 \left(\ln \frac{a_2}{u} \right) u \quad (0 < u \leq a_2),$$

for suitable constants a_1, a_2 .



Hyperbolic Conservation Laws, Figure 1

Modelling the density of cars by a conservation law

Example 2 (Gas dynamics) The Euler equations for a compressible, non-viscous gas in Eulerian coordinates take the form

$$\begin{cases} \frac{\partial}{\partial t} \rho + \operatorname{div}(\rho v) = 0 & \text{(conservation of mass),} \\ \frac{\partial}{\partial t}(\rho v_i) + \operatorname{div}(\rho v_i v) + \frac{\partial}{\partial x_i} p = 0 & i = 1, 2, 3, \\ \frac{\partial}{\partial t} E + \operatorname{div}((E + p)v) = 0 & \text{(conservation of energy).} \end{cases}$$

Here ρ is the mass density, $v = (v_1, v_2, v_3)$ is the velocity vector, E is the energy density, and p the pressure. In turn, the energy can be represented as a sum

$$E = \rho \frac{|v|^2}{2} + \rho e,$$

where the first term accounts for the kinetic energy while e is the internal energy density (related to the temperature). The system is closed by an additional equation $p = p(\rho, e)$, called the equation of state, depending on the particular gas under consideration [18]. Notice that here we are neglecting small viscous forces, as well as heat conductivity. Calling $\{e_1, e_2, e_3\}$ the standard basis of unit vectors in \mathbb{R}^3 , one has $\partial p / \partial x_i = \operatorname{div}(p e_i)$. Hence all of the above equations can be written in the standard divergence form (1).

Example 3 (Isentropic gas dynamics in Lagrangian variables) Consider a gas in a tube. Particles of the gas will be labeled by a one-dimensional variable y determined by their position in a reference configuration with constant unit density. Using this Lagrangian coordinate, we denote by $u(y, t)$ the velocity of the particle y at time t , and by $v(y, t) = \rho^{-1}(y, t)$ its specific volume.

The so-called p -system of isentropic gas dynamics [57] consists of the two conservation laws

$$v_t - u_x = 0, \quad u_t + p_x = 0. \quad (4)$$

The system is closed by an equation of state $p = p(v)$ expressing the pressure as a function of the specific volume.

A typical choice here is $p(v) = kv^{-\gamma}$, with $\gamma \in [1, 3]$. In particular $\gamma \approx 1.4$ for air.

In general, p is a decreasing function of v . Near a constant state v_0 , one can approximate p by a linear function, say $p(v) \approx p(v_0) - c^2(v - v_0)$. Here $c^2 = -p'(v_0)$. In this case the Eq. (4) reduces to the familiar wave equation

$$v_{tt} - c^2 v_{xx} = 0.$$

Shocks and Weak Solutions

A single conservation law in one space dimension is a first order partial differential equation of the form

$$u_t + f(u)_x = 0. \quad (5)$$

Here u is the *conserved quantity* while f is the *flux*. As long as the function u is continuously differentiable, using the chain rule the equation can be rewritten as

$$u_t + f'(u)u_x = 0, \quad (6)$$

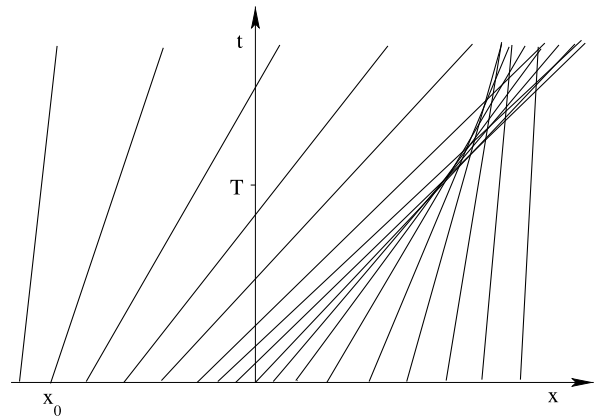
According to (6), in the x - t -plane, the directional derivative of the function u in the direction of the vector $\mathbf{v} = (f'(u), 1)$ vanishes.

At time $t = 0$, let an initial condition $u(x, 0) = \bar{u}(x)$ be given. As long as the solution u remains smooth, it can be uniquely determined by the so-called method of characteristics. For every point x_0 , consider the straight line (see Fig. 2)

$$x = x_0 + f'(\bar{u}(x_0))t.$$

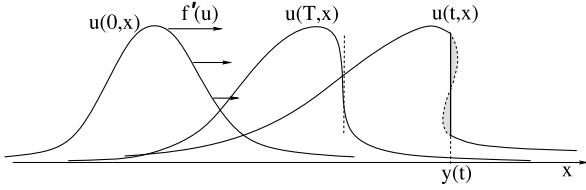
On this line, by (6) the value of u is constant. Hence

$$u(x_0 + f'(\bar{u}(x_0))t, t) = \bar{u}(x_0)$$



Hyperbolic Conservation Laws, Figure 2

Solving a conservation law by the method of characteristics. The function $u = u(x, t)$ is constant along each characteristic line



Hyperbolic Conservation Laws, Figure 3
Shock formation

for all $t \geq 0$. This allows to construct a solution up to the first time T where two or more characteristic lines meet. Beyond this time, the solution becomes discontinuous. Figure 3 shows the graph of a typical solution at three different times. Points on the graph of u move horizontally with speed $f'(u)$. If this speed is not constant, the shape of the graph will change in time. In particular, there will be an instant T at which one of the tangent lines becomes vertical. For $t > T$, the solution $u(\cdot, t)$ contains a shock. The position $y(t)$ of this shock can be determined by imposing that the total area of the region below the graph of u remains constant in time.

In order to give a meaning to the conservation law (6) when $u = u(x, t)$ is discontinuous, one can multiply both sides of the equation by a test function φ and integrate by parts. Assuming that φ is continuously differentiable and vanishes outside a bounded set, one formally obtains

$$\iint \{u\varphi_t + f(u)\varphi_x\} dx dt = 0. \quad (7)$$

Since the left hand side of the above equation does not involve partial derivatives of u , it remains meaningful for a discontinuous function u . A locally integrable function u is defined to be a *weak solution* of the conservation law (6) if the integral identity (7) holds true for every test function φ , continuously differentiable and vanishing outside a bounded set.

Hyperbolic Systems in One Space Dimension

A system of n conservation laws in one space dimension can be written as

$$\begin{cases} \frac{\partial}{\partial t} u_1 + \frac{\partial}{\partial x} [f_1(u_1, \dots, u_n)] = 0, \\ \dots \\ \frac{\partial}{\partial t} u_n + \frac{\partial}{\partial x} [f_n(u_1, \dots, u_n)] = 0. \end{cases} \quad (8)$$

For convenience, this can still be written in the form (5), but keeping in mind that now $u = (u_1, \dots, u_n) \in \mathbb{R}^n$ is a vector and that $f = (f_1, \dots, f_n)$ is a vector-valued function. As in the case of a single equation, a vector function

$u = (u_1, \dots, u_n)$ is called a *weak solution* to the system of conservation laws (8) if the integral identity (7) holds true, for every continuously differentiable test function φ vanishing outside a bounded set.

Consider the $n \times n$ Jacobian matrix of partial derivatives f at the point u :

$$A(u) \doteq Df(u) = \begin{pmatrix} \partial f_1 / \partial u_1 & \cdots & \partial f_1 / \partial u_n \\ \vdots & \ddots & \vdots \\ \partial f_n / \partial u_1 & \cdots & \partial f_n / \partial u_n \end{pmatrix}.$$

Using the chain rule, the system (8) can be written in the quasilinear form

$$u_t + A(u)u_x = 0. \quad (9)$$

We say that the above system is *strictly hyperbolic* if every matrix $A(u)$ has n real, distinct eigenvalues, say $\lambda_1(u) < \cdots < \lambda_n(u)$. In this case, one can find a basis of right eigenvectors of $A(u)$, denoted by $r_1(u), \dots, r_n(u)$, such that, for $i = 1, \dots, n$,

$$A(u)r_i(u) = \lambda_i(u)r_i(u), \quad |r_i(u)| = 1.$$

Example 4 Assume that the flux function is linear: $f(u) = Au$ for some constant matrix $A \in \mathbb{R}^{n \times n}$. If $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ are the eigenvalues of A and r_1, \dots, r_n are the corresponding eigenvectors, then any vector function of the form

$$u(x, t) = \sum_{i=1}^n g_i(x - \lambda_i t) r_i$$

provides a weak solution to the system (8). Here it is enough to assume that the functions g_i are locally integrable, not necessarily continuous.

Example 3 (continued) For the system (4) describing isentropic gas dynamics, the Jacobian matrix of partial derivatives of the flux is

$$Df = \begin{pmatrix} -\partial u / \partial v & -\partial u / \partial u \\ \partial p / \partial v & \partial p / \partial u \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ p'(v) & 0 \end{pmatrix}.$$

Assuming that $p'(v) < 0$, this matrix has the two real distinct eigenvalues $\pm \sqrt{-p'(v)}$. Therefore, the system is strictly hyperbolic.

Entropy Admissibility Conditions

Given two states $u^-, u^+ \in \mathbb{R}^n$ and a speed λ , consider the piecewise constant function defined as

$$u(x, t) = \begin{cases} u^- & \text{if } x < \lambda t, \\ u^+ & \text{if } x > \lambda t. \end{cases}$$

Then one can show that the discontinuous function u is a weak solution of the hyperbolic system (8) if and only if it satisfies the Rankine–Hugoniot equations

$$\lambda(u^+ - u^-) = f(u^+) - f(u^-). \quad (10)$$

More generally, consider a function $u = u(x, t)$ which is piecewise smooth in the x - t -plane. Assume that these discontinuities are located along finitely many curves $x = \gamma_\alpha(t)$, $\alpha = 1, \dots, N$, and consider the left and right limits

$$u_\alpha^-(t) = \lim_{x \rightarrow \gamma_\alpha(t)^-} u(x, t), \quad u_\alpha^+(t) = \lim_{x \rightarrow \gamma_\alpha(t)^+} u(x, t).$$

Then u is a weak solution of the system of conservation laws if and only if it satisfies the quasilinear system (9) together with the Rankine–Hugoniot equations

$$\gamma'_\alpha(t)(u_\alpha^+(t) - u_\alpha^-(t)) = f(u_\alpha^+(t)) - f(u_\alpha^-(t))$$

along each shock curve. Here $\gamma'_\alpha = d\gamma_\alpha/dt$.

Given an initial condition $u(x, 0) = \bar{u}(x)$ containing jumps, however, it is well known that the Rankine–Hugoniot conditions do not determine a unique weak solution. Several “admissibility conditions” have thus been proposed in the literature, in order to single out a unique physically relevant solution. A basic criterion relies on the concept of entropy: A continuously differentiable function $\eta: \mathbb{R}^n \mapsto \mathbb{R}$ is called an *entropy* for the system of conservation laws (8), with *entropy flux* $q: \mathbb{R}^n \mapsto \mathbb{R}$ if

$$D\eta(u) \cdot Df(u) = Dq(u).$$

If $u = (u_1, \dots, u_n)$ is a smooth solution of (8), not only the quantities u_1, \dots, u_n are conserved, but the additional conservation law $\eta(u)_t + q(u)_x = 0$ holds as well. Indeed,

$$D\eta(u)u_t + Dq(u)u_x = D\eta(u)[-Df(u)u_x] + Dq(u)u_x = 0.$$

On the other hand, if the solution u is not smooth but contains shocks, the quantity $\eta = \eta(u)$ may no longer be conserved. The admissibility of a shock can now be characterized by requiring that certain entropies be increasing (or decreasing) in time. More precisely, a weak solution u of (8) is said to be *entropy-admissible* if the inequality

$$\eta(u)_t + q(u)_x \leq 0$$

holds in the sense of distributions, for every pair (η, q) , where η is a convex entropy and q is the corresponding flux. Calling u^-, u^+ the states to the left and right of the shock, and λ its speed, the above condition implies

$$\lambda[\eta(u^+) - \eta(u^-)] \geq q(u^+) - q(u^-).$$

Various alternative conditions have been studied in the literature, in order to characterize the physically admissible shocks. For these we refer to Lax [39], or Liu [46].

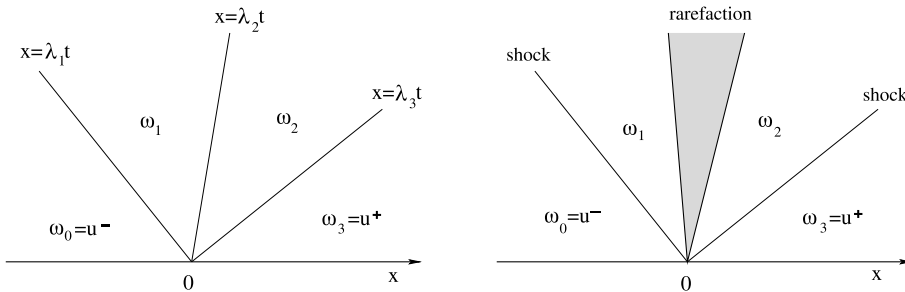
The Riemann Problem

Toward the construction of general solutions for the system of conservation laws (8), a basic building block is the so-called *Riemann problem* [55]. This amounts to choosing a piecewise constant initial data with a single jump at the origin:

$$u(x, 0) = \begin{cases} u^- & \text{if } x < 0, \\ u^+ & \text{if } x > 0. \end{cases} \quad (11)$$

In the special case where the system is linear, i. e. $f(u) = Au$, the solution is piecewise constant in the x - t -plane. It contains $n + 1$ constant states $u^- = \omega_0, \omega_1, \dots, \omega_n = u^+$ (see Fig. 4, left). Each jump $\omega_i - \omega_{i-1}$ is an eigenvector of the matrix A , and is located along the line $x = \lambda_i t$, whose speed equals the corresponding eigenvalue λ_i .

For nonlinear hyperbolic systems of n conservation laws, assuming that the amplitude $|u^+ - u^-|$ of the jump



Hyperbolic Conservation Laws, Figure 4

Solutions of a Riemann problem. Left: the linear case. Right: a nonlinear example

is sufficiently small, the general solution was constructed in a classical paper of Lax [38], under the additional hypothesis

- (H) For each $i = 1, \dots, n$, the i th field is either *genuinely nonlinear*, so that $D\lambda_i(u) \cdot r_i(u) > 0$ for all u , or *linearly degenerate*, with $D\lambda_i(u) \cdot r_i(u) = 0$ for all u .

The solution is self-similar: $u(x, t) = U(x/t)$. It still consists of $n + 1$ constant states $\omega_0 = u^-$, $\omega_1, \dots, \omega_n = u^+$ (see Fig. 4, right). Each couple of adjacent states ω_{i-1}, ω_i is separated either by a *shock* satisfying the Rankine Hugoniot equations, or else by a *centered rarefaction*. In this second case, the solution u varies continuously between ω_{i-1} and ω_i in a sector of the t - x -plane where the gradient u_x coincides with an i -eigenvector of the matrix $A(u)$.

Further extensions, removing the technical assumption (H), were obtained by T. P. Liu [46] and by S. Bianchini [3].

Global Solutions

Approximate solutions to a more general Cauchy problem can be constructed by patching together several solutions of Riemann problems. In the Glimm scheme [24], one works with a fixed grid in the x - t plane, with mesh sizes Δx , Δt . At time $t = 0$ the initial data is approximated by a piecewise constant function, with jumps at grid points (see Fig. 5, left). Solving the corresponding Riemann problems, a solution is constructed up to a time Δt sufficiently small so that waves generated by different Riemann problems do not interact. By a random sampling procedure, the solution $u(\Delta t, \cdot)$ is then approximated by a piecewise constant function having jumps only at grid points. Solving the new Riemann problems at every one of these points, one can prolong the solution to the next time interval $[\Delta t, 2\Delta t]$, etc. . .

An alternative technique for constructing approximate solutions is by wave-front tracking (Fig. 5, right). This method was introduced by Dafermos [17] in the scalar case and later developed by various authors [7,19,29]. It now

provides an efficient tool in the study of general $n \times n$ systems of conservation laws, both for theoretical and numerical purposes.

The initial data is here approximated with a piecewise constant function, and each Riemann problem is solved approximately, within the class of piecewise constant functions. In particular, if the exact solution contains a centered rarefaction, this must be approximated by a *rarefaction fan*, containing several small jumps. At the first time t_1 where two fronts interact, the new Riemann problem is again approximately solved by a piecewise constant function. The solution is then prolonged up to the second interaction time t_2 , where the new Riemann problem is solved, etc. . .

The main difference is that with the Glimm scheme one specifies a priori the nodal points where the Riemann problems are to be solved. On the other hand, in a solution constructed by wave-front tracking the locations of the jumps and of the interaction points depend on the solution itself. Moreover, no restarting procedure is needed.

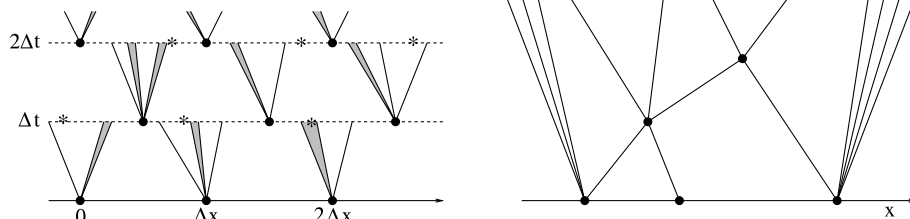
In the end, both algorithms produce a sequence of approximate solutions, whose total variation remains uniformly bounded. We recall here that the total variation of a function $u: \mathbb{R} \mapsto \mathbb{R}^n$ is defined as

$$\text{Tot.Var. } \{u\} \doteq \sup \sum_{i=1}^N |u(x_i) - u(x_{i-1})| ,$$

where the supremum is taken over all $N \geq 1$ and all N -tuples of points $x_0 < x_1 < \dots < x_N$. For functions of several variables, a more general definition can be found in [22]. Relying on a compactness argument, one can then show that these approximations converge to a weak solution to the system of conservation laws. Namely, one has:

Theorem 1 *Let the system of conservation laws (8) be strictly hyperbolic. Then, for every initial data \bar{u} with sufficiently small total variation, the initial value problem*

$$u_t + f(u)_x = 0 \quad u(x, 0) = \bar{u}(x) \quad (12)$$



Hyperbolic Conservation Laws, Figure 5

Left: the Glimm scheme. Right: a front tracking approximation

has a unique entropy-admissible weak solution, defined for all times $t \geq 0$.

The existence part was first proved in the famous paper of Glimm [24], under the additional hypothesis (H), later removed by Liu [46]. The uniqueness of the solution was proved more recently, in a series of papers by the present author and collaborators, assuming that all shocks satisfy suitable admissibility conditions [8,10]. All proofs are based on careful analysis of solutions of the Riemann problem and on the use of a quadratic interaction functional to control the formation of new waves. These techniques also provided the basis for further investigations of Glimm and Lax [25] and Liu [45] on the asymptotic behavior of weak solutions as $t \rightarrow \infty$.

It is also interesting to compare solutions with different initial data. In this direction, we observe that a function of two variables $u(x, t)$ can be regarded as a map $t \mapsto u(\cdot, t)$ from a time interval $[0, T]$ into a space $L^1(\mathbb{R})$ of integrable functions. Always assuming that the total variation remains small, the distance between two solutions u, v at any time $t > 0$ can be estimated as

$$\|u(t) - v(t)\|_{L^1} \leq L \|u(0) - v(0)\|_{L^1},$$

where L is a constant independent of time.

Estimates on the rate of convergence of Glimm approximations to the unique exact solutions are available. For every fixed time $T \geq 0$, letting the grid size $\Delta x, \Delta t$ tend to zero keeping the ratio $\Delta t/\Delta x$ constant, one has the error estimate [11]

$$\lim_{\Delta x \rightarrow 0} \frac{\|u^{\text{Glimm}}(T, \cdot) - u^{\text{exact}}(T, \cdot)\|_{L^1}}{\sqrt{\Delta x} \cdot |\ln \Delta x|} = 0.$$

An alternative approximation procedure involves the addition of a small viscosity. For $\varepsilon > 0$ small, one considers the viscous initial value problem

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{xx}^\varepsilon, \quad u^\varepsilon(x, 0) = \bar{u}(x). \quad (13)$$

For initial data \bar{u} with small total variation, the analysis in [6] has shown that the solutions u^ε have small total variation for all times $t > 0$, and converge to the unique weak solution of (12) as $\varepsilon \rightarrow 0$.

Hyperbolic Systems in Several Space Dimensions

In several space dimensions there is still no comprehensive theory for systems of conservation laws. Much of the literature has been concerned with three main topics: (i) Global solutions to a single conservation law. (ii) Smooth solutions to a hyperbolic system, locally in time. (iii) Particular solutions to initial or initial-boundary value problems.

Scalar Conservation Laws

The single conservation law on \mathbb{R}^m

$$u_t + \operatorname{div} f(u) = 0$$

has been extensively studied. The fundamental works of Volpert [59] and Kruzhkov [36] have established the global existence of a unique, entropy-admissible solution to the initial value problem, for any initial data $u(x, 0) = \bar{u}(x)$ measurable and globally bounded. This solution can be obtained as the unique limit of vanishing viscosity approximations, solving

$$u_t^\varepsilon + \operatorname{div} f(u^\varepsilon) = \varepsilon \Delta u^\varepsilon, \quad u^\varepsilon(x, 0) = \bar{u}.$$

As in the one-dimensional case, solutions which are initially smooth may develop shocks and become discontinuous in finite time. Given any two solutions u, v , the following key properties remain valid also in the presence of shocks:

- (i) If at the initial time $t = 0$ one has $u(x, 0) \leq v(x, 0)$ for all $x \in \mathbb{R}^m$, then $u(x, t) \leq v(x, t)$ for all x and all $t \geq 0$.
- (ii) The L^1 distance between any two solutions does not increase in time. Namely, for any $0 \leq s \leq t$ one has

$$\|u(t) - v(t)\|_{L^1(\mathbb{R}^m)} \leq \|u(s) - v(s)\|_{L^1(\mathbb{R}^m)}.$$

Alternative approaches to the analysis of scalar conservation laws were developed by Crandall [16] using nonlinear semigroup theory, and by Lions, Perthame and Tadmor [44] using a kinetic formulation. Regularity results can be found in [30].

Smooth Solutions to Hyperbolic Systems

Using the chain rule, one can rewrite the system of conservation laws (2) in the quasi-linear form

$$u_t + \sum_{\alpha=1}^m A^\alpha(u) u_{x_\alpha} = 0. \quad (14)$$

Various definitions of hyperbolicity can be found in the literature. Motivated by several examples from mathematical physics, the system (14) is said to be *symmetrizable hyperbolic* if there exists a positive definite symmetric matrix $S = S(u)$ such that all matrices $S^\alpha(u) = SA^\alpha$ are symmetric. In particular, this condition implies that each $n \times n$ matrix $A^\alpha(u)$ has real eigenvalues and admits a basis of linearly independent eigenvectors. As shown in [23], if a system of conservation laws admits a strictly convex entropy $\eta(u)$, such that the Hessian matrix of second derivatives

$D_u^2 \eta(u)$ is positive definite at every point u , then the system is symmetrizable.

A classical theorem states that, for a symmetrizable hyperbolic system with smooth initial data, the initial value problem has a unique smooth solution, locally in time. This solution can be prolonged in time up to the first time where the spatial gradient becomes unbounded at one or more points. In this general setting, however, it is not known whether the solution can be extended beyond this time of shock formation.

Special Solutions

In two space dimensions, one can study special solutions which are independent of time, so that $u(x_1, x_2, t) = U(x_1, x_2)$. In certain cases, one can regard one of the variables, say x_1 as a new time and derive a one-dimensional hyperbolic system of equations for U involving the remaining one-dimensional space variable x_2 .

Another important class of solutions relates to two-dimensional Riemann problems. Here the initial data, assigned on the x_1 - x_2 plane, is assumed to be constant along rays through the origin. Taking advantage of this self-similarity, the solution can be written in the form $u(x_1, x_2, t) = U(x_1/t, x_2/t)$. This again reduces the problem to an equation in two independent variables [35]. Even for the equation of gas dynamics, a complete solution to the Riemann problem is not available. Several particular cases are analyzed in [42,61].

Several other examples, in specific geometries have been analyzed. A famous problem is the reflection of a shock hitting a wedge-shaped rigid obstacle [15,51].

Numerical Methods

Generally speaking, there are three major classes of numerical methods suitable for partial differential equations: finite difference methods (FDM), finite volume methods (FVM) and finite element methods (FEM). For conservation laws, one also has semi-discrete methods, such as the method of lines, and conservative front tracking methods. The presence of shocks and the rich structure of shock interactions cause the main difficulties in numerical computations.

To illustrate the main idea, consider a uniform grid in the x - t -plane, with step sizes Δx and Δt . Consider the times $t_n = n\Delta t$ and let $I_i = [x_{i-1/2}, x_{i+1/2}]$ be a cell. We wish to compute an approximate value for the cell averages \bar{u}_i over I_i . Integrating the conservation law over the rectangle $I_i \times [t_n, t_{n+1}]$ and dividing by Δx , one obtains

$$\bar{u}_i^{n+1} = \bar{u}_i^n + \frac{\Delta t}{\Delta x} [F_{i-1/2} - F_{i+1/2}],$$

where $F_{i+1/2}$ is the average flux

$$F_{i+1/2} = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2})) dt.$$

FVM methods seek a suitable approximation to this average flux $F_{i+1/2}$. First order methods, based on piecewise constant approximations, are usually stable, but contain large numerical diffusion which smears out the shock profile. High order methods are achieved by using polynomials of higher degree, but this produces numerical oscillations around the shock. The basic problem is how to accurately capture the approximated solution near shocks, and at the same time retain stability of the numerical scheme. A common technique is to use a high order scheme on regions where the solution is smooth, and switch to a lower order method near a discontinuity. Well-known methods of this type include the Godunov methods and the MUSCL schemes, wave propagation methods [40,41], the central difference schemes [52,58] and the ENO/WENO schemes [56].

The conservative front tracking methods combine the FDM/FVM with the standard front tracking [26,27]. Based on a high order FDM/FVM, the methods in addition track the location and the strength of the discontinuities, and treat them as moving boundaries. The complexity increases with the number of fronts.

In the FEM setting, the discontinuous Galerkin's methods are widely used [13,14]. The method uses finite element discretization in space, with piecewise polynomials approximation, but allows the approximation to be discontinuous at cell boundaries.

Some numerical methods can be directly extended to the multi-dimensional case, but others need to use a dimensional splitting technique, which introduces additional diffusion. The performance of numerical algorithms is usually tested with some benchmark problem, and little is known theoretically, apart from the case of a scalar conservation law. Moreover, it remains a challenging problem to construct efficient high order numerical methods for systems of conservation laws, both in one and in several space dimensions.

Future Directions

In spite of extensive research efforts, the mathematical theory of hyperbolic conservation laws is still largely incomplete.

For hyperbolic systems in one space dimension, a major challenge is to study the existence and uniqueness of solutions to problems with large initial data. In this direction, some counterexamples show that, for particular sys-

tems, solutions can become unbounded in finite time [31]. However, it is conjectured that for many physical systems, endowed with a strictly convex entropy, such pathological behavior should not occur. In particular, the so-called “p-system” describing isentropic gas dynamics (4) should have global solutions with bounded variation, for arbitrarily large initial data [60].

It is worth mentioning that, for large initial data, the global existence of solutions is known mainly in the scalar case [36,59]. For hyperbolic systems of two conservation laws, global existence can still be proved, relying on a compensated compactness argument [20]. This approach, however, does not provide information on the uniqueness of solutions, or on their continuous dependence on the initial data.

Another major open problem is to theoretically analyze the convergence of numerical approximations. Error bounds on discrete approximations are presently available only in the scalar case [37]. For solutions to hyperbolic systems of n conservation laws, proofs of the convergence of viscous approximations [6], semidiscrete schemes [4], or relaxation schemes [5] have always relied on a priori bounds on the total variation. On the other hand, the counterexample in [2] shows that in general one cannot have any a priori bounds on the total variation of approximate solutions constructed by fully discrete numerical schemes. Understanding the convergence of these discrete approximations will likely require a new approach.

At present, the most outstanding theoretical open problem is to develop a fundamental existence and uniqueness theory for hyperbolic systems in several space dimensions. In order to achieve an existence proof, a key step is to identify the appropriate functional space where to construct solutions. In the one-dimensional case, solutions are found in the space BV of functions with bounded variation. In several space dimensions, however, it is known that the total variation of an arbitrary small solution can become unbounded almost immediately [54]. Hence the space BV does not provide a suitable framework to study the problem. For a special class of systems, a positive result and a counterexample, concerning global existence and continuous dependence on initial data can be found in [1] and in [9], respectively.

Bibliography

Primary Literature

1. Ambrosio L, Bouchut F, De Lellis C (2004) Well-posedness for a class of hyperbolic systems of conservation laws in several space dimensions. *Comm Part Diff Equat* 29:1635–1651
2. Baiti P, Bressan A, Jenssen HK (2006) An instability of the Godunov scheme. *Comm Pure Appl Math* 59:1604–1638
3. Bianchini S (2003) On the Riemann problem for non-conservative hyperbolic systems. *Arch Rat Mech Anal* 166:1–26
4. Bianchini S (2003) BV solutions of the semidiscrete upwind scheme. *Arch Ration Mech Anal* 167:1–81
5. Bianchini S (2006) Hyperbolic limit of the Jin-Xin relaxation model. *Comm Pure Appl Math* 59:688–753
6. Bianchini S, Bressan A (2005) Vanishing viscosity solutions to nonlinear hyperbolic systems. *Ann Math* 161:223–342
7. Bressan A (1992) Global solutions to systems of conservation laws by wave-front tracking. *J Math Anal Appl* 170:414–432
8. Bressan A (2000) *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem*. Oxford University Press, Oxford
9. Bressan A (2003) An ill posed Cauchy problem for a hyperbolic system in two space dimensions. *Rend Sem Mat Univ Padova* 110:103–117
10. Bressan A, Liu TP, Yang T (1999) L^1 stability estimates for $n \times n$ conservation laws. *Arch Ration Mech Anal* 149:1–22
11. Bressan A, Marson A (1998) Error bounds for a deterministic version of the Glimm scheme. *Arch Rat Mech Anal* 142:155–176
12. Chen GQ, Zhang Y, Zhu D (2006) Existence and stability of supersonic Euler flows past Lipschitz wedges. *Arch Ration Mech Anal* 181:261–310
13. Cockburn B, Shu CW (1998) The local discontinuous Galerkin finite element method for convection diffusion systems. *SIAM J Numer Anal* 35:2440–2463
14. Cockburn B, Hou S, Shu C-W (1990) The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case. *Math Comput* 54:545–581
15. Courant R, Friedrichs KO (1948) *Supersonic Flow and Shock Waves*. Wiley Interscience, New York
16. Crandall MG (1972) The semigroup approach to first-order quasilinear equations in several space variables. *Israel J Math* 12:108–132
17. Dafermos C (1972) Polygonal approximations of solutions of the initial value problem for a conservation law. *J Math Anal Appl* 38:33–41
18. Dafermos C (2005) *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 2nd edn. Springer, Berlin, pp 325–626
19. DiPerna RJ (1976) Global existence of solutions to nonlinear hyperbolic systems of conservation laws. *J Differ Equ* 20:187–212
20. DiPerna R (1983) Convergence of approximate solutions to conservation laws. *Arch Ration Mech Anal* 82:27–70
21. Euler L (1755) *Principes généraux du mouvement des fluides*. *Mém Acad Sci Berlin* 11:274–315
22. Evans LC, Gariepy RF (1992) *Measure Theory and Fine Properties of Functions*. C.R.C. Press, Boca Raton, pp viii–268
23. Friedrichs KO, Lax P (1971) Systems of conservation laws with a convex extension. *Proc Nat Acad Sci USA* 68:1686–1688
24. Glimm J (1965) Solutions in the large for nonlinear hyperbolic systems of equations. *Comm Pure Appl Math* 18:697–715
25. Glimm J, Lax P (1970) Decay of solutions of systems of nonlinear hyperbolic conservation laws. *Am Math Soc Memoir* 101:xvii–112

26. Glimm J, Li X, Liu Y (2001) Conservative Front Tracking in One Space Dimension. In: *Fluid flow and transport in porous media: mathematical and numerical treatment*, (South Hadley, 2001), pp 253–264. Contemp Math 295, Amer Math Soc, Providence
27. Glimm J, Grove JW, Li XL, Shyue KM, Zeng Y, Zhang Q (1998) Three dimensional front tracking. *SIAM J Sci Comput* 19: 703–727
28. Goodman J, Xin Z (1992) Viscous limits for piecewise smooth solutions to systems of conservation laws. *Arch Ration Mech Anal* 121:235–265
29. Holden H, Risebro NH (2002) *Front tracking for hyperbolic conservation laws*. Springer, New York
30. Jabin P, Perthame B (2002) Regularity in kinetic formulations via averaging lemmas. *ESAIM Control Optim Calc Var* 8: 761–774
31. Jenssen HK (2000) Blowup for systems of conservation laws. *SIAM J Math Anal* 31:894–908
32. Jiang G-S, Levy D, Lin C-T, Osher S, Tadmor E (1998) High-resolution non-oscillatory central schemes with non-staggered grids for hyperbolic conservation laws. *SIAM J Numer Anal* 35:2147–2168
33. Jin S, Xin ZP (1995) The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm Pure Appl Math* 48:235–276
34. Kawashima S, Matsumura A (1994) Stability of shock profiles in viscoelasticity with non-convex constitutive relations. *Comm Pure Appl Math* 47:1547–1569
35. Keyfitz B (2004) Self-similar solutions of two-dimensional conservation laws. *J Hyperbolic Differ Equ* 1:445–492
36. Kruzhkov S (1970) First-order quasilinear equations with several space variables. *Math USSR Sb* 10:217–273
37. Kuznetsov NN (1976) Accuracy of some approximate methods for computing the weak solution of a first order quasilinear equation. *USSR Comp Math Math Phys* 16:105–119
38. Lax P (1957) Hyperbolic systems of conservation laws II. *Comm Pure Appl Math* 10:537–566
39. Lax P (1971) Shock waves and entropy. In: Zarantonello E (ed) *Contributions to Nonlinear Functional Analysis*. Academic Press, New York, pp 603–634
40. Leveque RJ (1990) *Numerical methods for conservation laws. Lectures in Mathematics*. Birkhäuser, Basel
41. Leveque RJ (2002) *Finite volume methods for hyperbolic problems*. Cambridge University Press, Cambridge
42. Li J, Zhang T, Yang S (1998) The two dimensional problem in gas dynamics. Pitman, Longman Essex
43. Lighthill MJ, Whitham GB (1955) On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proc Roy Soc Lond A* 229:317–345
44. Lions PL, Perthame E, Tadmor E (1994) A kinetic formulation of multidimensional scalar conservation laws and related equations. *JAMS* 7:169–191
45. Liu TP (1977) Linear and nonlinear large-time behavior of solutions of general systems of hyperbolic conservation laws. *Comm Pure Appl Math* 30:767–796
46. Liu TP (1981) Admissible solutions of hyperbolic conservation laws. *Mem Am Math Soc* 30(240):iv–78
47. Liu TP (1985) Nonlinear stability of shock waves for viscous conservation laws. *Mem Am Math Soc* 56(328):v–108
48. Liu TP (1987) Hyperbolic conservation laws with relaxation. *Comm Math Pys* 108:153–175
49. Majda A (1984) *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*. Springer, New York
50. Metivier G (2001) Stability of multidimensional shocks. *Advances in the theory of shock waves*. Birkhäuser, Boston, pp 25–103
51. Morawetz CS (1994) Potential theory for regular and Mach reflection of a shock at a wedge. *Comm Pure Appl Math* 47: 593–624
52. Nessyahu H, Tadmor E (1990) Non-oscillatory central differencing for hyperbolic conservation laws *J Comp Phys* 87:408–463
53. Perthame B (2002) *Kinetic Formulation of Conservation Laws*. Oxford Univ. Press, Oxford
54. Rauch J (1986) BV estimates fail for most quasilinear hyperbolic systems in dimensions greater than one. *Comm Math Phys* 106:481–484
55. Riemann B (1860) Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. *Gött Abh Math Cl* 8:43–65
56. Shu CW (1998) Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. *Advanced numerical approximation of nonlinear hyperbolic equations*. (Cetraro, 1997), pp 325–432. *Lecture Notes in Mathematics*, vol 1697. Springer, Berlin
57. Smoller J (1994) *Shock Waves and Reaction-Diffusion Equations*, 2nd edn. Springer, New York
58. Tadmor E (1998) Approximate solutions of nonlinear conservation laws. In: *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. *Lecture Notes in Mathematics*, vol 1697. (1997 C.I.M.E. course in Cetraro, Italy) Springer, Berlin, pp 1–149
59. Volpert AI (1967) The spaces BV and quasilinear equations. *Math USSR Sb* 2:225–267
60. Young R (2003) Isentropic gas dynamics with large data. In: *Hyperbolic problems: theory, numerics, applications*. Springer, Berlin, pp 929–939
61. Zheng Y (2001) *Systems of Conservation Laws. Two-dimensional Riemann problems*. Birkhäuser, Boston
62. Zumbrun K (2004) Stability of large-amplitude shock waves of compressible Navier–Stokes equations. With an appendix by Helge Kristian Jenssen and Gregory Lyng. *Handbook of Mathematical Fluid Dynamics*, vol III. North-Holland, Amsterdam, pp 311–533

Books and Reviews

- Benzoni-Gavage S, Serre D (2007) *Multidimensional hyperbolic partial differential equations. First-order systems and applications*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford
- Boillat G (1996) *Nonlinear hyperbolic fields and waves*. In: *Recent Mathematical Methods in Nonlinear Wave Propagation* (Montecatini Terme, 1994), pp 1–47. *Lecture Notes in Math*, vol 1640. Springer, Berlin
- Chen GQ, Wang D (2002) The Cauchy problem for the Euler equations for compressible fluids. In: *Handbook of Mathematical Fluid Dynamics*, vol I. North-Holland, Amsterdam, pp 421–543
- Courant R, Hilbert D (1962) *Methods of Mathematical Physics*, vol II. John Wiley & Sons - Interscience, New York
- Garavello M, Piccoli B (2006) *Traffic Flow on Networks*. American Institute of Mathematical Sciences, Springfield
- Godlewski E, Raviart PA (1996) *Numerical approximation of hyperbolic systems of conservation laws*. Springer, New York

- Gurtin ME (1981) An Introduction to Continuum Mechanics. Mathematics in Science and Engineering, 158. Academic Press, New York, pp xi–265
- Hörmander L (1997) Lectures on Nonlinear Hyperbolic Differential Equations. Springer, Berlin
- Jeffrey A (1976) Quasilinear Hyperbolic Systems and Waves. Research Notes in Mathematics, vol 5. Pitman Publishing, London, pp vii–230
- Kreiss HO, Lorenz J (1989) Initial-boundary value problems and the Navier–Stokes equations. Academic Press, San Diego
- Kröner D (1997) Numerical schemes for conservation laws. In: Wiley-Teubner Series Advances in Numerical Mathematics. John Wiley, Chichester
- Landau LD, Lifshitz EM (1959) Fluid Mechanics. translated from the Russian by Sykes JB, Reid WH, Course of Theoretical Physics, vol 6. Pergamon press, London, Addison-Wesley, Reading, pp xii–536
- Li T-T, Yu W-C (1985) Boundary value problems for quasilinear hyperbolic systems. Duke University Math. Series, vol 5. Durham, pp vii–325
- Li T-T (1994) Global classical solutions for quasilinear hyperbolic systems. Wiley, Chichester
- Lu Y (2003) Hyperbolic conservation laws and the compensated compactness method. Chapman & Hall/CRC, Boca Raton
- Morawetz CS (1981) Lecture Notes on Nonlinear Waves and Shocks. Tata Institute of Fundamental Research, Bombay
- Rozhdestvenski BL, Yanenko NN (1978) Systems of quasilinear equations and their applications to gas dynamics. Nauka, Moscow. English translation: American Mathematical Society, Providence
- Serre D (2000) Systems of Conservation Laws, I Geometric structures, oscillations, and initial-boundary value problem, II Hyperbolicity, entropies, shock waves. Cambridge University Press, pp xii–263
- Whitham GB (1999) Linear and Nonlinear Waves. Wiley-Interscience, New York

Hyperbolic Dynamical Systems

VITOR ARAÚJO^{1,2}, MARCELO VIANA³

¹ CMUP, Porto, Portugal

² IM-UFRJ, Rio de Janeiro, Brazil

³ IMPA, Rio de Janeiro, Brazil

Article Outline

[Glossary](#)

[Definition](#)

[Introduction](#)

[Linear Systems](#)

[Local Theory](#)

[Hyperbolic Behavior: Examples](#)

[Hyperbolic Sets](#)

[Uniformly Hyperbolic Systems](#)

[Attractors and Physical Measures](#)

[Obstructions to Hyperbolicity](#)

[Partial Hyperbolicity](#)

[Non-Uniform Hyperbolicity – Linear Theory](#)

[Non-Uniformly Hyperbolic Systems](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Homeomorphism, diffeomorphism A *homeomorphism* is a continuous map $f: M \rightarrow N$ which is one-to-one and onto, and whose inverse $f^{-1}: N \rightarrow M$ is also continuous. It may be seen as a global continuous change of coordinates. We call f a *diffeomorphism* if, in addition, both it and its inverse are smooth. When $M = N$, the iterated n -fold composition $f \circ \dots \circ f$ is denoted by f^n . By convention, f^0 is the identity map, and $f^{-n} = (f^n)^{-1} = (f^{-1})^n$ for $n \geq 0$.

Smooth flow A flow $f^t: M \rightarrow M$ is a family of diffeomorphisms depending in a smooth fashion on a parameter $t \in \mathbb{R}$ and satisfying $f^{s+t} = f^s \circ f^t$ for all $s, t \in \mathbb{R}$. This property implies that f^0 is the identity map. Flows usually arise as solutions of autonomous differential equations: let $t \mapsto \phi^t(v)$ denote the solution of

$$\dot{X} = F(X), \quad X(0) = v, \quad (1)$$

and assume solutions are defined for all times; then the family ϕ^t thus defined is a flow (at least as smooth as the vector field F itself). The vector field may be recovered from the flow, through the relation $F(X) = \partial_t \phi^t(X) |_{t=0}$.

C^k topology Two maps admitting continuous derivatives are said to be C^1 -close if they are uniformly close, and so are their derivatives. More generally, given any $k \geq 1$, we say that two maps are C^k -close if they admit continuous derivatives up to order k , and their derivatives of order i are uniformly close, for every $i = 0, 1, \dots, k$. This defines a topology in the space of maps of class C^k .

Foliation A foliation is a partition of a subset of the ambient space into smooth submanifolds, that one calls leaves of the foliation, all with the same dimension and varying continuously from one point to the other. For instance, the trajectories of a vector field F , that is, the solutions of Eq. (1), form a 1-dimensional foliation (the leaves are curves) of the complement of the set of zeros of F . The main examples of foliations in the context of this work are the families of stable and unstable manifolds of hyperbolic sets.

Attractor A subset Λ of the ambient space M is *invariant* under a transformation f if $f^{-1}(\Lambda) = \Lambda$, that is,

a point is in Λ if and only if its image is. Λ is invariant under a flow if it is invariant under f^t for all $t \in \mathbb{R}$. An *attractor* is a compact invariant subset Λ such that the trajectories of all points in a neighborhood U converge to Λ as times goes to infinity, and Λ is *dynamically indecomposable* (or *transitive*): there is some trajectory dense in Λ . Sometimes one asks convergence only for points in some “large” subset of a neighborhood U of Λ , and dynamical indecomposability can also be defined in somewhat different ways. However, the formulations we just gave are fine in the uniformly hyperbolic context.

Limit sets The ω -limit set of a trajectory $f^n(x)$, $n \in \mathbb{Z}$ is the set $\omega(x)$ of all accumulation points of the trajectory as time n goes to $+\infty$. The α -limit set is defined analogously, with $n \rightarrow -\infty$. The corresponding notions for continuous time systems (flows) are defined analogously. The *limit set* $L(f)$ (or $L(f^t)$, in the flow case) is the closure of the union of all ω -limit and all α -limit sets. The *non-wandering set* $\Omega(f)$ (or $\Omega(f^t)$, in the flow case) is that set of points such that every neighborhood U contains some point whose orbit returns to U in future time (then some point returns to U in past time as well). When the ambient space is compact all these sets are non-empty. Moreover, the limit set is contained in the non-wandering set.

Invariant measure A probability measure μ in the ambient space M is *invariant* under a transformation f if $\mu(f^{-1}(A)) = \mu(A)$ for all measurable subsets A . This means that the “events” $x \in A$ and $f(x) \in A$ have equally probable. We say μ is invariant under a flow if it is invariant under f^t for all t . An invariant probability measure μ is *ergodic* if every invariant set A has either zero or full measure. An equivalently condition is that μ can not be decomposed as a convex combination of invariant probability measures, that is, one can not have $\mu = a\mu_1 + (1-a)\mu_2$ with $0 < a < 1$ and μ_1, μ_2 invariant.

Definition

In general terms, a smooth dynamical system is called hyperbolic if the tangent space over the asymptotic part of the phase space splits into two complementary directions, one which is contracted and the other which is expanded under the action of the system. In the classical, so-called uniformly hyperbolic case, the asymptotic part of the phase space is embodied by the limit set and, most crucially, one requires the expansion and contraction rates to be uniform. Uniformly hyperbolic systems are now fairly well understood. They may exhibit very complex behavior

which, nevertheless, admits a very precise description. Moreover, uniform hyperbolicity is the main ingredient for characterizing structural stability of a dynamical system. Over the years the notion of hyperbolicity was broadened (non-uniform hyperbolicity) and relaxed (partial hyperbolicity, dominated splitting) to encompass a much larger class of systems, and has become a paradigm for complex dynamical evolution.

Introduction

The theory of uniformly hyperbolic dynamical systems was initiated in the 1960s (though its roots stretch far back into the 19th century) by S. Smale, his students and collaborators, in the west, and D. Anosov, Ya. Sinai, V. Arnold, in the former Soviet Union. It came to encompass a detailed description of a large class of systems, often with very complex evolution. Moreover, it provided a very precise characterization of structurally stable dynamics, which was one of its original main goals.

The early developments were motivated by the problem of characterizing structural stability of dynamical systems, a notion that had been introduced in the 1930s by A. Andronov and L. Pontryagin. Inspired by the pioneering work of M. Peixoto on circle maps and surface flows, Smale introduced a class of *gradient-like* systems, having a finite number of periodic orbits, which should be structurally stable and, moreover, should constitute the majority (an open and dense subset) of all dynamical systems. Stability and openness were eventually established, in the thesis of J. Palis. However, contemporary results of M. Levinson, based on previous work by M. Cartwright and J. Littlewood, provided examples of open subsets of dynamical systems all of which have an infinite number of periodic orbits.

In order to try and understand such phenomenon, Smale introduced a simple geometric model, the now famous “horseshoe map”, for which infinitely many periodic orbits exist in a robust way. Another important example of structurally stable system which is not gradient like was R. Thom’s so-called “cat map”. The crucial common feature of these models is hyperbolicity: the tangent space at each point splits into two complementary directions such that the derivative contracts one of these directions and expands the other, at uniform rates.

In global terms, a dynamical system is called *uniformly hyperbolic*, or Axiom A, if its limit set has this hyperbolicity property we have just described. The mathematical theory of such systems, which is the main topic of this paper, is now well developed and constitutes a main paradigm for the behavior of “chaotic” systems. In our presentation we

go from local aspects (linear systems, local behavior, specific examples) to the global theory (hyperbolic sets, stability, ergodic theory). In the final sections we discuss several important extensions (strange attractors, partial hyperbolicity, non-uniform hyperbolicity) that have much broadened the scope of the theory.

Linear Systems

Let us start by introducing the phenomenon of hyperbolicity in the simplest possible setting, that of linear transformations and linear flows. Most of what we are going to say applies to both discrete time and continuous time systems in a fairly analogous way, and so at each point we refer to either one setting or the other. In depth presentations can be found in e.g. [6,8].

The general solution of a system of linear ordinary differential equations

$$\dot{X} = AX, \quad X(0) = v,$$

where A is a constant $n \times n$ real matrix and $v \in \mathbb{R}^n$ is fixed, is given by

$$X(t) = e^{tA} \cdot v, \quad t \in \mathbb{R},$$

where $e^{tA} = \sum_{n=0}^{\infty} (tA)^n/n!$. The linear flow is called *hyperbolic* if A has no eigenvalues on the imaginary axis. Then the *exponential* matrix e^A has no eigenvalues with norm 1. This property is very important for a number of reasons.

Stable and Unstable Spaces

For one thing it implies that all solutions have well-defined asymptotic behavior: they either converge to zero or diverge to infinity as time t goes to $\pm\infty$. More precisely, let

- E^s (*stable subspace*) be the subspace of \mathbb{R}^n spanned by the generalized eigenvector associated to eigenvalues of A with negative real part.
- E^u (*unstable subspace*) be the subspace of \mathbb{R}^n spanned by the generalized eigenvector associated to eigenvalues of A with positive real part

Then these subspaces are complementary, meaning that $\mathbb{R}^n = E^s \oplus E^u$, and every solution $e^{tA} \cdot v$ with $v \notin E^s \cup E^u$ diverges to infinity both in the future and in the past. The solutions with $v \in E^s$ converge to zero as $t \rightarrow +\infty$ and go to infinity as $t \rightarrow -\infty$, and analogously when $v \in E^u$, reversing the direction of time.

Robustness and Density

Another crucial feature of hyperbolicity is *robustness*: any matrix that is close to a hyperbolic one, in the sense that corresponding coefficients are close, is also hyperbolic. The stable and unstable subspaces need not coincide, of course, but the dimensions remain the same. In addition, hyperbolicity is *dense*: any matrix is close to a hyperbolic one. That is because, up to arbitrarily small modifications of the coefficients, one may force all eigenvalues to move out of the imaginary axis.

Stability, Index of a Fixed Point

In addition to robustness, hyperbolicity also implies *stability*: if B is close to a hyperbolic matrix A , in the sense we have just described, then the solutions of $\dot{X} = BX$ have essentially the same behavior as the solutions of $\dot{X} = AX$. What we mean by “essentially the same behavior” is that there exists a global continuous change of coordinates, that is, a homeomorphism $h: \mathbb{R}^n \rightarrow \mathbb{R}^n$, that maps solutions of one system to solutions of the other, preserving the time parametrization:

$$h(e^{tA} \cdot v) = e^{tB} \cdot h(v) \quad \text{for all } t \in \mathbb{R}.$$

More generally, two hyperbolic linear flows are conjugated by a homeomorphism h if and only if they have the same *index*, that is, the same number of eigenvalues with negative real part. In general, h can not be taken to be a diffeomorphism: this is possible if and only if the two matrices A and B are obtained from one another via a change of basis. Notice that in this case they must have the same eigenvalues, with the same multiplicities.

Hyperbolic Linear Systems

There is a corresponding notion of hyperbolicity for discrete time linear systems

$$X_{n+1} = CX_n, \quad X_0 = v,$$

with C a $n \times n$ real matrix. Namely, we say the system is *hyperbolic* if C has no eigenvalue in the unit circle. Thus a matrix A is hyperbolic in the sense of continuous time systems if and only if its exponential $C = e^A$ is hyperbolic in the sense of discrete time systems. The previous observations (well-defined behavior, robustness, denseness and stability) remain true in discrete time. Two hyperbolic matrices are conjugate by a homeomorphism if and only if they have the same index, that is, the same number of eigenvalues with norm less than 1, and they both either preserve or reverse orientation.

Local Theory

Now we move on to discuss the behavior of non-linear systems close to fixed or, more generally, periodic trajectories. By non-linear system we understand the iteration of a diffeomorphism f , or the evolution of a smooth flow f^t , on some manifold M . The general philosophy is that the behavior of the system close to a hyperbolic fixed point very much resembles the dynamics of its linear part.

A fixed point $p \in M$ of a diffeomorphism $f: M \rightarrow M$ is called *hyperbolic* if the linear part $Df_p: T_p M \rightarrow T_p M$ is a hyperbolic linear map, that is, if Df_p has no eigenvalue with norm 1. Similarly, an equilibrium point p of a smooth vector field F is *hyperbolic* if the derivative $DF(p)$ has no pure imaginary eigenvalues.

Hartman–Grobman Theorem

This theorem asserts that if p is a hyperbolic fixed point of $f: M \rightarrow M$ then there are neighborhoods U of p in M and V of 0 in the tangent space $T_p M$ such that we can find a homeomorphism $h: U \rightarrow V$ such that

$$h \circ f = Df_p \circ h,$$

whenever the composition is defined. This property means that h maps orbits of $Df(p)$ close to zero to orbits of f close to p . We say that h is a (local) *conjugacy* between the non-linear system f and its linear part Df_p . There is a corresponding similar theorem for flows near a hyperbolic equilibrium. In either case, in general h can not be taken to be a diffeomorphism.

Stable Sets

The *stable set* of the hyperbolic fixed point p is defined by

$$W^s(p) = \{x \in M: d(f^n(x), f^n(p)) \xrightarrow{n \rightarrow +\infty} 0\}.$$

Given $\beta > 0$ we also consider the *local stable set* of size $\beta > 0$, defined by

$$W_\beta^s(p) = \{x \in M: d(f^n(x), f^n(p)) \leq \beta \text{ for all } n \geq 0\}.$$

The image of W_β^s under the conjugacy h is a neighborhood of the origin inside E^s . It follows that the local stable set is an embedded topological disk, with the same dimension as E^s . Moreover, the orbits of the points in $W_\beta^s(p)$ actually converges to the fixed point as time goes to infinity. Therefore,

$$z \in W^s(p) \Leftrightarrow f^n(z) \in W_\beta^s(p) \text{ for some } n \geq 0.$$

Stable Manifold Theorem

The stable manifold theorem asserts that $W_\beta^s(p)$ is actually a smooth embedded disk, with the same order of differentiability as f itself, and it is tangent to E^s at the point p . It follows that $W^s(p)$ is a smooth submanifold, injectively immersed in M . In general, $W^s(p)$ is not embedded in M : in many cases it has self-accumulation points. For these reasons one also refers to $W^s(p)$ and $W_\beta^s(p)$ as *stable manifolds* of p . Unstable manifolds are defined analogously, replacing the transformation by its inverse.

Local Stability

We call *index* of a diffeomorphism f at a hyperbolic fixed point p the index of the linear part, that is, the number of eigenvalues of Df_p with negative real part. By the Hartman–Grobman theorem and previous comments on linear systems, two diffeomorphisms are locally conjugate near hyperbolic fixed points if and only if the stable indices and they both preserve/reverse orientation. In other words, the index together with the sign of the Jacobian determinant form a complete set of invariants for local topological conjugacy.

Let g be any diffeomorphism C^1 -close to f . Then g has a unique fixed point p_g close to p , and this fixed point is still hyperbolic. Moreover, the stable indices and the orientations of the two diffeomorphisms at the corresponding fixed points coincide, and so they are locally conjugate. This is called *local stability* near of diffeomorphisms hyperbolic fixed points. The same kind of result holds for flows near hyperbolic equilibria.

Hyperbolic Behavior: Examples

Now let us review some key examples of (semi)global hyperbolic dynamics. Thorough descriptions are available in e.g. [6,8,9].

A Linear Torus Automorphism

Consider the linear transformation $A: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by the following matrix, relative to the canonical base of the plane:

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

The 2-dimensional torus \mathbb{T}^2 is the quotient $\mathbb{R}^2/\mathbb{Z}^2$ of the plane by the equivalence relation

$$(X_1, y_1) \sim (x_2, y_2) \Leftrightarrow (x_1 - x_2, y_1 - y_2) \in \mathbb{Z}^2.$$

Since A preserves the lattice \mathbb{Z}^2 of integer vectors, that is, since $A(\mathbb{Z}^2) = \mathbb{Z}^2$, the linear transformation defines an invertible map $f_A: \mathbb{T}^2 \rightarrow \mathbb{T}^2$ in the quotient space, which is an example of linear automorphism of \mathbb{T}^2 . We call affine line in \mathbb{T}^2 the projection under the quotient map of any affine line in the plane.

The linear transformation A is hyperbolic, with eigenvalues $0 < \lambda_1 < 1 < \lambda_2$, and the corresponding eigenspaces E^1 and E^2 have irrational slope. For each point $z \in \mathbb{T}^2$, let $W_i(z)$ denote the affine line through z and having the direction of E^i , for $i = 1, 2$:

- distances along $W_1(z)$ are multiplied by $\lambda_1 < 1$ under forward iteration of f_A
- distances along $W_2(z)$ are multiplied by $1/\lambda_2 < 1$ under backward iteration of f_A .

Thus we call $W_1(z)$ *stable manifold* and $W_2(z)$ *unstable manifold* of z (notice we are not assuming z to be periodic). Since the slopes are irrational, stable and unstable manifolds are dense in the whole torus. From this fact one can deduce that the periodic points of f_A form a dense subset of the torus, and that there exist points whose trajectories are dense in \mathbb{T}^2 . The latter property is called *transitivity*.

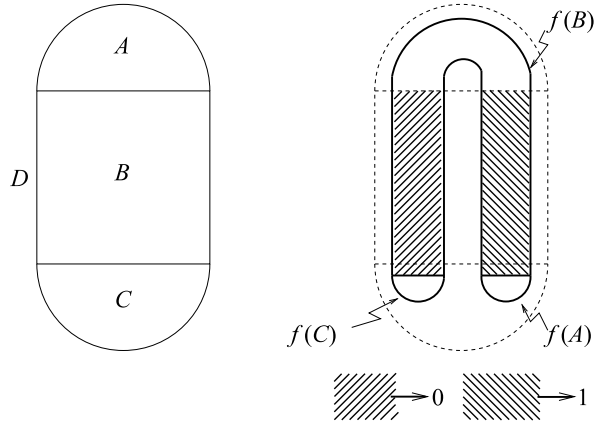
An important feature of this systems is that its behavior is (globally) stable under small perturbations: given any diffeomorphism $g: \mathbb{T}^2 \rightarrow \mathbb{T}^2$ sufficiently C^1 -close to f_A , there exists a homeomorphism $h: \mathbb{T}^2 \rightarrow \mathbb{T}^2$ such that $h \circ g = f_A \circ h$. In particular, g is also transitive and its periodic points form a dense subset of \mathbb{T}^2 .

The Smale Horseshoe

Consider a stadium shaped region D in the plane divided into three subregions, as depicted in Fig. 1: two half disks, A and C , and a square, B . Next, consider a map $f: D \rightarrow D$ mapping D back inside itself as described in Fig. 1: the intersection between B and $f(B)$ consists of two rectangles, R_0 and R_1 , and f is affine on the pre-image of these rectangles, contracting the horizontal direction and expanding the vertical direction.

The set $\Lambda = \bigcap_{n \in \mathbb{Z}} f^n(B)$, formed by all the points whose orbits never leave the square B is totally disconnected, in fact, it is the product of two Cantor sets. A description of the dynamics on Λ may be obtained through the following coding of orbits. For each point $z \in \Lambda$ and every time $n \in \mathbb{Z}$ the iterate $f^n(z)$ must belong to either R_0 or R_1 . We call *itinerary* of z the sequence $\{s_n\}_{n \in \mathbb{Z}}$ with values in the set $\{0, 1\}$ defined by $f^n(z) \in R_{s_n}$ for all $n \in \mathbb{Z}$. The itinerary map

$$\Lambda \rightarrow \{0, 1\}^{\mathbb{Z}}, \quad z \mapsto \{s_n\}_{n \in \mathbb{Z}}$$



Hyperbolic Dynamical Systems, Figure 1
Horseshoe map

is a homeomorphism, and conjugates f restricted to Λ to the so-called *shift map* defined on the space of sequences by

$$\{0, 1\}^{\mathbb{Z}} \rightarrow \{0, 1\}^{\mathbb{Z}}, \quad \{s_n\}_{n \in \mathbb{Z}} \mapsto \{s_{n+1}\}_{n \in \mathbb{Z}}.$$

Since the shift map is transitive, and its periodic points form a dense subset of the domain, it follows that the same is true for the horseshoe map on Λ .

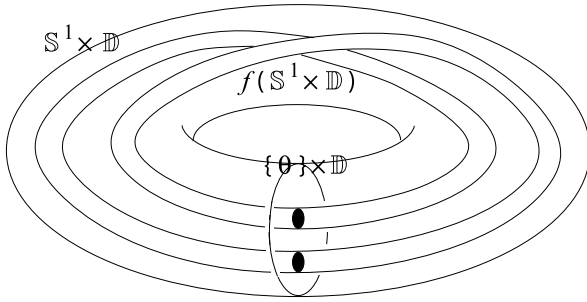
From the definition of f we get that distances along horizontal line segments through points of Λ are contracted at a uniform rate under forward iteration and, dually, distances along vertical line segments through points of Λ are contracted at a uniform rate under backward iteration. Thus, horizontal line segments are local stable sets and vertical line segments are local unstable sets for the points of Λ .

A striking feature of this system is the stability of its dynamics: given any diffeomorphism g sufficiently C^1 -close to f , its restriction to the set $\Lambda_g = \bigcap_{n \in \mathbb{Z}} g^n(B)$ is conjugate to the restriction of f to the set $\Lambda = \Lambda_f$ (and, consequently, is conjugate to the shift map). In addition, each point of Λ_g has local stable and unstable sets which are smooth curve segments, respectively, approximately horizontal and approximately vertical.

The Solenoid Attractor

The *solid torus* is the product space $\mathbb{S}^1 \times \mathbb{D}$, where $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$ is the circle and $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ is the unit disk in the complex plane. Consider the map $f: \mathbb{S}^1 \times \mathbb{D} \rightarrow \mathbb{S}^1 \times \mathbb{D}$ given by

$$(\theta, z) \mapsto (2\theta, \alpha z + \beta e^{i\theta/2}),$$



Hyperbolic Dynamical Systems, Figure 2
The solenoid attractor

$\theta \in \mathbb{R}/\mathbb{Z}$ and $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta < 1$. The latter condition ensures that the image $f(\mathbb{S}^1 \times \mathbb{D})$ is strictly contained in $\mathbb{S}^1 \times \mathbb{D}$. Geometrically, the image is a long thin domain going around the solid torus twice, as described in Fig. 2. Then, for any $n \geq 1$, the corresponding iterate $f^n(\mathbb{S}^1 \times \mathbb{D})$ is an increasingly thinner and longer domain that winds 2^k times around $\mathbb{S}^1 \times \mathbb{D}$. The maximal invariant set

$$\Lambda = \bigcap_{n \geq 0} f^n(\mathbb{S}^1 \times \mathbb{D})$$

is called *solenoid attractor*. Notice that the forward orbit under f of every point in $\mathbb{S}^1 \times \mathbb{D}$ accumulates on Λ . One can also check that the restriction of f to the attractor is transitive, and the set of periodic points of f is dense in Λ .

In addition Λ has a dense subset of periodic orbits and also a dense orbit. Moreover every point in a neighborhood of Λ converges to Λ and this is why this set is called an *attractor*.

Hyperbolic Sets

The notion we are now going to introduce distillates the crucial feature common to the examples presented previously. A detailed presentation is given in e.g. [8,10]. Let $f: M \rightarrow M$ be a diffeomorphism on a manifold M . A compact invariant set $\Lambda \subset M$ is a *hyperbolic set* for f if the tangent bundle over Λ admits a decomposition

$$T_\Lambda M = E^u \oplus E^s,$$

invariant under the derivative and such that $\|Df^{-1}|_{E^u}\| < \lambda$ and $\|Df|_{E^s}\| < \lambda$ for some constant $\lambda < 1$ and some choice of a Riemannian metric on the manifold. When it exists, such a decomposition is necessarily unique and continuous. We call E^s the stable bundle and E^u the unstable bundle of f on the set Λ .

The definition of hyperbolicity for an invariant set of a smooth flow containing no equilibria is similar, except that one asks for an invariant decomposition $T_\Lambda M =$

$E^u \oplus E^0 \oplus E^s$, where E^u and E^s are as before and E^0 is a line bundle tangent to the flow lines. An invariant set that contains equilibria is hyperbolic if and only if it consists of a finite number of points, all of them hyperbolic equilibria.

Cone Fields

The definition of hyperbolic set is difficult to use in concrete situations, because, in most cases, one does not know the stable and unstable bundles explicitly. Fortunately, to prove that an invariant set is hyperbolic it suffices to have some approximate knowledge of these invariant subbundles. That is the contents of the invariant cone field criterion: a compact invariant set is hyperbolic if and only if there exists some continuous (not necessarily invariant) decomposition $T_\Lambda M = E^1 \oplus E^2$ of the tangent bundle, some constant $\lambda < 1$, and some cone field around E^1

$$C_a^1(x) = \{v = v_1 + v_2 \in E_x^1 \oplus E_x^2 : \|v_2\| \leq a\|v_1\|\}, \quad x \in \Lambda,$$

which is

- (a) forward invariant: $Df_x(C_a^1(x)) \subset C_{\lambda a}^1(f(x))$ and
- (b) expanded by forward iteration: $\|Df_x(v)\| \geq \lambda^{-1}\|v\|$ for every $v \in C_a^1(x)$

and there exists a cone field $C_b^2(x)$ around E^2 which is backward invariant and expanded by backward iteration.

Robustness

An easy, yet very important consequence is that hyperbolic sets are robust under small modifications of the dynamics. Indeed, suppose Λ is a hyperbolic set for $f: M \rightarrow M$, and let $C_a^1(x)$ and $C_b^2(x)$ be invariant cone fields as above. The (non-invariant) decomposition $E^1 \oplus E^2$ extends continuously to some small neighborhood U of Λ , and then so do the cone fields. By continuity, conditions (a) and (b) above remain valid on U , possibly for a slightly larger constant λ . Most important, they also remain valid when f is replaced by any other diffeomorphism g which is sufficiently C^1 -close to it. Thus, using the cone field criterion once more, every compact set $K \subset U$ which is invariant under g is a hyperbolic set for g .

Stable Manifold Theorem

Let Λ be a hyperbolic set for a diffeomorphism $f: M \rightarrow M$. Assume f is of class C^k . Then there exist $\varepsilon_0 > 0$ and $0 < \lambda < 1$ and, for each $0 < \varepsilon \leq \varepsilon_0$ and $x \in \Lambda$, the *local stable manifold of size ε*

$$W_\varepsilon^s(x) = \{y \in M : \text{dist}(f^n(y), f^n(x)) \leq \varepsilon \text{ for all } n \geq 0\},$$

and the *local unstable manifold* of size ε

$$W_\varepsilon^u(x) = \{y \in M : \text{dist}(f^{-n}(y), f^{-n}(x)) \leq \varepsilon \\ \text{for all } n \geq 0\}$$

are C^k embedded disks, tangent at x to E_x^s and E_x^u , respectively, and satisfying

- $f(W_\varepsilon^s(x)) \subset W_\varepsilon^s(f(x))$ and $f^{-1}(W_\varepsilon^u(x)) \subset W_\varepsilon^u(f^{-1}(x))$;
- $\text{dist}(f(x), f(y)) \leq \lambda \text{dist}(x, y)$ for all $y \in W_\varepsilon^s(x)$
- $\text{dist}(f^{-1}(x), f^{-1}(y)) \leq \lambda \text{dist}(x, y)$ for all $y \in W_\varepsilon^u(x)$
- $W_\varepsilon^s(x)$ and $W_\varepsilon^u(x)$ vary continuously with the point x , in the C^k topology.

Then, the *global stable and unstable manifolds* of x ,

$$W^s(x) = \bigcup_{n \geq 0} f^{-n}(W_\varepsilon^s(f^n(x))) \\ \text{and } W^u(x) = \bigcup_{n \geq 0} f^n(W_\varepsilon^u(f^{-n}(x))),$$

are smoothly immersed submanifolds of M , and they are characterized by

$$W^s(x) = \{y \in M : \text{dist}(f^n(y), f^n(x)) \rightarrow 0 \text{ as } n \rightarrow \infty\} \\ W^u(x) = \{y \in M : \text{dist}(f^{-n}(y), f^{-n}(x)) \rightarrow 0 \\ \text{as } n \rightarrow \infty\}.$$

Shadowing Property

This crucial property of hyperbolic sets means that possible small “errors” in the iteration of the map close to the set are, in some sense, unimportant: to the resulting “wrong” trajectory, there corresponds a nearby genuine orbit of the map. Let us give the formal statement. Recall that a hyperbolic set is compact, by definition.

Given $\delta > 0$, a δ -pseudo-orbit of $f: M \rightarrow M$ is a sequence $\{x_n\}_{n \in \mathbb{Z}}$ such that

$$\text{dist}(x_{n+1}, f(x_n)) \leq \delta \quad \text{for all } n \in \mathbb{Z}.$$

Given $\varepsilon > 0$, one says that a pseudo-orbit is ε -shadowed by the orbit of a point $z \in M$ if $\text{dist}(f^n(z), x_n) \leq \varepsilon$ for all $n \in \mathbb{Z}$. The *shadowing lemma* says that for any $\varepsilon > 0$ one can find $\delta > 0$ and a neighborhood U of the hyperbolic set Λ such that every δ -pseudo-orbit in U is ε -shadowed by some orbit in U . Assuming ε is sufficiently small, the shadowing orbit is actually unique.

Local Product Structure

In general, these shadowing orbits need not be inside the hyperbolic set Λ . However, that is indeed the case if Λ is

a *maximal invariant set*, that is, if it admits some neighborhood U such that Λ coincides with the set of points whose orbits never leave U :

$$\Lambda = \bigcap_{n \in \mathbb{Z}} f^{-n}(U).$$

A hyperbolic set is a maximal invariant set if and only if it has the local product structure property stated in the next paragraph.

Let Λ be a hyperbolic set and ε be small. If x and y are nearby points in Λ then the local stable manifold of x intersects the local unstable manifold of y at a unique point, denoted $[x, y]$, and this intersection is transverse. This is because the local stable manifold and the local unstable manifold of every point are transverse, and these local invariant manifolds vary continuously with the point. We say that Λ has *local product structure* if there exists $\delta > 0$ such that $[x, y]$ belongs to Λ for every $x, y \in \Lambda$ with $\text{dist}(x, y) < \delta$.

Stability

The shadowing property may also be used to prove that hyperbolic sets are stable under small perturbations of the dynamics: if Λ is a hyperbolic set for f then for any C^1 -close diffeomorphism g there exists a hyperbolic set Λ_g close to Λ and carrying the same dynamical behavior.

The key observation is that every orbit $f^n(x)$ of f inside Λ is a δ -pseudo-orbit for g in a neighborhood U , where δ is small if g is close to f and, hence, it is shadowed by some orbit $g^n(z)$ of g . The correspondence $h(x) = z$ thus defined is injective and continuous.

For any diffeomorphism g close enough to f , the orbits of x in the maximal g -invariant set $\Lambda_g(U)$ inside U are pseudo-orbits for f . Therefore the shadowing property above enables one to bijectively associate g -orbits of $\Lambda_g(U)$ to f -orbits in Λ . This provides a homeomorphism $h: \Lambda_g(U) \rightarrow \Lambda$ which conjugates g and f on the respective hyperbolic sets: $f \circ h = h \circ g$. Thus *hyperbolic maximal sets are structurally stable*: the persistent dynamics in a neighborhood of these sets is the same for all nearby maps.

If Λ is a hyperbolic maximal invariant set for f then its hyperbolic continuation for any nearby diffeomorphism g is also a maximal invariant set for g .

Symbolic Dynamics

The dynamics of hyperbolic sets can be described through a symbolic coding obtained from a convenient discretization of the phase space. In a few words, one partitions the

set into a finite number of subsets and assigns to a generic point in the hyperbolic set its itinerary with respect to this partition. Dynamical properties can then be read out from a shift map in the space of (admissible) itineraries. The precise notion involved is that of Markov partition.

A set $R \subset \Lambda$ is a *rectangle* if $[x, y] \in R$ for each $x, y \in R$. A rectangle is *proper* if it is the closure of its interior relative to Λ . A *Markov partition* of a hyperbolic set Λ is a cover $\mathcal{R} = \{R_1, \dots, R_m\}$ of Λ by proper rectangles with pairwise disjoint interiors, relative to Λ , and such

$$W^u(f(x)) \cap R_j \subset f(W^u(x) \cap R_i) \\ \text{and } f(W^s(x) \cap R_i) \subset W^s(f(x)) \cap R_j$$

for every $x \in \text{int}_\Lambda(R_i)$ with $f(x) \in \text{int}_\Lambda(R_j)$. The key fact is that *any maximal hyperbolic set Λ admits Markov partitions with arbitrarily small diameter*.

Given a Markov partition \mathcal{R} with sufficiently small diameter, and a sequence $\mathbf{j} = (j_n)_{n \in \mathbb{Z}}$ in $\{1, \dots, m\}$, there exists at most one point $x = h(\mathbf{j})$ such that

$$f^n(x) \in R_{j_n} \quad \text{for each } n \in \mathbb{Z}.$$

We say that \mathbf{j} is admissible if such a point x does exist and, in this case, we say x admits \mathbf{j} as an itinerary. It is clear that $f \circ h = h \circ \sigma$, where σ is the shift (left-translation) in the space of admissible itineraries. The map h is continuous and surjective, and it is injective on the residual set of points whose orbits never hit the boundaries (relative to Λ) of the Markov rectangles.

Uniformly Hyperbolic Systems

A diffeomorphism $f: M \rightarrow M$ is *uniformly hyperbolic*, or satisfies the *Axiom A*, if the non-wandering set $\Omega(f)$ is a hyperbolic set for f and the set $\text{Per}(f)$ of periodic points is dense in $\Omega(f)$. There is an analogous definition for smooth flows $f^t: M \rightarrow M$, $t \in \mathbb{R}$. The reader can find the technical details in e. g. [6,8,10].

Dynamical Decomposition

The so-called “spectral” decomposition theorem of Smale allows for the global dynamics of a hyperbolic diffeomorphism to be decomposed into elementary building blocks. It asserts that the non-wandering set splits into a finite number of pairwise disjoint *basic pieces* that are compact, invariant, and dynamically indecomposable. More precisely, the non-wandering set $\Omega(f)$ of a uniformly hyperbolic diffeomorphism f is a finite pairwise disjoint union

$$\Omega(f) = \Lambda_1 \cup \dots \cup \Lambda_N$$

of f -invariant, transitive sets Λ_i , that are compact and maximal invariant sets. Moreover, the α -limit set of every orbit is contained in some Λ_i and so is the ω -limit set.

Geodesic Flows on Surfaces with Negative Curvature

Historically, the first important example of uniform hyperbolicity was the geodesic flow G^t on Riemannian manifolds of negative curvature M . This is defined as follows.

Let M be a compact Riemannian manifold. Given any tangent vector v , let $\gamma_v: \mathbb{R} \rightarrow TM$ be the geodesic with initial condition $v = \gamma_v(0)$. We denote by $\dot{\gamma}_v(t)$ the velocity vector at time t . Since $\|\dot{\gamma}_v(t)\| = \|v\|$ for all t , it is no restriction to consider only unit vectors. There is an important volume form on the unit tangent bundle, given by the product of the volume element on the manifold by the volume element induced on each fiber by the Riemannian metric. By integration of this form, one obtains the *Liouville measure* on the unit tangent bundle, which is a finite measure if the manifold itself has finite volume (including the compact case). The *geodesic flow* is the flow $G^t: T^1M \rightarrow T^1M$ on the unit tangent bundle T^1M of the manifold, defined by

$$G^t(v) = \dot{\gamma}_v(t).$$

An important feature is that this flow leaves invariant the Liouville measure. By Poincaré recurrence, this implies that $\Omega(G) = T^1M$.

A major classical result in Dynamics, due to Anosov, states that *if M has negative sectional curvature then this measure is ergodic for the flow*. That is, any invariant set has zero or full Liouville measure. The special case when M is a surface, had been dealt before by Hedlund and Hopf.

The key ingredient to this theorem is to prove that the geodesic flow is uniformly hyperbolic, in the sense we have just described, when the sectional curvature is negative. In the surface case, the stable and unstable invariant subbundles are differentiable, which is no longer the case in general in higher dimensions. This formidable obstacle was overcome by Anosov through showing that the corresponding invariant foliations retain, nevertheless, a weaker form of regularity property, that suffices for the proof. Let us explain this.

Absolute Continuity of Foliations

The invariant spaces E_x^s and E_x^u of a hyperbolic system depend continuously, and even Hölder continuously, on the base point x . However, in general this dependence is not differentiable, and this fact is at the origin of several important difficulties. Related to this, the families of stable

and unstable manifolds are, usually, not differentiable foliations: although the leaves themselves are as smooth as the dynamical system itself, the holonomy maps often fail to be differentiable. By holonomy maps we mean the projections along the leaves between two given cross-sections to the foliation.

However, Anosov and Sinai observed that if the system is at least twice differentiable then these foliations are *absolutely continuous*: their holonomy maps send zero Lebesgue measure sets of one cross-section to zero Lebesgue measure sets of the other cross-section. This property is crucial for proving that any smooth measure which is invariant under a twice differentiable hyperbolic system is ergodic. For dynamical systems that are only once differentiable the invariant foliations may fail to be absolutely continuous. Ergodicity still is an open problem.

Structural Stability

A dynamical system is *structurally stable* if it is equivalent to any other system in a C^1 neighborhood, meaning that there exists a global homeomorphism sending orbits of one to orbits of the other and preserving the direction of time. More generally, replacing C^1 by C^r neighborhoods, any $r \geq 1$, one obtains the notion of C^r structural stability. Notice that, in principle, this property gets weaker as r increases.

The Stability Conjecture of Palis–Smale proposed a complete geometric characterization of this notion: for any $r \geq 1$, C^r structurally stable systems should coincide with the hyperbolic systems having the property of strong transversality, that is, such that the stable and unstable manifolds of any points in the non-wandering set are transversal. In particular, this would imply that the property of C^r structural stability does not really depend on the value of r .

That hyperbolicity and strong transversality suffice for structural stability was proved in the 1970s by Robbin, de Melo, Robinson. It is comparatively easy to prove that strong transversality is also necessary. Thus, the heart of the conjecture is to prove that structurally stable systems must be hyperbolic. This was achieved by Mañé in the 1980s, for C^1 diffeomorphisms, and extended about ten years later by Hayashi for C^1 flows. Thus a C^1 diffeomorphism, or flow, on a compact manifold is structurally stable if and only if it is uniformly hyperbolic and satisfies the strong transversality condition.

Ω -stability

A weaker property, called Ω -stability is defined requiring equivalence only restricted to the non-wandering set.

The Ω -Stability Conjecture of Palis–Smale claims that, for any $r \geq 1$, Ω -stable systems should coincide with the hyperbolic systems with no cycles, that is, such that no basic pieces in the spectral decomposition are cyclically related by intersections of the corresponding stable and unstable sets.

The Ω -stability theorem of Smale states that these properties are sufficient for C^r Ω -stability. Palis showed that the no-cycles condition is also necessary. Much later, based on Mañé’s aforementioned result, he also proved that for C^1 diffeomorphisms hyperbolicity is necessary for Ω -stability. This was extended to C^1 flows by Hayashi in the 1990s.

Attractors and Physical Measures

A hyperbolic basic piece Λ_i is a *hyperbolic attractor* if the stable set

$$W^s(\Lambda_i) = \{x \in M : \omega(x) \subset \Lambda_i\}$$

contains a neighborhood of Λ_i . In this case we call $W^s(\Lambda_i)$ the *basin* of the attractor Λ_i , and denote it $B(\Lambda_i)$. When the uniformly hyperbolic system is of class C^2 , a basic piece is an attractor if and only if its stable set has positive Lebesgue measure. Thus, the union of the basins of all attractors is a full Lebesgue measure subset of M . This remains true for a residual (dense G_δ) subset of C^1 uniformly hyperbolic diffeomorphisms and flows.

The following fundamental result, due to Sinai, Ruelle, Bowen shows that, no matter how complicated it may be, the behavior of typical orbits in the basin of a hyperbolic attractor is well-defined at the statistical level: *any hyperbolic attractor Λ of a C^2 diffeomorphism (or flow) supports a unique invariant probability measure μ such that*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=0}^{n-1} \varphi(f^j(z)) = \int \varphi d\mu \quad (2)$$

for every continuous function φ and Lebesgue almost every point $x \in B(\Lambda)$. The standard reference here is [3].

Property (2) also means that the Sinai–Ruelle–Bowen measure μ may be “observed”: the weights of subsets may be found with any degree of precision, as the sojourn-time of any orbit picked “at random” in the basin of attraction:

$$\mu(V) = \text{fraction of time the orbit of } z \text{ spends in } V$$

for typical subsets V of M (the boundary of V should have zero μ -measure), and for Lebesgue almost any point $z \in B(\Lambda)$. For this reason μ is called a *physical measure*.

It also follows from the construction of these physical measures on hyperbolic attractors that they depend continuously on the diffeomorphism (or the flow). This *statistical stability* is another sense in which the asymptotic behavior is stable under perturbations of the system, distinct from structural stability.

There is another sense in which this measure is “physical” and that is that μ is the zero-noise limit of the stationary measures associated to the stochastic processes obtained by adding small random noise to the system. The idea is to replace genuine trajectories by “random orbits” $(z_n)_n$, where each z_{n+1} is chosen ε -close to $f(z_n)$. We speak of *stochastic stability* if, for any continuous function φ , the random time average

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=0}^{n-1} \varphi(z_j)$$

is close to $\int \varphi \, d\mu$ for almost all choices of the random orbit.

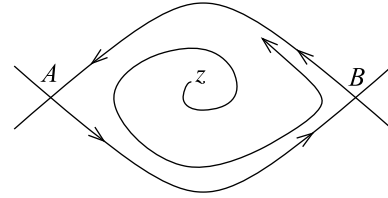
One way to construct such random orbits is through randomly perturbed iterations, as follows. Consider a family of probability measures ν_ε in the space of diffeomorphisms, such that each ν_ε is supported in the ε -neighborhood of f . Then, for each initial state z_0 define $z_{n+1} = f_{n+1}(z_n)$, where the diffeomorphisms f_n are independent random variables with distribution law ν_ε . A probability measure η_ε on the basin $B(\Lambda)$ is *stationary* if it satisfies

$$\eta_\varepsilon(E) = \int \eta_\varepsilon(g^{-1}(E)) \, d\nu_\varepsilon(g).$$

Stationary measures always exist, and they are often unique for each small $\varepsilon > 0$. Then stochastic stability corresponds to having η_ε converging weakly to μ when the noise level ε goes to zero.

The notion of stochastic stability goes back to Kolmogorov and Sinai. The first results, showing that uniformly hyperbolic systems are stochastically stable, on the basin of each attractor, were proved in the 1980s by Kifer and Young.

Let us point out that physical measures need not exist for general systems. A simple counter-example, attributed to Bowen, is described in Fig. 3: time averages diverge over any of the spiraling orbits in the region bounded by the saddle connections. Notice that the saddle connections are easily broken by arbitrarily small perturbations of the flow. Indeed, no robust examples are known of systems whose time-averages diverge on positive volume sets.



Hyperbolic Dynamical Systems, Figure 3
A planar flow with divergent time averages

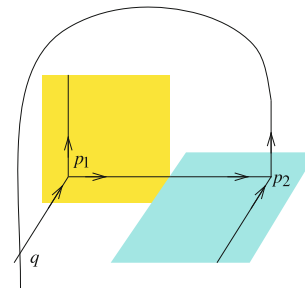
Obstructions to Hyperbolicity

Although uniform hyperbolicity was originally intended to encompass a residual or, at least, dense subset of all dynamical systems, it was soon realized that this is not the case: many important examples fall outside its realm. There are two main mechanisms that yield robustly non-hyperbolic behavior, that is, whole open sets of non-hyperbolic systems.

Heterodimensional Cycles

Historically, the first such mechanism was the coexistence of periodic points with different Morse indices (dimensions of the unstable manifolds) inside the same transitive set. See Fig. 4. This is how the first examples of C^1 -open subsets of non-hyperbolic diffeomorphisms were obtained by Abraham, Smale on manifolds of dimension $d \geq 3$. It was also the key in the constructions by Shub and Mañé of non-hyperbolic, yet robustly transitive diffeomorphisms, that is, such that every diffeomorphism in a C^1 neighborhood has dense orbits.

For flows, this mechanism may assume a novel form, because of the interplay between regular orbits and singularities (equilibrium points). That is, robust non-hyperbolicity may stem from the coexistence of regular and singular orbits in the same transitive set. The first, and very striking example was the geometric Lorenz attractor

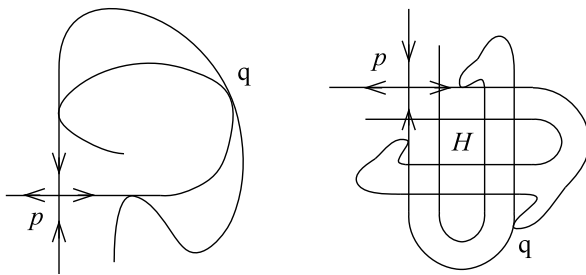


Hyperbolic Dynamical Systems, Figure 4
A heterodimensional cycle

proposed by Afraimovich, Bykov, Shil'nikov and Guckenheimer, Williams to model the behavior of the Lorenz equations, that we shall discuss later.

Homoclinic Tangencies

Of course, heterodimensional cycles may exist only in dimension 3 or higher. The first robust examples of non-hyperbolic diffeomorphisms on surfaces were constructed by Newhouse, exploiting the second of these two mechanisms: homoclinic tangencies, or non-transverse intersections between the stable and the unstable manifold of the same periodic point. See Fig. 5.



Hyperbolic Dynamical Systems, Figure 5
Homoclinic tangencies

It is important to observe that individual homoclinic tangencies are easily destroyed by small perturbations of the invariant manifolds. To construct open examples of surface diffeomorphisms with *some* tangency, Newhouse started from systems where the tangency is associated to a periodic point inside an invariant hyperbolic set with rich geometric structure. This is illustrated on the right hand side of Fig. 5. His argument requires a very delicate control of distortion, as well as of the dependence of the fractal dimension on the dynamics. Actually, for this reason, his construction is restricted to the C^r topology for $r \geq 2$. A very striking consequence of this construction is that these open sets exhibit *coexistence of infinitely many periodic attractors*, for each diffeomorphism on a residual subset. A detailed presentation of his result and consequences is given in [9].

Newhouse's conclusions have been extended in two ways. First, by Palis, Viana, for diffeomorphisms in any dimension, still in the C^r topology with $r \geq 2$. Then, by Bonatti, Díaz, for C^1 diffeomorphisms in any dimension larger or equal than 3. The case of C^1 diffeomorphisms on surfaces remains open. As a matter of fact, in this setting it is still unknown whether uniform hyperbolicity is dense in the space of all diffeomorphisms.

Partial Hyperbolicity

Several extensions of the theory of uniform hyperbolicity have been proposed, allowing for more flexibility, while keeping the core idea: splitting of the tangent bundle into invariant subbundles. We are going to discuss more closely two such extensions.

On the one hand, one may allow for one or more invariant subbundles along which the derivative exhibits mixed contracting/neutral/expanding behavior. This is generically referred to as *partial hyperbolicity*, and a standard reference is the book [5]. On the other hand, while requiring all invariant subbundles to be either expanding or contraction, one may relax the requirement of uniform rates of expansion and contraction. This is usually called *non-uniform hyperbolicity*. A detailed presentation of the fundamental results about this notion is available e.g. in [6]. In this section we discuss the first type of condition. The second one will be dealt with later.

Dominated Splittings

Let $f: M \rightarrow M$ be a diffeomorphism on a closed manifold M and K be any f -invariant set. A continuous splitting $T_x M = E_1(x) \oplus \cdots \oplus E_k(x)$, $x \in K$ of the tangent bundle over K is *dominated* if it is invariant under the derivative Df and there exists $\ell \in \mathbb{N}$ such that for every $i < j$, every $x \in K$, and every pair of unit vectors $u \in E_i(x)$ and $v \in E_j(x)$, one has

$$\frac{\|Df_x^\ell \cdot u\|}{\|Df_x^\ell \cdot v\|} < \frac{1}{2}, \quad (3)$$

and the dimension of $E_i(x)$ is independent of $x \in K$ for every $i \in \{1, \dots, k\}$. This definition may be formulated, equivalently, as follows: there exist $C > 0$ and $\lambda < 1$ such that for every pair of unit vectors $u \in E_i(x)$ and $v \in E_j(x)$, one has

$$\frac{\|Df_x^n \cdot u\|}{\|Df_x^n \cdot v\|} < C\lambda^n \quad \text{for all } n \geq 1.$$

Let f be a diffeomorphism and K be an f -invariant set having a dominated splitting $T_K M = E_1 \oplus \cdots \oplus E_k$. We say that the splitting and the set K are

- *partially hyperbolic* the derivative either contracts uniformly E_1 or expands uniformly E_k : there exists $\ell \in \mathbb{N}$ such that

$$\text{either } \|Df^\ell|_{E_1}\| < \frac{1}{2} \quad \text{or } \|(Df^\ell|_{E_k})^{-1}\| < \frac{1}{2}.$$

- *volume hyperbolic* if the derivative either contracts volume uniformly along E_1 or expands volume uniformly

along E_k : there exists $\ell \in \mathbb{N}$ such that

$$\text{either } |\det(Df^\ell | E_1)| < \frac{1}{2} \quad \text{or} \quad |\det(Df^\ell | E_k)| > 2.$$

The diffeomorphism f is *partially hyperbolic/volume hyperbolic* if the ambient space M is a partially hyperbolic/volume hyperbolic set for f .

Invariant Foliations

An crucial geometric feature of partially hyperbolic systems is the existence of invariant foliations tangent to uniformly expanding or uniformly contracting invariant subbundles: *assuming the derivative contracts E^1 uniformly, there exists a unique family $\mathcal{F}^s = \{\mathcal{F}^s(x) : x \in K\}$ of injectively C^r immersed submanifolds tangent to E^1 at every point of K , satisfying $f(\mathcal{F}^s(x)) = \mathcal{F}^s(f(x))$ for all $x \in K$, and which are uniformly contracted by forward iterates of f .* This is called *strong-stable foliation* of the diffeomorphism on K . Strong-unstable foliations are defined in the same way, tangent to the invariant subbundle E_k , when it is uniformly expanding.

As in the purely hyperbolic setting, a crucial ingredient in the ergodic theory of partially hyperbolic systems is the fact that strong-stable and strong-unstable foliations are absolutely continuous, if the system is at least twice differentiable.

Robustness and Partial Hyperbolicity

Partially hyperbolic systems have been studied since the 1970s, most notably by Brin, Pesin and Hirsch, Pugh, Shub. Over the last decade they attracted much attention as the key to characterizing robustness of the dynamics. More precisely, let Λ be a maximal invariant set of some diffeomorphism f :

$$\Lambda = \bigcap_{n \in \mathbb{Z}} f^n(U) \quad \text{for some neighborhood } U \text{ of } \Lambda.$$

The set Λ is *robust*, or *robustly transitive*, if its continuation $\Lambda_g = \bigcap_{n \in \mathbb{Z}} g^n(U)$ is transitive for all g in a neighborhood of f . There is a corresponding notion for flows.

As we have already seen, hyperbolic basic pieces are robust. In the 1970s, Mañé observed that the converse is also true when M is a surface, but not anymore if the dimension of M is at least 3. Counter-examples in dimension 4 had been given before by Shub. A series of results of Bonatti, Díaz, Pujals, Ures in the 1990s clarified the situation in all dimensions: robust sets always admit some dominated splitting which is volume hyperbolic; in general, this splitting needs not be partially hyperbolic, except when the ambient manifold has dimension 3.

Lorenz-like Strange Attractors

Parallel results hold for flows on 3-dimensional manifolds. The main motivation are the so-called Lorenz-like strange attractors, inspired by the famous differential equations

$$\begin{aligned} \dot{x} &= -\sigma x + \sigma y & \sigma &= 10 \\ \dot{y} &= \rho x - y - xz & \rho &= 28 \\ \dot{z} &= xy - \beta z & \beta &= 8/3 \end{aligned} \quad (4)$$

introduced by E. N. Lorenz in the early 1960s. Numerical analysis of these equations led Lorenz to realize that sensitive dependence of trajectories on the initial conditions is ubiquitous among dynamical systems, even those with simple evolution laws.

The dynamical behavior of (4) was first interpreted by means of certain geometric models, proposed by Guckenheimer, Williams and Afraimovich, Bykov, Shil'nikov in the 1970s, where the presence of strange attractors, both sensitive and fractal, could be proved rigorously. It was much harder to prove that the original Eqs. (4) themselves have such an attractor. This was achieved just a few years ago, by Tucker, by means of a computer assisted rigorous argument.

An important point is that Lorenz-like attractors cannot be hyperbolic, because they contain an equilibrium point accumulated by regular orbits inside the attractor. Yet, these strange attractors are robust, in the sense we defined above. A mathematical theory of robustness for flows in 3-dimensional spaces was recently developed by Morales, Pacifico, and Pujals. In particular, this theory shows that uniformly hyperbolic attractors and Lorenz-like attractors are the only ones which are robust. Indeed, they prove that *any robust invariant set of a flow in dimension 3 is singular hyperbolic*. Moreover, *if the robust set contains equilibrium points then it must be either an attractor or a repeller*. A detailed presentation of this and related results is given in [1].

An invariant set Λ of a flow in dimension 3 is *singular hyperbolic* if it is a partially hyperbolic set with splitting $E^1 \oplus E^2$ such that the derivative is volume contracting along E^1 and volume expanding along E^2 . Notice that one of the subbundles E^1 or E^2 must be one-dimensional, and then the derivative is, actually, either norm contracting or norm expanding along this subbundle. Singular hyperbolic sets without equilibria are uniformly hyperbolic: the 2-dimensional invariant subbundle splits as the sum of the flow direction with a uniformly expanding or contracting one-dimensional invariant subbundle.

Non-Uniform Hyperbolicity – Linear Theory

In its linear form, the theory of non-uniform hyperbolicity goes back to Lyapunov, and is founded on the multiplicative ergodic theorem of Oseledets. Let us introduce the main ideas, whose thorough development can be found in e. g. [4,6,7].

The *Lyapunov exponents* of a sequence $\{A^n, n \geq 1\}$ of square matrices of dimension $d \geq 1$, are the values of

$$\lambda(v) = \limsup_{n \rightarrow \infty} \frac{1}{n} \log \|A^n \cdot v\| \quad (5)$$

over all non-zero vectors $v \in \mathbb{R}^d$. For completeness, set $\lambda(0) = -\infty$. It is easy to see that $\lambda(cv) = \lambda(v)$ and $\lambda(v + v') \leq \max\{\lambda(v), \lambda(v')\}$ for any non-zero scalar c and any vectors v, v' . It follows that, given any constant a , the set of vectors satisfying $\lambda(v) \leq a$ is a vector subspace. Consequently, there are at most d Lyapunov exponents, henceforth denoted by $\lambda_1 < \dots < \lambda_{k-1} < \lambda_k$, and there exists a filtration $F_0 \subset F_1 \subset \dots \subset F_{k-1} \subset F_k = \mathbb{R}^d$ into vector subspaces, such that

$$\lambda(v) = \lambda_i \quad \text{for all } v \in F_i \setminus F_{i-1},$$

and every $i = 1, \dots, k$ (write $F_0 = \{0\}$). In particular, the largest exponent is given by

$$\lambda_k = \limsup_{n \rightarrow \infty} \frac{1}{n} \log \|A^n\|. \quad (6)$$

One calls $\dim F_i - \dim F_{i-1}$ the *multiplicity* of each Lyapunov exponent λ_i .

There are corresponding notions for continuous families of matrices $A^t, t \in (0, \infty)$, taking the limit as t goes to infinity in the relations (5) and (6).

Lyapunov Stability

Consider the linear differential equation

$$\dot{v}(t) = B(t) \cdot v(t), \quad (7)$$

where $B(t)$ is a bounded function with values in the space of $d \times d$ matrices, defined for all $t \in \mathbb{R}$. The theory of differential equations ensures that there exists a *fundamental matrix* $A^t, t \in \mathbb{R}$ such that

$$v(t) = A^t \cdot v_0$$

is the unique solution of (7) with initial condition $v(0) = v_0$.

If the Lyapunov exponents of the family $A^t, t > 0$ are all negative then the trivial solution $v(t) \equiv 0$ is asymptotically stable, and even exponentially stable. The stability

theorem of A. M. Lyapunov asserts that, under an additional regularity condition, stability is still valid for non-linear perturbations

$$w(t) = B(t) \cdot w + F(t, w),$$

with $\|F(t, w)\| \leq \text{const} \|w\|^{1+c}, c > 0$. That is, the trivial solution $w(t) \equiv 0$ is still exponentially asymptotically stable.

The regularity condition means, essentially, that the limit in (5) does exist, even if one replaces vectors v by elements $v_1 \wedge \dots \wedge v_l$ of any l th exterior power of $\mathbb{R}^d, 1 \leq l \leq d$. By definition, the norm of an l -vector $v_1 \wedge \dots \wedge v_l$ is the volume of the parallelepiped determined by the vectors v_1, \dots, v_l . This condition is usually tricky to check in specific situations. However, the multiplicative ergodic theorem of V. I. Oseledets asserts that, for very general matrix-valued stationary random processes, regularity is an almost sure property.

Multiplicative Ergodic Theorem

Let $f: M \rightarrow M$ be a measurable transformation, preserving some measure μ , and let $A: M \rightarrow \text{GL}(d, \mathbb{R})$ be any measurable function such that $\log \|A(x)\|$ is μ -integrable. The Oseledets theorem states that Lyapunov exponents exist for the sequence $A^n(x) = A(f^{n-1}(x)) \dots A(f(x))A(x)$ for μ -almost every $x \in M$. More precisely, for μ -almost every $x \in M$ there exists $k = k(x) \in \{1, \dots, d\}$, a filtration

$$F_x^0 \subset F_x^1 \subset \dots \subset F_x^{k-1} \subset F_x^k = \mathbb{R}^d,$$

and numbers $\lambda_1(x) < \dots < \lambda_k(x)$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|A^n(x) \cdot v\| = \lambda_i(x),$$

for all $v \in F_x^i \setminus F_x^{i-1}$ and $i \in \{1, \dots, k\}$. More generally, this conclusion holds for any vector bundle automorphism $\mathcal{V} \rightarrow \mathcal{V}$ over the transformation f , with $A_x: \mathcal{V}_x \rightarrow \mathcal{V}_{f(x)}$ denoting the action of the automorphism on the fiber of x .

The Lyapunov exponents $\lambda_i(x)$, and their number $k(x)$, are measurable functions of x and they are constant on orbits of the transformation f . In particular, if the measure μ is ergodic then k and the λ_i are constant on a full μ -measure set of points. The subspaces F_x^i also depend measurably on the point x and are invariant under the automorphism:

$$A(x) \cdot F_x^i = F_{f(x)}^i.$$

It is in the nature of things that, usually, these objects are *not* defined everywhere and they depend discontinuously on the base point x .

When the transformation f is invertible one obtains a stronger conclusion, by applying the previous result also to the inverse automorphism: assuming that $\log \|A(x)^{-1}\|$ is also in $L^1(\mu)$, one gets that there exists a decomposition

$$\mathcal{V}_x = E_x^1 \oplus \cdots \oplus E_x^k,$$

defined at almost every point and such that $A(x) \cdot E_x^i = E_{f(x)}^i$ and

$$\lim_{n \rightarrow \pm\infty} \frac{1}{n} \log \|A^n(x) \cdot v\| = \lambda_i(x),$$

for all $v \in E_x^i$ different from zero and all $i \in \{1, \dots, k\}$. These *Oseledets subspaces* E_x^i are related to the subspaces F_x^i through

$$F_x^j = \bigoplus_{i=1}^j E_x^i.$$

Hence, $\dim E_x^i = \dim F_x^i - \dim F_x^{i-1}$ is the multiplicity of the Lyapunov exponent $\lambda_i(x)$.

The angles between any two Oseledets subspaces decay sub-exponentially along orbits of f :

$$\lim_{n \rightarrow \pm\infty} \frac{1}{n} \log \angle \left(\bigoplus_{i \in I} E_{f^n(x)}^i, \bigoplus_{j \notin I} E_{f^n(x)}^j \right) = 0,$$

for any $I \subset \{1, \dots, k\}$ and almost every point. These facts imply the regularity condition mentioned previously and, in particular,

$$\lim_{n \rightarrow \pm\infty} \frac{1}{n} \log |\det A^n(x)| = \sum_{i=1}^k \lambda_i(x) \dim E_x^i.$$

Consequently, if $\det A(x) = 1$ at every point then the sum of all Lyapunov exponents, counted with multiplicity, is identically zero.

Non-Uniformly Hyperbolic Systems

The Oseledets theorem applies, in particular, when $f: M \rightarrow M$ is a C^1 diffeomorphism on some compact manifold and $A(x) = Df_x$. Notice that the integrability conditions are automatically satisfied, for any f -invariant probability measure μ , since the derivative of f and its inverse are bounded in norm.

Lyapunov exponents yield deep geometric information on the dynamics of the diffeomorphism, especially when they do not vanish. We call μ a *hyperbolic measure* if all Lyapunov exponents are non-zero at μ -almost every point. By *non-uniformly hyperbolic system* we shall mean a diffeomorphism $f: M \rightarrow M$ together with some invariant hyperbolic measure.

A theory initiated by Pesin provides fundamental geometric information on this class of systems, especially existence of stable and unstable manifolds at almost every point which form absolutely continuous invariant laminations. For most results, one needs the derivative Df to be Hölder continuous: there exists $c > 0$ such that

$$\|Df_x - Df_y\| \leq \text{const} \cdot d(x, y)^c.$$

These notions extend to the context of flows essentially without change, except that one disregards the invariant line bundle given by the flow direction (whose Lyapunov exponent is always zero). A detailed presentation can be found in e. g. [6].

Stable Manifolds

An essential tool is the existence of invariant families of local stable sets and local unstable sets, defined at μ -almost every point. Assume μ is a hyperbolic measure. Let E_x^u and E_x^s be the sums of all Oseledets subspaces corresponding to positive, respectively negative, Lyapunov exponents, and let $\tau_x > 0$ be a lower bound for the norm of every Lyapunov exponent at x .

Pesin's stable manifold theorem states that, for μ -almost every $x \in M$, there exists a C^1 embedded disk $W_{\text{loc}}^s(x)$ tangent to E_x^s at x and there exists $C_x > 0$ such that

$$\begin{aligned} \text{dist}(f^n(y), f^n(x)) &\leq C_x e^{-n\tau_x} \cdot \text{dist}(y, x) \\ &\text{for all } y \in W_{\text{loc}}^s(x). \end{aligned}$$

Moreover, the family $\{W_{\text{loc}}^s(x)\}$ is invariant, in the sense that $f(W_{\text{loc}}^s(x)) \subset W_{\text{loc}}^s(f(x))$ for μ -almost every x . Thus, one may define global stable manifolds

$$W^s(x) = \bigcup_{n=0}^{\infty} f^{-n}(W_{\text{loc}}^s(x)) \quad \text{for } \mu\text{-almost every } x.$$

In general, the local stable disks $W^s(x)$ depend only measurably on x . Another key difference with respect to the uniformly hyperbolic setting is that the numbers C_x and τ_x can not be taken independent of the point, in general. Likewise, one defines local and global unstable manifolds, tangent to E_x^u at almost every point. Most important for the applications, both foliations, stable and unstable, are absolutely continuous.

In the remaining sections we briefly present three major results in the theory of non-uniform hyperbolicity: the entropy formula, abundance of periodic orbits, and exact dimensionality of hyperbolic measures.

The Entropy Formula

The entropy of a partition \mathcal{P} of M is defined by

$$h_\mu(f, \mathcal{P}) = \lim_{n \rightarrow \infty} \frac{1}{n} H_\mu(\mathcal{P}^n),$$

where \mathcal{P}^n is the partition into sets of the form $P = P_0 \cap f^{-1}(P_1) \cap \dots \cap f^{-n}(P_n)$ with $P_j \in \mathcal{P}$ and

$$H_\mu(\mathcal{P}^n) = \sum_{P \in \mathcal{P}^n} -\mu(P) \log \mu(P).$$

The *Kolmogorov–Sinai entropy* $h_\mu(f)$ of the system is the supremum of $h_\mu(f, \mathcal{P})$ over all partitions \mathcal{P} with finite entropy. The Ruelle–Margulis inequality says that $h_\mu(f)$ is bounded above by the averaged sum of the positive Lyapunov exponents. A major result of the theorem, Pesin’s entropy formula, asserts that if the invariant measure μ is smooth (for instance, a volume element) then the entropy actually coincides with the averaged sum of the positive Lyapunov exponents

$$h_\mu(f) = \int \left(\sum_{j=1}^k \max\{0, \lambda_j\} \right) d\mu.$$

A complete characterization of the invariant measures for which the entropy formula is true was given by F. Ledrappier and L. S. Young.

Periodic Orbits and Entropy

It was proved by A. Katok that periodic motions are always dense in the support of any hyperbolic measure. More than that, assuming the measure is non-atomic, there exist Smale horseshoes H_n with topological entropy arbitrarily close to the entropy $h_\mu(f)$ of the system. In this context, the *topological entropy* $h(f, H_n)$ may be defined as the exponential rate of growth

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log \# \{x \in H_n : f^k(x) = x\}.$$

of the number of periodic points on H_n .

Dimension of Hyperbolic Measures

Another remarkable feature of hyperbolic measures is that they are *exact dimensional*: the pointwise dimension

$$d(x) = \lim_{r \rightarrow 0} \frac{\log \mu(B_r(x))}{\log r}$$

exists at almost every point, where $B_r(x)$ is the neighborhood of radius r around x . This fact was proved by L. Barreira, Ya. Pesin, and J. Schmeling. Note that this means that the measure $\mu(B_r(x))$ of neighborhoods scales as $r^{d(x)}$ when the radius r is small.

Future Directions

The theory of uniform hyperbolicity showed that dynamical systems with very complex behavior may be amenable to a very precise description of their evolution, especially in probabilistic terms. It was most successful in characterizing structural stability, and also established a paradigm of how general “chaotic” systems might be approached. A vast research program has been going on in the last couple of decades or so, to try and build such a global theory of complex dynamical evolution, where notions such as partial and non-uniform hyperbolicity play a central part. The reader is referred to the bibliography, especially the book [2] for a review of much recent progress.

Bibliography

1. Araujo V, Pacifico MJ (2007) Three Dimensional Flows. In: XXV Brazilian Mathematical Colloquium. IMPA, Rio de Janeiro
2. Bonatti C, Díaz LJ, Viana M (2005) Dynamics beyond uniform hyperbolicity. In: Encyclopaedia of Mathematical Sciences, vol 102. Springer, Berlin
3. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. In: Lecture Notes in Mathematics, vol 470. Springer, Berlin
4. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic theory. In: Grundlehren der Mathematischen Wissenschaften Fundamental Principles of Mathematical Sciences, vol 245. Springer, New York
5. Hirsch M, Pugh C, Shub M (1977) Invariant manifolds. In: Lecture Notes in Mathematics, vol 583. Springer, Berlin
6. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. Cambridge University Press, Cambridge
7. Mañé R (1987) Ergodic theory and differentiable dynamics. Springer, Berlin
8. Palis J, de Melo W (1982) Geometric theory of dynamical systems. An introduction. Springer, New York
9. Palis J, Takens F (1993) Hyperbolicity and sensitive-chaotic dynamics at homoclinic bifurcations. Cambridge University Press, Cambridge
10. Shub M (1987) Global stability of dynamical systems. Springer, New York