Technische Universität München
Zentrum Mathematik

Efficient approximation methods for the global long-term behavior of dynamical systems – Theory, algorithms and examples

Péter Koltai

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Anuschirawan Taraz
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Oliver Junge
	2.	Univ.-Prof. Dr. Michael Dellnitz, Universität Paderborn
	3.	Assoc. Prof. Gary Froyland, Univ. of New South Wales, Sydney/Australien (schriftliche Beurteilung)

Die Dissertation wurde am 19.05.2010 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 27.09.2010 angenommen.

Acknowledgements

For their assistance in the development of this thesis, many people deserve thanks.

First of all, I would like to thank Oliver Junge, my supervisor, for his guidance. I appreciated his continuous interest in my progress, friendly criticism, positive attitude, and always having time for encouraging talks as I was an immature student, just as I appreciate it now.

Special thanks goes to Gary Froyland for pointing out to me the importance of posing the right questions, and for inviting me to the UNSW; to Gero Friesecke for numerous interesting discussions and ideas; and to Folkmar Bornemann for an inspiring lecture on spectral methods.

I am grateful to the people in the TopMath program for setting up the framework which enables young students getting close to mathematical research.

The members of the research unit M3 at the TUM deserve mentioning for creating a pleasant atmosphere to work in.

I would also like to thank all those who contributed to this thesis in other ways, and were not named individually.

Contents

Chapter 1

Introduction and motivation for the thesis

Introduction to the problem. Processes in nature where motion or change of states is involved are mathematically modeled by dynamical systems. Their complexity ranges from the relatively simple motion of a pendulum under gravitational influence to e.g. the very complex processes in the atmosphere. Moreover, in a given context, particular aspects of the system under consideration are of interest. To understand the local behavior, one could ask *"Are there states which stay unchanged forever, and are they stable?"* or *"Is there a periodic motion?"*. For example, the vertically hanging pendulum is in a stable fixed state, and (unless there is external forcing) certain motions of the pendulum are periodic; but there is no stable weather like "eternal sunshine", and the rain is not falling "each Monday" either. This motivates a global analysis, where reasonable questions would be *"What is the probability that it will be warmer than* 24 $^{\circ}C$ *tomorrow at noon?"* or *"How often is it going to rain next month?"*. Questions like the latter one motivate us to understand the *long-term behavior of dynamical systems.*

Approaching these questions numerically by the direct simulation of a long trajectory works well for many systems, however, there are important applications where this method is not robust, or it is even computationally intractable. It is well known that the condition number of the flow arising from an ODE scales exponentially in time. Therefore, a trajectory obtained from a long simulation may show completely different behavior than *any* real trajectory of the system. There are results which remedy this fact. For example, if the system is stochastically stable [Kif86, Zee88, Ben93], and

one can view numerical errors as small random perturbations of the system, then a computed trajectory will exhibit similar statistical properties as the original one (cf. "shadowing" [Guc83]; see also [Del99, Kif86]).[1] Until now, not many systems have been proven to be stochastically stable (e.g. Axiom A diffeomorphisms, see [Kif86]), and in the corresponding proofs there are strong assumptions on the perturbation as well. Also in favor of simulation is the fact that using symplectic integrators for Hamiltonian systems allows one to interpret a numerically computed trajectory as a real trajectory of a slightly perturbed Hamiltonian system [Hai96, Hai06]. On the other hand, certain Hamiltonian systems arising from molecular dynamics elucidate a further problem related to direct simulation: The chemical properties of many biomolecules depend on their *conformation* [Zho98]. A conformation is the large-scale geometrical "shape" of the molecule which persists for a large time compared with the timescale of the motion of a single atom in the molecule. Thus, conformational changes occur at much slower timescales compared to the elementary frequencies of the system. The typical scale difference for folding transitions in proteins ranges between 10^8 and 10^{16}. Clearly, the statistical analysis of such systems is not accessible via direct trajectory simulation, since the time step for any numerical integrator needs to be chosen smaller than the period of the fastest oscillation.

We can generalize the notion of conformations for arbitrary dynamical systems. Suppose, that there are two or more "macroscopic states", i.e. subsets in phase space in which trajectories tend to stay long before switching to another. These sets are called *almost invariant* [Del99]. They are a curse for methods which try to extract long-term statistical properties from trajectories of finite length, since they can "trap" orbits for a long time, and regions in the phase space may stay unvisited, if the length of the simulation is not sufficient. Since they govern the dynamics of a system over a large timescale, one is also interested in finding these almost invariant sets, and to quantize "how much almost invariant" they are.

Fortunately, there are other (mathematical) objects which allow the characterization of the long-term dynamical behavior *without* the simulation of long trajectories. Ergodicity theorems relate the temporal averages of observables over particular trajectories to spatial averages with respect to *invariant measures* or *invariant densities*. The

[1] Cases are known, where numerical errors are *not* random [Hig02], and the above reasoning does not hold.

latter turn out to be eigenfunctions at eigenvalue one of so-called *transfer operators* (cf. Section 2.2). Also, we will see that information about almost invariance can be drawn from eigenfunctions at eigenvalues near one of the same operator (cf. Section 2.2.2). Thus, we can approach the problem also by solving an *infinite dimensional eigenvalue problem in L^p*. Apart from special examples, typically no analytical solution can be obtained, hence we are led to the challenge of designing *efficient* numerical algorithms for the eigenfunction approximation of transfer operators at the desired eigenvalues. Such approaches using transfer operators are applied in many fields, e.g. in molecular dynamics [Deu96, Deu01, Deu04a], astrodynamics [Del05], and oceanography [Fro07, Del09].

So far, the method dedicated to *Ulam* [Ula60] received the most attention, due to its robustness and the ability to interpret the resulting discretization as a Markov chain related to the dynamical system (these two properties may have a lot in common). It considers the Galerkin projection of the transfer operator onto a space of piecewise constant functions, and uses the eigenfunctions of the discretized operator (also called the transition matrix) as approximation to the real ones. Despite the rather slow convergence (piecewise constant interpolation does not allow faster than linear convergence, in general; however, not even this can be achieved in most cases [Bos01]), and the unpleasant representation of the transition matrix entries (they are integrals of non-continuous functions, cf. Section 2.3), the method has justified its usage by performing well in various applications. It is also worth to note that the convergence of Ulam's method is still an open question for most systems except some specific ones (cf. Section 2.3).

This thesis. The aim of this thesis is *to design algorithms based on transfer operator methods which enable an efficient computation of the objects describing the long-term dynamical behavior — the invariant density and almost invariant sets.* A particular emphasis lies on the theoretical analysis of these methods, regarding their efficiency and convergence properties.

Chapters 3–6 are independent from each other, and can be read separately. At the beginning of each chapter we motivate the work presented in there, and give a brief outline. At its end, conclusions are drawn, and possible further developments are discussed. For a deeper introduction to the chapters, we refer the reader to the particular chapter itself. Here, we restrict ourselves to a brief overview.

Chapter 2 gives a background review on dynamical systems, transfer operators, Ulam's method, and classical molecular dynamics.

Chapter 3 investigates the intriguing idea whether discretizations of a transfer operator can be viewed as small random perturbations of the underlying dynamical system. This would allow a convergence analysis by the means of stochastic stability. Our result states that, using Galerkin projections, Ulam's method is the only one with this property. Unfortunately, the random perturbation equivalent with Ulam's method does not meet all the assumptions under which stochastic stability can currently be shown.

Chapter 4 presents a discretization method (the Sparse Ulam method), using *sparse grids*, for arbitrary systems on a d dimensional hyperrectangle, and considers the question if one can defeat the curse of dimension from which Ulam's method suffers. A detailed numerical analysis of the Sparse Ulam method, and a comparison with Ulam's method is given.

Chapter 5 discusses two methods for approximating the eigenfunctions of the transfer operator (semigroup) for time-continuous systems by discretizing the corresponding *infinitesimal generator*. It enables to omit expensive time-integration of the underlying ODE, which results in a computational speed-up of at least a factor ~ 10 compared to standard methods. The methods (a robust cell-to-cell approach, and a spectral method approach for smooth problems) are tested on various examples.

Chapter 6 has the main focus on molecular dynamics, and analyzes if there are suitable low-dimensional systems, obtained by *mean field theory*, able to describe the conformation changes in chain molecules. The theoretical framework is developed on time-discrete systems. Numerical experiments help to understand the behavior of the method for weakly coupled systems. Afterwards, the method is extended to time-continuous systems, and presented on a model of n-butane.

Chapter 2

Background

2.1 Dynamical systems

2.1.1 Time-discrete dynamical systems

Given a metric space (X, d) and the map $S : X \to X$, the pair (X, S) is a *discrete-time dynamical system*. The set X is called the *state space*, while one refers to S as the *dynamics*. It models a system with motion; being at an instance in state x, in the next instance the system is going to be in state $S(x)$. For a $x \in X$ the elements of the set $\{x, S(x), S^2(x), \dots\}$ are called iterates of x and the whole set is the (forward) orbit starting in x.

Some subsets of X may be emphasized by the dynamics. Such are *invariant sets*. A set $A \subset X$ is invariant if $S^{-1}(A) = A$. The dynamics on A is independent of $X \setminus A$ and $(A, S\,|_A)$ is a dynamical system as well. Take a system with the invariant set A and introduce some other dynamics $\widetilde{S}$, such that $d(S(x), \widetilde{S}(x))$ is small for all $x \in X$. In this sense the dynamics $\widetilde{S}$ is said to be near S. We cannot expect anymore that all orbits starting in A stay in A forever, nevertheless we expect the majority of the orbits to stay in A for many iterates before leaving it. This motivates the notion of *almost invariance*. We would also like to measure "how invariant" the set A remained. For this, assume that the phase space can be extended to a measure space $(X, \mathcal{B}, \mu)$, where $\mathcal{B}$ denotes the Borel-sigma algebra on X and μ is a finite measure; further let the map S be Borel measurable. The set A with $\mu(A) > 0$ is called *ρ-almost-invariant* w.r.t. μ, if

$$\frac{\mu(S^{-1}(A) \cap A)}{\mu(A)} = \rho. \tag{2.1}$$

In other words: choose $x \in A$ at random according to the distribution $\mu(\cdot)/\mu(A)$, then the probability that $S(x) \in A$ is ρ.

Another interesting behavior is the accumulation of states around some subset of the phase space. We call a compact set $A \subset X$ an *attractor* if the iterates of every bounded set $B \subset X$ are uniformly tending to A; i.e. $d(S^n(B), A) \to 0$ as $n \to \infty$.[1] Sometimes not all states in X tend to A. Nevertheless there can be local attractors which dominate the asymptotic behavior of a subset of the state space. The attractor A_Y relative to $Y \subset X$ is given by $A_Y = \bigcap_{n\in\mathbb{N}} S^n(Y)$. The *domain of attraction* of a relative attractor A is defined as $D := \{x \in X \mid d(S^n(x), A) \to 0 \text{ as } n \to \infty\}$.

A map S defines the successive state always precisely. However, sometimes the precise dynamics depend on unknown circumstances, which one would like to model by random variables. This leads to non-deterministic dynamics, which are given by stochastic transition functions.

Definition 2.1. *Let $(X, \mathcal{B}, \mu)$ be a probability space. The function $p : X \times \mathcal{B} \to [0, 1]$ is a stochastic transition function if*

(a) $p(x, \cdot)$ is a probability measure for all $x \in X$, and

(b) $p(\cdot, A)$ is measurable for all $A \in \mathcal{B}$.

Unless stated otherwise, for a compact state space X we have $\mu = m/m(X)$, where m denotes the Lebesgue measure.
Setting $p^{(1)}(x, A) = p(x, A)$, the i-step transition function for $i = 1, 2, \ldots$ is defined by

$$p^{(i+1)}(x, A) = \int_X p^{(i)}(y, A)\ p(x, \mathrm{d}y).$$

If $p(x, \cdot)$ is absolutely continuous to μ for all $x \in X$, the Radon–Nikodym theorem implies the existence of a nonnegative function $q : X \times X \to \mathbb{R}$ with $q(x, \cdot) \in L^1(X, \mu)$ and

$$p(x, A) = \int_A q(x, y)\ \mathrm{d}\mu(y).$$

The function q is called the (stochastic) transition density (function).

The intuition behind the transition function is that if we are in state x, the probability of being in A in the next instance is $p(x, A)$. If we set $p(x, \cdot) = \delta_{S(x)}(\cdot)$, where $\delta_{S(x)}$ denotes the Dirac measure centered in $S(x)$, we obtain the deterministic dynamics.

[1] For $x \in X$ and $A, B \subset X$ we define $d(x, A) = \inf_{y\in A} d(x, y)$ and $d(A, B) = \max\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\}$.

Example 2.2. One could model unknown perturbations of the deterministic dynamics S as follows. Assuming, the iterate of $x \in X = \mathbb{R}^d$ is near $S(x)$ and no further specification of the perturbation is known, we set $\varepsilon > 0$ as the perturbation size and distribute the image point uniformly in an ε-ball around $S(x)$. The transition density will be

$$q(x, y) = \frac{1}{\varepsilon^d m(B)} \chi_B \left(\frac{1}{\varepsilon} (y - S(x)) \right), \tag{2.2}$$

where B is the unit ball in $\mathbb{R}^d$ centered in zero and χ_B the characteristic function of B [Del99].

An analogous definition of invariant sets, as we had them for deterministic systems does not make sense. Because of the uncertainty we cannot expect that only the points of A may be mapped into A. Weakening the claim of invariance to forward invariance gives an alternative definition. A set $A \subset X$ is called invariant w.r.t. p if all points in A are mapped into A almost surely (a.s.); i.e. $A \subset \{x \in X \mid p(x, A) = 1\}$. A set A satisfying $\lim_{i\to\infty} p^{(i)}(x, A) = 0$ for all $x \in X$ is called transient. A generalization of almost invariance is straightforward. A set $A \subset X$ is ρ-almost-invariant w.r.t. the measure μ if $\mu(A) > 0$ and

$$\int_X p(x, A) \mathrm{d}\mu(x) = \rho\mu(A). \tag{2.3}$$

Indeed, this is a generalization, since with $p(x, \cdot) = \delta_{S(x)}(\cdot)$ we have (2.1).

2.1.2 Time-continuous dynamical systems

Time-continuous dynamical systems arise as flows of ordinary differential equations (ODEs). Let the vector field $v : X \to \mathbb{R}^d$ be given.[1] We assume v to be at least once continuously differentiable. Let S^t denote the solution operator (flow) of the ODE $\dot{x}(t) := dx(t)/dt = v(x(t))$. All objects and properties introduced for time-discrete systems are carried over one-to-one or with slight modifications. A set A is invariant if $A = S^{-t}(A)$ for all $t \geq 0$. The almost invariance ratio ρ of a set A will depend on t.

The theory of non-deterministic systems needs a more advanced probability theory. Some tools required for this will not be used in this thesis anymore. Thus, instead of introducing them rigorously, we aim to show the intuition behind the objects and

[1] We think of the phase space X as a subset of $\mathbb{R}^d$, $\mathbb{T}^d$ or $\mathbb{R}^{d-k} \times \mathbb{T}^k$, where $\mathbb{T}$ is the one dimensional unit torus.

for a precise introduction we refer to the books [Las94] and [Pav08]. We will consider stochastically perturbed flows, where the perturbation is going to be a *Brownian motion* (or *Wiener process*).

A stochastic process is a family of random variables $\{\xi(t)\}_{t\geq 0}$. It is called continuous, if its sample paths are almost surely continuous functions in t. The one dimensional (normalized) Brownian motion is a continuous stochastic process $\{W(t)\}_{t\geq 0}$ satisfying

1. $W(0) = 0$, and
2. for every s, t, $0 \leq s < t$, the random variable $W(t) - W(s)$ has the Gaussian density
$$\frac{1}{\sqrt{2\pi(t-s)}} \exp\left(\frac{-x^2}{2(t-s)}\right).$$

A multidimensional Brownian motion is given by $W(t) = (c_1 W_1(t), \ldots, c_d W_d(t))$, where the $W_i(t)$ are independent one dimensional Brownian motions and the $c_i \geq 0$. A noteworthy way of thinking of the Brownian motion is presented in [Nor97]. Consider a random walk on an equispaced grid. If we let the jump distance and the time step go to zero between two consecutive jumps (while they satisfy a fixed relation), the limiting process can be viewed as the Brownian motion. This also helps to understand that the sample paths of a Brownian motion are almost surely not differentiable w.r.t. time at any point.

We define the stochastically perturbed dynamics by the stochastic differential equation (SDE)
$$\dot{x} = v(x) + \varepsilon\xi, \qquad x(0) = x_0, \tag{2.4}$$
where $\varepsilon > 0$ and ξ is a random variable given by $\xi = \dot{W}$. As mentioned above, $\dot{W}$ almost surely does not exist at all, hence this is only a convenient formal notation for the "vector field" of a flow perturbed by (scaled) Brownian motion.[1] The stochastical term is also called *diffusion*, while v is called the *drift*. The solution of such a SDE is the stochastic process $\{x(t)\}_{t\geq 0}$.

The definitions of the dynamical objects and properties introduced so far are carried over from the non-deterministic case just as they did for the deterministic case. However, the diffusion is a rather special random perturbation. There will not exist any

[1] The mathematically correct notation would be an integral equation including stochastic integrals; see references above.

invariant set (not even forward invariant) for (2.4), since for times large enough there will be a nonzero probability of being anywhere in the phase space — independent of the starting position; see Theorem 2.11 below. Similarly, if A was an attractor of the system defined by $\dot{x} = v(x)$, we may only expect for the system defined by (2.4) that A is a region where the system is with high probability, if ε is small enough.

Unlike in the other cases, we still miss a characterization of the dynamics for non-deterministic time-continuous systems. It would be desirable to have the distributions of the solution random variables, $x(t)$. Since the next section is devoted to the statistical properties of dynamical systems, we discuss this issue there.

2.2 Transfer operators

Non-deterministic systems need a probabilistic treatment anyway, but we may also gain a deeper insight into deterministic systems by exploring their statistical properties. One of the main benefits is that the theory gives a characterization of the long-term behavior of dynamical systems, without involving long orbits. This is a desirable property for designing numerical methods, since long trajectory simulations are computationally intractable if iterating is an ill-conditioned problem.

2.2.1 Invariant measures and ergodicity

Let X be a metric space, $\mathcal{B}$ the Borel-σ-algebra and $S : X \to X$ a nonsingular transformation.[1] Further let $\mathcal{M}$ denote the space of all finite signed measures on $(X, \mathcal{B})$. We examine the action of the dynamics on distributions. For this, draw $x \in X$ at random according to the probability distribution μ. Then

$$\text{Prob}(S(x) \in A) = \text{Prob}(x \in S^{-1}(A)) = \mu(S^{-1}(A)) \quad \forall A \in \mathcal{B},$$

and hence $S(x)$ is distributed according to $\mu \circ S^{-1}$. The operator $\mathcal{P} : \mathcal{M} \to \mathcal{M}$, defined by

$$\mathcal{P}\mu(A) = \mu(S^{-1}(A)) \qquad \forall A \in \mathcal{B}, \tag{2.5}$$

is called the *Frobenius–Perron operator* (FPO) or the *transfer operator*. Probability measures which do not change under the dynamics, i.e. $\mathcal{P}\mu = \mu$ holds, are called

[1] The measurable transformation S is called nonsingular if $m(A) = 0$ implies $m(S^{-1}(A)) = 0$.

invariant. If the dynamics are irreducible w.r.t. the invariant measure μ in the sense that all invariant sets A satisfy $\mu(A) = 0$ or $\mu(A) = 1$, then μ is called *ergodic* (w.r.t. S). Ergodic measures play an important role in the long-term behavior of the system:

Theorem 2.3 (Birkhoff ergodic theorem [Bir31])**.** *Let μ be an ergodic measure. Then, for any $\phi \in L^1(\mu)$, the average of the observable ϕ along an orbit of S is equal almost everywhere to the average of f w.r.t. μ; i.e.*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \phi(S^k(x)) = \int_X \phi \mathrm{d}\mu \qquad \mu\text{-a.e.} \tag{2.6}$$

As an example, by setting $\phi = \chi_A$ we obtain the relative frequency of an orbit visiting A.

We define the change of observables under the dynamics. From now on, if not stated elsewise, $L^p = L^p(X, m)$. The operator $\mathcal{U} : L^\infty \to L^\infty$ defined by

$$\mathcal{U}f(x) = f\left(S\left(x\right)\right) \tag{2.7}$$

is called the *Koopman operator* w.r.t. S. It is closely related to the Frobenius–Perron operator, as we will see later on. Also, ergodicity may be characterized by means of the Koopman operator, see Theorem 4.2.1 [Las94].

Theorem 2.4. *The measure μ is ergodic if and only if all measurable functions f satisfying $\mathcal{U}f = f$ μ-almost-everywhere are constant functions.*

In cases where the ergodic measure is not absolutely continuous to m it could happen that (2.6) does not give any "physically relevant" information. For this, the notion of physical measures was introduced; see [You02]. We call an ergodic measure μ physical measure, if (2.6) holds for all $\phi \in C^0(X)$ and $x \in U \subset X$ with $m(U) > 0$. One can show, that if an ergodic measure μ is absolutely continuous w.r.t. m, then μ is a physical measure. This motivates us to make our considerations on the level of densities, or more generally on functions in L^1. By the nonsingularity of S and the Radon–Nikodym theorem one can define the FPO via (2.5) also on L^1, see [Las94].

Proposition 2.5. *Given a nonsingular transformation $S : X \to X$, the Frobenius–Perron operator $\mathcal{P} : L^1 \to L^1$ is given uniquely by*

$$\int_A \mathcal{P}u \ \mathrm{d}m = \int_{S^{-1}(A)} u \ \mathrm{d}m \qquad \forall A \in \mathcal{B}. \tag{2.8}$$

If, in addition, S is differentiable up to a set of measure zero, we have

$$\mathcal{P}u(x) = \sum_{y \in S^{-1}(x)} \frac{u(y)}{|\det(DS(y))|}. \tag{2.9}$$

The density of an absolutely continuous invariant measure is called the *invariant density.*

Remarks 2.6. We note some properties of the FPO:

(a) The FPO is the adjoint of the Koopman operator; i.e. it holds for all $u \in L^1$ and $f \in L^\infty$ that

$$\int_X \mathcal{P}u \; f \; \mathrm{d}m = \int_X u \; \mathcal{U}f \; \mathrm{d}m.$$

(b) The FPO is a *Markov operator*, because it is a linear operator with $\mathcal{P}u \geq 0$ and $\|\mathcal{P}u\|_{L^1} = \|u\|_{L^1}$ for all $u \in L^1$ with $u \geq 0$.

(c) By (b), $\|\mathcal{P}u\|_{L^1} \leq \|u\|_{L^1}$ for all $u \in L^1$, thus the spectrum of $\mathcal{P}$ lies in the unit disk.

We may also define the FPO $\mathcal{P} : \mathcal{M} \to \mathcal{M}$ associated with stochastic transition functions. It is given by

$$\mathcal{P}\mu(A) = \int_X p(x, A) \; \mathrm{d}\mu(x) \qquad \forall A \in \mathcal{B}. \tag{2.10}$$

If the transition function has a transition density q, we can define the FPO $\mathcal{P} : L^1 \to L^1$ associated with transition densities q.[1] From (2.10) we have

$$\mathcal{P}u(y) = \int_X q(x, y)u(x) \; \mathrm{d}m. \tag{2.11}$$

A measure (or a density) is called invariant, if it is a fixed point of $\mathcal{P}$. Following ergodic theorem for transition densities can be found in [Doo60].

Theorem 2.7. *Let p be a transition function with transition density function q. Assume that q is bounded on $X \times X$. Then X can be decomposed into a finite number of disjoint invariant sets $E_1, E_2, \ldots, E_k$ and a transient set $F = X \setminus \bigcup_{j=1}^{k} E_j$ such that for*

[1] We just write $\mathcal{P}$, if it is clear what the FPO is associated with. Otherwise, the notation $\mathcal{P}_S$ and $\mathcal{P}_q$ ($\mathcal{P}_p$) should make it clear.

E_j there is a unique probability measure μ_j (called ergodic measure) with $\mu_j(E_j) = 1$ and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} p^{(i)}(x, A) = \mu_j(A) \quad \textit{for all } A \in \mathcal{B} \textit{ and } x \in E_j.$$

Furthermore, every invariant measure of p is a convex combination of the μ_j. Finally, the μ_j are absolutely continuous to m.

The evolution of observables needs an appropriate generalization for the non-deterministic case. Given the current state, the next state is a random variable and we can merely give the expectation value of an observable w.r.t. its distribution. The Koopman operator is defined by

$$\mathcal{U}f(x) = \int_X f(y)p(x, \mathrm{d}y) = \int_X f(y)q(x, y)\ \mathrm{d}y,$$

for $f \in L^\infty$. One can see easily that $\mathcal{U}$ and $\mathcal{P} : L^1 \to L^1$ are adjoint.

For deterministic continuous-time systems S^t, the transfer operator $\mathcal{P}^t : L^1 \to L^1$ (and $\mathcal{P}^t : \mathcal{M} \to \mathcal{M}$ as well) is time-dependent, and an analogous definition to (2.8) is possible. Moreover, since the flow S^t of an autonomous system is a diffeomorphism for all $t \in \mathbb{R}$ (provided the right hand side v is smooth enough), we can give the FPO in an explicit form equivalent to (2.9):

$$\mathcal{P}^t u(x) = u\left(S^{-t}(x)\right) \left|\det\left(DS^{-t}(x)\right)\right|. \tag{2.12}$$

The Koopman operator $\mathcal{U}^t : L^\infty \to L^\infty$ is given by $\mathcal{U}^t f(x) = f(S^t(x))$. A density u is called invariant, if $\mathcal{P}^t u = u$ for all $t \geq 0$. The ergodicity is defined just as in the discrete time case. The ergodic theorem can be derived from Theorem 2.3, see Theorem 7.3.1 in [Las94].

Corollary 2.8. *Let μ be an ergodic measure w.r.t. S^t and let $\phi \in L^1$. Then*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \phi\left(S^t(x)\right) \mathrm{d}t = \int_X \phi\ \mathrm{d}\mu$$

for all $x \in X$ except for a set of μ-measure zero.

Assume, that the solution of the SDE (2.4), the random variable $x(t)$, has the density function $u : [0, \infty) \times X \to [0, \infty]$; i.e.

$$\mathrm{Prob}\left(x(t) \in A\right) = \int_A u(t, x)\ \mathrm{d}x.$$

There is no explicit representation of u in general, however, the following characterization is very useful. It summarizes results from Chapter 11 in [Las94].

Theorem 2.9 (Fokker–Planck equation)**.** *Under some regularity assumptions on v, the function u satisfies the so-called Fokker–Planck equation (or Kolmogorov forward eq.),*

$$\partial_t u = \frac{\varepsilon^2}{2}\Delta u - \operatorname{div}(uv) \qquad \text{for } t > 0, x \in X.^1 \tag{2.13}$$

Posing some further growth conditions on v, (2.13) with the initial condition $u(0,\cdot) = f \in L^1$ has a unique (generalized) solution, which is the density of $x(t)$, where $x(t)$ is the solution of (2.4) with $x(0)$ being a random variable with density f.
Thus, the the FPO $\mathcal{P}^t$ is the solution operator of the Fokker–Planck equation.

Remark 2.10. If the phase space X is compact and $v \in C^3(X, \mathbb{R}^d)$, the regularity and growth conditions of Theorem 2.9 are satisfied.

Similar statements hold for the Koopman operator as well: $\mathcal{U}^t$ is the solution operator of the partial differential equation (PDE)

$$\partial_t u = \frac{\varepsilon^2}{2}\Delta u + \nabla u \cdot v, \tag{2.14}$$

also called as *Kolmogorov backward* equation. Note, that the operators $\mathcal{L}$ and $\mathcal{L}^*$, where $\mathcal{L}u = \frac{\varepsilon^2}{2}\Delta u + \nabla u \cdot v$ and $\mathcal{L}^* u = \frac{\varepsilon^2}{2}\Delta - \operatorname{div}(uv)$, are adjoint on suitable spaces, just as $\mathcal{U}^t$ and $\mathcal{P}^t$.

The following results are derived easily from Theorem 6.16 in [Pav08]. The null space of an operator is denoted by $\mathcal{N}$.

Theorem 2.11. *Let $X = \mathbb{T}^d$. Then the following hold:*

(a) $\mathcal{N}(\mathcal{L}) = \operatorname{span}\{1\}$;

(b) there exists a unique invariant density u with $\mathcal{N}(\mathcal{L}^*) = \operatorname{span}\{u\}$, $\inf_{x\in X} u(x) > 0$,

(c) the spectrum of $\mathcal{L}$ and $\mathcal{L}^*$ lie strictly in the left half-plane, except the simple eigenvalue 0, and the spectrum of $\mathcal{U}$ and $\mathcal{P}$ lie strictly in the unit disk, except the simple eigenvalue 1;

(d) constants $C, \lambda > 0$ exists such that for any $h \in L^1$ with $\|h\|_{L^1} = 1$ one has

$$\left\|\mathcal{P}^t h - u\right\|_{L^1} \le Ce^{-\lambda t} \quad \forall t \ge 0;$$

[1] Here and in the following, ∂_t defines the derivative w.r.t. t, $\Delta = \partial^2_{x_1} + \ldots + \partial^2_{x_d}$ is the Laplace operator and $\operatorname{div}(\cdot)$ stands for the divergence operator. ∇u denotes the gradient of u.

(e) for all $\phi \in C^0$

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \phi(x(t))\mathrm{d}t = \int_X \phi u,$$

for all $x(0) = x_0$.

This theorem shows the big influence of the diffusion on the dynamics. There will be a unique invariant density, which is uniformly positive everywhere; i.e. by the diffusion every trajectory samples the whole phase space, by ergodicity, property (e). Compare property (a) with Theorem 2.4 and property (c) with Section 5.2.

2.2.2 Almost invariance and the spectrum of the transfer operator

The previous section showed, that the eigenfunction at the eigenvalue 1 (the invariant density) of the FPO tells us about the long-term dynamical behavior. We will see, how the other eigenfunctions at eigenvalues close to 1 connect to almost invariance.

The considerations here and the next result can be found in [Del99]. Let $\mathcal{P}$ be the transfer operator for a discrete-time system or $\mathcal{P} = \mathcal{P}^T$ for a continuous-time system with some fixed time $T > 0$. Suppose $\lambda < 1$ is a real eigenvalue of $\mathcal{P}$ with the real signed eigenmeasure ν. Then $\nu(X) = 0$. If ν is scaled so that $|\nu|$ is a probability measure, there exists a set $A \in \mathcal{B}$, such that $\nu(A) = 1/2$ and $\nu(X \setminus A) = -1/2$ by the Hahn-decomposition. Then, $\nu = |\nu|$ on A and $\nu = -|\nu|$ on $X \setminus A$. We have

Theorem 2.12 (Proposition 5.7 [Del99])**.** *Suppose that* ν *is scaled so that* $|\nu|$ *is a probability measure, and let* $A \subset X$ *be a set with* $\nu(A) = 1/2$*. Then*

$$\rho_1 + \rho_2 = \lambda + 1, \tag{2.15}$$

if A *is* ρ_1*-almost invariant and* $X \setminus A$ *is* ρ_2*-almost invariant w.r.t.* $|\nu|$*.*

Note, that (2.15) implies $\rho_1, \rho_2 > \lambda$, i.e. the eigenvalue is a lower estimate for the almost invariance w.r.t. $|\nu|$. The almost invariant sets are given as the supports of the positive and negative part of the measure.

Concerning the previous result, two things are unsatisfactory. First, the almost invariance is given w.r.t. the measure $|\nu|$, and there is no evidence, in general, if this is a physically relevant information. Second, if there are more than two almost invariant sets, it is not obvious how to extract them from the information given by the eigenpairs

with eigenvalues near 1. However, results exist on bounding almost invariance ratios in terms of transfer operator eigenvalues [Hui06].

An option for tackling these problems for conformation analysis in molecular dynamics is introduced on a solid mathematical basis in [Deu04b]. Similar ideas appeared in [Gav98, Gav06]. The considerations have been made for dynamical systems with finite state space (i.e. Markov chains).

Let $T \in \mathbb{R}^{n\times n}$ be the transition matrix of the Markov chain[1] on $\Omega = \{1, 2, \ldots, n\}$, i.e. $T_{ij} = \text{Prob}(j \to i)$. As T is a column stochastic matrix (and the FPO of the finite state dynamical system), it holds $e^\top T = e^\top$, with $e^\top = (1, \ldots, 1)$, and there is an invariant distribution $\pi \geq 0$ (componentwise) with $T\pi = \pi$. Assume, that $\pi > 0$ and that T is reversible, i.e. T is symmetrical w.r.t. the scalar product $\langle \cdot, \cdot \rangle_\pi = \langle \cdot, \text{diag}(\pi) \cdot \rangle$ (the discretization of the spatial transfer operator of Schütte (cf. Section 2.4.1) satisfies this property).

Let us consider uncoupled Markov chains first. Assume, there exists a disjoint partition $\Omega = \Omega_1 \cup \ldots \cup \Omega_k$, where the Ω_i are invariant, therefore T is block diagonal with the blocks T_i being individual stochastic matrices. Let χ_i be the characteristic vector of the set Ω_i, i.e.

$$\left(\chi_i\right)_j = \begin{cases} 1, & j \in \Omega_i, \\ 0, & \text{otherwise.} \end{cases}$$

Assuming that all T_i are irreducible, the left eigenspace of T at the eigenvalue 1 is spanned by the χ_i. We can interpret these vectors as an indicator, to which extent a state j belongs to the invariant set Ω_i. Here, the entries are either 1 or 0, but they will take values in $[0, 1]$ as almost invariance enters the stage. For now, assume, there are k linearly independent left eigenvectors, $X_1, \ldots, X_k$, of T given. We wish to compute the invariant sets, by finding the vectors χ_i. Hence, we search for the linear transformation $A \in \mathbb{R}^{k\times k}$, such that $\chi = XA$, where $\chi = (\chi_i, \ldots, \chi_k) \in \mathbb{R}^{n\times k}$ and $X = (X_1, \ldots, X_k) \in \mathbb{R}^{n\times k}$, the columnwise composition of the vectors to a matrix. The task is easy, since the X_i take at most k distinct values. Note, that if we plot the vectors $\left(\left(\chi_1\right)_j, \ldots, \left(\chi_k\right)_j\right)$ in $\mathbb{R}^k$ for all j, we get points in the vertices of the $(k-1)$-simplex σ_{k-1} with the unit canonical vectors as vertices. Doing the same with X gives the vertices of the linearly transformed simplex. Hence the linear transformation can be read from figure 2.1: the ith row of A^{-1} is the ith vertex of the latter simplex.

[1] See eg. [Nor97] for an introduction on the basic theory of Markov chains.

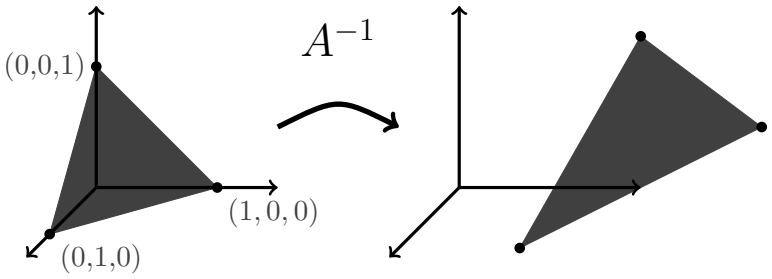


Figure 2.1: The linear transformation between the simplices - In the uncoupled case all points lie in the vertices, coupling makes them spread out.

Perturb the transition matrix T to obtain $\widetilde{T}(\varepsilon)$, an irreducible stochastic matrix[1]. Choose the perturbation in such a way, that it has the eigenvalues

$$\widetilde{\lambda}_1 = 1, \quad \widetilde{\lambda}_2 = 1 - \varepsilon, \quad \widetilde{\lambda}_3 = 1 - \mathcal{O}(\varepsilon), \quad \ldots, \quad \widetilde{\lambda}_k = 1 - \mathcal{O}(\varepsilon).$$

The eigenvectors at the first k eigenvalues perturb to $\widetilde{X}_1, \ldots, \widetilde{X}_k$, and we wish to compute the perturbed analogons of the χ_i, which we denote by $\widetilde{\chi}_i$. These characterize the almost invariant sets — the "leftovers" of the Ω_i. We do not aim a strict separation between the almost invariant sets, but think of the $(\widetilde{\chi}_i)_j$ as of the extent, that a given $j \in \Omega$ belongs to the ith almost invariant set, or ith *macroscopic* state. For this, it is natural to claim $\widetilde{\chi} \geq 0$ and $\sum_{i=1}^{k} (\widetilde{\chi}_i)_j = 1$ for all $j \in \{1, \ldots, n\}$. Again, we search for $\widetilde{A}$, such that $\widetilde{\chi} = \widetilde{X}\widetilde{A}$. Since the system has been perturbed, the points $((X_1)_j, \ldots, (X_k)_j) \in \mathbb{R}^k$ do not lie *in* the vertices of a simplex, but spread out, the same with $\widetilde{\chi}$. Hence, the transformation $\widetilde{A}$ will be defined by one simplex, which encloses the points $((X_1)_j, \ldots, (X_k)_j)$.

Theorem 2.13 (Theorem 2.1 [Deu04b])**.** *Three of the following four conditions are satisfiable:*

(a) $\sum_{i=1}^{k} \widetilde{\chi}_i = e$ *(partition of unity),*

(b) $(\widetilde{\chi}_i)_j \geq 0$ *for all* $i = 1, \ldots, k$ *and* $j \in 1, \ldots, n$ *(positivity),*

(c) $\widetilde{\chi} = \widetilde{X}\widetilde{A}$ *with a nonsingular* $\widetilde{A}$ *(regularity of the transformation),*

(d) for all $i = 1, \ldots, k$ *there exists a* $j \in \{1, \ldots, n\}$ *with* $(\widetilde{\chi}_i)_j = 1$ *(existence of a "center" of the almost invariant set).*

[1] All perturbed objects depend on ε, which dependence is omitted in the notation, from now on.

If all four conditions hold, the solution is unique up to permutation of the index set $\{1, \ldots, k\}$.

Having computed the $\widetilde{\chi}$, follwing information may be drawn of it. The probability of being in state i:

$$\widetilde{\pi}_i := \sum_{j=1}^{n} \pi_j (\widetilde{\chi}_i)_j = \left\langle \widetilde{\chi}_i, e \right\rangle_\pi ,$$

or the almost invariance (also called metastability, here) of the state i:

$$\widetilde{\rho}_i = \frac{\left\langle \widetilde{\chi}_i, \widetilde{T}^\top \widetilde{\chi}_i \right\rangle_\pi}{\widetilde{\pi}_i}.$$

Compared with Theorem 2.12 latter formula is of more physical relevance. It assumes, that the system ran for a time long enough to be at equilibrium (the distribution π), and computes the almost invariance ratio for the ith macroscopic state. The metastability can also be bounded by the eigenfunctions.

Theorem 2.14 (Theorem 2.2 [Deu04b])**.** *Given the transformation* $\widetilde{A}$ *with* $\|\widetilde{A}^{-1}\| = \mathcal{O}\left(\|\widetilde{X}^\top\|\right)$ *as* $\varepsilon \to 0$*, we have the bounds*

$$\sum_{i=1}^{k} \widetilde{\lambda}_i - \mathcal{O}\left(\varepsilon^2\right) \leq \sum_{i=1}^{k} \widetilde{\rho}_i < \sum_{i=1}^{k} \widetilde{\lambda}_i.$$

The theory allows an algorithmical approach. Conditions (a)–(c) can always be satisfied. The solution may not be unique, so we still have the freedom to optimize a parameter of choice, for example the metastability $\sum_i \widetilde{\rho}_i$. A vague visualization of the process is the following. Given the points $P_j = ((\widetilde{X}_1)_j, \ldots, (\widetilde{X}_k)_j) \in \mathbb{R}^k$, one chooses an as tight as possible enclosing simplex around them. The tightness refers to the property that $\|\widetilde{A}^{-1}\|$ is small. Then, the j, for which P_j is near to the ith vertex of the enclosing simplex, are going to build the core of the ith almost invariant set.

Summary: long-term behavior and spectral analysis. The previous sections showed how the long-term dynamical properties connect to the spectrum of the transfer operator. We are interested in these properties, and the major part of this thesis is devoted to the efficient computation of the associated objects: invariant densities and almost invariant sets.

Consider a naive approach computing some long orbits for the given system, and then trying to draw the desired information from these. While such an approach may work well in some cases, it fails in general. First, iterating a point for a long time is an ill conditioned problem; thus by the accumulation of rounding errors the numerical trajectory may not be even close to a real trajectory of the system. Second, if our trajectory is trapped in one almost invariant set, we may not explore important parts of the phase space. The transfer operator is given by one step of the dynamical system, and its numerical approximation does not involve long trajectory simulations either; see Section 2.3. Instead of long trajectories we will work with many short ones; this way of exploring the state space allows us to design more robust algorithms.

2.3 Ulam's method

In order to approximate the (most important) eigenfunctions of the Frobenius–Perron operator, we have to discretize the corresponding infinite dimensional eigenproblem. To this end, we project the L^1 eigenvalue problem $\mathcal{P}u = \lambda u$ into a finite dimensional subspace. Let $V_n \subset L^1$ (we write L^p instead of $L^p(X)$, if there is no ambiguity what is meant) be an *approximation subspace* of L^1 and let $\pi_n : L^1 \to V_n$ be some projection onto V_n. We then define the *discretized Frobenius-Perron operator* as

$$\mathcal{P}_n := \pi_n \mathcal{P}.$$

Ulam [Ula60] proposed to use spaces of piecewise constant functions as approximation spaces: Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be a disjoint partition of X. The X_i are usually rectangles and called boxes. Define $V_n := \operatorname{span}\{\chi_1, \ldots, \chi_n\}$, where χ_i denotes the characteristic function of X_i. Further, let

$$\pi_n h := \sum_{i=1}^{n} c_i \chi_i \qquad \text{with} \qquad c_i := \frac{1}{m(X_i)} \int_{X_i} h \, \mathrm{d}m,$$

yielding $\mathcal{P}_n : V_n^1 \subseteq V_n^1$ and $\mathcal{P}_n V_n^{1+} \subseteq V_n^{1+}$, where $V_n^1 := \left\{h \in V_n : \int |h| \, \mathrm{d}m = 1\right\}$ and $V_n^{1+} := \left\{h \in V_n^1 : h \geq 0\right\}$. Due to Brouwer's fixed point theorem there always exists an approximative invariant density $u_n = \mathcal{P}_n u_n \in V_n^{1+}$. The matrix representation of the linear map $\mathcal{P}_n|_{V_n} : V_n^1 \to V_n^1$ w.r.t. the basis of characteristic functions is given by the

transition matrix, P_n, with entries

$$P_{n,ij} = \frac{1}{m(X_i)} \int_{X_i} \mathcal{P}\chi_j \mathrm{d}m = \frac{m(X_j \cap S^{-1}(X_i))}{m(X_i)}. \tag{2.16}$$

Stochastic interpretation. The transition matrix, introduced as above, corresponds to a Galerkin projection w.r.t. the basis $B := \{\chi_1, \ldots, \chi_n\}$. From an applicational point of view it is very convenient to use this basis, since the coefficient representation of a function already yields the function values.

However, Ulam's discretization shows structural similarities to the Markov operator $\mathcal{P}$, which become obvious using the basis $B' := \{\chi_1/m(X_1), \ldots, \chi_n/m(X_n)\}$. Let P'_n denote the transition matrix w.r.t. B'. First, note that

$$P'_{n,ij} = \frac{m(X_j \cap S^{-1}(X_i))}{m(X_j)} = \int_{X_i} \mathcal{P}\frac{\chi_j}{m(X_j)} \mathrm{d}m, \tag{2.17}$$

which reads clearly as the probability that a point, sampled according to a uniform probability distribution in X_j, is mapped into X_i. Hence, $P'_{n,ij}$ is the *transition rate* from X_j to X_i and thus Ulam's method defines a finite state Markov chain on $\mathcal{X}_n$. This gives a nice probabilistic interpretation for the discretization, see [Fro96].

Indeed, the matrix P'_n is a stochastic matrix, i.e. P'_n is positive and $e^\top P'_n = e^\top$, with $e^\top = (1, \ldots, 1)$. The Markov operator $\mathcal{P}$ is approximated by a finite dimensional Markov operator $\mathcal{P}_n\,|_{V_n}$ which is represented by a stochastic matrix.

Remark 2.15. Let M_n denote the diagonal matrix with ith diagonal entry $m(X_i)$. We obtain from the basis change:

$$P'_n = M_n P_n M_n^{-1}.$$

The existence of a approximative invariant density $u_n \in V_n^{1+}$ follows now from the Perron–Frobenius theorem.

Not only a finite state Markov chain can be assigned to the discretized operator P'_n, but also a transition function $p_n : X \times \mathcal{B} \to [0, 1]$ on the whole state space, see the interpretation after (2.17):

$$p_n(x, A) = \sum_{j=1}^{n} \frac{m(A \cap X_i)}{m(X_i)} P'_{n,ij_x}, \tag{2.18}$$

where j_x is the unique (up to a set of measure zero, namely $\bigcup_i \partial X_i$) index with $x \in X_{j_x}$. The advantage of this viewpoint is that we can consider discretizations as small random

perturbations of the initial deterministic system, and extract connections between their statistical properties; cf. Chapter 3.

Convergence. Ulam conjectured [Ula60] that if $\mathcal{P}$ has a unique stationary density u, then the sequence $(u_n)_{n\in\mathbb{N}}$, with $\mathcal{P}_n u_n = u_n$, converges to u in L^1. It is still an open question under which conditions on S this is true in general. Li [Li76] proved the conjecture for expanding, piecewise continuous interval maps, Ding and Zhou [Din96] for the corresponding multidimensional case. The convergence rate was established in [Mur97, Bos01]. Froyland deals with the approximation of physical (or SRB) measures of Anosov systems in [Fro95].

In [Del99], Ulam's method was applied to a small random perturbation of S which might be chosen such that the corresponding transfer operator is compact on L^2. In this case, perturbation results (cf. [Osb75] and Section IV.3.5 in [Kat84]) for the spectrum of compact operators imply convergence.

Numerical computation of the eigenpairs. Consider (2.16) to see that $P_{n,ij} = 0$ if $S(X_j) \cap X_i = \emptyset$. Consequently, if S is Lipschitz continuous with a Lipschitz constant L_S and the partition elements X_i are congruent cubes, there can be at most L_S^d boxes X_i to intersect with $S(X_j)$. The partition being fine enough (i.e. $n \gg L_S$), this means that P_n is a sparse matrix — so the number of floating point operations (flops) required to compute a matrix-vector multiplication is $\mathcal{O}(n)$ for a large n. Moreover, we are interested only in the dominant part of the spectrum of P_n, hence Arnoldi type iterative eigenvalue solvers may be used, which require some (usually a problem-dependent number) matrix-vector multiplications to solve this problem. To sum up, having set up the transition matrix, the computational cost to compute the approximative eigenpairs is $\mathcal{O}(n)$.

Curse of dimension. If the dimension of state space is high and no further reduction is possible, problems arise concerning the computational tractability of Ulam's method. Suppose, for simplicity, that $X = [0,1]^d$. Divide X into m^d congruent cubes; there are m along each edge of X. Use the characteristic functions of these cubes to define the approximation space V_n. As one easily computes, for any given Lipschitz continuous function f holds $\|f - \pi_n f\|_{L^1} = \mathcal{O}\left(m^{-1}\right) = \|f - \pi_n f\|_{L^\infty}$. However, the costs of the

approximation are at least its storage costs; i.e. $\mathcal{O}\left(m^d\right)$. In other words, reaching the accuracy ε implies costs of $\mathcal{O}\left(\varepsilon^{-d}\right)$, exponential in the dimension d of the state space. This makes Ulam's method in dimensions $d \geq 4$ computationally inefficient or even untractable. The phenomenon is called the *curse of dimension*.

Computing the transition matrix. The computation of one matrix entry (2.16) requires a d-dimensional quadrature. A standard approach to this is Monte Carlo quadrature (also cf. [Hun94]), i.e.

$$P_{n,ij} \approx \frac{1}{K} \sum_{k=1}^{K} \chi_i \left(S(x_k)\right), \tag{2.19}$$

where the points $x_1, \ldots, x_K$ are chosen i.i.d from X_j according to a uniform distribution. In [Gud97], a recursive exhaustion technique has been developed in order to compute the entries to a prescribed accuracy. However, this approach relies on the availability of local Lipschitz estimates on S which might not be cheaply computable in the case that S is given as the time-T-map of a differential equation.

Number of sample points. Considering the Monte Carlo technique, we wish to estimate how many sample points are necessary that the error in the eigenfunctions (caused by the Monte Carlo quadrature) of the transition matrix goes to zero. One of the simplest results on bounding the error of eigenfunctions in terms of the error of the matrix is

Lemma 2.16 ([Qua00],pp. 203–204)**.** *For the (normed) eigenvectors x_k and $x_k(\varepsilon)$ of the matrices A resp. $A(\varepsilon) = A + \varepsilon E$ holds:*

$$\left\|x_k - x_k(\varepsilon)\right\|_2 \leq \frac{\varepsilon \left\|E\right\|_2}{\min_{j \neq k} |\lambda_j - \lambda_k|} + \mathcal{O}\left(\varepsilon^2\right).$$

In order to bound the norm of the difference matrix, first we have to estimate the error of the individual matrix entries. For simplicity, consider a uniform partition of X into n congruent cubes. Let P_n denote the transition matrix for this partition and let $\tilde{P}_n$ be its Monte Carlo approximation. According to the central limit theorem (and its error-estimate, the Berry–Esséen theorem [Fel71]) we have[1]

$$|\tilde{P}_{n,ij} - P_{n,ij}| \lesssim 1/\sqrt{K} \tag{2.20}$$

[1]We write $a(K) \lesssim b(K)$ if there is a constant $c > 0$ independent of K such that $a(K) \leq cb(K)$.

for the absolute error of the entries of $\tilde{P}$. Thereby, K denotes the number of Monte Carlo points.

Let $\Delta P_n := P_n - \tilde{P}_n$, i.e. the difference between the computed and the original transition matrix. The $\Delta P_{n,:j}$ denote its columns. In each column there are $\sim L_S$ entries, where L_S is the Lipschitz constant of S. Denote κ the number of all sample points, which are assumed to be distributed uniformly over X. Since the $m(X_i)$ are all equal, we have

$$\Delta P_{n,ij} \lesssim \sqrt{\frac{n}{\kappa}},$$

and for the columns

$$\left\|\Delta P_{n,:j}\right\|_2 \leq \left\|\Delta P_{n,:j}\right\|_1 \leq L_S\sqrt{\frac{n}{\kappa}}.$$

Using

$$\left\|\Delta P_n\right\|_2 = \sup_{\|x\|_2=1}\left|\sum_j x_j \Delta P_{n,:j}\right| \leq \sup_{\|x\|_2=1}\sum_j |x_j|\left\|\Delta P_{n,:j}\right\|_2 \leq \sqrt{\sum_j \left\|\Delta P_{n,:j}\right\|_2^2}, \quad (2.21)$$

we obtain

$$\left\|\Delta P_n\right\|_2 \lesssim \frac{L_S n}{\sqrt{\kappa}}.$$

By Lemma 2.16 we have for the error of the approximate eigenvector ($\Delta\lambda$ denotes the spectral gap at the eigenvalue in consideration)

$$\|\Delta f\|_{L^2} \lesssim \frac{L_S n}{\sqrt{\kappa}|\Delta\lambda|}, \quad (2.22)$$

and by the Hölder inequality on X holds

$$\|\Delta f\|_{L^1} \lesssim \frac{c_S n}{\sqrt{\kappa}|\Delta\lambda|},$$

where $c_S > 0$ depends only on the dynamical system (X and S). Consequently, one needs $\kappa/n^2 \to \infty$, if one would like to expect the algorithm to converge.

Remark 2.17. For the above bound to hold, it is necessary that the spectral gap $\Delta\lambda$ does not depend on n itself; or this dependence gets negligible as $n \to \infty$. This condition is not satisfied for certain dynamical systems, see [Jun04]. However, applying specific small stochastic perturbations to the dynamics, as it has been done e.g. in [Del99], makes the eigenvalue of interest to be isolated and of multiplicity one. We expect the above bound to work well in these cases.

2.4 Classical molecular dynamics

2.4.1 Short introduction

Simulation based analysis of physical, chemical, and even biological processes via classical molecular dynamics (MD) is a very attractive alternative to expensive and time-consuming experiments. In order to be able to predict accurately the outcome of these experiments just by computation, complicated MD models have arisen. Our aim here is to introduce the reader into the mathematical description of MD, by using a model, as simple as possible, which still captures the main property we would like to analyze with transfer operator methods: *conformation changes* (the term shall be explained below).

Transfer operator methods have been successfully applied for MD systems, even for molecules with a several hundred atoms [Deu96, Sch99, Deu01, Deu04b, Deu04a, Web07].

In situations when quantum effects can be neglected and no bond-breaking or bond-formation takes place, the dynamics of a molecule with N atoms moving around in $\mathbb{R}^3$ can be described by a Hamiltonian of form

$$H(q,p) = \frac{1}{2} p \cdot M(q)^{-1} p + V(q), \tag{2.23}$$

where $(q,p) \in \Omega \times \mathbb{R}^d \subset \mathbb{R}^{2d}$, Ω being the *configuration space*, the mass matrix $M(q)$ is a positive definite $d \times d$ matrix for all q, and $V : \mathbb{R}^d \to \mathbb{R}$ is a potential describing the atomic interactions. The first summand on the right hand side represents the kinetic energy of the molecule.

In the case when all degrees of freedom are explicitly included and cartesian coordinates are used, we have $d = 3N$ (where N is the number of atoms), $q = (q_1, \ldots, q_N) \in \mathbb{R}^{3N}$, $p = (p_1, \ldots, p_N)$, and $M = diag(m_i I_{3\times 3})$, where $q_i \in \mathbb{R}^3$ (i.e. the configuration space is $\mathbb{R}^{3N}$), $p_i \in \mathbb{R}^3$, $m_i > 0$ are the position, momentum, and mass of the ith atom. It will prove useful to work with the more general form (2.23), in which the kinetic energy is a quadratic form of p depending on q. This form arises when inner coordinates are used, which will play an important role below. For an N-atom chain molecule, the latter consist of the $(N-1)$ nearest neighbor bondlengths r_{ij}, the $(N-2)$ bond angles θ_{ijk} between any three successive atoms, and the $(N-3)$ torsion (also called "dihedral")

angles ϕ_{ijkl} between any four successive atoms. In order to accurately model conformation changes, V will have to contain at least nearest neighbor bond terms $V_{ij}(r_{ij})$, third neighbor angular terms $V_{ijk}(\theta_{ijk})$, and fourth neighbor torsion terms $V_{ijk\ell}(\phi_{ijk\ell})$. In practice the potentials could come either from a suitable semiempirical molecular force field model or from ab initio computations.

The Hamiltonian dynamics take the form

$$\dot{q} = \frac{\partial H}{\partial p}(q,p) = M(q)^{-1}p, \tag{2.24a}$$

$$\dot{p} = -\frac{\partial H}{\partial q}(q,p) = -\frac{\partial}{\partial q}\left(\frac{1}{2}p \cdot M(q)^{-1}p\right) - \nabla V(q). \tag{2.24b}$$

It will be convenient to denote the phase space coordinates by $z = (q,p) \in \Omega \times \mathbb{R}^d$ and the Hamiltonian vector field by

$$f := \begin{pmatrix} \frac{\partial H}{\partial p} \\ -\frac{\partial H}{\partial q} \end{pmatrix}, \tag{2.25}$$

so that (2.24) becomes

$$\dot{z} = f(z). \tag{2.26}$$

The change of probability densities under the dynamics is described by the Liouville equation associated to (2.24)

$$\partial_t u + f \cdot \nabla u = 0, \tag{2.27}$$

where $u = u(z,t)$, $u : \Omega \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, or, since the Hamiltonian vector field f is divergence-free, its equivalent form as continuity equation[1]

$$\partial_t u + \operatorname{div}(u\, f) = 0. \tag{2.28}$$

Compare (2.14) (set $\varepsilon = 0$) with (2.27) to see, that the FPO associated with the Hamiltonian H, and the Koopman operator associated with $-H$ coincide on $L^\infty(\Omega \times \mathbb{R}^d) \cap L^1(\Omega \times \mathbb{R}^d)$. This imples that the FPO associated with the system (2.24) is given by[2]

$$\mathcal{P}^t u = u \circ \Phi^{-t}, \tag{2.29}$$

where Φ^t is the time-t-map of (2.26). Note that for an arbitrary function $g : \mathbb{R} \to [0,\infty)$ of the Hamiltonian, the function $u(z) = g(H(z))$ satisfies $\nabla u(z) = g'(H(z))\nabla H(z)$.

[1] Compare with the Fokker–Planck equation (2.13) with zero diffusion.

[2] Compare with (2.12), and note that Φ^t is area preserving for all $t \in \mathbb{R}$, as $\operatorname{div}(f) = 0$.

Thus $f \cdot \nabla u = 0$ and u, normalized such that $\int u(z)\,\mathrm{d}z = 1$, is an invariant density. Of particular interest is the *canonical density*

$$h(z) = C \exp(-\beta H(z)), \tag{2.30}$$

$C = \int \exp(-\beta H(z))\,\mathrm{d}z$, where $\beta = 1/(k_B T)$ and k_B is Boltzmann's constant. This density describes the distribution of a (constant) large number of molecules at temperature T and of constant volume. Note that we have

$$h(z) = h(q,p) = C \exp\left(-\frac{\beta}{2} p \cdot M^{-1}(q) p\right) \exp\left(-\beta V(q)\right).$$

Finally, we note that (2.28) preserves the (expected value of) energy,

$$E(t) := \int H(z) u(z,t)\,\mathrm{d}z.$$

This is because by an integration by parts

$$\frac{d}{dt} E(t) = \int H(z)\Big(-\mathrm{div}(u(z,t) f(z))\Big)\,\mathrm{d}z = \int \nabla H(z) \cdot f(z) u(z,t)\,\mathrm{d}z$$

and the inner product $\nabla H(z) \cdot f(z)$ vanishes for all z, due to (2.25).

The spatial transfer operator. Molecular conformations should be thought of as almost invariant subsets of configuration space. Schütte [Sch99] introduced a corresponding spatial transfer operator by averaging (2.29) over the momenta: Let $h \in L^1(\Omega \times \mathbb{R}^d)$ be an invariant density[1] of (2.29) with $h(q,p) = h(q,-p)$, let $h_q(q) = \int h(q,p)\,\mathrm{d}p$, and consider the operator

$$T^t w(q) = \frac{1}{h_q(q)} \int w\left(\pi_q \Phi^{-t}(q,p)\right) h(q,p)\,\mathrm{d}p, \tag{2.31}$$

where $\pi_q(q,p) = q$ is the canonical projection onto the configuration space. It is designed to describe spatial fluctuations (i.e. fluctuations in the configuration space) inside an ensemble of molecules distributed according to h. Schütte [Sch99] showed that under suitable conditions, the spatial transfer operator is self-adjoint and quasi-compact on a suitably weighted L^2 space. Moreover, its eigenmodes with eigenvalue near one give information about almost invariant regions in the configuration space, cf. Section 2.2.2.

[1] Although the definition here works with *arbitrary* invariant densities, unless stated elsewise, we consider h to be the canonical density. Hence the same notation.

The spatial transfer operator T^t is strongly connected to a stochastic process, which can be sampled as follows. Given q_k, draw a random sample p_k according to the distribution $h(q_k,\cdot)/h_q(q_k)$. Set $q_{k+1} = \pi_q\Phi^t(q_k,p_k)$. The spatial tranfer operator is the transfer operator of this process on a suitably weighted L^1 space. This weighting makes numerical computations more complicated, hence we define a related operator, which we call *spatial transfer operator* as well:

$$\mathcal{S}^t w(q) = \int \mathcal{P}^t\left(w\frac{h(q,p)}{h_q(q)}\right)\mathrm{d}p. \tag{2.32}$$

This operator is the FPO on $L^1(\Omega)$ of the stochastic process described above. It is related to the transfer operator of Schütte by (note, that $h_q(q) > 0$ for all $q \in \Omega$)

$$\frac{1}{h_q}\mathcal{S}^t w = T^t\left(\frac{w}{h_q}\right).$$

Thus, if w is an eigenfunction of T^t, then $h_q w$ is an eigenfunction of $\mathcal{S}^t$ at the same eigenvalue. As we will see, we can draw from the eigenmodes of $\mathcal{S}^t$ *qualitatively* the same information about almost invariant sets as from the ones of T^t. Note also, that the spatial distribution of the ensemble, h_q, is a fixed point of $\mathcal{S}^t$, thus an invariant density of the process.[1]

Since we know how to sample the stochastic process, the discretization with Ulam's method is straightforward. Let us partition the configuration space Ω by using the boxes B_k. Let S_n^t denote the matrix representation of the corresponding Ulam discretization $\mathcal{S}_n^t$. Then we have

$$\begin{aligned} S_{n,ij} &= \frac{1}{m(B_i)}\int_{B_i}\int \mathcal{P}^t\left(\chi_j\frac{h(q,p)}{h_q(q)}\right)\mathrm{d}p\mathrm{d}q \\ &= \frac{m(B_j)}{m(B_i)}\mathrm{Prob}\left(\pi_q\Phi^t(q,p)\in B_i \Big| q\sim\frac{\chi_j}{m(B_j)}, p\sim\frac{h(q,\cdot)}{h_q(q)}\right) \\ &= \frac{1}{m(B_i)}\int_{B_j}\int \chi_i\left(\pi_q\Phi^t(q,p)\right)\frac{h(q,p)}{h_q(q)}\mathrm{d}p\mathrm{d}q. \end{aligned} \tag{2.33}$$

If the Hamiltonian is smooth, the integrand in $\int_{B_j}\ldots\mathrm{d}q$ is smooth as well, hence this integral may be very well approximated by a small number of evaluations of the integrand (e.g. by applying Gauss quadrature). The inner integral $\int\ldots\mathrm{d}p$ is evaluated by Monte Carlo quadrature.

[1] Schütte gives a thorough spectral analysis of the operator T^t in his work. In particular, conditions are given under which 1 is a simple and dominant eigenvalue of T^t, thus of $\mathcal{S}^t$ as well.

2.4.2 Example: n-butane

We consider a united atom model [Bro83] of the n-butane molecule CH_3-CH_2-CH_2-CH_3 (cf. Figure 2.2), viewing each CH_3, respectively, CH_2 group as a single particle. Consequently, the configuration of the model is described by six degrees of freedom: three bond lengths, two bond angles, and one torsion angle. We further simplify the

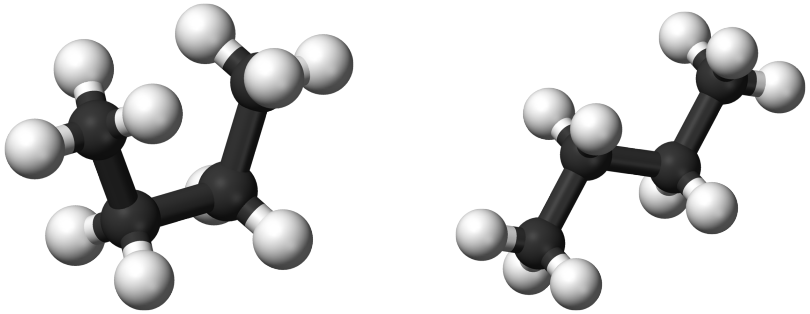


Figure 2.2: Cis- and trans-configuration of n-butane.

model by fixing the bond lengths at their equilibrium $r_0 = 0.153$ nm. This leaves us with the configuration variables θ_1, θ_2 and ϕ, the two bond angles and the torsion angle, respectively. For the bond angles we use the potential

$$V_2(\theta) = -k_\theta \left(\cos\left(\theta - \theta_0\right) - 1\right) \tag{2.34}$$

with $k_\theta = 65 \frac{\text{kJ}}{\text{mol}}$ and $\theta_0 = 109.47\ ^\circ$, and for the torsion angle we employ

$$\begin{aligned} V_3(\phi) \quad = \quad & K_\phi \left(1.116 - 1.462\cos\phi - 1.578\cos^2\phi + 0.368\cos^3\phi \right. \\ & \left. + 3.156\cos^4\phi + 3.788\cos^5\phi\right) \end{aligned}$$

with $K_\phi = 8.314 \frac{\text{kJ}}{\text{mol}}$; cf. Figure 2.3 (see also [Gri07]). There are three "potential wells", i.e. local minima of the potential, we expect the system to show rare transitions out of one well into another. The positions of these wells correspond to dominant (i.e. almost invariant) conformations. We wish to detect these with the eigenmodes of the spatial transfer operator.

We fix $m_p = 1.672 \cdot 10^{-24}$g as the mass of a proton and correspondingly $m_1 = 14\, m_p$ and $m_2 = 15\, m_p$ as the masses of CH_2 group and CH_3 group, respectively. With

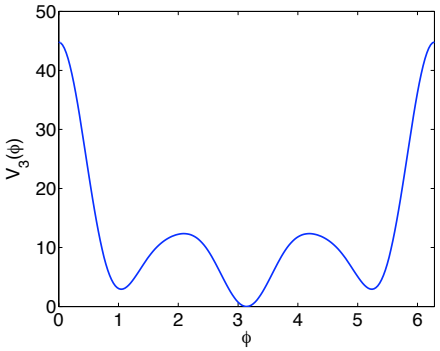


Figure 2.3: Potential of the torsion angle.

$q = (\theta_1, \theta_2, \phi)^\top \in [0, \pi] \times [0, \pi] \times [0, 2\pi] =: \Omega$ denoting the configuration of our model, the motion of our system is determined by the Hamiltonian

$$H(q, p) = \frac{1}{2} p^\top M(q)^{-1} p + V(q) \tag{2.35}$$

with $V(q) = V_2(q_1) + V_2(q_2) + V_3(q_3)$ and the mass matrix $M(q)$. The latter is computed by means of a coordinate transformation $q \mapsto \tilde{q}(q)$ to cartesian coordinates $\tilde{q} \in \mathbb{R}^{12}$ for the individual particles, assuming that there is no external influence on the molecule and its linear and angular momentum are zero: We have

$$\dot{\tilde{q}} = D\tilde{q}(q)\dot{q}$$

and consequently

$$M(q) = D\tilde{q}(q)^\top M D\tilde{q}(q),$$

where M denotes the (constant, diagonal) mass matrix of the Hamiltonian in cartesian coordinates.

Everything is set to compute the Ulam discretization of the spatial transfer operator. We consider an ensemble at temperature $T = 1000$K. Since transfer operator methods need only short trajectory simulations, we use $t = 5 \cdot 10^{-14}$s and the forward Euler method to integrate the system.[1]

We apply a $32 \times 32 \times 32$ uniform partition of the configuration space Ω, use in each box a three dimensional 8-node Gauss quadrature for the integral w.r.t. q, and for each

[1] The integration time t is chosen such that it is still small, but we can detect considerable motion in trajectory simulations. For such a short period of time the forward Euler method is sufficiently accurate for our purposes here. Of course, there are more suitable methods for integrating Hamilton systems [Hai06], e.g. the Verlet scheme.

q-node 8 p-samples, see (2.33). Having computed the approximate transition matrix, we compute the left and right eigenvectors. We visualize the latter by showing the θ_2-ϕ-marginals of the first 3 eigenfunctions in Figure 2.4. Note, that by the symmetry of the molecule, the θ_1-ϕ marginals have to look alike. Observe, that the sign structure of the second and third eigenvectors indicate almost invariant sets at $\phi \approx \pi/3$, $\phi \approx \pi$ and $\phi \approx 5\pi/3$ — just where the wells of the potential V_3 are. The compontents of

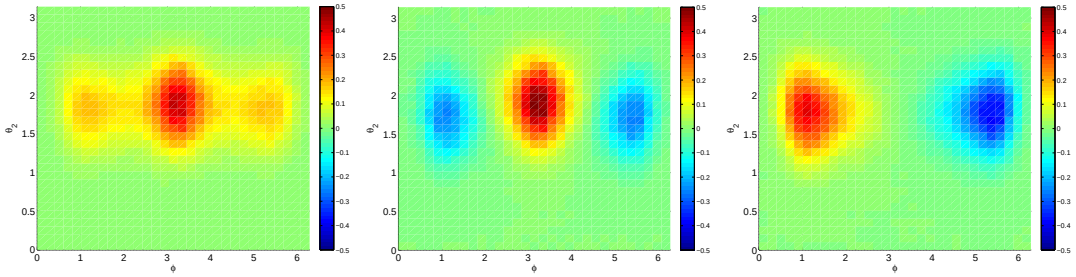


Figure 2.4: Dominant configurations of n-butane, analyzed via right eigenvectors - The θ_2-ϕ marginals of the first three eigenfunctions (from left to right) of the approximate spatial transfer operator $S_{32\times32\times32}$. The almost invariant sets can be extracted from the sign structure of the second and third eigenfunctions.

the second and third approximate left eigenvectors plotted in $\mathbb{R}^2$ are shown Figure 2.5 (left). According to Section 2.2.2, the points near the vertices of the simplex show the positions of the almost invariant sets. The corresponding areas in the configuration space are shown in the right plot of the same figure.

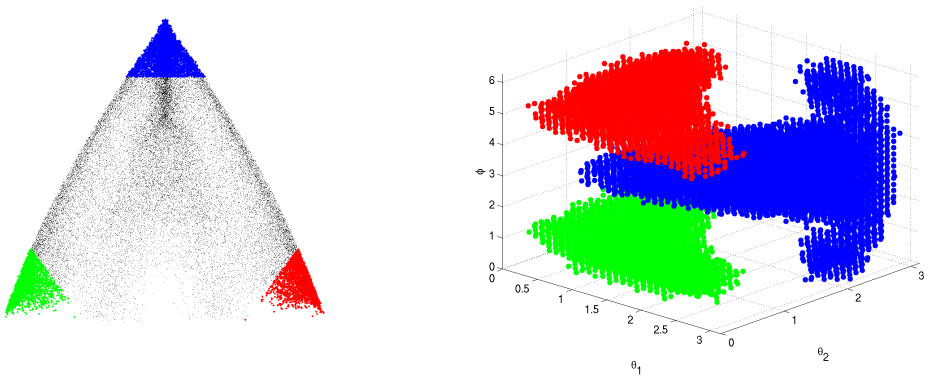


Figure 2.5: Dominant configurations of n-butane, analyzed via left eigenvectors - Left: the points $(v_{2,i}, v_{3,i}) \in \mathbb{R}^2$ for $i = 1, \ldots, 32^3$, where v_2 and v_3 are the second and third approximate left eigenvectors of the discretized spatial transfer operator. Right: the points near the "vertices" of the approximative simplex on the left correspond to boxes in the partition of the configuration space. The almost invariant configurations are seen easily here.

Chapter 3

Projection and perturbation

3.1 Small random perturbations

A lot of scientific interest has been devoted to the question how properties of (deterministic) dynamical systems change under perturbation of the system. There are two natural concepts of perturbation. The first is taking a deterministic system $\widetilde{S} : X \to X$ as a perturbation of the original one, $S : X \to X$, and comparing their (local) topological behavior. It is assumed that $\|S - \widetilde{S}\|$ is small for a suitable norm $\|\cdot\|$. These considerations are associated with the field of *structural stability*. Since we will not deal with this topic, the reader is referred to the textbook [Guc83]. The second concept is the notion of *stochastic stability*, which compares the original system S with nondeterministic ones "near" S in a way described below. It is an appropriate way of analyzing the robustness of statistical properties of dynamical systems. We use the following definition of small random perturbations:

Definition 3.1 (Small random perturbation,[Kif86])**.** *A family $p_\varepsilon : X \times \mathcal{B} \to [0,1]$ of stochastic transition functions is called a small random perturbation (s.r.p.) of the map $S : X \to X$, if*

$$\lim_{\varepsilon \to 0} \sup_{x \in X} \left| g(S(x)) - \int_X g(y) p(x, \mathrm{d}y) \right| = 0 \tag{3.1}$$

for all $g \in C^0(X)$.

One can also read this as "$p_\varepsilon(x, \cdot) \to \delta_{S(x)}(\cdot)$" as $\varepsilon \to 0$ uniformly in x, where δ_x is the Dirac delta function (or Dirac delta distribution) centered in x.

A first statement about the connection of the statistical properties of a dynamical system and its s.r.p. gives this theorem from Khas'minskii:

Proposition 3.2 ([Kha63]). *Let p_ε be a s.r.p. of S. For each ε let μ_ε be an invariant measure of p_ε. Let $\mu_{\varepsilon_i} \to \mu$ in the weak sense*[1] *for a sequence $\varepsilon_i \to 0$. Then μ is an invariant measure of S.*

The result raises the question, if there are such invariant measures μ of particular systems, where the above convergence holds for any arbitrary sequence $\varepsilon_i \to 0$ with the common limit μ (stochastic stability). Kifer gives a positive answer [Kif86] for axiom A C^2 diffeomorphisms, under some regularity assumptions on the s.r.p. In that case the limiting measure is a physical measure of the system. To omit technicalities, we only state the assumption which will play the most important role in our further considerations: the transition function p_ε should have a transition density function q_ε, and the support of $q_\varepsilon(x, \cdot)$ should vary *continuously* with x (see [Kif86], §2., Remark 1.).

If one could interpret discretizations of the transfer operators as s.r.p. of the corresponding dynamical system, there would be a chance to prove the convergence of approximative invariant measures to the invariant (physical) measure of the original system. To my best knowledge, this idea goes back to Góra [Gor84] and Froyland [Fro95], see also [Del99]. The current chapter is devoted to this question. More precisely, we will derive assumptions on the approximation space, which guarantee that the Galerkin projection of the transfer operator corresponds to a s.r.p. of the dynamical system in consideration.

3.2 On characterizing Galerkin discretizations as small random perturbations

The projection. Let X be a compact metric space and denote $L^p = L^p(X)$ for $1 \leq p \leq \infty$. Define linearly independent functionals $\ell_1, \ldots, \ell_n \in (L^1)'$, where $(L^1)'$ is the dual of L^1. Further let $V_n := \operatorname{span}\{\varphi_1, \ldots, \varphi_n\}$,[2] the φ_i are bounded, piecewise continuous[3] and linearly independent. Thus $V_n \subset L^\infty$ and $\dim V_n = n$. Let the

[1] A sequence of measures $\{\mu_n\}$ converges to the measure μ in the weak sense, if $\int g \, \mathrm{d}\mu_n \to \int g \, \mathrm{d}\mu$ for every continuous function g.

[2] We omit here the indication that the φ_i depend on n itself, although it may be the case.

[3] A function is piecewise continuous, if there is a finite partition of its domain, where the function is continuous on each partition element. Having numerical computations in mind, it certainly makes sense to work with bounded piecewise continuous functions.

projection $\pi_n : L^1 \to V_n$ be defined by

$$\ell_i(f - \pi_n f) = 0 \qquad \forall\, i = 1 \dots n.$$

It is unique if for every $\varphi \in V_n$ following implication holds: if $\ell_i(\varphi) = 0$ for all $i = 1 \dots n$, then $\varphi = 0$. Since $(L^1)'$ is isomorph to L^∞, there are $\psi_1, \dots, \psi_n \in L^\infty$ such that $\ell_i(f) = \int f\psi_i$ for every $f \in L^1$ and $i = 1 \dots n$.[1] The ψ_i are called *test functions*. For general ψ_i the projection is called *Petrov–Galerkin projection*, if $\psi_i = \varphi_i$, we call it a *Galerkin projection*. We are going to consider Galerkin projections here, nevertheless it should be clear from the derivation how can one construct the more general ones as well.

Setting $\pi_n f = \sum_{i=1}^n c_i \varphi_i$ and $\psi_i = \varphi_i$, by

$$b_j := \int f\ \varphi_j = \int \pi_n f\ \varphi_j = \sum_{i=1}^n c_i \underbrace{\int \varphi_i\ \varphi_j}_{=:A_{n,ji}},$$

the projection reads as $c = A_n^{-1} b$, where

$$A_n = \int \Phi_n\ \Phi_n^\top, \qquad b = \int \Phi_n f,$$

with $\Phi_n = (\varphi_1, \dots, \varphi_n)^\top$. Thus

$$\pi_n f = \Phi_n^\top A_n^{-1} \int \Phi_n f \tag{3.2}$$

Discretization as perturbation. We would like to find a stochastic transition density $q_n(x, y)$ such that $\mathcal{P}_{q_n} = \pi_n \mathcal{P}$ on L^1, $\mathcal{P}$ being the transfer operator associated with S. Recall, that $\mathcal{U}$ denotes the Koopman operator, which is adjoint to $\mathcal{P}$. Since

$$\mathcal{P}_{q_n} f(y) = \int q_n(x, y) f(x) \mathrm{d}x, \tag{3.3}$$

and

$$\pi_n \mathcal{P} f(y) = \Phi_n(y)^\top A_n^{-1} \int \Phi_n\ \mathcal{P} f = \Phi_n(y)^\top A_n^{-1} \int \underbrace{\mathcal{U}\Phi_n}_{=\Phi_n \circ S}\ f, \tag{3.4}$$

for all $f \in L^1$, we conclude

$$q_n(x, y) = \Phi_n(y)^\top A_n^{-1} \Phi_n(S(x)) = \bar{q}_n(S(x), y), \tag{3.5}$$

[1] If the set of integration is not indicated, the whole phase space X is meant to be integrated over.

where $\bar{q}_n(x,y) = \Phi_n(y)^\top A_n^{-1} \Phi_n(x)$. Note, that q_n is invariant under a change of the basis. Further, since A_n is symmetric positive definite (s.p.d.), A_n^{-1} is s.p.d. as well, which implies the symmetry of $\bar{q}_n$.

Equation (3.5) could be understood as well as

$$q_n(x,y) = \Phi_n(y)^\top A_n^{-1} \int \Phi_n \; \delta_{S(x)} = (\pi_n \delta_{S(x)})(y).$$

Topology of the approximating functions — some assumptions. Until now, the projection property (3.2), and everything derived from it, is meant to hold Lebesgue almost everywhere (a.e.). For later analysis we will need a stronger relation, which we obtain by extracting some topological features of the approximation space. These features appear to be evident if one has numerical applications in mind.

First of all, X should have a nonempty interior and $X = \mathrm{int}(X) \cup \partial X$. Further, recalling the piecewise continuity of Φ_n, there should be a finite collection of sets R_i^n and Γ_i^n, such that

(a) $R_i^n = \mathrm{int}(R_i^n) \cup \Gamma_i^n$ and $\mathrm{int}(R_i^n) \neq \emptyset$,

(b) $\Gamma_i^n \subset \partial R_i^n$,

(c) the R_i^n are disjoint with $\bigcup_i R_i^n = X$, and

(d) Φ_n is continuous on R_i^n.

Fix now some j, and recall the projection property (3.2)

$$\Phi_n(y)^\top A_n^{-1} \int \Phi_n \varphi_j = \varphi_j(y) \quad \text{for a.e. } y \in X, \tag{3.6}$$

where the integral does not depend on the L^∞–representative of Φ_n. If (3.6) holds Lebesgue a.e., it holds pointwise for a dense set $Y \subset X$. Let $y \in X$ be arbitrary and i such that $y \in R_i^n$. Then, by our assumptions, there is a sequence $\{y_k\} \subset R_i^n \cap Y$ such that $y_k \to y$. By the piecewise continuity of Φ_n, (3.6) holds for y as well, thus the projection property (3.2) (and all its consequences) holds pointwise in X.

Finally, we state that the Γ_i^n can be chosen in dependence on j (if necessary, by changing the values of φ_j on a zero-measure set) such that the basis function φ_j admits a maximum. It may be impossible, however, to choose a partition $\{R_i^n\}_i$ such that all φ_j admit their maxima at the same time. Nevertheless, changing the values of the φ_j on the zero–measure sets Γ_i^n is not decisive for the fact if q_n is a s.r.p. or not, but it will be important in the proof of Theorem 3.7.

First considerations. If we want q_n to be a stochastic transition density which is a s.r.p. of S, three requirements have to be fulfilled:

(i) $q_n \geq 0$ on $X \times X$,

(ii) $\int q_n(x, \cdot) = 1$ for all $x \in X$, and

(iii) q_n is the transition density of a transition function which is a s.r.p. in the sense of Definition 3.1.

Lemma 3.3. *Let S be onto. Then following holds:*

(i) $q_n \geq 0 \quad \Leftrightarrow \quad \bar{q}_n \geq 0$

(ii) $\int q_n(x, \cdot) = 1 \ \forall x \quad \Leftrightarrow \quad \mathbf{1} \in V_n$, *where* $\mathbf{1}(x) = 1$ *for all* $x \in X$.

(iii) If q_n is a stochastic transition density, the corresponding transition function is a small random perturbation of S, iff $\pi_n g \to g$ as $n \to \infty$, uniformly (in x) for all $g \in C^0$.

Proof. To (i): Trivial by (3.5) and the surjectivity of S.
To (ii): Substitute (3.5) in the claim, and see that it is equivalent with $\pi_n \mathbf{1} = \mathbf{1}$.
To (iii): As $n \to \infty$, we have

$$\begin{aligned} \sup_x \left| g(S(x)) - \int g(y) q_n(x, y) \mathrm{d}y \right| \to 0 \quad &\Leftrightarrow \quad \sup_x \left| g(S(x)) - \Phi_n(S(x))^\top A_n^{-1} \int \Phi_n \, g \right| \to 0 \\ &\Leftrightarrow \quad \sup_x \left| g(S(x)) - \pi_n g(S(x)) \right| \to 0 \\ &\Leftrightarrow \quad \left\| g - \pi_n g \right\|_{L^\infty} \to 0, \end{aligned}$$

where the last equivalence follows from the surjectivity of S. □

Remark 3.4. In some applications it may be the case that S is not onto, e.g. think of X as a finite box covering of an attractor of complicated geometry. In general, the covering certainly will not be congruent with the attractor and S cannot be onto. Note, however, that the conditions posed on V_n and π_n in Lemma 3.3 are still sufficient for the claims on q_n; only not necessary. In order to keep our analysis on the level of approximation space and the corresponding projection, we stick to these sufficient conditions. Otherwise, one would have to utilize specific geometrical properties of the phase space/attractor, which may differ from system to system.

3.3 The problem with nonnegativity

Let us fix the discretization parameter n and omit it as a subscript. This will ease the reading in the following.

By Lemma 3.3 the nonnegativity of q is equivalent with the nonnegativity of $\bar{q}$. For an Ulam type approximation, i.e. using characteristic functions over the partition elements X_i as basis functions, $\bar{q}(x, y) = 0$ if x and y are not contained in the same partition element X_i, meanwhile $\bar{q}(x, y) = 1/m(X_i)$ if both $x, y \in X_i$. The corresponding q is a stochastic density function and a s.r.p. of S. Indeed, all criteria of Lemma 3.3 are easily checked. Concerning (iii), note that continuous functions over a compact set are uniformly continuous, which allows the piecewise constant approximations to converge uniformly on X, if the box diameters tend to zero.[1] A pity that $\operatorname{supp}(\bar{q}(x, \cdot))$ does not depend continuously on x, hence the stochastic stability results from Kifer can not be applied here.

A simple example of continuous basis functions occurring often in applications are hat functions. Unfortunately, the resulting $\bar{q}$ are already nonnegative for a coarse discretization, and this gets only worse by increasing the resolution; cf. Figure 3.1.

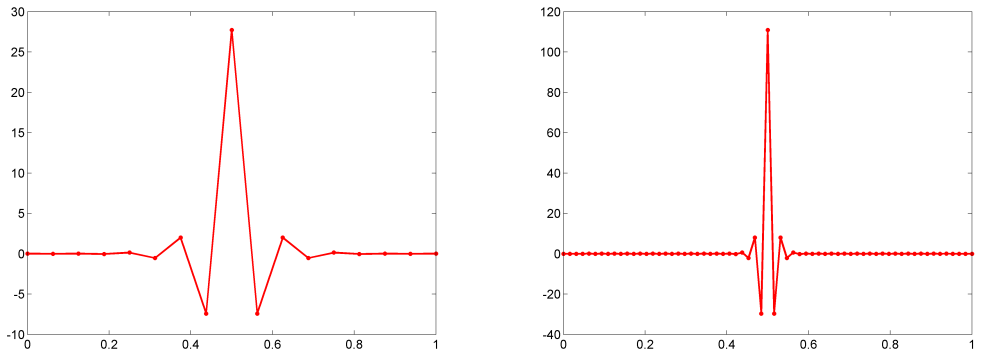


Figure 3.1: The transition density is not nonnegative - plotted is $\bar{q}(0.5, \cdot)$ for 17 basis functions (left) and 65 basis functions (right).

The result. It turns out, that $\bar{q}$ has negative parts not only for hat functions. We would like to characterize in the following the basis functions satisfying the nonnega-

[1] Froyland shows in [Fro95] that the operator $\pi_n \mathcal{P} \pi_n$ can be viewed as a s.r.p. of S. Note, that we work with $\pi_n \mathcal{P}$. The range and thus the invariant densities of the two operators are identical.

tivity requirements. For this, recall the projection property ($\pi\varphi = \varphi$ for $\varphi \in V$)

$$\int \bar{q}(x,y)\varphi(y)\mathrm{d}y = \varphi(x), \tag{3.7}$$

and that q ought to be a stochastic transition density; i.e. $\int \bar{q}(x,\cdot) = 1$ for all $x \in X$. By the symmetry of $\bar{q}$ it does not matter if $\bar{q}(\cdot,y)$ or $\bar{q}(x,\cdot)$ is the projection kernel. Now let $\varphi \in V$ be arbitrary and the R_i^n chosen such that $|\varphi|$ has a maximum place. By the piecewise continuity and boundedness of φ, further by the compactness of X, there will be (a not necessary unique) one, which we denote by x_0. It follows from (3.7) that

$$\left|\varphi(x_0)\right| = \left|\int \bar{q}(x_0,y)\varphi(y)\mathrm{d}y\right| \leq \underbrace{\left\|\bar{q}(x_0,\cdot)\right\|_{L^1}}_{=1} \max(|\varphi|) = \left|\varphi(x_0)\right|.$$

Equation can hold only if $|\varphi| \equiv \left|\varphi(x_0)\right|$ over $M_0 := \operatorname{supp}\left(\bar{q}\left(x_0,\cdot\right)\right)$ and $\varphi(y)$ has the same sign for all $y \in M_0$. Hence, $\varphi = \varphi(x_0)$ on M_0. With other words, all $x \in M_0$ are maximum places of $|\varphi|$. Continuing this argument, we obtain following:

Proposition 3.5. *Define* $M_0 := \operatorname{supp}\left(\bar{q}\left(x_0,\cdot\right)\right)$ *and*

$$M_k := \left\{x \in \operatorname{supp}\left(\bar{q}\left(z,\cdot\right)\right) \mid z \in M_{k-1}\right\}.$$

Then $\varphi(x) = \varphi(x_0)$ *for all* $x \in \bigcup_{k\in\mathbb{N}_0} M_k$.

We already know by (3.5) how $\bar{q}$ is obtained from a basis of V. Here is a result concerning the other direction.

Lemma 3.6. *There is a* $\mathbf{x} = (x_1,\ldots,x_n)$ *such that* $\left\{\bar{q}\left(x_i,\cdot\right)\right\}_{i=1,\ldots,n}$ *is a basis of* V. *The* x_i *may be chosen such that* $x_i \in \bigcup_k \operatorname{int}(R_k^n)$ *for every* $i = 1,\ldots,n$.

Proof. Since

$$\sum_{i=1}^{n} c_i\ \bar{q}(x_i,y) = \Phi(y)^\top A^{-1}\Phi(\mathbf{x})c,$$

with $\Phi(\mathbf{x}) = \left(\Phi(x_1)\middle|\ldots\middle|\Phi(x_n)\right) \in \mathbb{R}^{n\times n}$, the claim is equivalent with: there is an $\mathbf{x}$ such that the $\Phi(x_i)$ are linearly independent.

We construct the set $\left\{x_1,\ldots,x_n\right\}$ step by step. Choose x_1 arbitrary, such that $x_1 \in \operatorname{int}(R_k^n)$ for some k. From now on, the proof goes by induction. Assume, we have $x_1,\ldots,x_m$ with $m < n$ and $x_i \in \operatorname{int}(R_{k_i}^n)$. Assume further that there is no

$x \in \bigcup_k \operatorname{int}(R_k^n)$ such that the $\Phi(x_1), \ldots, \Phi(x_m), \Phi(x)$ are linearly independent. Thus, there are functions $c_1, \ldots, c_m : \bigcup_k \operatorname{int}(R_k^n) \to \mathbb{R}$ such that

$$\sum_{i=1}^{m} c_i(x)\Phi(x_i) = \Phi(x) \qquad \forall\, x \in \bigcup_k \operatorname{int}(R_k^n).$$

In other words, $\Phi(x)$ is in the range of the matrix $\mathbf{\Psi} \in \mathbb{R}^{n\times m}$ with $\mathbf{\Psi}_{ij} = \varphi_i(x_j)$ for all $x \in \bigcup_k \operatorname{int}(R_k^n)$. But the range is a closed subspace and Φ is continuous, hence $\Phi(x)$ is in the range of $\mathbf{\Psi}$ for $x \in \bigcup_k \Gamma_k^n$ as well. It follows, the c_i can be extended to the entire X and V is spanned by m functions (the c_i), which contradicts $\dim V = n$. The induction step is hereby complete, hence the proof as well. □

Now we are ready to prove the main result.

Theorem 3.7. *Assume V is spanned by such bounded, piecewise continuous functions that the corresponding $\bar{q}$ satisfies*

(i) $\bar{q} \geq 0$ on $X \times X$ and

(ii) $\int \bar{q}(x, \cdot) = 1$ for all $x \in X$.

Then V is spanned by characteristic functions.

Proof. By Lemma 3.6 there is a basis $\{\bar{q}(x_i, \cdot)\}_{i=1\ldots n}$ of V, where $x_i \in \bigcup_k \operatorname{int}(R_k^n)$ for $i = 1, \ldots, n$. Let i be arbitrary and denote (for simplicity) $z = x_i$. Then, by the basis representation formula, it holds

$$\int \bar{q}(x, y)\bar{q}(z, y)\mathrm{d}y = \bar{q}(z, x).$$

If necessary, change the φ_j on the Γ_k^n, so that $\bar{q}(z, \cdot)$ has a maximum, and let z^m denote a maximum place. This change affects each chosen basis function at most on a zero-measure set (since $x_i \notin \bigcup_k \Gamma_k^n$), hence linear independence is retained, and the basis property as well. Then

$$\int \bar{q}(z^m, y)\bar{q}(z, y)\mathrm{d}y = \bar{q}(z, z^m).$$

By $\bar{q} \geq 0$ and $\int \bar{q}(z^m, \cdot) = 1$ we have (recall the considerations at the beginning of the paragraph, in particular Proposition 3.5)

$$\bar{q}(z, y) = \bar{q}(z, z^m) \qquad \forall y \in \operatorname{supp}\left(\bar{q}(z^m, \cdot)\right). \tag{3.8}$$

By the symmetry of $\bar{q}$ is $\bar{q}(z^m, z) > 0$ and hence $z \in \operatorname{supp}\left(\bar{q}(z^m, \cdot)\right)$. Thus by (3.8) is z a maximum place of $\bar{q}(z, \cdot)$, and we can set $z^m = z$. Once more using (3.8), we have that $\bar{q}(z, \cdot)$ is constant over its whole support. □

The theorem tells us that if we would like to consider the Galerkin discretization of the transfer operator as a s.r.p. of the dynamical system, the chosen approximation space would consist of characteristic functions. We encounter the same problem as discussed before with Ulam's method: the continuous variation of the transition density function support.

3.4 The case $\mathcal{P}_n = \pi_n \mathcal{P} \pi_n$

It is also possible to consider, instead of $\mathcal{P}_n = \pi_n \mathcal{P}$, the operator $\mathcal{P}_n = \pi_n \mathcal{P} \pi_n$. The eigenmodes corresponding to the nonzero spectrum are the same for the both operators, in particular the modes at the largest eigenvalues. As one may easily see, the latter is the transfer operator associated with the transition function (2.18). Let us compute the transition density of this operator. Once again, we use the projection property (3.2).

$$\begin{aligned} \pi_n \mathcal{P} \pi_n f(z) &= \int \bar{q}_n(y,z) \mathcal{P} \int \bar{q}_n(x,y) f(x) \mathrm{d}x \mathrm{d}y \\ &= \int \mathcal{U} \bar{q}_n(y,z) \int \bar{q}_n(x,y) f(x) \mathrm{d}x \mathrm{d}y \\ &= \iint \bar{q}_n(S(y),z) \bar{q}_n(x,y) \mathrm{d}y f(x) \mathrm{d}x, \end{aligned}$$

where the compactness of X and the boundedness of $\bar{q}_n$ allows the change of the integral sequence. We obtain the transition density function

$$q_n(x,y) = \int \bar{q}_n(S(z),y) \bar{q}_n(x,z) \mathrm{d}z.$$

This may also be read as $q_n(\cdot, y) = \pi_n \bar{q}_n(S\cdot, y)$. Setting $S = \mathrm{Id}$, we are in the former case ($\mathcal{P}_n = \pi_n \mathcal{P}$), and see, that only piecewise constant functions may be interpreted as s.r.p's. Any more precise statement would require a deeper analysis of the interplay of the dynamics S and the approximation space V_n, which is not considered in this work.

However, this description of the discretized transfer operator gives us an option to show that (2.18) is a s.r.p. of S. The same has been proven earlier in [Fro95].

Proposition 3.8. *Ulam's method can be interpreted as a s.r.p. More precisely, the transition function* (2.18) *is a s.r.p. of S, provided S is continuous.*

Remark 3.9. The notion of s.r.p.'s used here was introduced in [Kif86] for *diffeomorphisms*, hence our assumption does not mean a serious restriction.

Proof. Let $g \in C^0$ be arbitrary. Then

$$\begin{aligned}\left|g(S(x)) - \int g(y) q_n(x,y)\mathrm{d}y\right| &= \left|g(S(x)) - \int g(y) \int \bar{q}_n(S(z),y)\bar{q}_n(x,z)\mathrm{d}z\mathrm{d}y\right| \\ &= \left|g(S(x)) - \pi_n\left(\left(\pi_n g\right)\circ S\right)(x)\right|,\end{aligned}$$

where the second equation follows by swapping the integration sequence (allowed, just as above). Thus, we need to show

$$\left\|g\circ S - \pi_n\left(\left(\pi_n g\right)\circ S\right)\right\|_{L^\infty} = \big\| \underbrace{\pi_n\left(\left(\pi_n g\right)\circ S\right) - \pi_n(g\circ S)}_{=:I_1} + \underbrace{\pi_n(g\circ S) - g\circ S}_{=:I_2} \big\|_{L^\infty} \to 0$$

as $n\to\infty$. Since the Ulam-type projection π_n is a $\|\cdot\|_{L^\infty}$-contraction, we have

$$\left\|I_1\right\|_{L^\infty} \le \left\|(\pi_n g)\circ S - g\circ S\right\|_{L^\infty} \le \left\|\pi_n g - g\right\|_{L^\infty} \to 0$$

as $n\to\infty$, because g is uniformly continuous on the compact phase space X. $\left\|I_2\right\|_{L^\infty} \to 0$ as $n\to\infty$ if $g\circ S$ is uniformly continuous as well. This follows from the continuity of S. □

3.5 A more general case

Note from the proof of Theorem 3.7, that except for the boundedness and piecewise continuity assumptions made on the basis functions (which we would not like to weaken), four conditions were used to end up with the (undesired) result:

- positivity of q_n;
- $\int q_n(x,\cdot) = 1$ for all x;
- projection property: $\int q_n(x,\cdot)\varphi(x) = \pi_n\mathcal{P}\varphi$ for all $\varphi \in V_n$; and
- symmetry: $\bar{q}_n(x,y) = \bar{q}_n(y,x)$ for all $x,y\in X$.

It is clear that the first three conditions are necessary if the Galerkin projection should be viewed as a s.r.p. However, we may wish to drop symmetry. The third condition tells us that it was also unnecessary strong to assume $\pi_n\mathcal{P} = \mathcal{P}_{q_n}$ on L^1; instead of this, for our purposes it would be sufficient to claim

- $\pi_n\mathcal{P} = \mathcal{P}_{q_n}$ on V_n, and

- $\mathcal{P}_{q_n}$ has a fixed point in V_n.

Thus, we also have the needed freedom to drop the symmetry of $\bar{q}$, since it was the consequence of $\pi_n\mathcal{P} = \mathcal{P}_{q_n}$ on L^1; cf. (3.3) and (3.4). We end up with the following task: find q_n with

(a) $q_n \geq 0$ a.e.,

(b) $\int q_n(x,\cdot) = 1$ a.e.,

(c) $\int q_n(x,\cdot)\varphi(x) = \pi_n\mathcal{P}\varphi$ for all $\varphi \in V_n$, and

(d) there is a $0 \leq \varphi_n^* \in V_n$ such that $\mathcal{P}_{q_n}\varphi_n^* = \varphi_n^*$.

Note, that the third assumption cannot be valid, if there is a dynamical system S and a positive function $\varphi \in V_n$ such that $\pi_n\mathcal{P}\varphi \ngeq 0$. Answering the question, if there is a q_n satisfying (a)–(d), may need further specifications of the approximation space and/or the dynamical system. This lies beyond the scope of this work, however, could be the topic of future investigations.

Remark 3.10. Another possibility to break symmetry, but still obtain an explicit representation of the transition density q_n, would be to consider Petrov–Galerkin discretizations. This would imply $q_n(x,y) = \bar{q}(S(x),y)$, with $\bar{q}(x,y) = \Psi_n(x)^\top A_n^{-1}\Phi_n(y)$, where $\Psi_n = (\psi_1,\ldots,\psi_n)^\top$ and $A_n = \int \Psi_n\Phi_n^\top$. To my knowledge, Petrov–Galerkin methods were only used in [Din91] to discretize transfer operators. Their approximation space consists of globally continuous piecewise linear and quadratic functions, the test functions are piecewise constant. Since this discretization leads to a Markov operator, as they show, it may be another interesting topic for a future work to investigate this from the point of view represented here.

Chapter 4

The Sparse Ulam method

4.1 Motivation and outline

If the set where the essential dynamical behavior of a system takes place is of nonzero Lebesgue measure in a high dimensional space, or if we have not enough knowledge about the system to ease our numerical computations by reducing the dimension of the computational domain, transfer operator methods will suffer from the curse of dimension; cf. Section 2.3. In such cases a more efficient approximation of the eigenfunctions of the transfer operator would be desirable. Of course, without any further assumptions on these functions this is hardly possible. However, in particular cases where the dynamics is subject to a (small) random perturbation, invariant densities and other dominant eigenfunctions of the FPO tend to show regularities like Lipschitz continuity. There should not occur any high oscillatory behavior in the eigenfunctions, since due to the random perturbation the system reaches states close to each other with almost the same probability. A similar statement on geometrical regularity is shown in [Jun04].

As approximation methods for regular scalar functions on high dimensional domains, *sparse grid* techniques have been very successfully used in different fields in the last decade. The idea goes back to [Smo63], where an efficient quadrature method was proposed for evaluating integrals of specific functions. Later, it was extended to interpolation and the solution of partial differential equations [Zen91], see also the comprehensive work [Bun04].

Sparse grid interpolation allows us to achieve a higher approximation accuracy by employing a smaller number of basis function. This is done by replacing the usual basis,

where all basis functions are "equal" (characteristic functions over boxes), and using a hierarchical basis instead. By comparing the approximation potential of the functions on the different levels of this hierarchy, the most "efficient" basis is constructed under the constraint, that the maximal number of all basis functions is given.

We propose to work with the transfer operator projected onto the sparse grid approximation spaces consisting of piecewise constant functions. The resulting method is derived by giving a short introduction to sparse grids in Section 4.2, and discussing some properties of the discretized operator in Section 4.3. A detailed analysis of the efficiency and numerical realization is given in Section 4.4; with particular focus on a comparison with Ulam's method. Section 4.5 includes two examples on which our method is tested and compared with Ulam's method. Finally, the conclusions are drawn in Section 4.6.

The results have partially been published in [Jun09].

4.2 Hierarchical Haar basis

We describe the Haar basis on the d-dimensional unit cube $[0,1]^d$, deriving the multi-dimensional basis functions from the one dimensional ones, see e.g. [Gri99]. Let

$$f_{\text{Haar}}(x) = -\text{sign}(x) \cdot (|x| \leq 1), \tag{4.1}$$

where $(|x| \leq 1)$ equals 1, if the inequality is true, otherwise 0. A basis function of the Haar basis is defined by the two parameters *level* i and *center (point)* j:

$$f_{i,j}(x) := \begin{cases} 1 & \text{if } i = 0, \\ 2^{\frac{i-1}{2}} \cdot f_{\text{Haar}}\left(2^i \left(x - x_{i,j}\right)\right) & \text{if } i \geq 1, \end{cases} \tag{4.2}$$

where

$$x_{i,j} := (2j+1)/2^i, \qquad j \in \{0, \ldots, 2^{i-1} - 1\}. \tag{4.3}$$

A d-dimensional basis function is constructed from the one dimensional ones using a tensor product construction:

$$\varphi_{\mathbf{k},\mathbf{l}}(x) := \prod_{i=1}^{d} f_{\mathbf{k}_i,\mathbf{l}_i}(x_i), \tag{4.4}$$

for $x = (x_1, \ldots, x_d) \in [0,1]^d$. Here $\mathbf{k} = (\mathbf{k}_1, \ldots, \mathbf{k}_d)$, $\mathbf{k}_i \in \{0, 1, 2, \ldots\}$, denotes the level of the basis function and $\mathbf{l} = (\mathbf{l}_1, \ldots, \mathbf{l}_d)$, $\mathbf{l}_i \in \{0, \ldots, 2^{\mathbf{k}_i} - 1\}$, its center.

Theorem 4.1 (Haar basis)**.** *The set*

$$H = \left\{ f_{i,j} \mid i \in \mathbb{N}_0, j \in \{0, \dots, 2^i - 1\} \right\}$$

is an orthonormal basis of $L^2([0,1])$*, the* Haar basis*. Similarly, the set*

$$H^d = \left\{ \varphi_{\mathbf{k},\mathbf{l}} \mid \mathbf{k} \in \mathbb{N}_0^d, \mathbf{l}_i \in \{0, \dots, 2^{\mathbf{k}_i} - 1\} \right\}$$

is an orthonormal basis of $L^2([0,1]^d)$*.*[1]

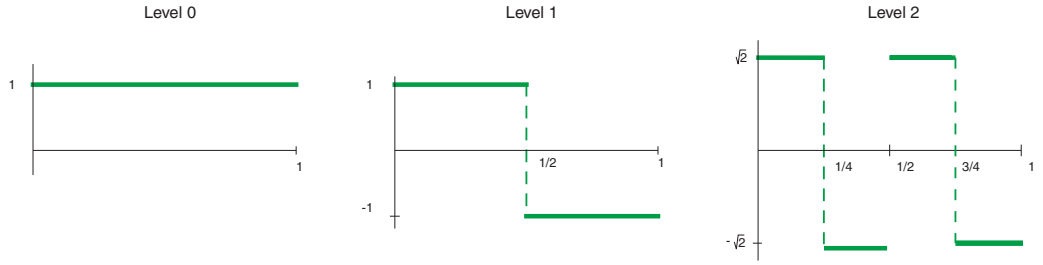


Figure 4.1: Haar basis - First three levels of the 1D Haar basis

Figure 4.1 shows the basis functions of the first three levels of the one dimensional Haar basis. It will prove useful to collect all basis functions of one level in one subspace:

$$W_{\mathbf{k}} := \operatorname{span} \left\{ \varphi_{\mathbf{k},\mathbf{l}} \mid \mathbf{l}_i \in \{0, \dots, 2^{\mathbf{k}_i} - 1\} \right\}, \qquad \mathbf{k} \in \mathbb{N}_0^d. \tag{4.5}$$

Consequently, $L^2 = L^2([0,1]^d)$ can be written as the infinite direct sum of the subspaces $W_{\mathbf{k}}$,

$$L^2 = \bigoplus_{\mathbf{k} \in \mathbb{N}_0^d} W_{\mathbf{k}}. \tag{4.6}$$

In fact, $L^1 = L^1([0,1]^d) = \bigoplus_{\mathbf{k} \in \mathbb{N}_0^d} W_{\mathbf{k}}$ holds as well, because L^2 is dense in L^1. Moreover, we have

$$\dim W_{\mathbf{k}} = \prod_{i=1}^{d} 2^{\max\{0, \mathbf{k}_i - 1\}} = 2^{\sum_{\mathbf{k}_i \neq 0} \mathbf{k}_i - 1}. \tag{4.7}$$

[1] The claim can be easily seen by observing that all Haar functions up to level ℓ span the space of piecewise constant functions over an equipartition of the unit interval into 2^ℓ subintervals. The union of these spaces for $\ell \to \infty$ is known to be dense in $L^2([0,1])$. Analogously follows the multidimensional case.

In order to get a finite dimensional approximation space most appropriate for our purposes, we are going to choose an optimal finite subset of the basis functions $\varphi_{\mathbf{k},\mathbf{l}}$. Since in general we do not have any a priori information about the function to be approximated, and since all basis functions in one subspace $W_{\mathbf{k}}$ deliver the same contribution to the approximation error, we will use either all or none of them. In other words, the choice for the approximation space is transferred to the level of subspaces $W_{\mathbf{k}}$.

4.2.1 Approximation properties

The choice of the optimal set of subspaces $W_{\mathbf{k}}$ relies in the contribution of each of these to the approximation error. The following statements give estimates on this.

Lemma 4.2. *Let $f \in C^1([0,1])$ and let $c_{i,j}$ be its coefficients with respect to the Haar basis, i.e. $f = \sum_{ij} c_{i,j} f_{i,j}$. Then for $i > 0$ and all j*

$$\left|c_{i,j}\right| \leq 2^{-\frac{3i+1}{2}} \left\|f'\right\|_\infty .$$

For $f \in C^1\left([0,1]^d\right)$ we analogously have for $\mathbf{k} \neq \mathbf{0}$ and all $\mathbf{l}$

$$\left|c_{\mathbf{k},\mathbf{l}}\right| \leq 2^{-\left(\sum_{\mathbf{k}_i \neq 0} 3\mathbf{k}_i + 1\right)/2} \prod_{\mathbf{k}_i \neq 0} \left\|\partial_i f\right\|_\infty .$$

Proof. For $i \geq 1$

$$\begin{aligned} 2^{\frac{1-i}{2}} c_{ij} &= \int_{x_j - 2^{-i}}^{x_j} f - \int_{x_j}^{x_j + 2^{-i}} f \\ &= \int_{x_j - 2^{-i}}^{x_j} \left(f(x_j) + \int_{x_j}^{x} f' \right) \mathrm{d}x - \int_{x_j}^{x_j + 2^{-i}} \left(f(x_j) + \int_{x_j}^{x} f' \right) \mathrm{d}x \end{aligned}$$

and thus

$$2^{\frac{1-i}{2}} \left|c_{ij}\right| \leq 2 \left\|f'\right\|_\infty \int_0^{2^{-i}} x \, \mathrm{d}x,$$

which yields the claimed estimate for the 1d case. The bound in the d-dimensional case follows similarly. □

Using this bound on the contribution of a single basis function to the approximation of a given function f, we can derive a bound on the total contribution of a subspace $W_{\mathbf{k}}$.

For $f_{\mathbf{k}} \in W_{\mathbf{k}}$

$$\left\|f_{\mathbf{k}}\right\|_{L^1} \leq 2^{-\sum_{\mathbf{k}_i\neq 0}(\mathbf{k}_i+1)} \prod_{\mathbf{k}_i\neq 0} \|\partial_i f\|_\infty, \tag{4.8}$$

$$\left\|f_{\mathbf{k}}\right\|_{L^2} \leq 2^{-\sum_{\mathbf{k}_i\neq 0}(\mathbf{k}_i+3)/2} \prod_{\mathbf{k}_i\neq 0} \|\partial_i f\|_\infty. \tag{4.9}$$

4.2.2 The optimal subspace

The main idea of the sparse grid approach is to choose *cost* and (approximation) *benefit* of the approximation subspace in an optimal way. We briefly sketch this idea here, for a detailed exposition see [Zen91, Bun04]. For a set $\mathbf{I} \subset \mathbb{N}_0^d$ of multiindices we define

$$W_{\mathbf{I}} = \bigoplus_{\mathbf{k}\in\mathbf{I}} W_{\mathbf{k}}.$$

Correspondingly, for $f \in L^1$, let $f_{\mathbf{I}} = \sum_{\mathbf{k}\in\mathbf{I}} f_{\mathbf{k}}$, where $f_{\mathbf{k}}$ is the L^2-orthogonal projection[1] of f onto $W_{\mathbf{k}}$. We define the *cost* $C(\mathbf{k})$ of a subspace $W_{\mathbf{k}}$ as its dimension,

$$C(\mathbf{k}) = \dim W_{\mathbf{k}} = 2^{\sum_{\mathbf{k}_i\neq 0} \mathbf{k}_i - 1}.$$

Since

$$\left\|f - f_{\mathbf{I}}\right\| \leq \sum_{\mathbf{k}\notin\mathbf{I}} \left\|f_{\mathbf{k}}\right\| = \sum_{\mathbf{k}\in\mathbb{N}_0^d} \left\|f_{\mathbf{k}}\right\| - \sum_{\mathbf{k}\in\mathbf{I}} \left\|f_{\mathbf{k}}\right\|, \tag{4.10}$$

the guaranteed increase in accuracy is bounded by the contribution of a subspace $W_{\mathbf{k}}$ which we add to the approximation space. We therefore define the *benefit* $B(\mathbf{k})$ of $W_{\mathbf{k}}$ as the upper bound on its L_1-contribution as derived above,

$$B(\mathbf{k}) = 2^{-\sum_{\mathbf{k}_i\neq 0}(\mathbf{k}_i+1)}. \tag{4.11}$$

Note that we omited the factor involving derivatives of f. The reason is that it does not affect the solution of the optimization problem (4.12)

Let $C(\mathbf{I}) = \sum_{\mathbf{k}\in\mathbf{I}} C(\mathbf{k})$ and $B(\mathbf{I}) = \sum_{\mathbf{k}\in\mathbf{I}} B(\mathbf{k})$ be the total cost and the total benefit of the approximation space $W_{\mathbf{I}}$. In order to find the optimal approximation

[1] Note that since all functions in $W_{\mathbf{k}}$ are piecewise constant and have compact support, this projection is well defined on L^1 as well.

space we are now solving the following optimization problem: Given a bound $c > 0$ on the total cost, find an approximation space $W_{\mathbf{I}}$ which solves

$$\max_{C(\mathbf{I}) \leq c} B(\mathbf{I}). \tag{4.12}$$

One can show (cf. [Bun04]) that $\mathbf{I} \subset \mathbb{N}_0^d$ is an optimal solution to (4.12) iff

$$\frac{C(\mathbf{k})}{B(\mathbf{k})} = \text{const} \qquad \text{for } \mathbf{k} \in \partial\mathbf{I}, \tag{4.13}$$

where the *boundary* $\partial\mathbf{I}$ is given by $\partial\mathbf{I} = \{\mathbf{k} \in \mathbf{I} \mid \mathbf{k}' \in \mathbf{I}, \mathbf{k}' \geq \mathbf{k} \ \Rightarrow \mathbf{k}' = \mathbf{k}\}$[1]. Using the definitions for cost and benefit as introduced above, we obtain

$$\frac{C(\mathbf{k})}{B(\mathbf{k})} = \frac{2^{\sum_{\mathbf{k}_i \neq 0}(\mathbf{k}_i - 1)}}{2^{-\sum_{\mathbf{k}_i \neq 0}(\mathbf{k}_i + 1)}} = 2^{2\sum_{\mathbf{k}_i \neq 0} \mathbf{k}_i} = 2^{2|\mathbf{k}|}, \tag{4.14}$$

where $|\mathbf{k}|$ means the 1-norm of the vector $\mathbf{k}$.
The optimality condition (4.13) thus translates into the simple condition

$$|\mathbf{k}| = \text{const} \quad \text{for } \mathbf{k} \in \partial\mathbf{I}. \tag{4.15}$$

As a result, the optimal approximation space is $W_{\mathbf{I}(N)}$ with

$$\mathbf{I}(N) = \left\{ \mathbf{k} \in \mathbb{N}_0^d \mid |\mathbf{k}| \leq N \right\}, \tag{4.16}$$

where the *level* $N = N(c) \in \mathbb{N}$ is depending on the chosen cost bound c. Figure 4.2 schematically shows the basis functions of the optimal subspace in $2D$ for $N = 3$.

Remark 4.3. Because of the orthogonality of the Haar-basis in L^2 one can take the squared contribution as the benefit in the L^2-case (resulting in equality in (4.10)). In this case we obtain the optimality condition

$$\sum_{\mathbf{k}_i \neq 0} (\mathbf{k}_i + 1) = \text{const} \quad \text{for } \mathbf{k} \in \partial\mathbf{I} \tag{4.17}$$

and correspondingly $W_{\mathbf{I}}$ with

$$\mathbf{I}(N) = \left\{ \mathbf{k} \in \mathbb{N}_0^d : \sum_{\mathbf{k}_i \neq 0} (\mathbf{k}_i + 1) \leq N \right\}, \tag{4.18}$$

$N = N(c)$, as the optimal approximation space.

[1] $\mathbf{k}' \geq \mathbf{k}$ is meant componentwise

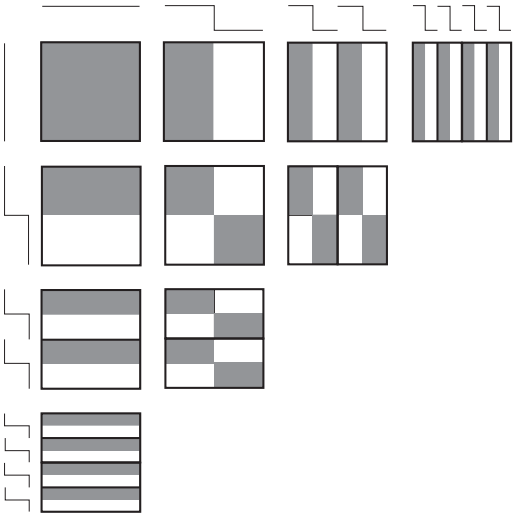


Figure 4.2: 3^{rd} **level sparse basis in two dimensions** - Shaded means value 1, white means value -1, thicker lines are support boundaries.

4.3 The discretized operator

Having chosen the optimal approximation space $V_N = W_{\mathbf{I}(N)}$ we now build the corresponding discretized Frobenius-Perron operator P_N. Since the sparse basis

$$B_N := \left\{ \varphi_{\mathbf{k},\mathbf{l}} \mid |\mathbf{k}| \leq N, \mathbf{l}_i \in \{0, \ldots, 2^{\mathbf{k}_i} - 1\} \right\} \tag{4.19}$$

is an L^2-orthogonal basis of V_N, the natural projection $\pi_N : L^2 \to V_N$ is given by

$$\pi_N f = \sum_{\varphi \in B_N} \left(\int f\varphi \right) \varphi. \tag{4.20}$$

As noted in the previous section, the above definition of π_N makes it well defined on L^1 as well. Choosing an arbitrary enumeration of the basis, the (*transition*) matrix P_N of the discretized Frobenius-Perron operator

$$\mathcal{P}_N = \pi_N \mathcal{P}$$

with respect to B_N has entries

$$P_{N,ij} = \int \varphi_i \, \mathcal{P}\varphi_j. \tag{4.21}$$

Writing $\varphi_i = \varphi_i^+ - \varphi_i^- = |\varphi_i| \cdot (\chi_i^+ - \chi_i^-)$, where $|\varphi_i|$ is the (constant) absolute value of the function over its support and χ_i^+ and χ_i^- are the characteristic functions on the supports of the positive and negative parts of φ_i, we obtain

$$P_{N,ij} = |\varphi_i||\varphi_j| \left(\int \chi_i^+ \ \mathcal{P}\chi_j^+ - \int \chi_i^- \ \mathcal{P}\chi_j^+ - \int \chi_i^+ \ \mathcal{P}\chi_j^- + \int \chi_i^- \ \mathcal{P}\chi_j^- \right), \tag{4.22}$$

which is, by (2.16)

$$P_{N,ij} = |\varphi_i||\varphi_j| \sum \pm m \left(X_j^\pm \cap S^{-1} \left(X_i^\pm \right) \right), \tag{4.23}$$

where $X_i^\pm = \operatorname{supp}(\varphi)_i^\pm$ and we add the 4 summands like in (4.22). These can be computed in the same way as presented in Section 2.3.

Remark 4.4. We note that

(a) if the ith basis function is the one corresponding to $\mathbf{k} = (0, \ldots, 0)$, then

$$P_{N,ij} = \delta_{ij}.$$

(b) The entries of P_N are bounded via

$$\left| P_{N,ij} \right| \leq \sqrt{\frac{m(X_j)}{m(X_i)}} \leq 2^{N/2}.$$

(c) If $P_N y = \lambda y$ for a $0 \neq y \in \mathbb{C}^{\dim V_N}$ with $\lambda \neq 1$, then $y_i = 0$ if the i^{th} basis function is the one corresponding to $\mathbf{k} = (0, \ldots, 0)$. This follows from

$$y_i \overset{(a)}{=} (e_i^\top P_N) y = e_i \lambda y = \lambda y_i. \tag{4.24}$$

It is straightforward to show that this property is shared by every Ulam type projection method with a constant function as element of the basis of the approximation space. This observation is useful for the reliable computation of an eigenvector at an eigenvalue close to one (since it is ill conditioned): (4.24) allows us to reduce the eigenproblem to the subspace orthogonal to the constant function.

Properties (a)–(c) are valid for the numerical realization as well.

4.3.1 Convergence

As has been pointed out in Section 2.3, statements about the convergence of Ulam's method exist in certain cases. For certain random perturbations of S we obtain the convergence of the Sparse Ulam method by applying the same arguments as for Ulam's method, [Del99], and the following lemma. An open question is, if in general, the convergence of Ulam's method implies convergence of Sparse Ulam and vice versa.

Lemma 4.5. $\left\|\pi_N f - f\right\|_{L^p} \overset{n\to\infty}{\longrightarrow} 0$ *for* $f \in L^2$, $p = 1, 2$.

Proof. The convergence in the L^2-norm is trivial.

Since X is bounded we have $L^2(X) \subset L^1(X)$. Moreover, there is a constant $c_2 > 0$ such that

$$\|h\|_{L^1} \leq c_2 \|h\|_{L^2} \qquad \forall h \in L^2(X).$$

Thus convergence in L^2-norm also implies the convergence in L^1 norm. □

4.3.2 Spectral properties of the operator

Unlike the transition matrix from Ulam's method, the one from the Sparse Ulam method, P_N, is not stochastic. Therefore, we cannot bound its spectrum in advance. Such bounds are desirable, to know, e.g. if the eigenvalues we are searching for are in fact the ones with the greatest magnitude. In this section we aim to find bounds on the spectrum of P_N.

Let V_N^U be the ("Ulam type") space spanned by characteristic functions over a full equipartition of $[0,1]^d$ with a resolution 2^N in each dimension, and let π_N^U denote the L^2-orthogonal projection onto V_N^U. Then $\pi_N = \pi_N \pi_N^U$ and hence $\mathcal{P}_N = \pi_N \mathcal{P}_N^U$, with $\mathcal{P}_N^U = \pi_N^U \mathcal{P}$ the Ulam matrix for the full grid. Thus, the Sparse Ulam transition matrix is the product of a projector $\Pi_N \in \mathbb{R}^{2^{dN} \times 2^{dN}}$, which is the matrix representation of $\pi_N : V_N^U \to V_N \subset V_N^U$, and a stochastic matrix $T \in \mathbb{R}^{2^{dN} \times 2^{dN}}$, the matrix representation of $\mathcal{P}_N^U$ (for both operators, the underlying basis is chosen to be the set of characteristic functions of the partition elements). We determine the projector:

Lemma 4.6. *Let* X_i *denote the partition elements of the full grid box covering, and choose* $x_i \in X_i$ *arbitrary. With* $\mathbf{x} = (x_1, \ldots, x_{2^{dN}})^\top$ *and* $m_X := m(X_i)$ *we have*

$$\Pi_{N,ki} = m_X \sum_{j=1}^{dim\ V_N} \varphi_j(x_i)\varphi_j(x_k).$$

Alternatively,

$$\pi_N = R_N R_N^\top$$

with $R_{N,ij} = \sqrt{m_X}\varphi_j(x_i)$. *The columns of* R_N *are mutually perpendicular.*

Proof. Projecting one characteristic function (χ_i) onto V_N yields

$$\begin{aligned}\Pi_{N,ki} &= \sum_j \underbrace{\left\langle \chi_i, \varphi_j \right\rangle_{L^2}}_{=\varphi_j(x_i)m(X_i)} \varphi_j(x_k) \\ &= m(X_i)\sum_j \varphi_j(x_i)\varphi_j(x_k).\end{aligned}$$

Because the partition elements are all congruent, $m(X_i) = m_X\ \forall i$. This gives the first claim. The second follows by the L^2-orthogonality of the basis functions and the proper scaling. □

It follows

Corollary 4.7. *For the projection* Π_N *the following equations hold:*

(a) $\Pi_N = \Pi_N^\top$.

(b) $\Pi_N^2 = \Pi_N$.

(c) $\Pi_N e = e$ *(the constant function is projected to itself).*

Properties (a) and (b) say, that Π_N is an orthogonal projector [Tre97]. Our first observation based on numerical experiments led us to

Conjecture 4.8 (Norm of Π_N)**.** *For* $d = 2$ *it holds* $\left\|\Pi_N\right\|_1 = \left\|\Pi_N\right\|_\infty = 1 + N/2$.

The second observation, also based on numerical experiments, is important. Its validity would mean, that the spectrum of the Sparse Ulam transition matrix lies in the unit disk.

Conjecture 4.9 (Spectrum of P_N)**.** *For any stochastic matrix* T *we have* $\sigma(\Pi_N T) \subset B_1(0)$.

It is interesting, that for an *arbitrary* projection Π the properties (a)–(c) of Corollary 4.7 are not sufficient to obtain $\sigma(\Pi T) \subset B_1(0)$.

Example 4.10. Define $v_1 = (1,1,1,1)^\top$, $v_2 = (1,3,1,0)^\top$ and $v_3 = (0,3,3,1)^\top$. Let Π be the orthogonal projector onto the subspace of $\mathbb{R}^4$ spanned by v_1, v_2 and v_3. Let

$$T = \begin{pmatrix} T_1 & \\ & T_2 \end{pmatrix}, \qquad T_1 = T_2 = \begin{pmatrix} 1 & 0.5 \\ 0 & 0.5 \end{pmatrix}.$$

Then there is a $\lambda \in \sigma(\Pi T)$ with $\lambda \geq 1.006$.

Therefore, if Conjecture 4.9 is valid, it has to be a consequence of the special structure of the Sparse Ulam discretization. A deeper analysis of this problem exploiting spatial pattern of the basis functions in V_N could be the subject of future work.

4.4 Numerical computation and complexity

In this section, we collect basic statements about the complexity of both methods.

4.4.1 Cost and accuracy

We defined the total cost of an approximation space as its dimension and the accuracy via its *contribution* or *benefit*, see (4.11). In this section we derive a recurrence formula for these numbers, depending on the *level* of the optimal subspaces and the system dimension.

Let $C(N,d)$ be the dimension of $W_{\mathbf{I}(N)}$ in phase space dimension d. Then

$$C(N,d) = C(N,d-1) + \sum_{k=1}^{N} C(N-k,d-1)2^{k-1}, \tag{4.25}$$

since if $\mathbf{k} = (*,\ldots,*,0)$, then the last dimension does not affect the number of basis functions, and the total number of basis function's for such $\mathbf{k}$'s is $C(N,d-1)$. If $\mathbf{k} = (*,\ldots,*,\mathbf{k}_d)$ with $\mathbf{k}_d > 0$, then the number of basis functions with such $\mathbf{k}$'s is $C(N-\mathbf{k}_d,d-1)2^{\mathbf{k}_d-1}$, because there are $2^{\mathbf{k}_d-1}$ one-dimensional basis functions of level $\mathbf{k}_d$ possible for the tensor product in the last dimension. For $d=1$ we simply deal with the standard Haar basis, so $C(N,1) = 2^N$.

Lemma 4.11.

$$C(N,d) \doteq \frac{N^{d-1}\ 2^{N-d+1}}{(d-1)!}, \tag{4.26}$$

where $\doteq$ means the leading order term in N.

Proof. By induction on d. The claim holds clearly for $d = 1$. Assume, it holds for $d-1$. By considering the recurrence formula (4.25), we see that $C(N,d) = p(N)\ 2^N$, where p is a polynomial of order less or equal to d. Consequently,

$$\begin{aligned} C(N,d) &\doteq \frac{N^{d-2}\ 2^{N-d+2}}{(d-2)!} + \sum_{k=1}^{N} \frac{(N-k)^{d-2}\ 2^{N-k-d+2}}{(d-2)!} 2^{k-1} \\ &= \frac{N^{d-2}\ 2^{N-d+2}}{(d-2)!} + \frac{2^{N-d+1}}{(d-2)!} \sum_{k=1}^{N} (N-k)^{d-2} \\ &\doteq \frac{N^{d-2}\ 2^{N-d+2}}{(d-2)!} + \frac{2^{N-d+1}}{(d-2)!} \frac{N^{d-1}}{d-1} \\ &\doteq \frac{N^{d-1}\ 2^{N-d+1}}{(d-1)!} \end{aligned}$$

□

According to (4.10), the approximation error $\left\|f - f_{\mathbf{I}}\right\|$ is bounded by $\sum_{\mathbf{k}\notin\mathbf{I}} \left\|f_{\mathbf{k}}\right\|$, i.e.

$$\left\|f - f_{\mathbf{I}}\right\| \leq \sum_{|\mathbf{k}|>N} \left\|f_{\mathbf{k}}\right\|,$$

if we use the optimal approximation space $W_{\mathbf{I}(N)}$. By (4.8) this means

$$\left\|f - f_{\mathbf{I}}\right\| \leq \sum_{|\mathbf{k}|>N} \left[2^{-\sum_{\mathbf{k}_i\neq 0}(\mathbf{k}_i+1)} \prod_{\mathbf{k}_i\neq 0} \left\|\partial_i f\right\|_\infty \right]$$

Again, the constants $\prod_{\mathbf{k}_i\neq 0} \left\|\partial_i f\right\|_\infty$ only depend on the function to be approximated. Thus, without a priori knowledge about f we need to assume that they can be bounded by some common constant and accordingly define the discretization error of the Nth level sparse basis as

$$E(N,d) = \sum_{|\mathbf{k}|>N} 2^{-\sum_{\mathbf{k}_i\neq 0}(\mathbf{k}_i+1)}. \tag{4.27}$$

Let $E(-n,d)$ for $n \in \mathbb{N}, n > 0$ represent the error of the empty basis and $\mathbf{k} = (\tilde{\mathbf{k}}, \mathbf{k}_d)$

with $\tilde{\mathbf{k}} \in \mathbb{N}_0^{d-1}$. Then

$$\begin{aligned} E(N,d) &= \sum_{|\mathbf{k}|>N} 2^{-\sum_{\mathbf{k}_i\neq 0}(\mathbf{k}_i+1)} \\ &= \sum_{\mathbf{k}_d=0}^{\infty} 2^{-(\mathbf{k}_d+1)(\mathbf{k}_d\neq 0)} \sum_{|\tilde{\mathbf{k}}|>N-\mathbf{k}_d} 2^{-\sum_{\tilde{\mathbf{k}}_i\neq 0}(\tilde{\mathbf{k}}_i+1)} \\ &= \sum_{\mathbf{k}_d=0}^{\infty} 2^{-(\mathbf{k}_d+1)(\mathbf{k}_d\neq 0)} E(N-\mathbf{k}_d, d-1), \end{aligned}$$

where the expression $(\mathbf{k}_i \neq 0)$ has the value 1, if it is true, otherwise 0. By splitting the sum, this leads to the recurrence formula

$$E(N,d) = E(N,d-1) + \sum_{k=1}^{N} E(N-k,d-1)2^{-k-1} + \underbrace{\sum_{k=N+1}^{\infty} 2^{-k-1} E(-1,d-1)}_{=2^{-N-1}E(-1,d-1)}. \quad (4.28)$$

We easily compute that $E(N,1) = 2^{-N-1}$ for $N \geq 0$ and $E(-1,d) = (3/2)^d$.

Lemma 4.12.

$$E(N,d) \doteq \frac{N^{d-1}\ 2^{-N-d}}{(d-1)!}, \quad (4.29)$$

where, again, $\doteq$ means the leading order term in N.

Proof. By induction on d. The claim holds for $d = 1$, assume it holds for $d-1$. Then

$$\begin{aligned} E(N,d) &\doteq \frac{N^{d-2}2^{-N-d+1}}{(d-2)!} + \sum_{k=1}^{N} \frac{(N-k)^{d-2}2^{-N+k-d+1}}{(d-2)!} 2^{-k-1} + \left(\frac{3}{2}\right)^{d-1} 2^{-N-1} \\ &\doteq \frac{N^{d-2}2^{-N-d+1}}{(d-2)!} + \frac{2^{-N-d}}{(d-2)!} \sum_{k=1}^{N} (N-k)^{d-2} \\ &\doteq \frac{2^{-N-d}}{(d-2)!} \frac{N^{d-1}}{d-1} \end{aligned}$$

□

An asymptotic estimate. In order to be able to give more precise asymptotic estimates we define beyond the estimation sign $\sim$ ($a_n \sim b_n$ iff $a_n \lesssim b_n$ and $b_n \lesssim a_n$) another one. By abusing the common sign $\approx$, we say $a_n \approx b_n$ iff $a_n/b_n \to 1$ as $n \to \infty$. Since this is meant in limit, there should be no confusion with the original meaning of the sign.

Let us fix the dimension d and define

$$C(N) := \frac{N^{d-1}\ 2^{N-d+1}}{(d-1)!} \approx C(N,d), \tag{4.30}$$

$$E(N) := \frac{N^{d-1}\ 2^{-N-d}}{(d-1)!} \approx E(N,d). \tag{4.31}$$

We prescribe the accuracy ε and let $N(\varepsilon)$ be the smallest solution to $E(N) \leq \varepsilon$. Further, we define $C(\varepsilon) := C(N(\varepsilon))$, the (approximative) costs to achieve the desired accuracy. We would like to derive an asymptotic estimate for $C(\varepsilon)$ as $\varepsilon \to 0$.

First we take the logarithm of (4.31):

$$N = (d-1)\log_2 N + \log_2 \varepsilon^{-1} + \text{const.}$$

Using $N - (d-1)\log_2 N \approx N$ we get

$$N \approx \log_2 \varepsilon^{-1}. \tag{4.32}$$

Dividing (4.30) by (4.31) we obtain

$$2^{N+1} = \sqrt{\frac{C(\varepsilon)}{\varepsilon}}.$$

Substituting this and (4.32) into (4.30) we have[1]

$$C(\varepsilon) \approx \frac{1}{(d-1)!\ 2^d} \left(\log_2 \varepsilon^{-1}\right)^{d-1} \sqrt{\frac{C(\varepsilon)}{\varepsilon}},$$

which we may solve for $C(\varepsilon)$:

$$C(\varepsilon) \approx \frac{1}{\left((d-1)!\ 2^d\right)^2} \left(\log_2 \varepsilon^{-1}\right)^{2d-2} \varepsilon^{-1}. \tag{4.33}$$

Remark 4.13. It is important that (4.33) is an asymptotic estimate as $N \to \infty$. It does not say anything about the behavior of the complexity in d. Moreover, it gives the false intuition, that $C(\varepsilon) \to 0$ as $d \to \infty$. In fact, as d gets bigger, the more smaller part $\mathbf{I}(N)$ is of the "full" index set $\left\{\mathbf{k} \mid \mathbf{k}_i \leq N\right\}$ (think of the d dimensional simplex in the d dimensional cube), where latter has the approximation potential $\mathcal{O}\left(2^{-N}\right)$ independent of d. So, the approximation error of V_N is increasing in d.

[1] Note that $a_n^k \approx b_n^k$ for $k \in \mathbb{N}$ if $a_n \approx b_n$, since $(a_n/b_n)^k \to 1$ if $a_n/b_n \to 1$.

Comparison with Ulam's method. We now compare the expressions for the asymptotic behavior of cost and discretization error in dependence of the discretization level N and the problem dimension d in Lemmata 4.11 and 4.12 to the corresponding expressions for the standard Ulam basis, i.e. the span of the characteristic functions on a uniform partition of the unit cube into cubes of edge length 2^{-M} in each coordinate direction — this is $\bigoplus_{\|\mathbf{k}\|_\infty \leq M} W_{\mathbf{k}}$. This space consists of $(2^M)^d$ basis functions, the discretization error is $\mathcal{O}\left(2^{-M}\right)$.

We thus have — up to constants — the following asymptotic expressions for cost and error of the sparse and the standard basis:

	cost	error
sparse basis	$(N/2)^{d-1}\, 2^{N-d}$	$(N/2)^{d-1}\, 2^{-N-d}$
standard basis	2^{dM}	2^{-M}

Table 4.1: Cost and accuracy comparison

To highlight the main difference compare the cost (approximation space dimension) estimate $n_{\mathrm{Ulam}} = \mathcal{O}\left(\varepsilon^{-d}\right)$ for Ulam's method (cf. Section 2.3) with (4.33); $n_{\mathrm{SpU}} = \mathcal{O}\left(\left(\log_2 \varepsilon^{-1}\right)^{2d-2} \varepsilon^{-1}\right)$. The dimension appears in the exponent only for the logarithmic term, which indicates the partial overcoming of the curse of dimension.

Since we neglected lower order terms in the estimate (4.33), the only conclusion we can draw from this is that from a certain accuracy requirement on, the sparse basis is more efficient than the standard one.

However, the number of basis functions is not the only cost source to look at: we also have to assembly the discretized operator, and for this, compute the matrix entries.

When using Monte Carlo quadrature in order to approximate the entries of the transition matrix in both methods, the overall computation breaks down into the following three steps:

1. mapping the sample points,
2. constructing the transition matrix,
3. solving the eigenproblem.

While steps 1. and 3. are identical for both methods, step 2. differs significantly. This is due to the fact that in contrast to Ulam's method, the basis functions of the sparse hierarchical tensor basis have global and non-disjoint supports.

4.4.2 Number of sample points

Applying Monte Carlo approximation to (4.22), we obtain

$$\tilde{p}_{ij} = |\varphi_i||\varphi_j| \left(\frac{m\left(X_j^+\right)}{K_j^+} \sum_{k=1}^{K_j^+} \chi_i^+\left(S(x_k^+)\right) - \chi_i^-\left(S(x_k^+)\right) \right. \tag{4.34}$$

$$\left. - \frac{m\left(X_j^-\right)}{K_j^-} \sum_{k=1}^{K_j^-} \chi_i^+(S(x_k^-)) - \chi_i^-(S(x_k^-)) \right), \tag{4.35}$$

where the sample points $x_k^\pm$ are chosen i.i.d. from a uniform distribution on X_j^+ and X_j^-, respectively. In fact, since the union of the supports of the basis functions in one subspace $W_{\mathbf{k}}$ covers all of X, we can reuse the same set of κ sample points and their images for each of the subspaces $W_{\mathbf{k}}$ (i.e. $\binom{N+d}{d}$ times). Note that the number $K_j^\pm$ of test points chosen in $X_j^\pm$ now varies with j since the supports of the various basis functions are of different size: on average, $K_j^\pm = \kappa m(X_j^\pm)$. Now, we estimate the total number of sample points needed to approximate the discretized operator (and its eigenfunctions) to a desired accuracy.

Error estimation. We proceed as in Section 2.3. Recall that on $\operatorname{supp}\left(\varphi_j\right)$ $\left|\varphi_j\right| = 1/\sqrt{m_j}$, with $m_j := m(X_j)$ holds. Then

$$P_{N,ij} = |\varphi_i||\varphi_j| \underbrace{\sum \pm m(X_j^\pm \cap S^{-1}(X_i^\pm))}_{=:M_{ij}}$$

and for the error:

$$\Delta P_{N,ij} = |\varphi_i||\varphi_j|\Delta M_{ij}.$$

With $i(\mathbf{l}) := \{i \mid \varphi_i \in W_{\mathbf{l}}\}$ we have, that $\{S^{-1}(X_i) \mid i \in i(\mathbf{l})\}$ is a disjoint partition of X, thus

$$\sum_{i(\mathbf{l})} |M_{ij}| = m_j \quad \forall j.$$

Further,

$$\Delta M_{ij} \sim \frac{M_{ij}}{\sqrt{\kappa m_j}}.$$

While in an Ulam type basis consisting of characteristic functions of congruent boxes all basis functions are a priori equivalent, this does not hold in the Sparse Ulam case. They have supports of different size and a rescaled basis may perform better in our error analysis. Thus, introduce a rescaled basis $\{\bar{\varphi}_j\}$ with $\bar{\varphi}_j = c_j\varphi_j$ and $c_j > 0$. Since all $\varphi_j \in W_{\mathbf{l}}$ are handled equivalently, they should have a common scaling factor $c_{\mathbf{l}}$ as well. The corresponding transition matrix writes as

$$\bar{P}_{N,ij} = \frac{c_j}{c_i} P_{ij},$$

same for $\Delta\bar{P}_N$. Hence for its columns

$$\left\|\Delta\bar{P}_{N,:j}\right\|_2^2 \sim \sum_i \left(|\varphi_j||\varphi_i|\frac{c_j}{c_i}\frac{M_{ij}}{\kappa m_j}\right)^2 = \frac{c_j^2}{\kappa m_j^2}\sum_{\mathbf{l}\in\mathbf{I}}\frac{1}{m_{\mathbf{l}}c_{\mathbf{l}}^2}\underbrace{\sum_{i(\mathbf{l})} M_{ij}^2}_{\leq(\sum_{i(\mathbf{l})}|M_{ij}|)^2\leq m_j^2} \leq \frac{c_j^2}{\kappa}\sum_{\mathbf{l}}\frac{1}{m_{\mathbf{l}}c_{\mathbf{l}}^2}$$

holds, and so (using (2.21))

$$\left\|\Delta\bar{P}_N\right\|_2 \leq \frac{1}{\sqrt{\kappa}}\left(\sum_{\mathbf{l}}\frac{1}{m_{\mathbf{l}}}c_{\mathbf{l}}^{-2}\sum_{\mathbf{l}}\frac{1}{m_{\mathbf{l}}}c_{\mathbf{l}}^2\right)^{1/2}.$$

For the orthonormal basis, i.e. $c_{\mathbf{l}} = 1$, we obtain

$$\|\Delta f\|_{L^2} \leq \frac{\sum\frac{1}{m_{\mathbf{l}}}}{\sqrt{\kappa}|\Delta\lambda|} = \frac{n_{\mathrm{SpU}}}{\sqrt{\kappa}|\Delta\lambda|}. \tag{4.36}$$

Compare estimate (4.36) with the corresponding one for Ulam's method, (2.22): they are the same up to a constant factor and by $n_{\mathrm{SpU}} \ll n_{\mathrm{Ulam}}$ we expect the Sparse Ulam method to get along with a less amount of sample points. Once again, the (asymptotic) invariance of the spectral gap in n is crucial, see Remark 2.17.

Is there any scaling, which gives a better estimate? In the new basis, a coefficient representation v of the function f means a norm

$$\|f\|_{L^2} = \|Cv\|_2,$$

with $C = diag(c_i)$. Thus,

$$\frac{\|\Delta f\|_{L^2}}{\|f\|_{L^2}} \leq \underbrace{\frac{\max c_j}{\min c_j}}_{=:\Lambda(c)}\frac{\|\Delta v\|_2}{\|v\|_2}.$$

Using the error estimate from above, we seek for a c s.t.

$$\mathcal{E}(c) := \Lambda(c) \sum \frac{1}{m_{\mathbf{l}}} c_{\mathbf{l}}^{-2} \sum \frac{1}{m_{\mathbf{l}}} c_{\mathbf{l}}^{2} = \min!$$

By the Cauchy–Schwarz inequality,

$$\mathcal{E}(c) \geq \sum \frac{1}{\sqrt{m_{\mathbf{l}}}} c_{\mathbf{l}}^{-1} \cdot \frac{1}{\sqrt{m_{\mathbf{l}}}} c_{\mathbf{l}} = \sum \frac{1}{m_{\mathbf{l}}} = n_{\mathrm{SpU}},$$

with equation iff $c_{\mathbf{l}} = 1$ for all $\mathbf{l}$. The orthonormal basis is the best choice.

Comparison with Ulam's method. Since the error estimates concerning the Monte Carlo method are very similar for the two methods, it is easy to draw the conclusion, which method needs fewer sample points. If ε is the error of the approximation space, it is a rational choice to set $\Delta f = \mathcal{O}(\varepsilon)$ as well. Taking the estimates for n_{Ulam} and n_{SpU}, substituting them into (2.22) respectively (4.36), we obtain

$$\begin{aligned} \kappa_{\mathrm{Ulam}} &= \mathcal{O}\left(\varepsilon^{-2d-2}\right), \\ \kappa_{\mathrm{SpU}} &= \mathcal{O}\left(\left(\log_2 \varepsilon^{-1}\right)^{4d-4}\right) \varepsilon^{-3}. \end{aligned}$$

Once more, these expressions allow us a qualitative comparison, how many sample points the two methods need. The dominance of the Sparse Ulam method is well highlighted by the formulas.

Generating the sample points. We discussed the number of sample points needed for the Sparse Ulam method, if they are uniformly distributed. To ensure the uniform distribution, a quasi-Monte Carlo sampling is used. First, we partition the state space into segments, and then draw a given number of (uniform) random sample points in each segment. In general, the segments are chosen to be congruent, hence the same number of samples will be drawn in each of them. The number of segments (usually chosen to be m^d with some $m \in \mathbb{N}$) is determined such that it is not too large (not more than 10^6), and that in each segment there are $\sim$100 sample points.

4.4.3 Number of index computations

While in Ulam's method each sample point is used in the computation of one entry of the transition matrix only, this is not the case in the Sparse Ulam method. In fact,

each sample point (and its image) is used in the computation of $|\mathbf{I}(N)|^2$ matrix entries, namely one entry for each pair $(W_{\mathbf{k}}, W_{\mathbf{l}})$ of subspaces.

Correspondingly, for each sample point x (and its image) and for each $\mathbf{k} \in \mathbf{I}(N)$, we have to compute the index $\mathbf{l}$ of the basis function $\varphi_{\mathbf{k},\mathbf{l}} \in W_{\mathbf{k}}$ whose support contains x. Since (cf. the previous section) the required number of sample points to achieve accuracy TOL is $\mathcal{O}\left(\frac{(\dim V_N)^2}{\mathrm{TOL}^2}\right)$ and $|\mathbf{I}(N)| = \binom{N+d}{d} \approx \frac{N^d}{d!}$ (cf. (4.16)), this leads to

$$\kappa \cdot |\mathbf{I}(N)| \lesssim \frac{N^d}{d!} \left(\frac{\dim V_N}{\mathrm{TOL}}\right)^2$$

of these computations (for reasonable d). In contrast, in Ulam's method the corresponding number is

$$\kappa \cdot 1 = \frac{\left(2^{dM}\right)^2}{\mathrm{TOL}^2} = \left(\frac{\dim V_M}{\mathrm{TOL}}\right)^2.$$

Note that for the Sparse Ulam method the number of index computations is not staying proportional to the (squared) dimension of the approximation space. However, it is still scaling much more mildly with d than for Ulam's method.

4.4.4 The transition matrix is full

The matrix which represents the discretized transfer operator in Ulam's method is *sparse*: the supports of the basis functions are disjoint, and thus $P_{n,ij} \neq 0$ only if $S(X_j) \cap X_i \neq \emptyset$. Hence, for a sufficiently fine partition, the number of partition elements X_i which are intersected by the image $S(X_j)$ is determined by the local expansion of S. This is a fixed number related to a Lipschitz estimate on S and so the matrix of the discretized transfer operator with respect to the standard Ulam basis is sparse for sufficiently large n. Unfortunately this property is not shared by the matrix with respect to the sparse basis as the following considerations show.

The main reason for this is that the supports of the basis functions in the sparse basis are not localized, cf. the thin and long supports of the basis of $W_{\mathbf{k}}$ for $\mathbf{k} = (N, 0, \ldots, 0)$. This means that the occupancy of the transition matrix strongly depends on the *global behavior* of the dynamical system S. Let

$$B_{\mathbf{k}} := \left\{\varphi_{\mathbf{k},\mathbf{l}} \mid \mathbf{l}_i \in \{0, \ldots, 2^{\mathbf{k}_i} - 1\}\right\}$$

denote the basis of $W_{\mathbf{k}}$ and let

$$\mathrm{nnz}(\mathbf{k}, \mathbf{l}) = \left|\left\{(i,j) \mid S(\mathrm{supp}(\varphi_i)) \cap \mathrm{supp}(\varphi_j) \neq \emptyset, \varphi_i \in B_{\mathbf{k}}, \varphi_j \in B_{\mathbf{l}}\right\}\right|$$

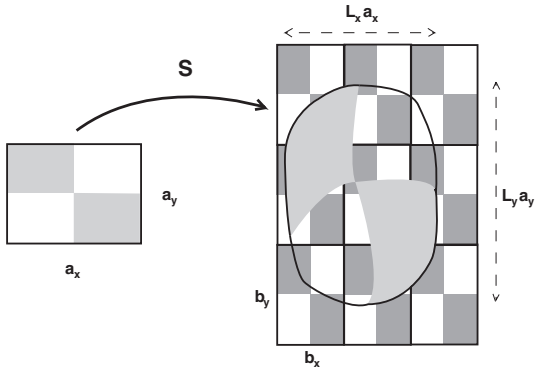


Figure 4.3: Modeling the matrix occupancy two dimensions - shaded and colorless (white) show the function values ($\pm|\varphi|$), thicker black lines the support boundaries

be the number of nonzero matrix entries which arise from the interaction of the basis functions from the subspaces $W_{\mathbf{k}}$ and $W_{\mathbf{l}}$ if $W_{\mathbf{k}}$ is mapped. We define the *matrix occupancy* of a basis $B_{\mathbf{I}} = \bigcup_{\mathbf{k}\in\mathbf{I}} B_{\mathbf{k}}$ as

$$\operatorname{nnz}(B_{\mathbf{I}}) = \sum_{\mathbf{k},\mathbf{l}\in\mathbf{I}} \operatorname{nnz}(\mathbf{k},\mathbf{l}). \tag{4.37}$$

In order to estimate $\operatorname{nnz}(\mathbf{k},\mathbf{l})$ we employ upper bounds L_i, $i = 1,\ldots,d$, for the Lipschitz-constants of S, cf. Figure 4.3. We obtain

Proposition 4.14.

$$\operatorname{nnz}(\mathbf{k},\mathbf{l}) \le \left|B_{\mathbf{k}}\right| \prod_{i=1}^{d} \left\lceil \frac{L_i \cdot 2^{-\mathbf{k}_i+1-(\mathbf{k}_i=0)}}{2^{-\mathbf{l}_i+1-(\mathbf{l}_i=0)}} \right\rceil. \tag{4.38}$$

Proof. Since we have used upper bounds for the Lipschitz constants, one mapped box has at most the extension $L_i \cdot 2^{-\mathbf{k}_i+1-(\mathbf{k}_i=0)}$ in the ith dimension. Consequently, its support intersects with at most

$$\left\lceil \frac{L_i \cdot 2^{-\mathbf{k}_i+1-(\mathbf{k}_i=0)}}{2^{-\mathbf{l}_i+1-(\mathbf{l}_i=0)}} \right\rceil$$

supports of basis functions from $W_{\mathbf{l}}$. □

Remark 4.15. Numerical experiments suggest that the above bound approximates the matrix occupancy quite well. However, it could be improved: (4.21) shows that a matrix entry still can be zero even if $\operatorname{supp}(\varphi_i)$ and $\operatorname{supp}(\mathcal{P}\varphi_j)$ intersect. This is e.g. the case if $\operatorname{supp}(\mathcal{P}\varphi_j)$ is included in a subset of $\operatorname{supp}(\varphi_i)$, where φ_i is constant (i.e. does

not change sign). The property $\|\mathcal{P}f\|_{L^1} = \|f\|_{L^1}$ for $f \geq 0$ and positivity of $\mathcal{P}$ imply $P_{N,ij} = 0$, since $\left\|\varphi_j^+\right\|_{L^1} = \left\|\varphi_j^-\right\|_{L^1}$.

An asymptotic estimate. Let us examine $\mathrm{nnz}(\mathbf{k}, \mathbf{l})$ for $\mathbf{k} = (0, \ldots, 0, N)$ and $\mathbf{l} = (N, 0, \ldots, 0)$. By taking all Lipschitz-constants $L_i = 1$ we get

$$\mathrm{nnz}(\mathbf{k}, \mathbf{l}) \gtrsim 2^{2N},$$

since $|B_{\mathbf{k}}| = 2^{N-1}$ and the image of each basis function from $B_{\mathbf{k}}$ intersects with each basis function from $B_{\mathbf{k}}$. Since $|B_N| \sim N^{d-1}2^N$, we get

$$2^{2N} \lesssim \mathrm{nnz}(B_N) \lesssim N^{2d-2}2^{2N}. \tag{4.39}$$

The exponential term dominates the polynomial one for large N, so asymptotically we will *not* get a sparse matrix.

Does this affect the calculations regarding efficiency made above? As already mentioned, the costs of Ulam's method are proportional to the dimension of the approximation space, $\mathcal{O}\left(\varepsilon^{-d}\right)$. Assuming that the Sparse Ulam method has the same error, its worst-case cost is $\left(\log_2 \varepsilon^{-1}\right)^{4d-4} \varepsilon^{-2}$. Clearly, this means — similarly to Section 4.4.1 — partially overcoming the curse of dimensionality. Even in the most optimistic case, i.e. the costs are $\mathcal{O}\left(2^{2N}\right)$, we have at least $\mathcal{O}\left(\varepsilon^{-2}\left(\log_2 \varepsilon^{-1}\right)^{2d-2}\right)$ costs, so the Sparse Ulam method is more efficient (concerning the number of flops for a matrix-vector multiplication) than Ulam's only if $d \geq 3$.

However, the fact that the transition matrix with respect to the sparse basis is not sparse posts another obstacle: the memory requirements for storing the matrix grow faster with the dimension d of phase space than one would desire. Figure 4.4 shows a comparison of the estimated number of nonzero entries (for the Sparse Ulam method, the number is obtained by taking the geometric mean of the two bounds in (4.39)) in dependence of d. Clearly, for $d > 5$ the storage requirements render computations on standard workstations impossible.

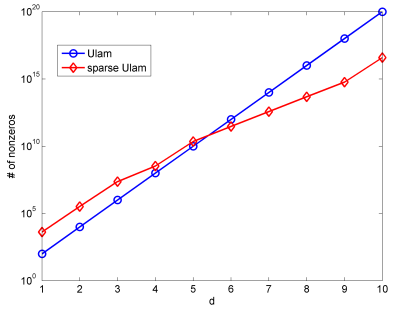


Figure 4.4: Estimated number of nonzero entries in the matrix representation of the discretized operator - in dependence of the dimension of phase space for $\varepsilon = 0.01$

4.5 Numerical examples

4.5.1 A 3d expanding map

We compare both methods by approximating the invariant density of a simple three dimensional map. Let $S_i : [0,1] \to [0,1]$ be given by

$$\begin{aligned} S_1(x) &= 1 - 2|x - 1/2|, \\ S_2(x) &= \begin{cases} 2x/(1-x), & x < 1/3 \\ (1-x)/(2x), & \text{else,} \end{cases}, \\ S_3(x) &= \begin{cases} 2x/(1-x^2), & x < \sqrt{2} - 1 \\ (1-x^2)/(2x), & \text{else,} \end{cases} \end{aligned}$$

and $S : [0,1]^3 \to [0,1]^3$ be the tensor product map $S(x) = \big(S_1(x_1), S_2(x_2), S_3(x_3)\big)^\top$, where $x = (x_1, x_2, x_3)^\top$. This map is expanding and its unique invariant density is given by (cf. [Din96])

$$f_1(x) = \frac{8}{\pi(1 + x_3^2)(1 + x_2)^2}.$$

We approximate f_1 by Ulam's method on an equipartition of 2^{3M} boxes for $M = 4, 5, 6$ as well as by the Sparse Ulam method on levels $N = 4, 5, 6$. Each discretization was computed several times for an increasing number of sample points, until no improvement was visible any more; i.e. the accuracy limit of the approximation space was reached. Figure 4.5 shows the L^1-error for both methods in dependence of the number of sample points (left) as well as the number of index computations (right). Identical discretizations, computed with different number of sample points, are connected. While the Sparse Ulam method requires almost three orders of magnitude fewer sample points

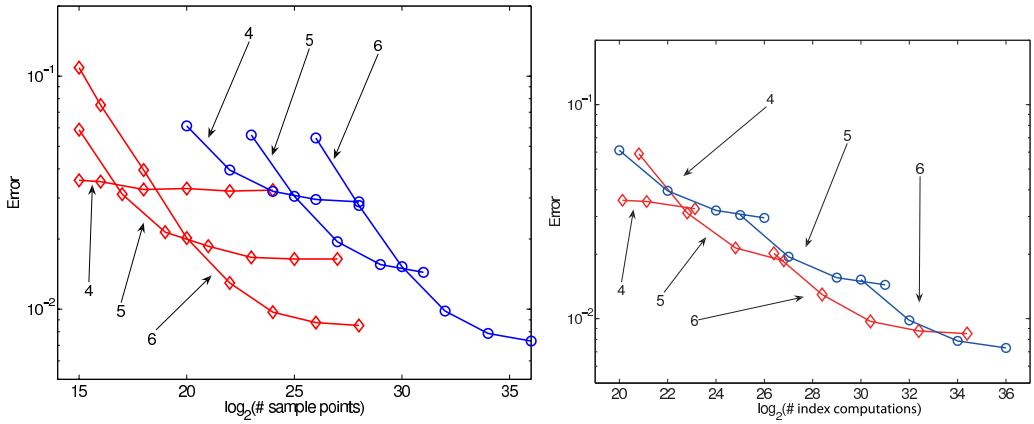


Figure 4.5: Left: L^1-error of the approximate invariant density in dependence on the number of sample points for levels $N, M = 4, 5, 6$. Right: Corresponding number of index computations. Ulam's method: blue circles; Sparse Ulam method: red diamonds.

than Ulam's method, the number of index computations is roughly comparable. This is in good agreement with our theoretical considerations in sections 4.4.2 and 4.4.3.

In Figure 4.6 we show the dependence of the L^1-error on the number of nonzeros in the transition matrices for levels $M, N = 3, \ldots, 6$. Again, the Sparse Ulam method is ahead of Ulam's method by almost an order of magnitude.

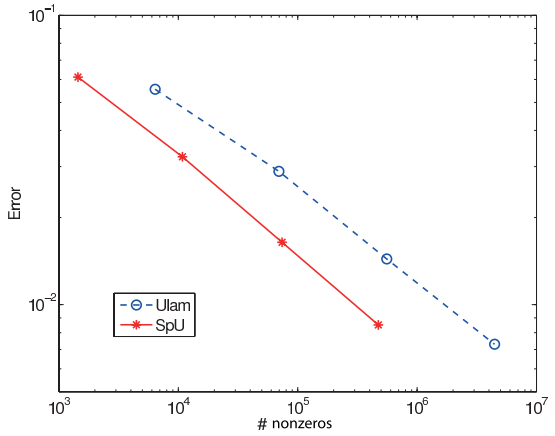


Figure 4.6: L^1-error of the approximate invariant densities in dependence on the number of nonzeros in the transition matrices.

4.5.2 A 4d conservative map

In a second numerical experiment, we approximate a few dominant eigenfunctions of the transfer operator for an area preserving map. Since the information on almost invariant sets does not change [Fro05] (but the eigenproblem becomes easier to solve) we here consider the symmetrized transition matrix $\frac{1}{2}(P + P^\top)$, cf. also [Jun04].

Consider the so called *standard map* $S_\rho : [0,1]^2 \to [0,1]^2$,

$$(x_1, x_2)^\top \mapsto \left(x_1 + x_2 + \rho \sin(2\pi x_1) + 0.5, x_2 + \rho \sin\left(2\pi x_1\right)\right)^\top \mod 1,$$

where $0 < \rho < 1$ is a parameter. This map is *area preserving*, i.e. the Lebesgue measure is invariant w.r.t. S_ρ. Figure 4.7 shows approximations of the eigenfunctions at the second largest eigenvalue of S_ρ for $\rho = 0.3$ (left) and $\rho = 0.6$ (right) computed via Ulam's method on an equipartition of $2^{2\cdot 6}$ boxes (i.e. for $M = 6$).

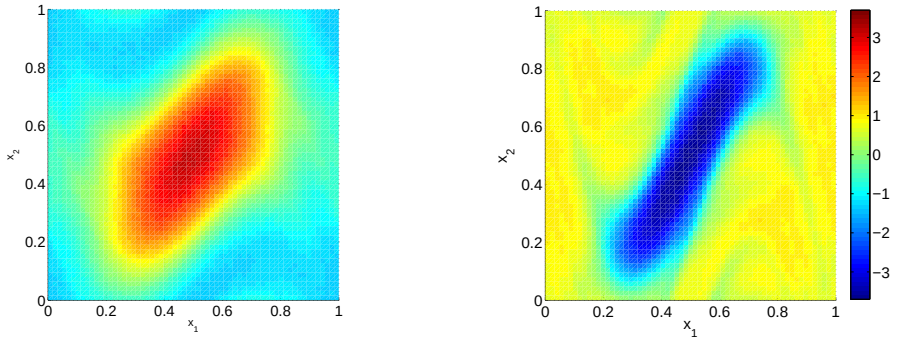


Figure 4.7: Eigenfunction of the symmetrized transition matrix at the second largest eigenvalue for the standard map. Left: $\rho = 0.3$, $\lambda_2 = 0.97$, right: $\rho = 0.6$, $\lambda_2 = 0.93$.

We now define $S : [0,1]^4 \to [0,1]^4$ by

$$S = S_{\rho_1} \otimes S_{\rho_2},$$

with $\rho_1 = 0.3$ and $\rho_2 = 0.6$. Note that the eigenfunctions of S are tensor products of the eigenfunctions of the S_{ρ_i}. This is reflected in figures 4.8 and 4.9 where we show the eigenfunctions at the two largest eigenvalues, computed by the Sparse Ulam method on level $N = 8$, using 2^{24} sample points overall. Clearly, each of these two is a tensor product of the (2d-) eigenfunction at the second largest eigenvalue with the (2d-) invariant (i.e. constant) density.

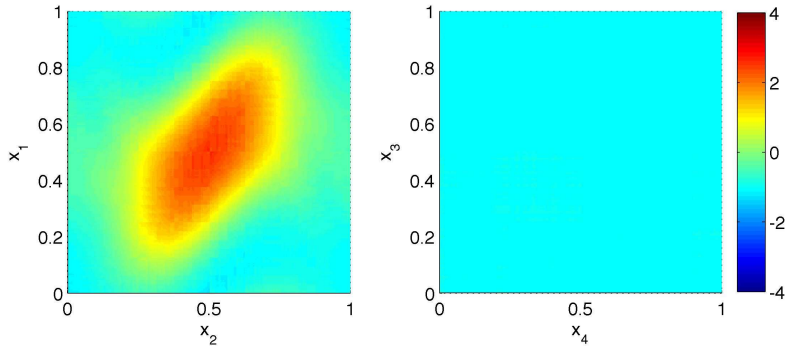


Figure 4.8: Approximate eigenfunction at $\lambda_2 = 0.97$. Left: $f_2(\cdot, \cdot, x_3, x_4)$ for fixed x_3, x_4, right: $f_2(x_1, x_2, \cdot, \cdot)$ for fixed x_1, x_2.

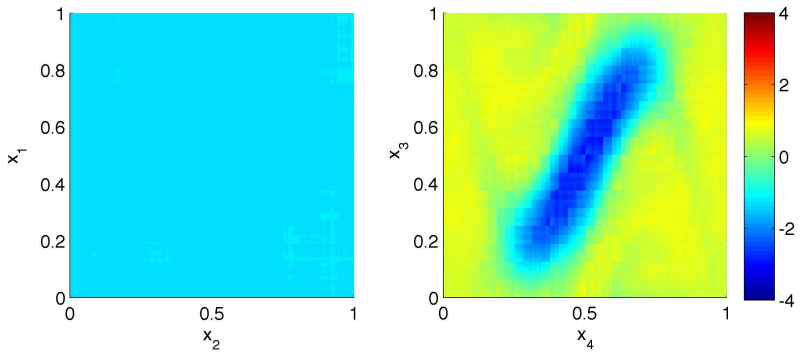


Figure 4.9: Approximate eigenfunction at $\lambda_2 = 0.93$.

Figure 4.10 shows an eigenfunction for which both factors of the tensor product are non-constant. The resolution of this eigenfunction seems worse than for those with one constant factor. In fact, for an approximation of an eigenfunction which is constant with respect to, say, x_3 and x_4 it suffices to consider subspaces W_ℓ with $\ell = (\ell_1, \ell_2, 0, 0)$. Since level N in 2d allows a better approximation than level N in 4d (see Remark 4.13), functions varying in merely two dimensions can be better approximated than one varying in all four.

As we know S to be conservative, its invariant density is the constant function. We compute the L^1-error of the approximative invariant density and compare it with a computation for the same system made with Ulam's method on a uniform $32 \times 32 \times 32 \times 32$ partition with 100 Monte Carlo sample points per box. The following table compares the accuracy and the cost factors of the two methods. Note that the Sparse

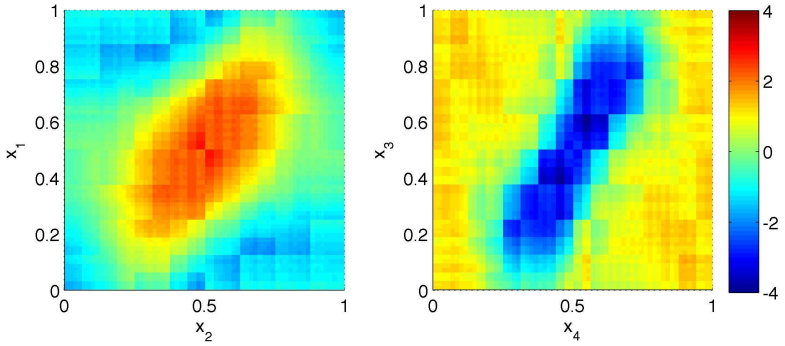


Figure 4.10: Approximate eigenfunction at $\lambda = 0.80$.

Ulam method yields a ten times smaller error, and requires ten times fewer sampling points.

	error	# basis functions	# nonzeros	# samples
Sparse Ulam	$7.8{\cdot}10^{-5}$	10496	$\approx 10^8$	$\approx 1.7{\cdot}10^7$
Ulam's method	$9.8{\cdot}10^{-4}$	1048576	$\approx 5{\cdot}10^7$	$\approx 10^8$

Table 4.2: The Sparse Ulam method and Ulam's method for the four dimensional standard map - comparison of the accuracy of the approximative invariant density and of some cost indicators (number of basis functions, number of nonzero entries in the transition matrix, and overall number of sample points).

Remark 4.16. The standard map for the parameter values given here has infinitely many periodic orbits, hence the associated transfer operator has the eigenvalue one with infinite multiplicity. From this, it is not clear which invariant densities are approximated by our numerical methods. Therefore, we applied in this example a componentwise additive random perturbation with uniform distribution on $[-0.05, 0.05]$. This random perturbation ensured that the eigenvalues of the transfer operator are isolated and of multiplicity one. Note, that the invariant density stays unchanged under this perturbation.

4.6 Conclusions and outlook

While there are $\mathcal{O}\left(\varepsilon^{-d}\right)$ basis functions needed in the standard basis consisting of characteristic functions to achieve the approximation error ε, using piecewise constant

sparse grid functions we need a number of $\mathcal{O}\left(\left(\log_2 \varepsilon^{-1}\right)^{2d-2} \varepsilon^{-1}\right)$ functions. The term $(\log_2 \varepsilon^{-1})^d$ is growing slow enough, such that the sparse grid approximation method allows us to overcome the curse of dimension — partly.

Consider expressions (4.8) and (4.9) (the ith derivative indicates how strongly is the function varying in the ith direction), and the fact that the sparse grid approximation spaces V_N include only basis functions which do not allow a good spatial resolution in many directions at a time (cf. Figure 4.2). They lead us to the conclusion, that those functions can be particularly well approximated in V_N which do not vary strongly in many dimensions. This is reflected by the eigenfunctions of the discretized transfer operator in the examples above. Figures 4.8, 4.9 and 4.10 emphasize this very well. If there is an eigenfunction varying strongly in all directions, the Sparse Ulam method will be unable to detect it, unless the level N gets very large. This, in turn, leads to computational inefficiency.

A thorough algorithmical analysis showed that not only the number of basis functions can be decreased significantly in comparison to Ulam's method, but other costs as well. The computationally most expensive step is the mapping of the sample points — from which the Sparse Ulam method requires far less than Ulam's; cf. Section 4.4.2.

Unfortunately, the geometry of the basis function supports has the side-effect that the transition matrix of the Sparse Ulam method is not sparse, but fully occupied. Hence, storage of the matrix, and manipulation with it have a complexity quadratic in the dimension of V_N. Clearly, this is the main bottleneck of this method. As long as basis functions with "widespread" supports are applied, this seems inevitable.

Another issue, not discussed in this work, is the one of handling more complex geometries. Our considerations here were restricted to the unit cube as phase space. It is straightforward to extend the method for rectangular phase spaces, but more complex geometries need some other treatment. How to "cover" the phase space with basis functions? How does the geometry of the phase space influence the approximation properties of V_N? As a first step towards answering these questions, we suggest to consult the existing literature on sparse grid methods for partial differential equations.

Further work could be done to detect the spectral properties of the discretized operator $\mathcal{P}_N$, to verify if Conjecture 4.9 holds. Also, more developed sampling techniques could make the numerical computation $\widetilde{\mathcal{P}}_N$ to inherit properties of the operator $\mathcal{P}_N$, if desired so.

To sum up, we expect the Sparse Ulam method to be a very efficient method for analyzing chaotic systems on a high dimensional phase space with regular geometry, where the eigenfunctions of the associated transfer operator (e.g. because of random perturbations) are sufficiently smooth and varying strongly only in several dimensions.

Chapter 5

Approximation of the infinitesimal generator

5.1 Motivation and outline

The general analysis of continuous-time systems with transfer operator methods involves the associated FPO $\mathcal{P}^t$, where $t > 0$ is some characteristic time of the system, such that significant motion can be observed. Assuming that the system is autonomous (i.e. its vector field does not depend on the time t), $\mathcal{P}^t$ is also the FPO associated with the time-t-map S^t of the system. Any numerical approximation of the transfer operator needs the computation of the time-t-map, hence the numerical integration of the underlying ODE, say $\dot{x} = v(x)$, with vector field v. This, in turn, requires several evaluations of the vector field v.

Now, if we consider Ulam's method on a partition of the phase space into n boxes, where the transition rates are computed by Monte Carlo quadrature, we need a total number $\mathcal{O}\left(n^2\right)$ of sample points, as shown in Section 2.3. All these have to be integrated for time t, which results in typically $k \sim 10-100$ vector field evaluations *for each sample*. For a large n, the size of k makes a big difference in the computational costs.

However, for autonomous systems the vector field v carries all the information needed to obtain $\mathcal{P}^t$ for *any* $t \geq 0$. The long-term dynamical behavior, which we wish to compute, is encoded in the eigenpairs of the transfer operator. One could ask the question, if there is a possibility to obtain these eigenmodes without time integration. The answer is given by Theorem 5.6 below, which states that one may get eigenpairs

of $\mathcal{P}^t$ for any $t > 0$ by computing eigenpairs of just *one* operator $\mathcal{A}$, the so-called *infinitesimal generator* of $\mathcal{P}^t$. From a computational point of view, we expect that $\mathcal{A}$ is numerically cheaper to compute than $\mathcal{P}^t$, because we can avoid the time integration. If the associated eigenvalue problem is also similarly cheap to solve than the one for $\mathcal{P}^t$, such a method has obvious advantages compared to Ulam's method, for example.

In this chapter we introduce two methods for the approximation of the infinitesimal generator of $\mathcal{P}^t$. Further, we discuss their advantages, numerical computation, convergence properties, limitations, and problems arising in their implementation. All these are shown on several examples. The mathematical tools relating the operators $\mathcal{A}$ and $\mathcal{P}^t$ belong to the field of semigroups of operators, hence we start with a brief introduction on this in Section 5.2. Sections 5.3, 5.4 and 5.5 are dealing with the first discretization method, which is the spatial discretization from the *upwind scheme*, a well-known numerical technique to approximate solutions of hyperbolic conservation laws; cf. [Kro97, LeV02] and references in them. However, the idea of applying this method for the approximation of the long-term dynamical behavior is new, and goes back to Froyland [Fro]. Also, to the best knowledge of the author, there exists no previous work which applies semigroup theory in order to analyze the convergence properties of such discretizations. Introducing the second method in Section 5.6, we exploit the exponential convergence speed of spectral methods to obtain a powerful discretization of the infinitesimal generator for smooth vector fields and tensor product phase space geometry. The methods are demonstrated on several numerical examples in Section 5.7.

Parts of the results in this chapter are intended to be published in [Fro]. Lemmas 5.9, 5.11 and 5.13, and their proofs are due to Gary Froyland. An earlier attempt to discretize the infinitesimal generator has been made in the honours thesis [Sta07].

5.2 Semigroups of operators

Definition 5.1. *Let $(Y, \|\cdot\|)$ be a Banach space. A one parameter family $\{\mathcal{T}^t\}_{t\geq 0}$ of bounded linear operators is called a semigroup on Y, if*

(a) $\mathcal{T}^t = I$ (I denoting the identity on Y),

(b) $\mathcal{T}^{t+s} = \mathcal{T}^t\mathcal{T}^s$ for all $t, s \geq 0$.

Further, if $\|\mathcal{T}^t\| \leq 1$*, the family is called a semigroup of contractions.*
If

$$\lim_{t\to 0} \|\mathcal{T}^t f - f\| = 0 \qquad \textit{for every } f \in Y,$$

$\mathcal{T}^t$ *is a continuous semigroup (*C_0 *semigroup).*

The transfer operator $\mathcal{P}^t$, the FPO associated with the ODE $\dot{x} = v(x)$ on the phase space X, is a C_0 semigroup of contractions on $L^1(X)$.[1] See [Las94] for a proof on this (especially Remark 7.6.2 to see the continuity). Now we introduce the central object we are going to work with.

Definition 5.2 (Infinitesimal generator)**.** *For a semigroup* $\mathcal{T}^t$ *we define the operator* $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to Y$ *as*

$$\mathcal{A}f = \lim_{t\to 0} \frac{\mathcal{T}^t f - f}{t}, \qquad f \in \mathcal{D}(\mathcal{A}),$$

with $\mathcal{D}(\mathcal{A}) \subset Y$ *being the linear subspace of* Y *where the above limit exists; called the domain of* $\mathcal{A}$*. The operator* $\mathcal{A}$ *is called the infinitesimal generator of the semigroup.*
Further, if $\mathcal{A}$ *is the infinitesimal generator of the semigroup* $\mathcal{T}^t$*, we write* $\mathcal{A} \in G(M, \omega)$ *if* $\left\|\mathcal{T}^t\right\| \leq Me^{\omega t}$*.*

We also have

Proposition 5.3 ([Paz83])**.** *Let* $\mathcal{T}^t$ *be a* C_0 *semigroup,* $\mathcal{A}$ *its infinitesimal generator and* $f \in \mathcal{D}(\mathcal{A})$*. Then* $u(t) = \mathcal{T}^t f$ *is the unique solution of*

$$\begin{aligned} \frac{\mathrm{d}u(t)}{\mathrm{d}t} &= \mathcal{A}u(t) \quad \textit{for } t > 0, \\ u(0) &= f. \end{aligned}$$

For $\mathcal{P}^t$, the infinitesimal generator turns out to be

$$\mathcal{A}_{PF} f = -\mathrm{div}(fv),$$

(provided the v_i are continuously differentiable, what we assume from now on), see [Las94]. Therefore, $C^1 \subset \mathcal{D}(\mathcal{A})$.

The intuition that $\mathcal{T}^t = e^{t\mathcal{A}}$ is strong, however *false* in general. If $\mathcal{A}$ is a bounded operator, this equation holds indeed. For unbounded ones, there are several results for the representation of the semigroup by exponential formulas. We shall use the following one later.

[1] We omit the indication of X from now on if we write function spaces, like L^p, C^1, etc.

Theorem 5.4 (Theorem 1.8.1 [Paz83]). *Let $\mathcal{T}^t$ be a C_0 semigroup on Y. Let*

$$A(h)f = \frac{\mathcal{T}^h f - f}{h},$$

then for every $f \in Y$ we have

$$\mathcal{T}^t f = \lim_{h \searrow 0} e^{tA(h)} f$$

and the limit is uniform in t for t in bounded intervals.

The approximation of the infinitesimal generator can be related to the approximation of the corresponding semigroups by

Theorem 5.5 (Theorem 3.4.5 [Paz83]). *Let $\mathcal{A}_n \in G(M, \omega)$ and assume*

(a) As $n \to \infty$, $\mathcal{A}_n f \to \mathcal{A} f$ for every $f \in D$, where D is a dense subset of Y.

(b) There exists a λ_0 with $Re\,\lambda_0 > \omega$ for which $(\lambda_0 I - \mathcal{A})D$ is dense in Y.

Then, the closure $\bar{\mathcal{A}}$ of $\mathcal{A}$ is in $G(M, \omega)$. If $\mathcal{T}_n(t)$ and $\mathcal{T}(t)$ are the C_0 semigroups generated by $\mathcal{A}_n$ and $\bar{\mathcal{A}}$ respectively, then

$$\lim_{n \to \infty} \mathcal{T}_n(t) f = \mathcal{T}(t) f \qquad \text{for all } t \geq 0, f \in Y,$$

and this limit is uniform in t for t in bounded intervals.

Theorem 2.2.4 from [Paz83] shows the connection between the eigenvalues of the semigroup operators and their infinitesimal generator:

Theorem 5.6 (Spectral mapping theorem). *Let $\mathcal{T}^t$ be a C_0 semigroup and let $\mathcal{A}$ be its infinitesimal generator. Then*

$$e^{t\sigma(\mathcal{A})} \subset \sigma\left(\mathcal{T}^t\right) \subset e^{t\sigma(\mathcal{A})} \cup \{0\},$$

where $\sigma(\cdot)$ denotes the point spectrum of the operator. The corresponding eigenvectors are identical.

This has important consequences for invariant densities:

Corollary 5.7. *The function f is an invariant density of $\mathcal{P}^t$ for all $t \geq 0$ if and only if $\mathcal{A}_{PF} f = 0$.*

Since $\mathcal{P}^t$ is a contraction, we have

Corollary 5.8. *The eigenvalues of $\mathcal{A}_{PF}$ lie in the closed left complex half plane.*

From now on we drop the subscripts and write $\mathcal{A}$ for the infinitesimal generator of $\mathcal{P}^t$ as well. It should always be clear from the context, which semigroup is meant.

5.3 The Ulam type approach for the nondiffusive case

5.3.1 The method

Let us consider $X = \mathbb{T}^d$, the d dimensional unit torus, and let a time-continuous dynamical system S^t be given by the ODE $\dot{x} = v(x)$. Assume v to be twice continuously differentiable.[1] The corresponding transfer operator is denoted by $\mathcal{P}^t$, its infinitesimal generator by $\mathcal{A}$. We partition X into d dimensional connected, positive volume subsets $\{X_1, \ldots, X_n\}$. Typically, each X_i will be a hyperrectangle to simplify computations. We always assume that the X_i are closed sets, i.e. $\bar{X}_i = X_i$.

We wish to give a numerical approximation of the infinitesimal generator, analogous to Ulam's discretization. First, we wish to deal with the deterministic case, hence we consider the system without diffusion, i.e. $\varepsilon = 0$.

Let $V_n = \operatorname{span}\{\chi_1, \ldots, \chi_n\}$, χ_i denoting χ_{X_i}, the characteristic function of X_i. When we give a matrix representation of an operator acting on V_n, we always refer to the basis $\{\chi_i\}_{i=1,\ldots,n}$, unless stated elsewise. For any fixed time t, one may form the Ulam approximation of $\mathcal{P}^t$, namely the operator $\mathcal{P}_n^t : V_n \to V_n$ with matrix representation $P_{n,ij} := m(X_j \cap S^{-t}X_i)/m(X_i)$.

We wish to construct an operator $\mathcal{A}_n : V_n \to V_n$ that is close in some sense to the operator $\mathcal{A}$. Motivated by Ulam's method, one would like to form $\pi_n \mathcal{A} \pi_n$, which unfortunately does not exist, because $V_n \nsubseteq \mathcal{D}(\mathcal{A})$. Recall, that $\mathcal{A}$ is the time derivative of $\mathcal{P}^t$. Instead of differentiating w.r.t. time and then doing the projection, we swap the order of these operations. Let us build the Ulam approximation $\mathcal{P}_n^t$ first, which will *not* be a semigroup any more, still, for fixed t it approximates $\mathcal{P}^t$. Taking the time derivative at $t = 0$, our candidate approximate operator is

$$\mathcal{A}_n u := \lim_{t \to 0} \left(\frac{\pi_n \mathcal{P}^t \pi_n u - \pi_n u}{t} \right).$$

We conclude from the following lemma that $\mathcal{D}(\mathcal{A}_n) = L^1$. The lemma also emphasizes the intuition behind the definition of the discretized generator: if $\mathcal{P}_n^t$ defines a finite state Markov chain on the X_i, then $\mathcal{A}_n$ is the generator of a Markov jump process,[2] which stays "near" $\mathcal{P}_n^t$ (the meaning of "near" will be elucidated in Proposition 5.22).

[1] I.e. $v_i \in C^2(X, \mathbb{R}^d)$ for $i = 1, \ldots, d$. Apart from Lemma 5.20, simple continuous differentiability suffices as well.

[2] A matrix $A \in \mathbb{R}^{n \times n}$ is said to generate a Markov jump process on the finite state space $\{1, \ldots, n\}$, if $P(t) = e^{tA}$ is a (column-)stochastic matrix for all $t \geq 0$ and $\operatorname{Prob}(x(t+s) = i |\, x(s) = j) = P(t)_{ij}$

Lemma 5.9. *The matrix representation of* $\mathcal{A}_n : V_n \to V_n$ *is*

$$A_{n,ij} = \begin{cases} \lim_{t\to 0} \frac{m(X_j \cap S^{-t} X_i)}{t \cdot m(X_i)}, & i \neq j; \\ \lim_{t\to 0} \frac{m(X_j \cap S^{-t} X_j) - m(X_j)}{t \cdot m(X_j)}, & \text{otherwise.} \end{cases} \tag{5.1}$$

Proof. We consider the action of $\mathcal{P}^t$ on χ_j.

$$\begin{aligned} \lim_{t\to 0} \pi_n \frac{\mathcal{P}^t \chi_j - \chi_j}{t} &= \lim_{t\to 0} \sum_{i=1}^{n} \frac{1}{m(X_i)} \left(\int_{X_i} \frac{\mathcal{P}^t \chi_j - \chi_j}{t} \, \mathrm{d}m \right) \chi_i \\ &= \lim_{t\to 0} \sum_{i\neq j} \frac{1}{m(X_i)} \left(\int_{X_i} \frac{\mathcal{P}^t \chi_j}{t} \, \mathrm{d}m \right) \chi_i + \lim_{t\to 0} \frac{1}{m(X_j)} \left(\int_{X_j} \frac{\mathcal{P}^t \chi_j - \chi_j}{t} \, \mathrm{d}m \right) \chi_j \\ &= \lim_{t\to 0} \sum_{i\neq j} \frac{1}{m(X_i)} \left(\int_{S^{-t} X_i} \frac{\chi_j}{t} \, \mathrm{d}m \right) \chi_i \\ &\quad + \lim_{t\to 0} \frac{1}{m(X_j)} \left(\int_{S^{-t} X_j} \frac{\chi_j}{t} \, \mathrm{d}m - \int_{X_j} \frac{\chi_j}{t} \, \mathrm{d}m \right) \chi_j \\ &= \sum_{i\neq j} \lim_{t\to 0} \frac{m(X_j \cap S^{-t}(X_i))}{t \cdot m(X_i)} \chi_i + \lim_{t\to 0} \frac{m(X_j \cap S^{-t} X_j) - m(X_j)}{t \cdot m(X_j)} \chi_j \end{aligned}$$

Thus under right multiplication we obtain (5.1). The question, if the limits exist, is answered below by Lemma 5.11. □

Remark 5.10. Lemma 5.9 states, that $A_{n,ij}$, $i \neq j$, is the *outflow rate of uniformly distributed mass* from X_j into X_i.

The following lemma shows the main advantage of this discretization. It allows us the construction of A_n *without the computation of the flow* S^t, which is the numerically most expensive step in other methods used so far.

Lemma 5.11. *For* $x \in \partial X_j$, *define* $\mathbf{n}_j(x)$ *to be the unit normal vector pointing out of* X_j. *The sets* X_j *should be chosen such that* $\mathbf{n}_j$ *exists almost everywhere on* ∂X_j *(measured by the* $d-1$ *dimensional Lebesgue measure on* ∂X_j*). The matrix representation*

for all $i, j \in \{1, \ldots, n\}$ and $s, t \geq 0$. One can think of the jump process as a stochastic process jumping at random times from one state to another. We will not need the mathematical background of these processes, hence we refer to [Nor97] for more details. However, analogously as the Ulam discretization is connected with discrete-time Markov chains, the viewpoint of jump processes enables us to give a physical meaning to the discretization of the generator. The justification that $\mathcal{A}_n$ generates a Markov jump process on the set of boxes is given later in Remark 5.14.

of $\mathcal{A}_n : V_n \to V_n$ *is*

$$A_{n,ij} = \begin{cases} (1/m(X_i)) \int_{X_j \cap X_i} \max\{v(x)\cdot \mathbf{n}_j(x), 0\} \, \mathrm{d}m_{d-1}(x), & i \neq j; \\ -\sum_{k \neq i} \frac{m(X_k)}{m(X_i)} A_{n,ki}, & i = j. \end{cases} \tag{5.2}$$

Proof. From (5.1) we have for $i \neq j$ that $A_{n,ij} = \lim_{t\to 0} \frac{m(X_j \cap S^{-t} X_i)}{t m(X_i)}$. Denoting $M_{ij}(t) = m(X_j \cap S^{-t} X_i)$ we have that $A_{n,ij} = M'_{ij}(0)/m(X_i)$ where the prime denotes differentiation with respect to t. The quantity $M'_{ij}(0)$ is simply the rate of flux out of X_j through the face $X_j \cap X_i$ into X_i and so $M'_{ij}(0) = \int_{X_j \cap X_i} \max\{v(x)\cdot \mathbf{n}_j(x), 0\} \, \mathrm{d}m_{d-1}(x)$.

For the diagonal elements $A_{n,jj}$ we have $A_{n,jj} = \lim_{t\to 0} \frac{m(X_j \cap S^{-t} X_j) - m(X_j)}{t m(X_j)}$. Note that $m(X_j) - m(X_j \cap S^{-t} X_j) = m(X_j \setminus S^{-t} X_j)$. Clearly $X_j \setminus S^{-t} X_j = X_j \cap \bigcup_{k\neq j} S^{-t} X_k = \bigcup_{k \neq j} X_j \cap S^{-t} X_k$ modulo sets of Lebesgue measure zero. Thus, $m(X_j) - m(X_j \cap S^{-t} X_j) = m(X_j \setminus S^{-t} X_j) = \sum_{k\neq j} m(X_j \cap S^{-t} X_k)$. It follows that $A_{n,jj} = -\sum_{k\neq j} \frac{m(X_k)}{m(X_j)} A_{n,kj}$. □

In one dimension, (5.2) has a particularly simple form.

Corollary 5.12. *Let* $X = \mathbb{T}^1$*, and consider the flow generated by* $\dot{x} = v(x)$*. Assume without loss that* $v \geq 0$ *on* X*.*[1] *Denote by* $\{x_0, x_1, \ldots, x_n\}$ *the endpoints of the subintervals* $\{X_1, \ldots, X_n\}$ *in the partition of* X*. Then the matrix representation of* $\mathcal{A}_n : V_n \to V_n$ *is*

$$A_{n,ij} = \begin{cases} -v(x_j)/m(X_j), & i = j; \\ v(x_j)/m(X_i), & i = j+1; \\ 0, & \textit{otherwise.} \end{cases} \tag{5.3}$$

We remark that (5.3) is the matrix arising in finite difference methods using backward differences (clearly, it would be forward differences if $v \leq 0$). Finally, we show that our constructions (5.2) and (5.3) always provide a solution to the system $\mathcal{A}_n u = 0$ for some $u \in V_n$.

Lemma 5.13. *There exists a nonnegative, nonzero* $u \in V_n$ *so that* $\mathcal{A}_n u = 0$*.*

Proof. Let $M_{n,ij} = m(X_i)\delta_{ij}$ and note that $Q_n := M_n A_n M_n^{-1}$ satisfies

$$Q_{n,ij} = \begin{cases} (1/m(X_j)) \int_{X_i \cap X_j} \max\{v(x)\cdot \mathbf{n}_j(x), 0\} \, \mathrm{d}m_{d-1}(x), & i \neq j; \\ -\sum_{i \neq j} Q_{n,ij}, & \text{otherwise.} \end{cases} \tag{5.4}$$

[1] If $v \not\geq 0$ and $v \not\leq 0$, we have one or more stable fixed points, and every trajectory converges to one of them. Hence, there is no interesting statistical behavior to analyze.

Let $c = \max_{1\le i\le n} \sum_{i\neq j} Q_{n,ij}$. The matrix $\hat{Q}_n := Q_n + cI$ is nonnegative with all column sums equal to c. By the Perron–Frobenius theorem, the largest eigenvalue of $\hat{Q}_n$ is c (of multiplicity[1] possibly greater than 1) and there is a corresponding left/right eigenvector pair u_n, v_n that may be chosen to be nonnegative. Clearly u_n, v_n are left/right eigenvectors of Q_n corresponding to the eigenvalue 0 and $M_n u_n, M_n v_n$ are nonnegative left/right eigenvectors corresponding to 0 for A_n. □

Remark 5.14. Note, that the existence of an eigenvector (not necessarily non-negative) to eigenvalue zero follows already from $(1,\dots,1)A_n = 0$, see (5.2). Furthermore, it can be shown easily by the same formula that A_n generates a Markov jump process [Nor97] on the set of boxes $\{X_1,\dots,X_n\}$, i.e. e^{tA_n} is (column-)stochastic for all $t \ge 0$.

Algorithm 5.15 (Ulam type discretization of the generator)**.**

1. Partition X into positive volume connected sets $\{X_1,\dots,X_n\}$. Typically each X_i will be a hyperrectangle.

2. Compute

$$A_{n,ij} = \begin{cases} (1/m(X_i)) \int_{X_j\cap X_i} \max\{v(x)\cdot \mathbf{n}_j(x), 0\}\, \mathrm{d}m_{d-1}(x), & i \neq j, \\ -\sum_{k\neq i} \frac{m(X_k)}{m(X_i)} A_{n,ki}, & i = j, \end{cases}$$

   where some numerical quadrature method is used to estimate the integral.

3. Estimates of invariant densities for S^t lie in right null space of A_n. Let $A_n w = 0$; the existence of such a w is guaranteed by Lemma 5.13. Then $u := \sum_{i=1}^n w_i \chi_i$ satisfies $\mathcal{A}_n u = 0$.

4. Left and right eigenvectors of A_n corresponding to small (in magnitude) real eigenvalues $\lambda < 0$ provide information about almost invariant sets.

Remark 5.16. Note, that the discretized generator A_n is a sparse matrix, since $A_{n,ij} = 0$ if X_i and X_j are not adjacent.

5.3.2 Convergence

The main results in this section are Theorem 5.21, which states the pointwise convergence in L^1 of the semigroup generated by $\mathcal{A}_n$ to $\mathcal{P}^t$; and Proposition 5.22 which shows

[1] in this 1D situation, $\hat{Q}_n$ is primitive (irreducible and there exists k such that $\hat{Q}_n^k > 0$) and the eigenvalue c has algebraic and geometric multiplicity 1.

the asymptotic closeness of the semigroup generated by $\mathcal{A}_n$ and the Ulam discretization $\pi_n\mathcal{P}^t$ in t. We will use Theorem 5.5 to show the first result. For this, some preparation is needed. The next lemma states that our approximation to the infinitesimal generator is a meaningful one.

Lemma 5.17. *Let $X = \mathbb{T}^d$, and let all boxes of the underlying discretization be congruent with edge length $1/n$. Then for all $u \in C^1$ we have $\mathcal{A}_n u \to \mathcal{A}u$ in the L^1-norm as $n \to \infty$.*

Proof. Fix $u \in C^1$. Note $u \in \mathcal{D}(\mathcal{A})$. Since the defining limits of $\mathcal{A}u$ and $\mathcal{A}_n u$ exist, we may write

$$\mathcal{A}_n u - \mathcal{A}u = \lim_{t\to 0} \frac{\pi_n\mathcal{P}^t u - \pi_n u}{t} - \frac{\mathcal{P}^t u - u}{t} + \frac{\pi_n\mathcal{P}^t(\pi_n - I)u}{t}.$$

The second summand tends to $\mathcal{A}u$, the first to $\pi_n\mathcal{A}u$ as $t \to 0$. Latter follows by the continuity of π_n. We also have $\pi_n\mathcal{A}u \to \mathcal{A}u$ as $n \to \infty$, hence it remains to show

$$\lim_{n\to\infty}\lim_{t\to 0}\frac{1}{t}\pi_n\mathcal{P}^t(\pi_n - I)u = 0.$$

Let x_i denote the center of the box X_i. Fix the index i. Let $u = \tilde{u} + \delta u$, where $\tilde{u}(x) = u(x_i) + Du(x_i)(x - x_i)$, the local linearization of u. Since $u \in C^1$, it holds $\delta u(x) = o(n^{-1})$ for $|x - x_i| = \mathcal{O}(n^{-1})$, as $n \to \infty$.[1] Now define $\tilde{v}(x) \equiv v(x_i)$ and let $\widetilde{\mathcal{P}}^t$ be the associated FPO. Let $\pi_{n,i}$ denote the L^2-orthogonal projection onto the constant functions over the box X_i, i.e.

$$\pi_{i,n}h = \left(\frac{1}{m(X_i)}\int_{X_i} h\right)\chi_i = \left(n^d\int_{X_i} h\right)\chi_i.$$

Then $\pi_n = \sum_j \pi_{n,j}$. We have

$$\frac{1}{t}\pi_{n,i}\mathcal{P}^t(\pi_n - I)u = \underbrace{\frac{1}{t}\pi_{n,i}\mathcal{P}^t(\pi_n - I)\tilde{u}}_{(I)} + \underbrace{\frac{1}{t}\pi_{n,i}\mathcal{P}^t(\pi_n - I)\delta u}_{(II)}. \tag{5.5}$$

We investigate the summands separately:
To (I). By the linearity of $\tilde{u}$ and the congruency of the boxes, one has $(\pi_n - I)\tilde{u}\,|_{X_j}(x) = -Du(x_i)(x - x_j)$. Thus, $\int_{X_j}(\pi_n - I)\tilde{u} = 0$ for every j and the function $(\pi_n - I)\tilde{u}$ is periodic in each coordinate with period $\frac{1}{n}$. By this, each translation of the function

[1] We say $f(x) = o(g(x))$ as $x \to 0$, if $f(x)/g(x) \to 0$ as $x \to 0$.

$(\pi_n - I)\tilde{u}$ has integral zero over each box. Since the transfer operator $\widetilde{\mathcal{P}}^t$ corresponding to the constant flow $\tilde{v}$ is merely a translation, we have

$$\pi_{n,i}\widetilde{\mathcal{P}}^t(\pi_n - I)\tilde{u} = 0. \tag{5.6}$$

Let $\widetilde{S}^{-t}$ be the flow associated with the vector field $-\tilde{v}$. Then $S^{-t}(x) - \widetilde{S}^{-t}(x) = \mathcal{O}(tn^{-1})$ as $t \to 0$ and $n \to \infty$, uniformly in x with $\left|x - x_i\right| = \mathcal{O}\left(n^{-1}\right)$. This implies for the symmetrical difference of the sets:

$$S^{-t}X_i \,\Delta\, \widetilde{S}^{-t}X_i \subset B_\varepsilon\left(\partial\widetilde{S}^{-t}X_i\right),$$

where $\varepsilon = \mathcal{O}\left(tn^{-1}\right)$ and $B_\varepsilon(\cdot)$ denotes the ε neighborhood of a set. From this we have

$$m\left(S^{-t}X_i \,\Delta\, \widetilde{S}^{-t}X_i\right) \le m\left(B_\varepsilon\left(\partial\widetilde{S}^{-t}X_i\right)\right) \le \mathcal{O}\left(tn^{-1}\right) m_{d-1}\left(\partial\widetilde{S}^{-t}X_i\right) = \mathcal{O}\left(tn^{-d}\right),$$

since the perimeter of X_i is $\mathcal{O}\left(n^{1-d}\right)$ and the translation $\widetilde{S}^{-t}$ does not change this. Recall $\int_{X_j}\mathcal{P}^t u = \int_{S^{-t}X_j} u$. Thus, for an arbitrary $h \in C^1$ we have

$$\left|\int_{X_i}\mathcal{P}^t h - \int_{X_i}\widetilde{\mathcal{P}}^t h\right| \le \int_{S^{-t}X_i \,\Delta\, \widetilde{S}^{-t}X_i} |h| = \|h\|_\infty\, \mathcal{O}(tn^{-d}).$$

Set $h = (\pi_n - I)\tilde{u}$. Since $\int_{X_i}\widetilde{\mathcal{P}}^t(\pi_n - I)\tilde{u} = 0$, $\left\|(\pi_n - I)\tilde{u}\right\|_\infty = \mathcal{O}(n^{-1})$, and since $\frac{1}{t}\pi_{i,n}\mathcal{P}^t(\pi_n - I)\tilde{u} = \frac{1}{t}n^d\int_{X_i}\mathcal{P}^t(\pi_n - I)\tilde{u}$, the first summand in (5.5) is $\mathcal{O}(n^{-1})$ as $n \to \infty$. To (II). Considering the second summand, note, that $\int_{X_i}(\pi_n - I)h = 0$ for all $h \in L^1$. We have

$$\begin{aligned}
\frac{1}{t}n^d\int_{X_i}\mathcal{P}^t(\pi_n - I)\delta u &= \frac{1}{t}n^d\left(\int_{X_i}\mathcal{P}^t(\pi_n - I)\delta u - \int_{X_i}(\pi_n - I)\delta u\right) \\
&\xrightarrow{t\to 0} n^d\,\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{X_i}\mathcal{P}^t(\pi_n - I)\delta u\right)\Big|_{t=0} \\
&= -n^d\int_{\partial X_i} g_i\ \mathbf{n}_i\cdot v \\
&= o(1) \qquad \text{as } n \to \infty,
\end{aligned}$$

where

$$g_i(x) := \begin{cases} \lim\limits_{\substack{y\to x\\ y\in X_i}}(\pi_n - I)\delta u(y), & \text{if } \mathbf{n}_i(x)\cdot v(x) \ge 0, \\ \lim\limits_{\substack{y\to x\\ y\in X_j}}(\pi_n - I)\delta u(y), & \text{otherwise, with } x \in \partial X_j. \end{cases}$$

The second equation follows from the fact that the derivative is simply the rate of flux across ∂X_i. The function $(\pi_n - I)\delta u$ is merely piecewise differentiable (is $C^1(X_j)$

for each j). That makes the definition of g_i necessary: we have to look at what does the flow "drag" into X_i and what is "dragged" outside. The last equation follows by $(\pi_n - I)\delta u(x) = o(n^{-1})$ as $n \to \infty$, uniformly in x for $|x - x_i| = \mathcal{O}\left(n^{-1}\right)$.

Thus, we showed

$$\lim_{n\to\infty} \lim_{t\to 0} n^d \int_{X_i} \frac{\mathcal{P}^t(\pi_n - I)u}{t} = 0.$$

All approximations were uniform in i, since the first derivatives of u are uniformly continuous by the compactness of X. Thus $\lim_{t\to 0} \frac{1}{t}\pi_n \mathcal{P}^t(\pi_n - I)u \to 0$ as $n \to \infty$. □

Remark 5.18. The assumption that the boxes are congruent is crucial. The discretized operator $\mathcal{P}_n^t = \pi_n \mathcal{P}^t$ from Ulam's method converges pointwise if the diameter of the largest box tends to zero. This is not sufficient here. We give a counterexample:

Take $X = \mathbb{T}^1$ the unit circle and $v \equiv 1$ the constant flow. Let V_n (n even) be associated with the box covering of $\mathbb{T}^1$, where the box numbering is from left to right and each box with odd number is an interval of length $\frac{4}{3n}$, each one with even number is an interval of length $\frac{2}{3n}$. Then

$$A_n = \frac{3n}{4} \begin{pmatrix} -1 & & & & 1 \\ 2 & -2 & & & \\ & 1 & -1 & & \\ & & 2 & \ddots & \\ & & & \ddots & -2 \end{pmatrix}.$$

Let $f(x) = \sin(2\pi x)$. Then $\mathcal{A}f(x) = -2\pi\cos(2\pi x)$. As Figure 5.1 shows, $\mathcal{A}_n f$ (red) does not converge to $\mathcal{A}f$ (blue). As an interesting observation, we note that the corresponding semigroup does seem to converge in this example; i.e. $\exp(t\mathcal{A}_n)f \to \mathcal{P}^t f$ as $n \to \infty$ for a fixed $t > 0$.

Nevertheless, we may weaken the assumption about the congruency of the boxes. However, they still have to tend to a uniform shape in the following sense: are $b_{i,1}, \ldots, b_{i,d}$ the edge lengths of the ith box, it should hold

$$\frac{\max_{i=1,\ldots,n} b_{i,j}}{\min_{i=1,\ldots,n} b_{i,j}} \to 1, \qquad \text{for } j = 1, \ldots, d \tag{5.7}$$

as $n \to \infty$. Then

Corollary 5.19. *By weakening the congruency assumption on the boxes to* (5.7) *the claim of Lemma 5.20 still holds.*

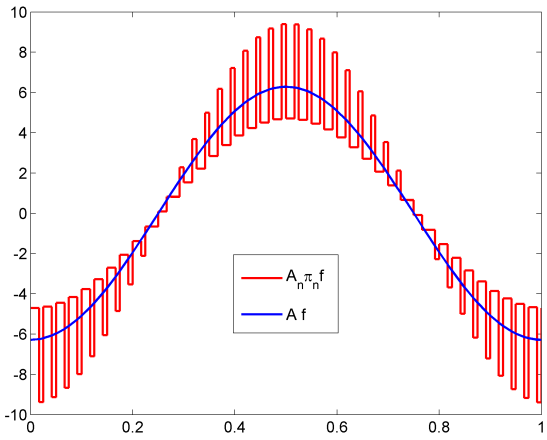


Figure 5.1: Improper convergence of the approximative infinitesimal generator on a non-uniform grid - This computation with grid size $n = 80$ highlights the problem: $\mathcal{A}_n f$ (red) converges on the subintervals of different size against different multiples of $\mathcal{A}f$ (blue).

Sketch of proof. The proof of Lemma 5.17 still applies by changing (5.6) to

$$\pi_{n,i}\widetilde{\mathcal{P}}^t(\pi_n - I)\tilde{u} = o(tn^{-1}),$$

which can be shown by considering that $\widetilde{P}$ is just a translation, and the edges lengths of the boxes differ by $o(n^{-1})$, cf. (5.7). □

Lemma 5.20. *For a $\lambda > 0$ sufficiently large, one has $(\lambda - \mathcal{A})^{-1} u \in C^1$ for all $u \in C^1$.*

Proof. We have from Remark 1.5.4 in [Paz83] that

$$(\lambda - \mathcal{A})^{-1}u(x) = \int_0^\infty e^{-\lambda t}\mathcal{P}^t u(x)\mathrm{d}t. \tag{5.8}$$

By Lebesgue's dominated convergence theorem, it is sufficient for the differentiability of the right hand side w.r.t. x that

$$e^{-\lambda t}\left|D\mathcal{P}^t u(x)\right| \leq h(t) \qquad \text{uniformly in } x,$$

for an integrable h. Here and in the following D denotes the derivative w.r.t. x. Recall the explicit representation of the FPO,

$$\mathcal{P}^t u(x) = u\left(S^{-t}x\right)\ \left|\det DS^{-t}(x)\right|.$$

For autonomous flows the above determinant is nonzero for all t and x. So, it will not change sign, since it is continuous as a function of t. By this, we drop the absolute value, since $DS^0 = I$ with positive determinant. We compute

$$D\mathcal{P}^t u(x) = Du\left(S^{-t}x\right) DS^{-t}(x) \det\left(DS^{-t}(x)\right) + u\left(S^{-t}x\right) \det'\left(DS^{-t}(x)\right) D^2S^{-t}(x).$$

Note, that the determinant is just a polynomial in the entries of the matrix. Thus, to bound $\left|D\mathcal{P}^t u\right|$, we need bounds on the derivatives DS^{-t} and D^2S^{-t} of the flow. For this, derive the variational equation for the flow through x of the differential equation $\dot{x} = v(x)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} DS^{-t}x = -Dv(S^{-t}x)DS^{-t}x,$$

or with $W_1(t) := DS^{-t}x : \dot{W}_1(t) = -Dv(S^{-t}x)W_1(t)$. For $W_2(t) = D^2S^{-t}x$, we obtain

$$\dot{W}_2(t) = -D^2v(S^{-t}x)W_1(t)^2 - Dv(S^{-t}x)W_2(t).$$

We do not care about the exact tensor structures of the particular derivatives, just note that they are multilinear functions. Gronwall's inequality gives

$$\left\|W_1(t)\right\|_\infty \leq e^{\lambda_1 t},$$

where $\lambda_1 = \left\|Dv(S^{-t}\cdot)\right\|_\infty$. By this, applying Gronwall's inequality on the ODE for $W_2(t)$, we obtain

$$\left\|W_2(t)\right\|_\infty \leq e^{\lambda_2 t},$$

with a suitable $\lambda_2 > 0$. The determinant is a polynomial in the entries of the matrix, consequently $\left|\det(DS^{-t}(x))\right| \leq ce^{d\lambda_1 t}$ for a suitable $c > 0$ and for all $x \in X$. Similar holds for $\left|\det'\left(DS^{-t}(x)\right)\right|$. $Du\left(S^{-t}x\right)$ and $u\left(S^{-t}x\right)$ are uniformly bounded, since $u \in C^1$ and X is compact. Thus, we can conclude that there are constants $C, \Lambda > 0$, Λ independent on u, such that

$$\left|D\mathcal{P}^t u(x)\right| \leq Ce^{\Lambda t} \qquad \text{uniformly in } x.$$

Setting $\lambda > \Lambda$, $|h(t)| \leq Ce^{(\Lambda-\lambda)t}$ is integrable over $[0, \infty]$, hence the right hand side of (5.8) is differentiable w.r.t. x, so is $(\lambda - \mathcal{A})^{-1}u$. □

Theorem 5.21. *The operator $\mathcal{A}_n$ generates a C_0 semigroup $\mathcal{R}_n^t := \exp\left(t\mathcal{A}_n\right) = I - \pi_n + \exp(t\mathcal{A}_n\,|_{V_n})\pi_n$. For all $u \in L^1$ and $t \geq 0$ we have $\mathcal{R}_n^t u \to \mathcal{P}^t u$ in L^1, uniformly in t on bounded intervals.*

Proof. We use Theorem 5.5 with $D = C^1$. By the Hille–Yosida theorem (Theorem 1.3.1 [Paz83]), $\mathcal{A}$ is a closed operator. Since we showed in Lemma 5.17, that $\mathcal{A}_n u \to \mathcal{A}u$ as $n \to \infty$ for all $u \in C^1$, it remains to show:

(a) $\mathcal{A}_n \in G(1,0)$, i.e. $\mathcal{A}_n$ generates a semigroup, which is uniformly bounded by 1 in the operator norm.

(b) There is a λ with Re $\lambda > 0$ such that $(\lambda - \mathcal{A})C^1$ is dense in L^1.

To (a). The range of $\mathcal{A}_n$ is in V_n and $\mathcal{A}_n = \mathcal{A}_n\pi_n$. Both π_n and $\mathcal{A}_n|_{V_n}$ are bounded operators with $\|\pi_n\|_{L^1} = 1$ and $\left\|e^{t\mathcal{A}_n|_{V_n}}\right\|_{L^1} \leq 1$ (see Remark 5.14), hence $\mathcal{R}_n^t = e^{t\mathcal{A}_n}$ exists and $\|\mathcal{R}_n^t\|_{L^1} \leq 1$. This implies $\mathcal{A}_n \in G(1,0)$.
Moreover, by

$$\mathcal{R}_n^t = e^{t\mathcal{A}_n} = (I - \pi_n) + (I + \mathcal{A}_n + \frac{1}{2}\mathcal{A}_n^2 + \ldots)\pi_n$$

we have $\mathcal{R}_n^t = I - \pi_n + \exp\left(t\mathcal{A}_n|_{V_n}\right)\pi_n$.
To (b). By Lemma 5.20 one has $C^1 \subset (\lambda - \mathcal{A})C^1$. Since C^1 is dense in L^1, this completes the proof. □

After the convergence results for $n \to \infty$ we present a result which gives closeness of $\mathcal{R}_n^t$ and $\mathcal{P}^t$ for small times.

Proposition 5.22. *As $t \to 0$ it holds*

$$\mathcal{R}_n^t u - \pi_n \mathcal{P}^t u = \mathcal{O}(t^2) \tag{5.9}$$

for all $u \in V_n$.

Proof. First we give an expansion of $\pi_n\mathcal{P}^t u$ in t. For this, define

$$A(h)g := \frac{\mathcal{P}^h g - g}{h}.$$

By Theorem 5.4 we have

$$\mathcal{P}^t u = \lim_{h\to 0} e^{tA(h)}u \tag{5.10}$$

uniformly on bounded t-intervals, hence by $\pi_n u = u$ and the continuity of π_n,

$$\pi_n\mathcal{P}^t u = \lim_{h\to 0} \pi_n e^{tA(h)}u = u + t \lim_{h\to 0} \pi_n A(h)u + \lim_{h\to 0} r(t,h).$$

The first limit on the right hand side exists, and is equal to $\mathcal{A}_n u$. Therefore, the second limit must exist as well, and because of the uniform convergence in (5.10) and uniform boundedness of the term $t\pi_n A(h)u$ in t and h, $r(t,h)$ is uniformly bounded as well; $\|r(t,h)\| \leq C$. Moreover, since $r(t,h)$ is the remainder in the expansion of the exponential function, it holds $\|r(t,h)\| \leq C(h)t^2$ as $t \to 0$. Together with the previous bound we have $C(h) \leq \tilde{C} < \infty$. This implies

$$\lim_{h\to 0} r(t,h) = \mathcal{O}(t^2),$$

which gives

$$\pi_n \mathcal{P}^t u = u + t\mathcal{A}_n u + \mathcal{O}(t^2).$$

Since

$$\mathcal{R}_n^t \,|_{V_n} = e^{t\mathcal{A}_n}|_{V_n} = I_{V_n} + t\mathcal{A}_n|_{V_n} + \mathcal{O}(t^2),$$

the proof is completed. □

Remark 5.23 (Connections with the upwind scheme). Clearly, $\mathcal{A}_n$ is the spatial discretization from the so-called upwind scheme in finite volume methods; cf. [LeV02]. The scheme is known to be stable. Stability of finite volume schemes is often related to "numerical diffusion" in them; cf. Section 5.7.1. Our derivation allows the understanding of stability in a similar way. We shoved in Proposition 5.22 that $\mathcal{P}_n^t$ is the transition matrix of a Markov process near the Markov jump process generated by $\mathcal{A}_n$ for small $t > 0$. The discretized FPO $\mathcal{P}_n^t$ can be related to a non-deterministic dynamical system, which, after mapping the initial point, adds some uncertainty to produce a uniform distribution of the image point in the box where it landed; see Chapter 3 and [Fro96]. This uncetrainty resulting from the numerical discretization, equivalently to the numerical diffusion in the upwind scheme, can be viewed as the reason for robust behavior — stability.

5.4 The Ulam type approach for the diffusive case

5.4.1 The method

We still assume that $X = \mathbb{T}^d$ is partitioned by congruent cubes with edge length $1/n$. We introduce a small uncertainty to the dynamics. Latter will be governed by the SDE

$$\dot{x} = v(x) + \varepsilon \dot{W},$$

where W denotes the Brownian motion; cf. Section 2.1.2. The associated transfer operator $\mathcal{Q}^t$ (we use another symbol instead of $\mathcal{P}^t$ to emphasize that the underlying dynamics is non-deterministic; and the dependence of the semigroup on the diffusion parameter ε is dropped in the notation) is the evolution operator of the Fokker–Planck equation

$$\partial_t u = \frac{\varepsilon^2}{2}\Delta u - \operatorname{div}(uv) =: \mathcal{A}^{(\varepsilon)} u.$$

This equation has a classical solution for sufficiently smooth data. More importantly, for $t > 0$, $\mathcal{Q}^t$ is a compact operator on C^0 and on L^1, see [Zee88]. Compactness of the

semigroup is a desirable property and can be used to show convergence of numerical methods, like Ulam's method [Del99].

Unfortunately, here it is not possible to discretize the infinitesimal generator by considering the exact box-to-box flow rates, since

$$\lim_{t\to 0} \frac{\pi_n \mathcal{Q}^t \pi_n u - \pi_n u}{t}$$

may not exist in L^1. This can be seen by the simple one dimensional example with zero flow (only diffusion) and $u = \chi_i$ for an arbitrary subinterval X_i. The diffusion smears out the discontinuity of χ_i with an infinite flow rate, hence the above limit does not exist. We have to deal with the diffusion differently. Define the discrete Laplace operator $\Delta_n : L^1 \to V_n$ as

$$\Delta_n u := \sum_i n^2 \left(\sum_{j\in\mathcal{N}(i)} \Big(u_j - u_i\Big) \right) \chi_i, \quad \text{where} \quad \pi_n u = \sum_i u_i \chi_i, \tag{5.11}$$

and

$$\mathcal{N}(i) := \Big\{ j \neq i \mid m_{d-1}(X_i \cap X_j) \neq 0 \Big\},$$

with m_{d-1} being the $d-1$ dimensional Lebesgue measure. The set $\mathcal{N}(i)$ contains the indices of the neighboring boxes to i which have a common ($d-1$ dimensional) face. This is not only the usual discretization from finite differences, but it also restores some of the lost intuition, that the discretization may be viewed in terms of flow rates. It tells us, that the flow rate between adjacent boxes is proportional to the mass difference between them. This is a known property of the diffusion, since $\Delta u = \operatorname{div}(\nabla u)$. The matrix representation D_n of Δ_n satisfies

$$D_{n,ij} = \begin{cases} n^2 & j \in \mathcal{N}(i), \\ -2dn^2 & j = i, \\ 0 & \text{otherwise.} \end{cases}$$

We still denote by $\mathcal{P}^t$ the transfer operator of the deterministic system ($\varepsilon = 0$) and by $\mathcal{A}_n$ its discretized generator. The discretized generator of the diffusive system is now defined as:

$$\mathcal{A}_n^{(\varepsilon)} u := \frac{\varepsilon^2}{2} \Delta_n u + \mathcal{A}_n u. \tag{5.12}$$

Remark 5.24. A slight modification has to be applied, if the boxes are not cubes, but hyperrectangles with edge length h_k along the kth coordinate direction. The mass loss of box i (to box j, which is adjacent to i along the kth coordinate direction) is proportional to the mass difference between the two boxes and the surface of their common face, however, inversely proportional to h_k and the volume of box i. Thus, (5.11) turns to

$$\Delta_n u := \sum_i \left(\sum_{j \in \mathcal{N}(i)} h_{k(j)}^{-2} \left(u_j - u_i \right) \right) \chi_i,$$

where $k(j)$ is the direction along which X_i and X_j are adjacent.

5.4.2 Convergence

Pointwise convergence of the approximative generator and the corresponding semigroup. It is easy to check that $\Delta_n u \to \Delta u$ in L^1 as $n \to \infty$ for every $u \in C^2$. Since for $u \in C^2 \subset C^1$ also $\mathcal{A}_n u \to \mathcal{A} u$ holds, we have $\mathcal{A}_n^{(\varepsilon)} u \to \mathcal{A}^{(\varepsilon)} u$ for $u \in C^2$. To show the convergence of the semigroup corresponding to the approximative generator to the transfer operator semigroup by Theorem 5.5, we just need the following:

Lemma 5.25. *Assume $v \in C^\infty(X, \mathbb{R}^d)$. Then, for a $\lambda > 0$ sufficiently large $(\lambda - \mathcal{A}^{(\varepsilon)})C^2$ is dense in L^1.*

Proof. From Theorem 9.9 in [Agm65] we have $C^\infty \subset (\lambda - \mathcal{A}^{(\varepsilon)})C^\infty$ for a sufficiently large λ. Since C^∞ is contained in C^2 and dense in L^1, the claim follows immediately. □

Corollary 5.26. *Assume $v \in C^\infty(X, \mathbb{R}^d)$. Then, the semigroup generated by the approximative generator $\mathcal{A}_n^{(\varepsilon)}$ converges to $\mathcal{Q}^t$ pointwise in L^1 as $n \to \infty$ and uniformly in t for t from bounded intervals.*

Convergence of eigenfunctions. We recall, that our aim with the discretization of the infinitesimal generator is the approximation of its eigenmodes, from which we extract the information about the long-term behavior of the corresponding system. Therefore, the most desired convergence results are of the following form.

Conjecture 5.27. *Fix $\varepsilon > 0$. Let $\mathcal{A}^{(\varepsilon)} u = \lambda u$ for some $\|u\| = 1$. Then, for n sufficiently large there are λ_n, u_n, with $\|u_n\| = 1$, such that $\mathcal{A}_n^{(\varepsilon)} u_n = \lambda_n u_n$, and $\lambda_n \to \lambda$ and $\|u_n - u\| \to 0$ as $n \to \infty$.*

We sketch here a possible proof. The missing link is Conjecture 5.28, for which we do not have a proof.

Fix $t > 0$ and consider $\mathcal{Q}^t$ and $\mathcal{Q}_n^t$, the semigroups generated by $\mathcal{A}^{(\varepsilon)}$ and $\mathcal{A}_n^{(\varepsilon)}$ respectively. Since the range of $\mathcal{Q}_n^t$ is *not* V_n,[1] it is advantageous to work with $\widetilde{\mathcal{Q}}_n^t = \mathcal{Q}_n^t \pi_n$ instead, which is *no* semigroup, however. Because the range of $\mathcal{A}_n^{(\varepsilon)}$ is a subset of V_n, $\widetilde{\mathcal{Q}}_n^t$ and $\mathcal{A}_n^{(\varepsilon)}$ share the same eigenfunctions. The corresponding eigenvalues transform as $\lambda\left(\mathcal{Q}^t\right) \mapsto \lambda(\mathcal{A}) = \frac{1}{t}\log\left(\lambda\left(\mathcal{Q}^t\right)\right)$, which is a Lipschitz continuous transformation for $\lambda(\mathcal{Q}^t)$ near one. Hence, it is equivalent to state Conjecture 5.27 with replacing the generators by the corresponding operators $\mathcal{Q}^t$ and $\widetilde{\mathcal{Q}}_n^t$ (for the fixed time $t > 0$).

The advantage of doing this is that $\mathcal{Q}^t$ and $\widetilde{\mathcal{Q}}_n^t$ are compact operators, and these are better understood from the perspective of spectral approximation. We would like to use the results from [Osb75]. There are two assumptions which have to hold:

1. Pointwise convergence of $\widetilde{\mathcal{Q}}_n^t$ to $\mathcal{Q}^t$ in L^1 as $n \to \infty$.

2. Collective compactness of the sequence $\left\{\widetilde{\mathcal{Q}}_n^t\right\}_{n\in\mathbb{N}}$; i.e. that the set $\left\{\widetilde{\mathcal{Q}}_n^t u \middle| \|u\|_{L^1} \le 1, n \in \mathbb{N}\right\}$ is relatively compact.

The first assumption follows from Corollary 5.26 and that $\pi_n \to I$ pointwise as $n \to \infty$. Concerning the second one, we would like to show that the total variation of the functions $\widetilde{\mathcal{Q}}_n^t u$, where $\|u\|_{L^1} \le 1$, is bounded from above independently on n. This would imply the relative compactness by Theorem 1.19 in [Giu84]. One can see easily that if the following conjecture holds, we have the (in n uniform) boundedness of the total variation.

Conjecture 5.28. *For simplicity, assume, that every box covering consists of congruent boxes with edge length $1/n$. For every $t > 0$ there is a $K(t) > 0$ such that for any $f \in V_n$ with $\|f\|_{L^1} \le 1$, $u := \widetilde{\mathcal{Q}}_n^t f$ satisfies*

$$\frac{\left|u_i - u_j\right|}{1/n} \le K(t) \qquad \text{for all } j \in \mathcal{N}(i), \tag{5.13}$$

and the bound is independent on $n \in \mathbb{N}$.

[1] It holds merely that the range of $\left(\mathcal{Q}_n^t - I\right)$ is a subset of V_n. Compare with the representation of $\mathcal{R}_n^t$ in Theorem 5.21.

Inequality (5.13) bounds the "discrete derivatives" of the piecewise constant functions from $\widetilde{\mathcal{Q}}_n^t f \in V_n$. So, we expect (5.13) to hold, since the diffusion "smears out" any rough behavior in the initial functions f; just as it was exploited for the continuous case in [Zee88]. On analogy to the proof of Zeeman, we are able to show (5.13) for $X = \mathbb{T}^1$ and pure diffusion by using discrete Fourier transformation; however, more general results have to be found. The author is confident that results on this exist, but there is none known to him yet.

5.5 How to handle boundaries?

In this section we would like to adjust the above introduced Ulam type infinitesimal generator approach to cases where the phase space of interest has a boundary. Additional complications arise if there is no box covering which is identical with the phase space; then the latter has to be a real subset of the former. We motivate the cases with examples, but their numerical study is postponed to a later section.

If results could be shown for the diffusive case similar to Conjecture 5.28, the convergence of eigenfunctions and eigenvalues could be obtained in a similar manner like in Section 5.4.2.

5.5.1 Nondiffusive case

Our motivating example is the Lorenz system, cf. Section 5.7. For the given parameter values the system has an attractor of complicated geometry which has zero Lebesgue measure [Tuc99]. Hence, measures supported on the attractor are not absolutely continuous to the Lebesgue measure, which makes a comparison with the computed densities hard. Moreover, the covering is bigger than the attractor itself, whereby it will not be an invariant set, in general.

Keeping this example in mind, we consider a general system with the attractor X, a closed set $\widetilde{X} \supset X$ with nonempty interior and a piecewise smooth boundary. Further, let $\mathcal{X}_n$ be a covering partition of $\widetilde{X}$ containing congruent hyperrectangles, such that $\widetilde{X} \subset \operatorname{int}(X_n^+)$ with $X_n^+ := \bigcup_{X_i \in \mathcal{X}_n} X_i$; cf. Figure 5.2

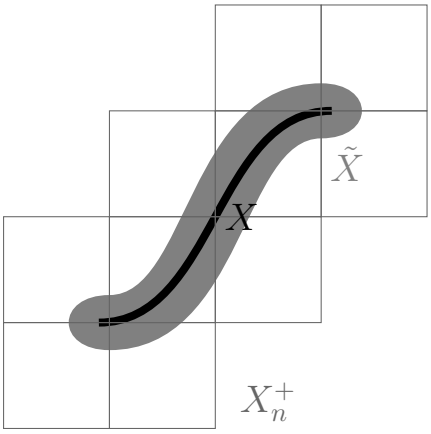


Figure 5.2: Handling complicated geometry - X is the set of interest, $\tilde{X}$ the regular neighborhood and X_n^+ its box covering.

No outflow on $\partial\tilde{X}$. Assume that $\mathbf{n}\cdot v \leq 0$ on $\partial\tilde{X}$; i.e. there is no outflow out of $\tilde{X}$. We may restrict the transfer operator $\mathcal{P}^t : L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^d)$ onto $L^1(\widetilde{X})$. For this, we extend $u \in L^1(\widetilde{X})$ to $L^1(\mathbb{R}^d)$ by

$$Eu(x) = \begin{cases} u(x), & x \in \widetilde{X}, \\ 0, & \text{otherwise}, \end{cases}$$

and set

$$\widetilde{\mathcal{P}}^t u = (\mathcal{P}^t Eu) \mid_{\widetilde{X}} .$$

Since there is no flow outwards of $\widetilde{X}$, it holds

$$\operatorname{supp}\left(\mathcal{P}^t Eu\right) \subset \widetilde{X}$$

and $\widetilde{\mathcal{P}}^t$ is a semigroup. We also have mass conservation:

$$\int_{\widetilde{X}} \widetilde{\mathcal{P}}^t u = \int_{\widetilde{X}} u.$$

Lemmas 5.17 and 5.20 apply with some slight changes (see corollaries 5.29 and 5.30), such that pointwise convergence of the approximative semigroup to $\widetilde{\mathcal{P}}^t$ follows by Theorem 5.5, analogously as in Theorem 5.21. The trick is to extend the considerations to $\mathbb{R}^d$:

Corollary 5.29. *Let* $C^1_{\widetilde{X}}(\mathbb{R}^d) := \left\{ f \in C^1(\mathbb{R}^d) \middle| \operatorname{supp}(f) \subset \widetilde{X} \right\}$. *We have* $\mathcal{A}_n u \to \mathcal{A} u$ *as* $n \to \infty$ *for* $u \in C^1_{\widetilde{X}}(\mathbb{R}^d)$.

Proof. Since there is no outflow out of $\widetilde{X}$, we have $\operatorname{supp}\left(\mathcal{P}^t E u\right) \subset \widetilde{X} \subset X_n^+$ for $t > 0$. Every function in $C^1_{\widetilde{X}}(\mathbb{R}^d)$ has uniformly continuous derivatives. Now we may reason exactly as in the proof of Lemma 5.17. □

Corollary 5.30. *For λ large enough, we have $C^1_{\widetilde{X}}(\mathbb{R}^d) \subset (\lambda - \mathcal{A})\ C^1_{\widetilde{X}}(\mathbb{R}^d)$, thus the latter set is dense in $L^1(\widetilde{X})$.*

Proof. The proof follows the lines of the one of Lemma 5.20: for $u \in C^1_{\widetilde{X}}(\mathbb{R}^d)$ we show that

$$(\lambda - \mathcal{A})^{-1} u = \int_0^\infty e^{-\lambda t} \mathcal{P}^t u \, \mathrm{d}t$$

exists and is differentiable, then a simple argument leads to the inclusion.
• Existence/differentiability: The Gronwall estimates hold uniformly in x, since u, Du, v, Dv and $D^2 v$ are all uniformly bounded on the compact set $\widetilde{X}$. If $S^{-t_0} x \notin \widetilde{X}$ for a $t_0 > 0$, then $u(S^{-t}x) = 0$ and $Du(S^{-t}x) = 0$ for all $t \geq t_0$ and the Gronwall estimate still applies.
• Inclusion: By the existence and differentiability, the above equation will hold pointwise. If $x \notin \widetilde{X}$, then $S^{-t}x \notin \widetilde{X}$ for all $t > 0$, hence $(\lambda - \mathcal{A})^{-1} u(x) = 0$, and we conclude $(\lambda - \mathcal{A})^{-1} u \in C^1_{\widetilde{X}}(\mathbb{R}^d)$. □

Including outflow on $\partial\tilde{X}$. The case where we have to take also outflow in consideration is more subtle. The restriction of the transfer operator to $\widetilde{X}$ is no semigroup anymore, since mass could leave $\widetilde{X}$ and then enter again at another place on the boundary. Our discretization is, however, constructed in a way that it cannot keep track of such mass fractions; if something leaves $\widetilde{X}$, it is lost.

We do not wish to construct adequate semigroups, which could be approximated by the one generated by $\mathcal{A}_n$, just conjecture the following:

Conjecture 5.31. *We expect $\mathcal{R}_n^t u \to \mathcal{P}^t u$ in L^1 as $n \to \infty$ for all $u \in L^1$ with*

$$\operatorname{supp}(u) \subset \left\{ x \in \widetilde{X} \middle| S^t x \in \widetilde{X}\ \forall t \geq 0 \right\},$$

i.e. for functions, which support stays completely inside $\widetilde{X}$ for all times.

5.5.2 Diffusive case

Absorbing boundary. Take the guiding example from the former section, but add now a small amount of diffusion to the dynamics. If the attracting effect of X is strong (or the diffusion is small) enough, after a sufficiently long time the majority of the

mass will be concentrated in a small neighborhood of the attractor X. We would like to restrict the significant dynamics to a bounded set which we can handle numerically.

Let $\widetilde{X} \supset X$ be an arbitrary set with a smooth boundary. We think of $\widetilde{X}$ as a set so large that only an insignificant amount of mass leaves $\widetilde{X}$, provided the initial mass was distributed closely around X. Then we may pose absorbing boundary conditions: what hits the boundary, gets lost. To this correspond homogeneous Dirichlet boundary conditions in the Fokker–Planck equation:

$$\partial_t u = \mathcal{A}^{(\varepsilon)} u, \qquad u(t,\cdot)\big|_{\partial\widetilde{X}} = 0 \quad \forall t > 0, \quad u(0,x) = u_0(x), \tag{5.14}$$

where $\mathcal{A}^{(\varepsilon)} := \frac{\varepsilon^2}{2}\Delta + \mathcal{A}$. Under the given assumptions, and by assuming that $v \in C^1(\widetilde{X}, \mathbb{R}^d)$, we have that $\mathcal{A}^{(\varepsilon)}$ generates a compact C_0 semigroup of contractions on $L^1(\widetilde{X})$, see [Ama83].[1]

Just as in the previous section, consider a tight box covering $\mathcal{X}_n$ of $\widetilde{X}$ (i.e. there is no $X_i \in \mathcal{X}_n$ with $X_i \cap \widetilde{X} = \emptyset$). Let $\mathcal{X}_n^b := \left\{ X_i \in \mathcal{X}_n \middle| \exists j \in \mathcal{N}(i) \cup \{i\} : X_j \cap \partial\widetilde{X} \neq \emptyset \right\}$ denote the set of boundary (and boundary-near) boxes, called the boundary covering. We call $X_n^\partial := \bigcup_{X_i \in \mathcal{X}_n^b} X_i$ the boundary layer. Boxes which are not in the boundary covering, have all their ($d-1$ dimensional) face neighbors in $\mathrm{int}(\widetilde{X})$, hence $\mathcal{A}_n^{(\varepsilon)} u$, defined as in (5.12), makes sense on these boxes for every $u \in L^1(\widetilde{X})$. Define $\mathcal{A}_n^{(\varepsilon)} : L^1(\widetilde{X}) \to V_n$ by

$$\mathcal{A}_n^{(\varepsilon)} u = \begin{cases} \frac{\varepsilon^2}{2}\Delta_n u + \mathcal{A}_n u, \text{ as in (5.12)}, & \text{on } X_n^+ \setminus X_n^\partial, \\ 0, & \text{on } X_n^\partial \cap \widetilde{X}. \end{cases}$$

We obtain

Theorem 5.32. *Assume $v \in C^\infty(\widetilde{X}, \mathbb{R}^d)$. Let $\mathcal{Q}_n^t$ denote the semigroup generated by $\mathcal{A}_n^{(\varepsilon)}$, defined above. Then we have the following convergences in L^1 as $n \to \infty$:*

(a) $\mathcal{A}_n^{(\varepsilon)} u \to \mathcal{A}^{(\varepsilon)} u$ for all $u \in C_0^2(\widetilde{X}) := \left\{ g \in C^2(\widetilde{X}) \middle| g\big|_{\partial\widetilde{X}} = 0 \right\}$; and

(b) $\mathcal{Q}_n^t u \to \mathcal{Q}^t u$ for all $u \in L^1(\widetilde{X})$ and for any fixed $t > 0$.

Proof. To (a). The proof of Lemma 5.17 is based on local estimates, and that argumentation applies here for all boxes in $\mathcal{X}_n \setminus \mathcal{X}_n^b$ too. Since the function $u \in C_0^2(\widetilde{X})$ has

[1] The generated semigroup is even *analytic* (in the time variable t). A semigroup $\{\mathcal{T}^t\}_{t\geq 0}$ is called compact, if T^t is a compact operator for every $t > 0$. The analyticity of the semigroup is also shown by Theorem 7.3.10 in [Paz83].

uniformly bounded derivatives, the local estimates imply the global one by the uniformity, and we have $\mathcal{A}_n u \to \mathcal{A}u$ on $\widetilde{X}$, because $m(X_n^\partial) \to 0$ as $n \to \infty$. Also $\Delta_n u \to \Delta u$ as $n \to \infty$ on $X_n^+ \setminus X_n^\partial$. This can be seen easily by Taylor expansions, considering the fact that $u \in C_0^2$ and that the operator Δ_n takes information from first-neighbor-boxes, which are still completely in $\mathrm{int}(\widetilde{X})$ for $\mathcal{X}_n \setminus \mathcal{X}_n^b$. Once again, the measure of the sets X_n^∂ tends to zero as $n \to \infty$, hence the convergence in L^1 follows.

To (b). This goes analogously to the proof of Theorem 5.21. From the theory of stochastic matrix semigroups and their generators we have that $\mathcal{A}_n^{(\varepsilon)} \in G(1,0)$, and we need to show that $(\lambda - \mathcal{A}^{(\varepsilon)})C_0^2(\widetilde{X})$ is dense in $L^1(\widetilde{X})$ for a sufficiently large $\lambda > 0$. Theorem 9.9 and Section 10 in [Agm65] shows that the Dirichlet boundary value problem

$$(\lambda - \mathcal{A}^{(\varepsilon)})w = h, \qquad w\,|_{\partial\widetilde{X}} = 0$$

has a unique solution $w \in C^\infty(\widetilde{X})$, $w\,|_{\partial\widetilde{X}} = 0$, provided $\partial\widetilde{X}$ is smooth, the coefficients of $\mathcal{A}^{(\varepsilon)}$ are smooth, and $h \in C^\infty(\widetilde{X})$. Since C^∞ is dense in L^1, and the former conditions are satisfied, the claim follows. □

Remark 5.33. Perhaps a more extensive literature study would show that the smoothness condition $v \in C^\infty(\widetilde{X}, \mathbb{R}^d)$ can be weakened. The same holds for the results in Section 5.4.2.

Reflecting boundary. Let X be a phase space which can be perfectly partitioned by boxes. In some cases an absorbing boundary does not make physically sense. Such a case would be a fluid flow in a fixed container. The vector field on the boundary is tangential to it, and the portion of mass transport caused by diffusion is reflected on the boundary. This is modeled by *reflecting* boundary conditions in the Fokker–Planck equation:

$$\partial_t u = \frac{\varepsilon^2}{2}\Delta u - \mathrm{div}(uv), \qquad \mathbf{n}\cdot\nabla u = 0 \text{ on } \partial X.^1 \tag{5.15}$$

Amann shows [Ama83] that if $v \in C^1(X, \mathbb{R}^d)$ and ∂X is a C^3 boundary, then (5.15) defines a compact C_0 semigroup of contractions on L^1.

The boundary condition, of course, has to be respected by the discretization. The definition of the drift is consistent with the boundary condition; there is no flow on the face of the box which is a part of the boundary, since the flow is tangential. Diffusion

[1] These are called *natural* boundary conditions. The general condition would be $\mathbf{n}\cdot(\frac{\varepsilon^2}{2}\nabla u - uv) = 0$, i.e. no probability flow is allowed transversely to the boundary, but by $\mathbf{n}\cdot v = 0$ this reduces to the condition given here.

occurs only between boxes of the phase space. Using the definition (5.11) for Δ_n (note the difference in the adjacency of boxes between the current phase space, which has a boundary, and between $\mathbb{T}^d$) to obtain $\mathcal{A}_n^{(\varepsilon)}$, we have:

Lemma 5.34. *Define*

$$C^2_{\mathbf{n}}(X) := \left\{ f \in C^2(X) \mid \nabla f \cdot \mathbf{n} = 0 \text{ on } \partial X \right\}.$$

Then $\mathcal{A}_n^{(\varepsilon)} u \to \mathcal{A}^{(\varepsilon)} u$ *as* $n \to \infty$ *for all* $u \in C^2_{\mathbf{n}}(X)$.

To prove this, one has to deal with the boundary terms. A Taylor expansion and considering the fact that normal derivatives are zero leads to the desired result. We omit the details. The previous lemma with the following one gives the convergence of the corresponding operator semigroups. Once again, this is a consequence of Theorem 5.5.

Lemma 5.35. *Assume that* ∂X *is uniformly* C^3. *Then there is a* $\lambda > 0$ *such that* $(\lambda - \mathcal{A}^{(\varepsilon)}) C^2_{\mathbf{n}}(X)$ *is dense in* $L^1(X)$.

Proof. From [Lun95] Proposition 3.1.23 and Theorem 3.1.25 we have that for all $f \in C^1$

$$(\lambda - \mathcal{A}^{(\varepsilon)}) u = f, \qquad (\nabla u \cdot \mathbf{n}) \mid_{\partial X} = 0$$

is solvable and $u \in C^2_{\mathbf{n}}$. Since C^1 is dense in L^1, the claim follows. □

5.6 The spectral method approach

The Ulam type approximation method for the infinitesimal generator performs very well for general systems, see Section 5.7. However, by their poor approximation properties, the piecewise constant basis functions do not allow faster than linear convergence, in general. In some specific cases, as we will see, the eigenfunctions of the infinitesimal generator, which are to be approximated, are smooth enough, such that higher order approximation functions would allow faster convergence rates and even less vector field evaluations to obtain a high accuracy.

Extensive studies have been made using piecewise polynomials as approximation functions to discretize the Frobenius–Perron operator associated with interval maps, see, e.g. [Din93, Din91]. These *local* higher order approximations perform well in most cases, and the convergence theory of Ulam's method (see [Li76]) can be extended to them.

The aim of this section is to apply tools known as spectral methods for the numerical approximation of the eigenfunctions of the infinitesimal generator. These are *global* methods, in the sense that the approximation functions have global support. We have to note that spectral methods are a highly-developed field of numerical analysis, and have been used, e.g. for the approximation of eigenmodes of differential operators; cf. [Boy01, Tre00] and references in them. Once again, the novelty is their directed usage for smooth dynamical systems. We restrict our attention to cases which are interesting for us, and focus on the question if there is a gain by using these methods, and how to implement them.

We need to justify if the objects we intend to approximate are smooth, indeed. The following result is a consequence of Theorem 9.9 in [Agm65] (see also the considerations in Section 10 in the same textbook). The definitions of an elliptic operator and of a smooth (i.e. C^∞) boundary can be found in textbooks on partial differential equations, e.g. [Agm65],[Eva98]. Note, that the infinitesimal generator $\mathcal{A}^{(\varepsilon)}$ is strongly elliptic.

Theorem 5.36. *Let X be a (closed) subset of a Euclidean space with boundary of class C^∞ and*

$$\mathcal{L}u(x) = \sum_{j,k} a_{jk}(x)\partial_{x_j x_k} u(x) + \sum_j b_j(x)\partial_{x_j} u(x) + c(x)u(x)$$

be a strongly elliptic differential operator on X with $a_{jk}, b_j, c \in C^\infty(X)$. Then all eigenfunctions of $\mathcal{L}$ (equipped with homogeneous Dirichlet or with natural boundary conditions) are in $C^\infty(X)$.

This theorem applies for domains $X \subset \mathbb{R}^d$ with smooth boundary as well as for domains like $X = \mathbb{T}^{d-k} \times [0,1]^k$, $k \in \{0,1\}$ (for $k \geq 2$ the boundary of such domains is not smooth). We will have examples on such domains too.

Similar results may hold for the case when X is a compact C^∞ Riemannian manifold with C^∞ boundary. Some results on this are Theorems 4.4, 4.7 and 4.18 in [Aub82]. Unfortunately they cover merely the pure diffusion case $\mathcal{L} = \Delta$.

5.6.1 Spectral methods for smooth problems

Function approximation. Let $X = [-1,1]$ or $X = \mathbb{T}^1$ and $u \in C^\infty(X)$. We wish to approximate u to a possibly high accuracy in the $\|\cdot\|_\infty$ norm by using a small number of approximating functions. If $X = \mathbb{T}^1$, the Fourier basis is a natural choice:

$$F_k(x) := e^{2\pi i k x}, \qquad \mathfrak{B}_n^f := \left\{ F_{k-\lfloor \frac{n-1}{2} \rfloor}(x) \middle| \; k = 0, \ldots, n-1 \right\},$$

where $i = \sqrt{-1}$ and $\lfloor x \rfloor$ is the biggest integer smaller than x. In general, we choose n to be odd such that every imaginary mode has its counterpart (the zero mode is pure real) which allows real functions to have real Fourier interpolants.

For $X = [-1, 1]$, use the Chebyshev polynomials

$$T_k(x) := \cos\left(k \ \arccos(x)\right), \qquad \mathfrak{B}_n^c := \left\{T_k(x) \mid \ k = 0, \ldots, n-1\right\}.$$

It can be shown that T_k is a polynomial[1] of degree k. By writing $\mathfrak{B}_n$ we mean "$\mathfrak{B}_n^f$ or $\mathfrak{B}_n^c$, depending on X". Choose a set of test functions, $\Psi_n = \{\psi_k : X \to \mathbb{R} \mid \ k = 0, \ldots, n-1\}$, and define the (hopefully) unique function $u_n \in \mathrm{lin}(\mathfrak{B}_n)$ as the solution of the set of linear equations

$$\int_X (u - u_n)\psi_k = 0, \quad k = 0, \ldots, n-1. \tag{5.16}$$

If $\Psi_n = \mathfrak{B}_n$, the solution of (5.16) is unique and the u_n is called the *Galerkin projection* of u onto $\mathrm{lin}\mathfrak{B}_n$.

Define the nodes $x_k^{(n)} = k/n$ if $X = \mathbb{T}^1$, and $x_k^{(n)} = -\cos\left(\frac{k\pi}{n-1}\right)$ if $X = [-1, 1]$, $k = 0, \ldots, n-1$. Setting formally $\psi_k = \delta_{x_k^{(n)}}$, with δ_x being the Dirac delta function centered in x, (5.16) turns into an interpolation problem

$$u_n\big(x_k^{(n)}\big) - u\big(x_k^{(n)}\big) = 0, \quad k = 0, \ldots, n-1. \tag{5.17}$$

The solution to this is also unique, since the x_k are pairwise different; and u_n is called the *interpolant* of u. We have for both approximation methods:

Theorem 5.37 ([Boy01],[Tre00])**.** *For $u \in C^\infty(X)$, let u_n be the Galerkin projection or the interpolant w.r.t. the nodes introduced above. Then for each $k \in \mathbb{N}$ and $\nu \in \mathbb{N}_0$ there is a $c_{k,\nu} > 0$ such that*

$$\left\|u^{(\nu)} - u_n^{(\nu)}\right\|_\infty \leq c_{k,\nu} n^{-k} \quad \textit{for all } n \in \mathbb{N}, \tag{5.18}$$

i.e. the convergence rate is faster than algebraic for each derivative of u.[2] *To this is referred as spectral accuracy. If, in addition, u is analytic, one has $c, C_\nu > 0$ such that*

$$\left\|u^{(\nu)} - u_n^{(\nu)}\right\|_\infty \leq C_\nu e^{-cn} \quad \textit{for all } n \in \mathbb{N},$$

i.e. exponential convergence.

[1] See [Tre00], Chapter 8.

[2] The νth order derivatives of a function u are denoted by $u^{(\nu)}$.

Remark 5.38. (a) We can simply extend our considerations to arbitrary intervals $[a, b] \subset \mathbb{R}$. We just use the affine-linear transformations which map X to $[a, b]$ and vice versa.

(b) Theorem 5.37 also holds if X is a multidimensional domain obtained as an arbitrary tensor product of domains $\mathbb{T}^1$ and $[-1, 1]$, e.g. $X = \mathbb{T}^1 \times [-1, 1] \times \mathbb{T}^1$. The basis of the approximation space is obtained by building tensor products of the one dimensional ones. The interpolation is also done on a tensor product grid.

(c) The reason why we picked the Chebyshev polynomials instead of any other arbitrary polynomial basis is twofold. First, interpolation on the Chebyshev grid is a well-conditioned problem, unlike the interpolation w.r.t. an equispaced grid. Second, Chebyshev and Fourier approximations are strongly related via transforming $u : [-1, 1] \to \mathbb{R}$ into $U : \mathbb{T}^1 \to \mathbb{R}$ by $U(\theta) = u(\cos(2\pi\theta))$. For further details we refer the reader to [Tre00], Chapter 8.

Operator discretization. Having a way to approximate functions by the set of approximate functions, $\mathfrak{B}_n$, it is straightforward to define approximations of differential operators. Let $V_n = \mathrm{lin}(\mathfrak{B}_n)$ and $W_n = \mathrm{lin}(\Psi_n)$. We restrict our considerations to second order operators of the form:

$$\mathcal{L}u(x) = \sum_{j,k} a_{jk}(x)\partial_{x_j x_k} u(x) + \sum_j b_j(x)\partial_{x_j} u(x) + c(x)u(x),$$

where the coefficients a_{jk}, b_j and c are smooth functions. Then we define the linear operator $\mathcal{L}_n : V_n \to V_n$ by

$$\int_X (\mathcal{L}\phi - \mathcal{L}_n\phi)\psi = 0, \quad \text{for all } \phi \in V_n, \psi \in W_n,$$

which makes sense because $V_n \subset C^\infty(X)$. If the test functions $\psi \in W_n$ are Dirac delta functions, the discretization is called *collocation*, since

$$\mathcal{L}u\big(x_k^{(n)}\big) = \mathcal{L}_n u\big(x_k^{(n)}\big), \quad \text{for all } k = 0, \ldots, n-1. \tag{5.19}$$

In the case of $W_n = V_n$ we refer to it as the *Galerkin projection*. Just as in Chapter 3, both discretizations can be written as $\mathcal{L}_n = \pi_n \mathcal{L}$ with a projector $\pi_n : C^\infty \to V_n$ defined by

$$\int_X (u - \pi_n u)\psi = 0, \quad \text{for all } \psi \in W_n.$$

Spectral convergence of eigenfunctions. The spectral accuracy of the approximation carries over to the approximate eigenmodes as well.

Theorem 5.39. *Let $\mathcal{L}$ be as above, strongly elliptic and let $\mathcal{L}_n$ be its Galerkin projection onto V_n. Then there are sequences $\left\{\lambda_{j,n}\right\}_{n\in\mathbb{N}}$ and $\left\{w_{j,n}\right\}_{n\in\mathbb{N}}$, $w_{j,n}$ being normed to unity, such that $\mathcal{L}_n w_{j,n} = \lambda_{j,n} w_{j,n}$ and*

$$\left|\lambda_{j,n} - \lambda_j\right| = \mathcal{O}\left(n^{-k}\right)$$

as $n \to \infty$ for all $k \in \mathbb{N}$. Also, there is a $u_{j,n}$ with $\mathcal{L}u_{j,n} = \lambda_j u_{j,n}$ such that

$$\left\|u_{j,n} - w_{j,n}\right\|_{H^1} = \mathcal{O}\left(n^{-k}\right)$$

as $n \to \infty$ for all $k \in \mathbb{N}$. H^1 denotes the usual Sobolev space, see, e.g. [Eva98].

Sketch of the proof. The proof is exactly the same as in II.8 [Bab91], only applied for our setting. We just verify the assumptions made there for our case. We employ the same notation as in the above work. If we refer to equations in [Bab91], it is done by using the bracket [ ].

Set $H_1 = H_2 = H^1(X)$ (or $H_0^1(X)$ in the special case of homogeneous Dirichlet boundary conditions). Let $\mu > 0$ and $\mathcal{L}_\mu := \mathcal{L} + \mu I$, where I denotes the identity. By this we just shift the spectrum, the eigenfunctions remain the same. By [3.14], if μ is sufficiently large, $\mathcal{L}_\mu$ gives rise to a strongly elliptic bilinear form. Estimates [8.2]–[8.5] follow. Continuity of the form, [8.1], follows by standard estimates, [8.7] as well.

The approximation space is defined by $S_{1,h} = V_n$ with $h = 1/n$. [8.11]–[8.12] follow from ellipticity, [8.13] from the denseness of test functions in H^1. The crucial objects which control the spectral convergence are ε_h and ε_h^* from [8.21] and [8.22]. The generalized eigenfunctions are smooth[1] and they span a finite dimensional subspace. Hence the set of normed generalized eigenfunctions M and M^* is approximated uniformly with spectral accuracy,

$$\varepsilon_h = \mathcal{O}\left(h^k\right) \quad \text{and} \quad \varepsilon_h^* = \mathcal{O}\left(h^k\right) \quad \text{for all } k \in \mathbb{N}.$$

Theorems 8.1–8.4 in [Bab91] complete the proof. □

[1] Let α be the ascent of $\lambda - \mathcal{L}_\mu$, i.e. α is the smallest number with $N\left((\lambda - \mathcal{L}_\mu)^\alpha\right) = N\left((\lambda - \mathcal{L}_\mu)^{\alpha+1}\right)$. The generalized eigenvectors are those u which satisfy $(\lambda - \mathcal{L}_\mu)^\alpha u = 0$. Let $(\lambda - \mathcal{L}_\mu)^2 u = 0$ and define $v = (\lambda - \mathcal{L}_\mu)u$. Then $(\lambda - \mathcal{L}_\mu)v = 0$, hence v is eigenvector of $\mathcal{L}_\mu$ and thus smooth. Since $(\lambda - \mathcal{L}_\mu)u = v$ it follows from Theorem 9.9 [Agm65] that u is smooth as well. The general case follows by induction.

Remark 5.40. It could seem strange in the proof above that we need to shift $\mathcal{L}$ in order to be able to apply the convergence theory. The key fact is that the spectral theory of *compact* operators is used, and $\mathcal{L}_\mu^{-1}$ is compact on suitable Sobolev spaces, with a sufficiently large shift μ. The shift influences the constant in the $\mathcal{O}\left(n^{-k}\right)$ estimate. However, modifying the r.h.s. of the variationally posed eigenvalue problem [8.10] from $b(\cdot,\cdot)$ to $\mu b(\cdot,\cdot)$, the eigenvalues transform as $\lambda \mapsto \frac{\lambda+\mu}{\mu}$, hence remain at an order of magnitude 1 for large μ. Moreover, the proofs of Theorems 8.1–8.4 in [Bab91] tell us that the factor of change introduced by the shift in the constant of the $\mathcal{O}\left(n^{-k}\right)$ estimate tends to 1 for $\mu \to \infty$. Hence, the shift does not affect the spectral convergence rate.

Presumably, it is harder to obtain similar results for the collocation method, cf. the convergence theory of both methods (Galerkin and collocation) for boundary value problems in [Can07]. However, we may strengthen our intuition that collocation converges as well, if we consider the following (cf. [Boy01] Chapter 4). First, if we compute the integrals arising in the Galerkin method by Gauss quadrature (and we will have to use numerical integration, in general) we obtain the collocation method. Second, the approximation error of interpolation is at most a factor two worse than the one of the Galerkin projection.

Algorithm 5.41 (Spectral method discretization of the generator)**.**

1. Define the approximation space V_n, which is spanned by tensor product Chebyshev and/or Fourier polynomials.

2. Compute the matrix representation $A_n^{(\varepsilon)}$ of the discretized (Galerkin or collocation) infinitesimal generator $\mathcal{A}_n^{(\varepsilon)}$ by

$$\int_X \left(\mathcal{A}^{(\varepsilon)}\phi - \mathcal{A}_n^{(\varepsilon)}\phi\right)\psi = 0, \quad \text{for all } \phi \in V_n, \psi \in W_n,$$

   as described in the following sections.

3. Right eigenvectors of $A_n^{(\varepsilon)}$ correspond to eigenfunctions of $\mathcal{A}_n^{(\varepsilon)}$, which are considered as approximations to the eigenfunctions of $\mathcal{A}^{(\varepsilon)}$. In particular, the eigenfunction of $\mathcal{A}_n^{(\varepsilon)}$ at the eigenvalue with smallest magnitude approximates the invariant density.

4. Unlike for the Ulam type approach, *left* eigenvectors of $A_n^{(\varepsilon)}$, where $A_n^{(\varepsilon)}$ is obtained by the collocation method, do not correspond to eigenfunctions of the adjoint operator. If one would like to extract information about almost invariance using

the simplex-method (cf. Section 2.2.2), one has to discretize the adjoint operator; cf. (2.14). However, this is possible *without* additional vector field evaluations.

5.6.2 Implementation and numerical costs

For simplicity and better readability we show first how to implement the spectral discretizations of differential operators in one space dimension,

$$\mathcal{L}u(x) = a(x)u''(x) + b(x)u'(x) + c(x)u(x),$$

and proceed later to the multidimensional case. The main tools will be so-called differentiation matrices, $D_n^{(1)}$ and $D_n^{(2)}$, which realize the first and second derivatives of functions in V_n.

From an applicational point of view the most convenient is to work with the nodal evaluations. Mathematically, this corresponds to the basis $\mathfrak{E}_n$ of Lagrange polynomials $\ell_0, \ldots, \ell_{n-1}$ with $\ell_j\left(x_k^{(n)}\right) = \delta_{jk}$. The multiplication by the functions a, b and c is also very simple using this basis.

Fourier and Chebyshev collocation method. Given a smooth function u, differentiating the interpolant is a good approximation to u'. For this, we define u_n as the vector of point evaluations with $u_{n,j} = u\left(x_j^{(n)}\right)$. Denoting the interpolant of u by p_n, we define $D_n^{(1)}$ and $D_n^{(2)}$ by

$$u'\left(x_j^{(n)}\right) \approx \left(D_n^{(1)} u_n\right)_j := p_n'\left(x_j^{(n)}\right) \quad \text{for all } j = 0, \ldots, n-1$$

and

$$u''\left(x_j^{(n)}\right) \approx \left(D_n^{(2)} u_n\right)_j := p_n''\left(x_j^{(n)}\right) \quad \text{for all } j = 0, \ldots, n-1.$$

For the Fourier case holds $D_n^{(1)} D_n^{(1)} = D_n^{(2)}$, which is not true for the Chebyshev case. Also, there is a simple computation of $D_n^{(1)} u_n$ in the Fourier case (the methodology is extendable to the Chebyshev case as well, cf. Remark 5.38 (c)). Note:

- Differentiation in the frequency space is merely a diagonal scaling:

$$F_k'(x) = 2\pi i k F_k(x).$$

An additional constant factor is applied if $\mathbb{T}^1$ is scaled.

- By aliasing, the modes $-\frac{n-1}{2}, \ldots, -1$ are indistinguishable from the modes $\frac{n-1}{2}+1, \ldots, n-1$ on the given grid.

Hence $D_n^{(1)} h_n$ is easily computed in several steps:

1. Compute the fast Fourier transform (FFT) of u_n and assign the frequencies $-\frac{n-1}{2}, \ldots, \frac{n-1}{2}$ to the modes (by aliasing).

2. Apply a componentwise scaling to the vector, realizing the differentiation in the frequency space.

3. Assign the frequencies $0, \ldots, n$ to the modes (again, by aliasing) and apply the inverse FFT (IFFT) to get back to the physical space (nodal evaluations).

The following diagram emphasizes the computational steps.

$$\mathfrak{E}_n \xrightarrow{\text{FFT}} \mathfrak{B}_n \xrightarrow{\frac{\mathrm{d}}{\mathrm{d}x}} \mathfrak{B}_n \xrightarrow{\text{IFFT}} \mathfrak{E}_n$$

$D_n^{(2)}$ is computed in the same way. The computational cost is $\mathcal{O}(n \log n)$.

The matrix representation L_n of $\mathcal{L}_n : V_n \to V_n$ w.r.t. the basis $\mathfrak{E}_n$ is obtained as follows. Define

$$a_n = \left(a\left(x_0^{(n)}\right), \ldots, a\left(x_{n-1}^{(n)}\right) \right)^\top,$$

b_n and c_n analogously. Let $\operatorname{diag}(d)$ denote the diagonal matrix with the vector d on the diagonal. Then we have

$$L_n = \operatorname{diag}(a_n) D_n^{(2)} + \operatorname{diag}(b_n) D_n^{(1)} + \operatorname{diag}(c_n). \tag{5.20}$$

For the grids used here, both the Fourier and Chebyshev differentiation matrices can be given analytically and can be calculated in $\mathcal{O}\left(n^2\right)$ flops [Tre00].

Fourier Galerkin method. The Galerkin discretization is more subtle to set up. While (5.19) and (5.20) gives the matrix representation L_n of the discretized operator w.r.t. $\mathfrak{E}_n$ directly, here we have $L_n = M_n^{-1} \bar{L}_n$ w.r.t. $\mathfrak{B}_n$ with

$$\bar{L}_{n,jk} = \int_X \mathcal{L} F_k \; F_j, \qquad M_{n,jk} = \int_X F_j F_k,$$

where M_n is called the mass matrix. Since the coefficient functions a, b and c are arbitrary, we cannot set up $\bar{L}_n$ analytically, numerical quadrature is needed.

On the one hand we are facing the two problems (a) we would like to obtain L_n w.r.t. $\mathfrak{E}_n$ and (b) numerical approximation of the integrals. On the other hand we already have a simple way to approximate $\mathcal{L}$: collocation. Choosing $N > n$ sufficiently large, we expect by spectral accuracy, that $\mathcal{L}_N^{\mathrm{col}} u$ (obtained by collocation) is for all $u \in V_n$ far closer to $\mathcal{L}u$ than the approximation potential on the space V_n (note that $V_n \subset V_N$). So we could use $\pi_n^{\mathrm{gal}} \mathcal{L}_N^{\mathrm{col}}$ as the numerical approximation of $\mathcal{L}_n^{\mathrm{gal}}$. We would like $L_n = L_n^{\mathrm{gal}}$ w.r.t. the basis $\mathfrak{E}_n$, but the projection π_n^{gal} is easily implemented w.r.t. $\mathfrak{B}_{\mathfrak{n}}$. To sum up, we take following strategy to obtain L_n:

$$\mathfrak{E}_n \to \mathfrak{B}_n \overset{\text{embed}}{\longrightarrow} \mathfrak{B}_N \to \mathfrak{E}_N \overset{\mathcal{L}_N^{\mathrm{col}}}{\longrightarrow} \mathfrak{E}_N \to \mathfrak{B}_N \overset{\text{project}}{\longrightarrow} \mathfrak{B}_n \to \mathfrak{E}_n. \tag{5.21}$$

The transformations $\mathfrak{E} \leftrightarrow \mathfrak{B}$ are simple FFT/IFFT-pairs (one should not forget the rearranging; see above). The embedding and projecting needs some explanation, however. Generally, we consider the truncated Fourier series containing the frequencies $-\frac{n-1}{2}, \dots, 0, \dots \frac{n-1}{2}$. We respect this with the embedding, hence the amplitudes of the frequencies $-\frac{N-1}{2}, \dots, -\frac{n+1}{2}$ and $\frac{n+1}{2}, \dots, \frac{N-1}{2}$ are set to zero and the embedding $\mathfrak{B}_n \to \mathfrak{B}_N$ is complete. The projection is not more complicated either. Since the basis is orthogonal w.r.t. the L^2 scalar product, projection is nothing but throwing out unwanted frequencies.

Chebyshev Galerkin method. In fact, the strategy is exactly the same as for the Fourier Galerkin method, however the basis transformations $\mathfrak{E} \leftrightarrow \mathfrak{B}$ and the projection are not so simple.

The embedding is the extension of $T_0, \dots, T_n$ to $T_0, \dots, T_N$. The transformation $\mathfrak{B}_n \to \mathfrak{E}_n$ is given by $S_n \in \mathbb{R}^{n \times n}$ with

$$S_{n,jk} = T_{k-1}(x_{j-1}^{(n)}).$$

Now to the projection. The Chebyshev polynomials satisfy[1]

$$T_{m,n} := \int_{-1}^{1} T_m(x) T_n(x) \mathrm{d}x = -\frac{(m^2 + n^2 - 1)(1 + (-1)^{m+n})}{((m-n)^2 - 1)((m+n)^2 - 1)}.$$

Observe that if m and n don't share the same parity, $T_{m,n} = 0$. By transforming the problem onto the interval $[a, b]$ is $T_{m,n}$ multiplied by a factor $(b - a)/2$. The mass

[1] Computation made by Mathematica.

matrices M_N resp. M_n are given by $M_{N,jk} = T_{j,k}$ resp. $M_n = (M_N)_{1:n,1:n}$, where we are using the usual Matlab notation to indicate sub-matrices. Hence, the projection from $\mathfrak{B}_N$ to $\mathfrak{B}_n$ is given by the matrix

$$M_n^{-1}(M_N)_{1:n,1:N} = \begin{bmatrix} I_n & M_n^{-1}(M_N)_{1:n,n+1:N} \end{bmatrix}.$$

I_n denotes the identity. This gives by the diagram (5.21)

$$L_n = S_n \begin{bmatrix} I_n & M_n^{-1}(M_N)_{1:n,n+1:N} \end{bmatrix} S_M^{-1} L_n^{\mathrm{col}} (S_M)_{1:N,1:n}.$$

Extending to multiple dimensions. For multidimensional domains of tensor product structure (i.e. $X = \bigotimes_{j=1}^{d} X_j$, where either $X_j = [a_j, b_j] \subset \mathbb{R}$ or $X_j = \mathbb{T}^1$ for each j) there is a very simple extension of the above introduced methods. For notational simplicity we handle here the two dimensional case where the domain is $Y \times Z$, Y and Z being one dimensional, and we show it only for the collocation method. The methodology is then applicable for more dimensions and the Galerkin method without difficulties.

In multiple dimensions, we consider tensor product grids resp. tensor product basis functions. Let the one dimensional grids be given by $\mathbf{y} = \{y_1, \ldots, y_n\}$ and $\mathbf{z} = \{z_1, \ldots, z_m\}$. The grid points in the two dimensional grid are ordered by the "z-first-principle", i.e.[1]

$$(y_1, z_1), (y_1, z_2), \ldots, (y_1, z_m), (y_2, z_1), (y_2, z_2), \ldots, (y_n, z_m).$$

This implies, that any linear operation $\mathcal{L}$ on the y coordinate, given on the grid $\mathbf{y}$ by $L_{\mathbf{y}}$, is carried out on the full grid by $L_{\mathbf{y}} \otimes I_m$; and any linear operation $\mathcal{L}$ on the z coordinate, given on the grid $\mathbf{z}$ by $L_{\mathbf{z}}$, by $I_n \otimes L_{\mathbf{z}}$. I_n is the unit matrix in $\mathbb{R}^n$ and $A \otimes B$ denotes the Kronecker product of the matrices A and B.
For example, the divergence operator $\partial_y + \partial_z$ is discretized by

$$D_n^{(1)} \otimes I_m + I_n \otimes D_m^{(1)},$$

where $D_n^{(1)}$ and $D_m^{(1)}$ are the differentiation matrices derived earlier for the factor spaces.

If one would like to apply consecutively two linear operations on one coordinate, following identity may save computational resources: $(I_n \otimes L_{\mathbf{z}})(I_n \otimes K_{\mathbf{z}}) = I_n \otimes (L_{\mathbf{z}} K_{\mathbf{z}})$.

[1] Hence the global index of the point (y_j, z_k) is $(j-1)m + k$.

Discussion and computational costs. It should be emphasized once more that the collocation methods have a very simple implementation (see also [Tre00]). The computationally most expensive step is to evaluate the coefficients of $\mathcal{L}$. In our case $\mathcal{L}u = -\mathcal{A}^{(\varepsilon)}u = -\frac{\varepsilon^2}{2}\Delta u + \operatorname{div}(uv)$, so the coefficient evaluation reduces to the evaluation of the vector field v. This suggests to measure the costs of the assembling of the approximative operator in the number of the vector field evaluations. The collocation method uses one evaluation per node, i.e. $\mathcal{O}(n)$, where n is the dimension of the approximation space V_n.

The question may arise, that if we already have computed an accurate approximation $\mathcal{L}_N^{\text{col}}$ to the operator $\mathcal{L}$, why do we not just use it instead of the low-precision one, $\mathcal{L}_n^{\text{gal}}$?

Unlike the basis in the Ulam type approach, the basis of the approximation space for spectral methods consists of globally supported functions. Hence, the discretized operator will be a fully occupied matrix. By this, the eigenvalue and eigenvector computations cost at least a factor $\mathcal{O}(n)$ more in comparison to the sparse matrices of the Ulam type method. It is also worth to note, that for Ulam's method one searches for the largest eigenvalues of the discrete transfer operator. This is done by forward iteration. For the infinitesimal generator approach, we are seeking for the eigenvalues with the smallest magnitude, which is implemented by backward iteration. That means, we have to solve in each iteration step a system of linear equations. Iterative methods (e.g. GMRES) can solve a problem $Ax = b$ in $\mathcal{O}(\#\text{flops}(A\cdot x))$ flops. Still, it means a complexity of $\mathcal{O}\left(n^2\right)$ for our fully occupied matrices. Although, by spectral accuracy we expect to obtain fairly good results with a *small* number of ~ 10 basis functions in each dimension, the effect of the $\mathcal{O}\left(n^2\right)$ complexity should not be underestimated in higher dimensions.

So, while setting up the operator approximation is cheap, since a small number of vector field evaluations have to be used, solving the eigenproblem may be computationally expensive. In general, one expects Galerkin methods to do better than collocation methods with the same number of basis functions, since the projection uses global information (since the ψ_k are globally supported functions) in contrast to collocation, where we have the information merely from the nodes. If there are high oscillatory modes "hidden" from collocation, the Galerkin method may deal with them as well. Consequently, one is well advised to use Galerkin methods if collocation does not seem

to be accurate enough, and the approximation matrix is so big that we are already on the limit of our computational resources.

However, in all examples below we have obtained sufficiently accurate results by the collocation method.

5.6.3 Adjustments to meet the boundary conditions

The two dynamical boundary conditions (absorbing and reflecting) also equip the corresponding infinitesimal operator with boundary conditions (homogeneous Dirichlet or natural/Neumann). The discretization has to behold this as well. Since $\mathbb{T}^1$ has no boundary, boundaries arise only on directions where the Chebyshev grid is applied. The endpoints of the interval are Chebyshev nodes, that allows a comfortable treatment.

Homogeneous Dirichlet BC: Setting the function values to zero at the boundary is equivalent with erasing the rows and columns of the matrix L_n which correspond to these nodes. The eigenvectors of the resulting matrix L'_n correspond to values in the "inner" nodes, the nodes on the boundary have value zero.

Also, we could choose basis functions which satisfy the boundary conditions a priori. One possible way is explained below the Neumann boundary conditions. For the Dirichlet boundary we did not use this kind of approach in our examples. We refer the reader to Section 3.2 in [Boy01].

Natural/Neumann BC: Since we expect the vector field to be tangential at the boundary of the state space, the natural boundary conditions simplify to $\nabla u \cdot \mathbf{n} = 0$. The tensor product structure of the state space reduces this to $\partial_{x_j} u = 0$ on the boundary defined by $x_j = \text{const}$. Here we have two possible solutions: include the boundary conditions by setting up a generalized eigenvalue problem or use another set of of basis functions which satisfy the condition $\partial_{x_j} u = 0$.

The first idea includes the boundary conditions into the operator. The eigenvalue problem $L_n u = \lambda u$ is replaced with $L'_n u = \lambda K_n u$. Those rows of L_n which correspond to the boundary nodes are replaced by the corresponding row of the differentiation matrix which discretizes the operator ∂_{x_j}, hence we obtain L'_n. K_n is the identity matrix except the diagonal entries corresponding to the boundary nodes, which are set to zero. The modified rows enforce $\partial_{x_j} u = 0$ for the computed eigenfunctions.

A basis adapted to the boundary conditions. Once again, the ideas are represented in one dimension, and carried over easily to multidimensional tensor product spaces. We would like to use a subspace of V_n consisting of functions which a priori satisfy the boundary conditions $u'(x) \mid_{x=\pm 1} = 0$.

Our first aim is to find a *simple* linear combination of the Chebyshev polynomials T_k, such that the resulting functions form a basis of the desired space. We have from [Boy01]:

$$\frac{\mathrm{d}T_k}{\mathrm{d}x}(x)\bigg|_{x=1} = k^2, \qquad \frac{\mathrm{d}T_k}{\mathrm{d}x}(x)\bigg|_{x=-1} = k^2(-1)^{k-1}.$$

Possible simple combinations of basis functions are (for $k \geq 1$)

(a) $\tilde{T}_{2k+1} = \frac{1}{(2k+1)^2} T_{2k+1} - T_1$, $\tilde{T}_{2k} = \frac{1}{k^2} T_{2k} - T_2$,

(b) $\tilde{T}_k = (k-1)^2/(k+1)^2 T_{k+1} - T_{k-1}$.

The factors are chosen such that $\|\tilde{T}_k\|_\infty \nrightarrow 0$ and $\|\tilde{T}_k\|_\infty \nrightarrow \infty$ as $k \to \infty$. Choice (a) has the drawback, that the $\tilde{T}$ converge to T_1 or T_2. This ruins the condition of the approximation problem. Thus, we take choice (b), $\tilde{T}_k = (k-1)^2/(k+1)^2, \tilde{T}_{k+1} - T_{k-1}$. Note that

$$\|\tilde{T}_k\|_\infty \leq 2 \quad \text{and} \quad |\tilde{T}_k(\pm 1)| = \frac{1-(k-1)^2}{(k+1)^2} \sim \frac{4}{k} \quad \text{as } k \to \infty.$$

The basis functions $\tilde{T}_k$ get smaller closer to the boundary. Nevertheless, the number of basis functions is ~ 50 for spectral methods, so interpolating with this basis should stay well-conditioned.

Implementation: The usual approach to compute a differentiation matrix of dimension n would be to fix some interpolation (and evaluation) points, interpolate on this grid w.r.t. $\tilde{T}_1, \ldots, \tilde{T}_n$, and derive a (hopefully simple) analytic formula for the matrix. To omit a possibly complicated analysis, we take advantage of the known differentiation matrix for the full Chebyshev basis and use another approach instead: embed the subspace spanned by the $\tilde{T}_k$ into the span of the T_k and make the differentiation w.r.t. the known basis.

Note, $\operatorname{span}\left\{\tilde{T}_1, \ldots, \tilde{T}_n\right\} \subset \operatorname{span}\left\{T_0, \ldots, T_{n+1}\right\}$. Let $\left\{x_k\right\}_{k=0\ldots n+1}$ denote the points of the $(n+2)$-point Chebyshev grid. Further define

- $\mathfrak{E}_{\tilde{T}}$: Lagrange basis in the nodes $x_1, \ldots, x_n$.

- $\mathfrak{E}_T$: Lagrange basis in the nodes $x_0, \ldots, x_{n+1}$.
- $\mathfrak{B}_{\tilde{T}}$: Basis $\left\{\tilde{T}_1, \ldots, \tilde{T}_n\right\}$.
- $\mathfrak{B}_T$: Basis $\left\{T_0, \ldots, T_{n+1}\right\}$.
- $D_{\tilde{T}}, D_T$: Differentiation matrices on the spaces $\mathfrak{E}_{\tilde{T}}$ and $\mathfrak{E}_T$ respectively.

We would like to set up the differentiation matrix on $\mathfrak{E}_{\tilde{T}}$. We know the differentiation matrix on $\mathfrak{E}_T$, and the transformation $\mathfrak{B}_{\tilde{T}} \to \mathfrak{B}_T$ by the above definition of the $\tilde{T}_k$. The basis transformation $\mathfrak{E} \leftrightarrow \mathfrak{B}$ is given by matrices S and S^{-1} below. Hence, the computation follows the diagram:

$$\mathfrak{E}_{\tilde{T}} \xrightarrow{S_{\tilde{T}}^{-1}} \mathfrak{B}_{\tilde{T}} \xrightarrow{B_{\tilde{T}\to T}} \mathfrak{B}_T \xrightarrow{S_T} \mathfrak{E}_T \xrightarrow{\frac{\mathrm{d}}{\mathrm{d}x}} \mathfrak{E}_T \xrightarrow{\text{restrict}} \mathfrak{E}_{\tilde{T}},$$

where

$$S_{\tilde{T},ij} = \tilde{T}_j(x_i), \qquad S_{T,ij} = T_{j-1}(x_{i-1}) \qquad \text{and} \qquad B_{\tilde{T}\to T,ij} = \begin{cases} \frac{(j-1)^2}{(j+1)^2} & i = j+2, \\ -1 & i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Note: $S_{\tilde{T}} \in \mathbb{R}^{n\times n}, S_T \in \mathbb{R}^{(n+2)\times n+2}$ and $B_{\tilde{T}\to T} \in \mathbb{R}^{(n+2)\times n}$. Considering, that the restriction is just cutting off the first and last components, we have (using MATLAB notation)

$$D_{\tilde{T}} = \left(D_T S_T B_{\tilde{T}\to T} S_{\tilde{T}}^{-1}\right)_{2:n+1,:}.$$

Further simplifications can be made by realizing that $S_T B_{\tilde{T}\to T} S_{\tilde{T}}^{-1} : \mathfrak{E}_{\tilde{T}} \to \mathfrak{E}_T$ is the identity on the inner grid points, i.e.

$$S_T B_{\tilde{T}\to T} S_{\tilde{T}}^{-1} = \begin{bmatrix} w_1^\top \\ I_{n\times n} \\ w_2^\top \end{bmatrix}.$$

Using the partition $(D_T)_{2:n+1,:} = \begin{bmatrix} d_1 & \widetilde{D} & d_2 \end{bmatrix}$, where d_1 and d_2 are the first and last columns, respectively, we may write

$$D_{\tilde{T}} = d_1 w_1^\top + \widetilde{D} + d_2 w_2^\top.$$

5.7 Numerical examples

5.7.1 A flow on the circle

We start with a one dimensional example, a flow on the unit circle. The vector field is given by

$$v(x) = \sin(4\pi x) + 1.1,$$

$x \in \mathbb{T}^1 = [0, 1]$ with periodic boundary conditions, and we wish to compute the invariant density of the system. Recall that an invariant density $u \in L^1(\mathbb{T}^1)$ needs to fulfill $\mathcal{A}u = 0$, where $\mathcal{A}u = -(uv)'$. The unique solution to this equation is $u^*(x) = C/v(x)$, C being a normalizing constant (i.e. such that $\|u^*\|_{L^1} = 1$). We use three methods in order to approximate u^*:

1. the classical method of Ulam for the Frobenius-Perron operator (cf. Section 2.3) for $t = 0.01$,
2. Ulam's method for the generator and
3. the spectral method for the generator.

Figure 5.3 (left) shows the true invariant density (dashed black line), together with its approximations by Ulam's method for the generator (bars) on a partition with 16 intervals and the spectral method for the generator for 16 grid points (solid line). In Figure 5.3 (right) we compare the efficiency of the three methods in terms of how the L^1-error of the computed invariant density depends on the number of vector field evaluations.

Efficiency comparison

- **Ulam's method.** The error in Ulam's method decreases like $\mathcal{O}(n^{-1})$ for smooth invariant densities [Din93]. Thus, we need to compute the transition rates between the intervals to an accuracy of $\mathcal{O}(n^{-1})$ (since otherwise we cannot expect the approximate density to have a smaller error). To this end, we use a uniform grid of n sample points in each interval. In summary, this leads to $\mathcal{O}(n^2)$ evaluations of the vector field. For the numbers in Figure 5.3 we only counted each point once, i.e. we neglected the fact that for the time integration we have to perform several time steps per point.

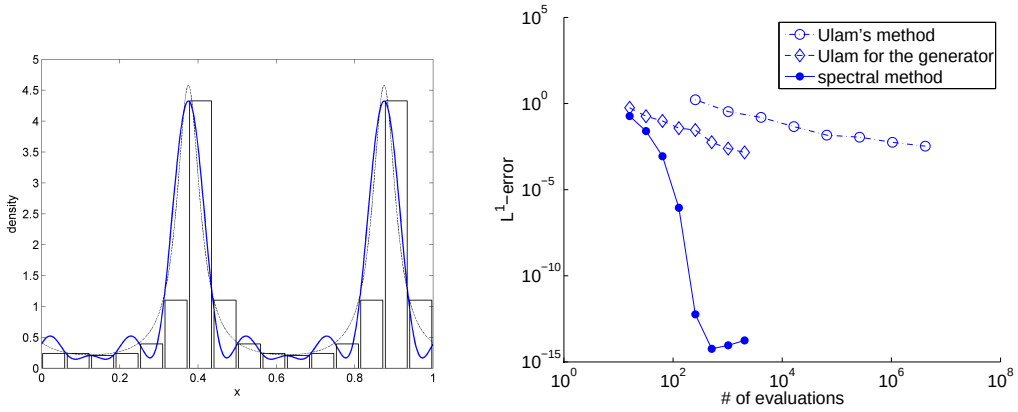


Figure 5.3: Left: true invariant density (dashed line), approximation by Ulam's method for the generator (bars) and approximation by the spectral method (solid line). Right: L^1-error of the approximate invariant density in dependence on the number vector field evaluations.

- **Ulam's method for the generator.** Here, only one evaluation of the vector field per interval is needed. On a partition with n intervals, this method then seems to yield an accuracy of $\mathcal{O}(n^{-1})$. Note, that from Corollary 5.12 it follows that the vector with components $1/v(x_i)$ is a right eigenvector of the transition matrix (5.3) for the generator at the eigenvalue 0. This fact shows the pointwise convergence of the invariant density of the discretization towards the real one.

- **Spectral method.** Choose n odd here. By the odd number of grid points every complex mode has also its conjugate in the approximation space, thus real data have a pure real interpolant. This helps to avoid instabilities in the imaginary direction.[1]

  Here, the vector field is evaluated once per grid point. As predicted by Theorem 5.39, the accuracy increases exponentially with n.

(Almost) cyclic behavior. It has been shown in [Del99] that complex eigenvalues of modulo (near) one of the transfer operator imply (almost) cyclic dynamical behavior. Similar holds for the generator as well.

Lemma 5.42. *Let $\mathcal{A}u = \lambda_A u$ and let u_{re} denote the real part of u. Let $t > 0$ be such that $e^{t\lambda_A} = \lambda_P \in \mathbb{R}$. Then $\mathcal{P}^t u_{re} = \lambda_P u_{re}$.*

[1] This problem is also known in numerical differentiation, see [Tre00], Chapter 3.

Proof. From the proof of Theorem 2.2.4 in [Paz83] we have $\mathcal{P}^t u = \lambda_P u$. If u_{im} denotes the imaginary part of u, we have by linearity: $\mathcal{P}^t u = \mathcal{P}^t u_{re} + i\, \mathcal{P}^t u_{im}$. Thus

$$\underbrace{\lambda_P u_{re}}_{\in\mathbb{R}} + i\ \underbrace{\lambda_P u_{im}}_{\in\mathbb{R}} = \underbrace{\mathcal{P}^t u_{re}}_{\in\mathbb{R}} + i\ \underbrace{\mathcal{P}^t u_{im}}_{\in\mathbb{R}}.$$

The claim follows immediately. □

Hence, having a non-real $\lambda_A \in \sigma(\mathcal{A})$ and a $t > 0$ with $1 \approx e^{t\lambda_A} \in \mathbb{R}$, then the real part of the corresponding eigenfunction yields a decomposition of the phase space into almost cyclic sets.

Let us test this on our example. The vector field v gives rise to a periodic flow with period $t_0 = \int_0^1 1/v(x)\,\mathrm{d}x \approx 2.1822$. Thus, we expect the infinitesimal generator to have pure imaginary eigenvalues with imaginary parts $2\pi k/t_0$, $k \in \mathbb{Z}$. For $k = 1, 2, 3$, the spectral method approach with $n = 63$ provides these eigenvalues with an error of 10^{-14}, 10^{-5} and 10^{-3}, respectively. The real parts of the computed eigenvalues are all at most 10^{-13}.

Making these computations with the Ulam type generator approach, we experience that the eigenvalues have not negligible negative real parts; however diminishing in magnitude, as n gets larger. This phenomenon is discussed in the following paragraph.

Numerical diffusion. Assume, for a moment, that $v \equiv \bar{v} > 0$, i.e. the flow is constant. Numerical diffusion arises, when the discretization $\mathcal{A}_n$ of the differential operator $\mathcal{A}u = -(uv)'$ is actually a higher order approximation of the differential operator $\mathcal{A}_\varepsilon u := \varepsilon u'' - (uv)'$ for some $\varepsilon > 0$. This is the case for the upwind method (the Ulam type generator approximation). To see this, let a uniform partition of $\mathbb{T}^1$ be given with box size $1/n$, and π_n the projection onto the space of piecewise constant functions over this partition. Let $u \in C^4(\mathbb{T}^1)$ and $u_n := \pi_n u$. Then it holds

$$\begin{aligned}
(\mathcal{A}_n u)_i &= n\bar{v}\left(u_{n,i-1} - u_{n,i}\right) \\
&= n\bar{v}\left(\frac{u_{n,i-1} - u_{n,i+1}}{2} + \frac{u_{n,i-1} - 2u_{n,i} + u_{n,i+1}}{2}\right) \\
&= \bar{v}\frac{u_{n,i-1} - u_{n,i+1}}{2n^{-1}} + \frac{\bar{v}}{2n}\frac{u_{n,i-1} - 2u_{n,i} + u_{n,i+1}}{n^{-2}},
\end{aligned}$$

hence $\mathcal{A}_n u = \pi_n \mathcal{A}_\varepsilon u + \mathcal{O}\left(n^{-2}\right)$ with $\varepsilon = \frac{\bar{v}}{2n}$, while $\mathcal{A}_n u = \pi_n \mathcal{A} u + \mathcal{O}\left(n^{-1}\right)$. That is why one expects quantities computed by $\mathcal{A}_n$ to reflect the actual behavior of $\mathcal{A}_\varepsilon$. For more details we refer to [LeV02], Section 8.6.1.

Since general flows are not constant, better models of the numerical diffusion can be gained by setting the diffusion term dependent on the spatial variable; i.e. $\varepsilon = \varepsilon(x)$.

Figure 5.4 shows a numerical justification of the above considerations. We compare the dependence of the real part of the second smallest eigenvalue of the Ulam type generator on the number of partition elements n, and the dependence of the real part of the second smallest eigenvalue of $\mathcal{A}_\varepsilon$ on ε, where $\mathcal{A}_\varepsilon$ is discretized by the spectral method (for $n = 151$ the computed eigenvalues are considered to be exact).

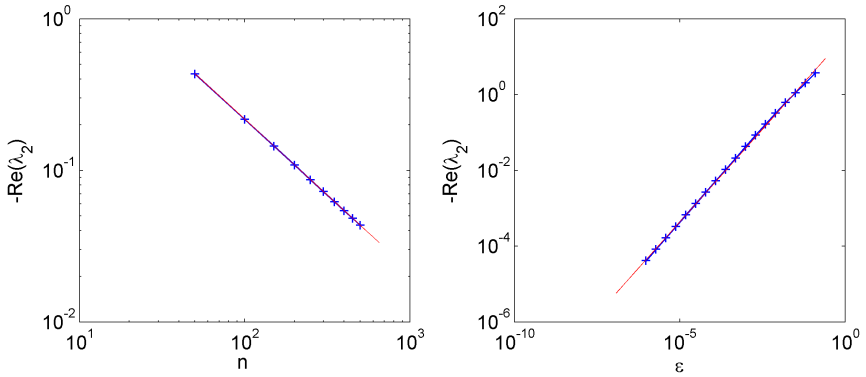


Figure 5.4: Dependence of the second smallest eigenvalue of the Ulam type generator approximation on the partition size n (left); and dependence of the second smallest eigenvalue of the infinitesimal generator on the diffusion parameter ε (right). The '+' signs indicate the computed values and the solid line is obtained by linear fitting of the data.

A linear fitting (indicated in the plots by red lines) gives $\varepsilon \sim 0.55 \cdot n^{-0.98}$, which is in very good correspondence with the theoretical prediction. Moreover, the slope equal to one in the right plot also suggests the asymptotics $\mathrm{Re}(\lambda_2) \sim c\varepsilon$ as $\varepsilon \to 0$.

5.7.2 An area-preserving cylinder flow

We consider an area-preserving flow on the cylinder, defined by interpolating a numerically given vector field as shown in Figure 5.5, which is a snapshot from a quasi-geostrophic flow, cf. [Tre90, Tre94]. The domain is periodic with respect to the x coordinate and the field is zero at the boundaries $y = 0$ and $y = 8 \cdot 10^5$.

Perturbing the model. Looking at the vector field we expect the system to have several fixed points in the interior of the domain, which are surrounded by periodic

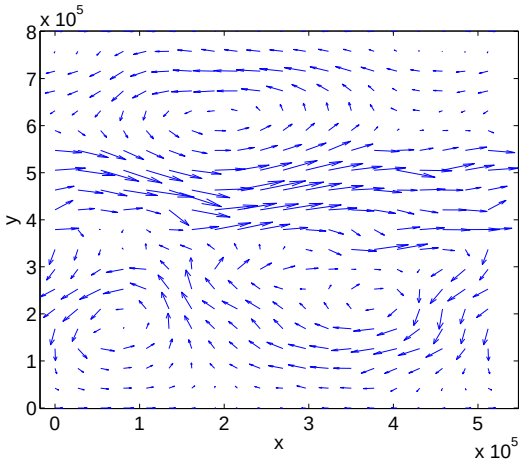


Figure 5.5: Vector field of the area-preserving cylinder flow

orbits. Hence, there will be a continuum of invariant sets; and we examine their robustness under random perturbations of the deterministic system.

For this, we choose the noise level ε such that the resulting diffusion coefficient $\varepsilon^2/2$ is larger than, but has the same order of magnitude as, the numerical diffusion present within Ulam's method for the generator. Since the estimate from Section 5.7.1 yields a numerical diffusion coefficient of ≈ 120, we choose $\varepsilon = \sqrt{2 \cdot 500}$ here.

Again, we apply the three methods discussed in Section 5.7.1 in order to compute approximate eigenfunctions of the transfer operator resp. the generator.

1. **Ulam's method:** For the simulation of the SDE (2.4) a fourth order Runge–Kutta method is used, where in every time step a properly scaled (by a factor $\sqrt{\tau}\cdot\varepsilon$, where τ is the time step) normally distributed random number is added. We use 1000 sample points per box and the integration time $T = 5\cdot 10^6$, which is realized by 20 steps of the Runge–Kutta method. Note, that the integrator does not know that the flow lines should not cross the lower and upper boundaries of the state space. Points that leave phase space are projected back along the y axis into the next boundary box. An adaptive step–size control could resolve this problem, however at the cost of even more right hand side evaluations. The domain is partitioned into 128×128 boxes.

2. **Ulam's method for the generator.** Again, we employ a partition of 128×128 boxes and approximate the edge integrals by the trapezoidal rule using nine nodes.

3. **Spectral method.** We employ 51 Fourier modes in the x coordinate (periodic boundary conditions) and the first 51 Chebyshev polynomials in the y coordinate, together with Neumann boundary conditions (the two approaches for handling the boundary conditions from Section 5.6.3 do not show significant differences).

Computing almost invariant sets. In Figure 5.6 we compare the approximate eigenvectors at the second, third and fourth relevant eigenvalue of the transfer operator (resp. generator) for the three different methods discussed in the previous sections. Clearly, they all give the same qualitative picture. Yet, the number of vector field evaluations differs significantly, as shown in the following table.

method	# of rhs evals
Ulam's method	$\approx 3{\cdot}10^8$
Ulam's method for the generator	$\approx 3{\cdot}10^5$
Spectral method for the generator	$\approx 3{\cdot}10^3$

Table 5.1: Number of vector field evaluations in order to set up the approximate operator or generator.

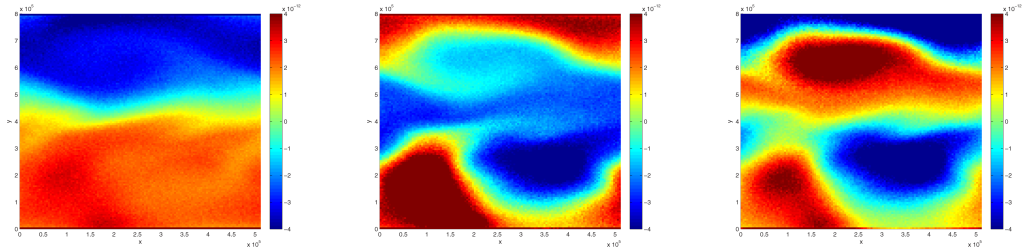
We list the corresponding eigenvalues in the next table. The ones of Ulam's method and the spectral method for the generator match well, while Ulam's method for the generator gives eigenvalues approximately $\frac{6}{5}$ times bigger in magnitude. As estimated above, the numerical diffusion is roughly $\frac{1}{5}$ of the applied artificial diffusion, which explains the difference between the eigenvalues.[1]

method	λ_2	λ_3	λ_4
Ulam's method ($\log(\lambda_i)/T$)	$-1.64{\cdot}10^{-8}$	$-0.91{\cdot}10^{-7}$	$-1.06{\cdot}10^{-7}$
Ulam's method for the generator	$-1.98{\cdot}10^{-8}$	$-1.03{\cdot}10^{-7}$	$-1.19{\cdot}10^{-7}$
spectral method for the generator	$-1.65{\cdot}10^{-8}$	$-0.91{\cdot}10^{-7}$	$-1.05{\cdot}10^{-7}$

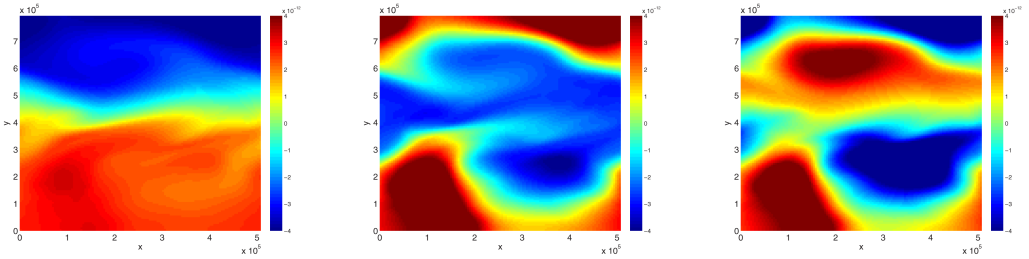
Table 5.2: Approximate eigenvalues.

For illustration, we apply the simplex method [Deu04b], also discussed in Section 2.2.2, on the current example to obtain the four most dominant almost invariant sets. The method is applicable, according to the theory, if the Markov (jump) chain

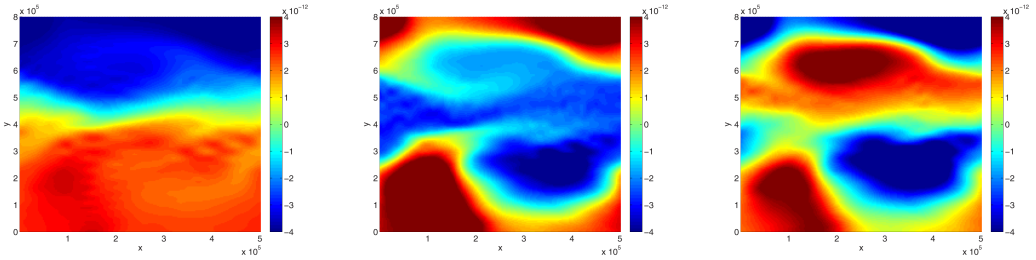
[1] This reasoning assumes that the eigenvalues vary linearly in the diffusion coefficient; cf. Section 5.7.1.



(a) Ulam's method



(b) Ulam's method for the generator



(c) Spectral method for the generator

Figure 5.6: From left to right: Eigenvectors at the second, third and fourth eigenvalue.

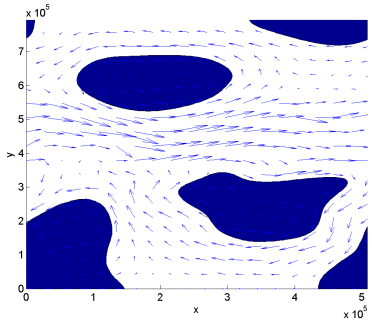


Figure 5.7: Almost invariant sets of the area-preserving flow - the sets most robust under random perturbation are neighborhoods of the steady states of the flow.

generated by the approximative generator is reversible. It is not the case here, however the method seems to work. The (left) eigenfunctions plotted in $\mathbb{R}^3$ do give an object which' convex hull is nearly a simplex. Cutting down the vertices and plotting the corresponding points in the phase space yields the four sets we already expected to be almost invariant, see Figure 5.7.

5.7.3 A volume-preserving three dimensional example: the ABC-flow

We consider the so-called ABC-flow [Arn65], given by

$$\begin{aligned} \dot{x} &= a\sin(2\pi z) + c\cos(2\pi y), \\ \dot{y} &= b\sin(2\pi x) + a\cos(2\pi z), \\ \dot{z} &= c\sin(2\pi y) + b\cos(2\pi x), \end{aligned}$$

on the 3 dimensional torus $\mathbb{T}^3$. The flow is volume-preserving, for $a = \sqrt{3}$, $b = \sqrt{2}$ and $c = 1$, it seems to exhibit complicated dynamics and invariant sets of complicated geometry [Dom86, Fro09].

This example serves to compare the performances of the Ulam type and the spectral type generator methods for a higher dimensional smooth problem. The methods are realized as follows.

1. **Ulam's method for the generator.** A $64 \times 64 \times 64$ box covering is used. To set up the approximative generator, surface integrals have to be computed, see (5.2). Since the vector field is smooth, a 3-by-3-point Gaussian quadrature rule

is used on each box face.[1] The numerical diffusion is estimated to be ≈ 0.013; and we do not add any extra diffusion.

2. **Spectral method.** By the smoothness of the vector field a small number of grid points should suffice to obtain an accurate result. We add extra diffusion $\frac{\varepsilon^2}{2} = 0.013$, and compute the six dominant eigenmodes of the generator obtained by the collocation spectral method on a $11 \times 11 \times 11$ and a $13 \times 13 \times 13$ grid, respectively. The eigenvalues differ by a relative magnitude of 10^{-3}. We deduce from this, that the spectral method approach converges so fast that the 13 grid points per dimension are sufficient.

Error of the invariant density. The ABC-flow is area-preserving, thus the invariant density is the constant one function. Table 5.3 shows the L^1-errors for the methods. Note, that both methods suffer from the curse of dimension; but, since the spectral

method	L^1-error
Ulam's method for the generator	$2{\cdot}10^{-9}$
spectral method for the generator	$8{\cdot}10^{-15}$

Table 5.3: L^1-error of the approximative invariant density of the ABC-flow.

method needs only a few degrees of freedom in each coordinate direction to approximate smooth functions well, the number of vector field evaluations (one per degree of freedom, in fact) stays low. Furthermore, we have the numerical diffusion in the Ulam type generator method, which cannot be controlled. The only way to make it smaller is decreasing the box diameters — and thus increasing the number of boxes and vector field evaluations. For the spectral method approach any kind of diffusion can be simply added artificially.

The disadvantage of the spectral method approach can be seen by looking at the matrix occupancies. The Ulam type method generates a sparse matrix, while the spectral generator gives a full matrix (cf. Section 5.6.2). This could make the eigenvalue problem computationally intractable, if too many basis functions are involved in the

[1] Contrary to v, the function on the box faces, $(v{\cdot}\mathbf{n}_j)^+$, does not have to be smooth, only continuous. Hence any other quadrature rule could perform at least similarly well. However, if the resolution is fine enough, $(v{\cdot}\mathbf{n}_j)^+$ will not change sign on the majority of box faces. Therefore, we expect the Gaussian quadrature rule to be a proper compromise between accuracy and efficiency.

approximation space. However, this problem is not present in this example, because of the small number of grid points. Trying to solve the large eigenvalue problem for the generator discretized by the Ulam type approach, one experiences difficulties. They are thoroughly discussed in Section 5.7.4.

The previous observations are summarized in Table 5.4.

method	# of rhs evals	nonzeros in A_n
Ulam's method for the generator	$\approx 7.1 \cdot 10^6$	$\approx 1.1 \cdot 10^6$
Spectral method for the generator	≈ 2200	$\approx 4.8 \cdot 10^6$

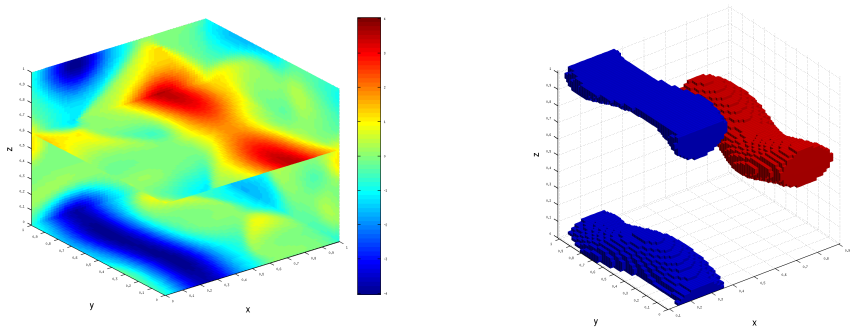
Table 5.4: Number of vector field evaluations in order to set up the approximate generator, and number of nonzeros in it's matrix representation.

Computing and visualizing almost invariant sets. We briefly compare the approximative eigenfunctions at the second dominant eigenvalue for the two methods; and the almost invariant sets extracted from them. To visualize the almost invariant sets we are inspired by the *thresholding* strategy introduced in [Fro03]. For simplicity, instead of finding an optimal threshold, we just heuristically set $c = 0.6 \left\|u_2\right\|_{L^\infty}$ (u_2 is the approximative eigenfunction at the second dominant eigenvalue), and draw the sets $\{u_2 > c\}$ and $\{u_2 < -c\}$. It turns out, that the half of the total mass of $|u_2|$ is supported on the sets shown in Figure 5.8, i.e. they can be seen as the "cores" of the actual almost invariant sets.

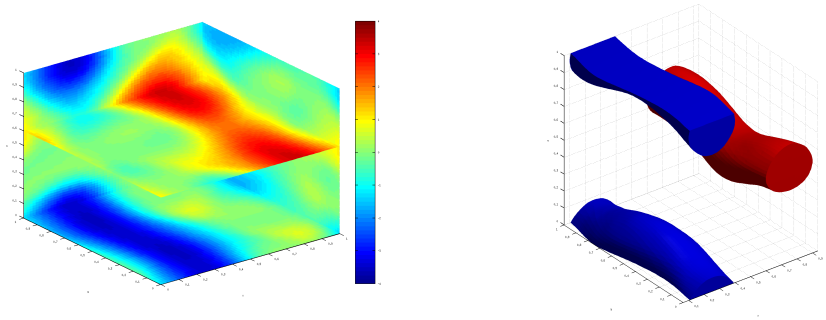
The reader may wish to compare these pictures with computations made with Ulam's method for the ABC-flow [Fro09].

5.7.4 A three dimensional example with complicated geometry: the Lorenz system

As a last example we consider a system where the effective dynamics is supported on a set of complicated geometry – and of not even full dimension. This is the well-known



(a) Ulam's method for the generator



(b) Spectral method for the generator

Figure 5.8: Eigenvector at the second eigenvalue of the approximative generator (left); and the almost invariant sets $\{u_2 > c\}$ (red) and $\{u_2 < -c\}$ (blue), extracted by thresholding (right).

Lorenz system [Lor63]

$$\begin{aligned} \dot{x} &= \sigma(y-x), \\ \dot{y} &= x(\varrho - z) - y, \\ \dot{z} &= xy - \beta z, \end{aligned}$$

with $\sigma = 10$, $\varrho = 28$ and $\beta = 8/3$. The effective dynamics happen on the attractor – a set of complicated geometry. Eigenfunctions of the transfer operator may be supported on the attractor, hence we do not expect the spectral approach to work well. Thus, we use the Ulam type approach for the generator.

A decade ago numerical techniques have been constructed to compute box coverings for attractors of complicated structure [Del97, Del96, Del98]. These techniques exploit the fact, that the set X to be computed is an attractor, hence each trajectory starting in its vicinity will be pulled to X in a fairly short time. In our approach time is not considered, we use only movement directions, *speed vectors*. Since the boundary of the box covering does not have to coincide with the boundary of the attractor, a tight box covering might not show desired results, because relatively big outflow rates in boundary boxes could occur. The simplest idea is (as discussed in Section 5.5.1) to use a rectangle big enough – in our case $[-30, 30] \times [-30, 30] \times [-10, 70]$.

This rectangle is not a forward invariant set, hence we may have outflow on its boundary. If it's so, there will be no invariant density, just an "almost invariant" one, corresponding to a negative eigenvalue close to zero.

Now we are ready to compute the discretized generator, and its left and right eigenvectors. We use a $128 \times 128 \times 128$ box covering. The attractor is then extracted by simple thresholding of the approximative (almost) invariant density u_1: where u_1 is strictly away from zero, we expect to have a small neighborhood of the attractor[1]. As one would expect from the presence of numerical diffusion, outside of the attractor u_1 drops exponentially. We cut off the invariant density at a threshold value $c = 5 \cdot 10^{-6}$, such that 96% of its mass is supported on $\{u_1 > c\}$. Having the attractor, we may restrict the other eigenfunctions on this set to yield the almost invariant sets in the attractor, see Figure 5.9.

[1]The finer the resolution, the smaller the diffusion introduced by the discretization, the tighter this neighborhood of the attractor is.

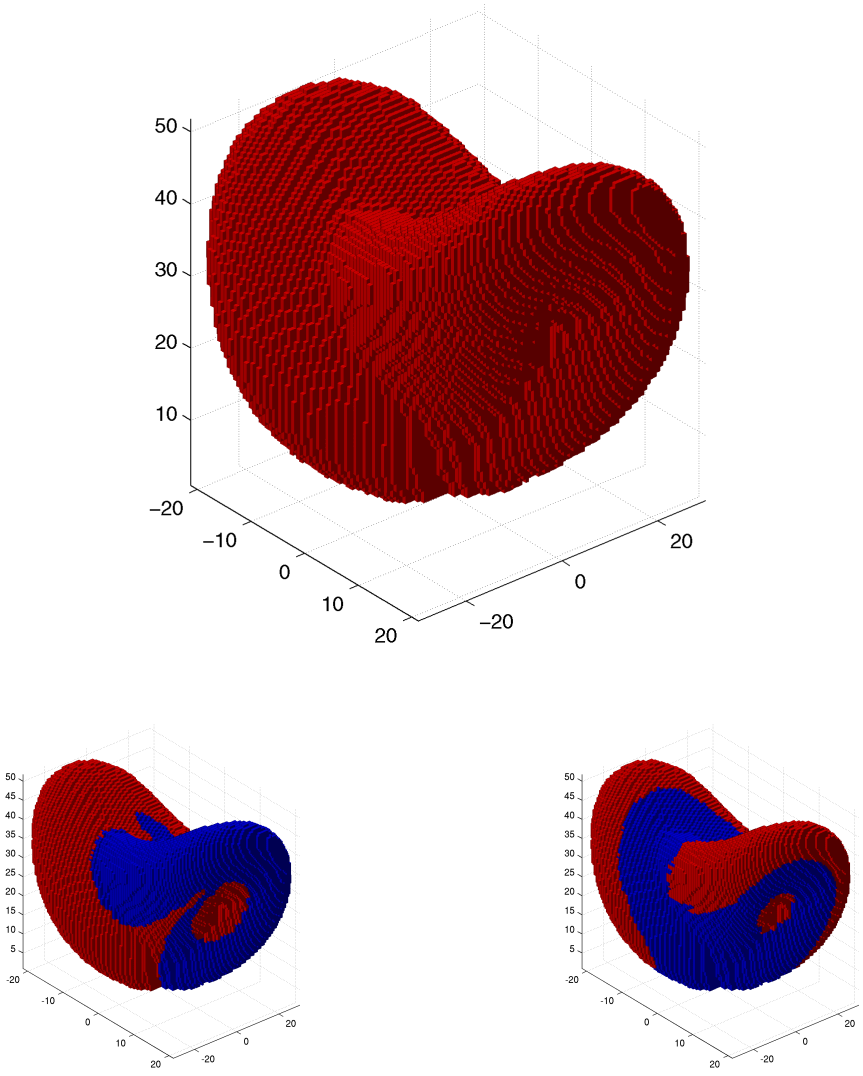


Figure 5.9: Approximation of the Lorenz attractor - and of almost invariant sets in it. The bottom graphs show the sign structure (red and blue) of the second and third left eigenfunction, respectively. The computation has been done on a uniform covering with 128 boxes in each direction.

Solving the large eigenvalue problem. Our computations have been done in Matlab, and we used the built-in solver `eigs` for solving the eigenvalue problem for the matrix A. It is an Arnoldi type iterative solver. We are interested in the eigenvalues with the smallest magnitude; these are computed by *backward* iteration. Hence, in each step a system of linear equations of the form $Ax = b$ has to be solved. Already for a resolution $64 \times 64 \times 64$ the matrix A is too big to compute a sparse LU - decomposition, that is just what `eigs` tries to do.

Consequently, we have to provide a program computing $A^{-1}b$ for an input vector b. Unfortunately A is not symmetric and the CG method is not applicable. We chose to use the GMRES method. This did not converge for random initial vectors and we were lead to the problem of finding good starting vectors.

Our strategy to obtain proper initial vectors is inspired by multigrid methods. The matrix A stems from a discretization of the operator $\mathcal{A}$ by the Ulam type approach. Take a coarser box partition (for example merge 2 boxes in each dimension to one big box — as we were doing it) and compute the matrix A_1 arising from the corresponding operator discretization. Project the vector b onto this coarse partition to obtain b_1. Compute $x_1 = A_1^{-1}b_1$ and embed it back to the fine partition (this can be done easily if the fine partition is obtained from the coarse one by subdividing boxes). So, if we can obtain a numerical solution to $A_1x_1 = b_1$, then we use the embedding of x_1 as initial vector for the GMRES iteration, if not, we apply the same strategy to the problem $A_1x_1 = b_1$, and so on. No later than the problem $A_kx_k = b_k$ is small enough to obtain a solution by direct LU - decomposition, we have a starting vector for the $(k-1)$st "inner" iteration. In other words, each problem $A_kx_k = b_k$ provides a starting vector for the GMRES iteration to solve the problem $A_{k-1}x_{k-1} = b_{k-1}$. In the end, we expect to get a numerical solution of $Ax = b$.

Of course, it is undesirable to compute all the coarser discretizations $A_1, A_2, \ldots$ just as we computed A (especially if the vector field v is expensive to evaluate). Fortunately, we may compute them directly by linearcombining entries of A at linear complexity. Denoting X_i^k the elements of the kth partition, and

$$\mathcal{I}_{ij}^k := \left\{ (i_\ell, j_\ell) \middle| X_{i_\ell}^{k-1} \subset X_i^k, X_{j_\ell}^{k-1} \subset X_j^k \text{ and } m_{d-1}(\partial X_{i_\ell}^{k-1} \cap \partial X_{j_\ell}^{k-1}) \neq 0 \right\},$$

(5.2) gives

$$A_{k,ij} = \frac{1}{m(X_i^k)} \sum_{(i_\ell, j_\ell) \in \mathcal{I}_{ij}^k} m(X_{i_\ell}^{k-1}) A_{k-1,i_\ell j_\ell}.$$

Figure 5.10 visualizes this.

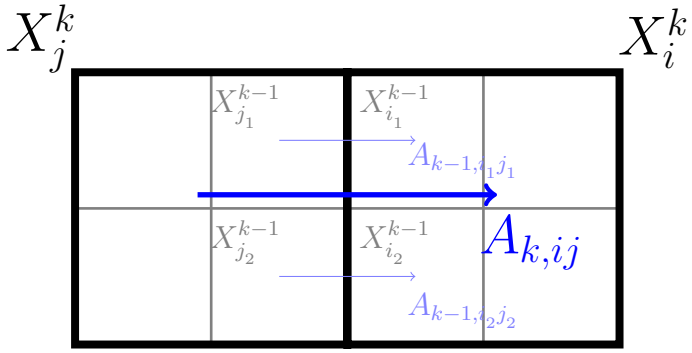


Figure 5.10: Collapsing of A - obtaining A_k from A_{k-1}.

The reader may be confused, that in order to compute eigenvectors at eigenvalues *zero* or *near zero* we try to solve problems $x = A^{-1}b$. These are of course ill conditioned, and despite strategies like above, the GMRES method (or the backward iteration itself) may not converge. In all these cases the following shifting strategy solved our problems. Take $\mu \approx |\lambda_2|$ and work with the matrix $(A - \mu I)$ instead of A. This is merely a shift of the eigenvalues; the eigenvectors stay unchanged. The spectrum of A is expected to lie in the left complex half plane, hence $(A - \mu I)$ is non-singular, and $(\lambda_1 - \mu)^{-1}, (\lambda_2 - \mu)^{-1}, \ldots$ are the dominant eigenvalues of $(A - \mu I)^{-1}$.

5.7.5 Computing the domain of attraction without trajectory simulation

To close the sequence of examples we demonstrate the usability of the infinitesimal generator approach to compute the domain of attraction of an asymptotically stable fixed point. The following system is fully artificial, made for the purpose of yielding complicated dynamics and an asymptotically stable fixed point (the origin) with bounded domain of attraction.[1]

$$\begin{aligned} \dot{x} &= (3x^2 + 3y^2)(3y^2 - 50y^4 + 2y + x) - y - 2x + 3x^2, \\ \dot{y} &= (3x^2 + 3y^2)(2x - 3x^2 + y) - (2y + 1)(3y^2 - 50y^4 + 2y + x). \end{aligned} \tag{5.22}$$

[1] I am grateful to Alexander Volf, who inspired the application of the infinitesimal generator in order to compute domains of attraction. Also, the system analyzed here is due to him.

The idea of using transition probabilities to compute the domain of attraction is exploited in [Gol04]. A different approach also for cell-to-cell mappings is shown in [Hsu87].

Consider a dynamical system governed by a SDE. We denote the solution random variable of the SDE by $X(t)$. Define an absorbing state x_0 (i.e. $X(t) = x_0$ implies $X(s) = x_0$ for all $s > t$), and the absorption probability function (APF) $p(x) := \text{Prob}\big(X(t) = x_0 \text{ for a } t > 0 \mid X(0) = x\big)$. For a fixed $t > 0$, let $q^t(x, \cdot)$ denote the density of $X(t)$, provided $X(0) = x$. Then it holds $\int q^t(x, y)p(y)\,\mathrm{d}y = p(x)$ for all x and all $t \geq 0$. In other words: $\mathcal{U}^t p = p$, the APF is the fixed point of the Koopman operator. Denoting the infinitesimal generator of $\mathcal{U}^t$ by $\mathcal{A}^*$, we have $\mathcal{A}^* p = 0$.

If the dynamical system is deterministic (i.e. $\varepsilon = 0$), p is 1 in the domain of attraction of x_0 and 0 outside of it. From an applicational point of view, mostly this case is of interest. Hence, we have to approximate nearly characteristic functions of a set of possibly complicated geometry. Therefore, the spectral method approach is not expected to work well (numerical experiments, not discussed here, confirm this). However, the Ulam type generator method turns out to perform properly. Define an analogous discretization of $\mathcal{U}^t$ as of $\mathcal{P}^t$:

$$\mathcal{A}_n^* f := \lim_{t \to 0} \pi_n \frac{\mathcal{U}^t \pi_n f - \pi_n f}{t}. \tag{5.23}$$

If we compute the approximate generator of the FPO, we have the approximate generator of $\mathcal{U}^t$ as well:

Proposition 5.43. *The operator $\mathcal{A}_n^*$ is the adjoint of $\mathcal{A}_n$.*

Proof. Deriving the entries of the matrix representation of $\mathcal{A}_n^*$ involve entirely the same computations as deriving the matrix representation of $\mathcal{A}_n|_{V_n}$. Using the adjointness of $\mathcal{P}^t$ and $\mathcal{U}^t$,

$$\int_{X_j} \mathcal{U}^t \chi_i = \int \chi_j \mathcal{U}^t \chi_i = \int \mathcal{P}^t \chi_j \chi_i = \int_{X_i} \mathcal{P}^t \chi_j,$$

the claim follows. □

Thus, if we are given a matrix representation A_n of $\mathcal{A}_n$, the *left* eigenvector (normed to one in the ∞-norm) at the eigenvalue 0 gives us the approximative absorption probabilities. We expect these values to be 1 in the interior of the domain of attraction, 0 outside, and between 0 and 1 near its boundary. This is due to the discretization,

which introduces numerical diffusion, that can be viewed as uncertainty in the dynamics: near the boundary there is a considerable probability, that trajectories starting in the domain of attraction, but near its boundary, do not tend to the absorbing state.

Figure 5.11 shows the left eigenvector to eigenvalue 0 of the Ulam type generator approximated on a 1001×1001 box covering of $[-1, 1]^2$. Note the regions along the boundary, where the absorption probabilities do not fall so steep. These may indicate

(a) trajectories which run a long way along the boundary before attracted to the origin, so that the diffusion has "much time" to drag trajectories out of the domain; or

(b) strong drift (large vector field values), which implies a big numerical diffusion.

We remark, that if not even the rough location of the domain of attraction is known, one may get a bound by making a coarse computation on a larger domain, and iterate this process on more and more tight approximate regions.

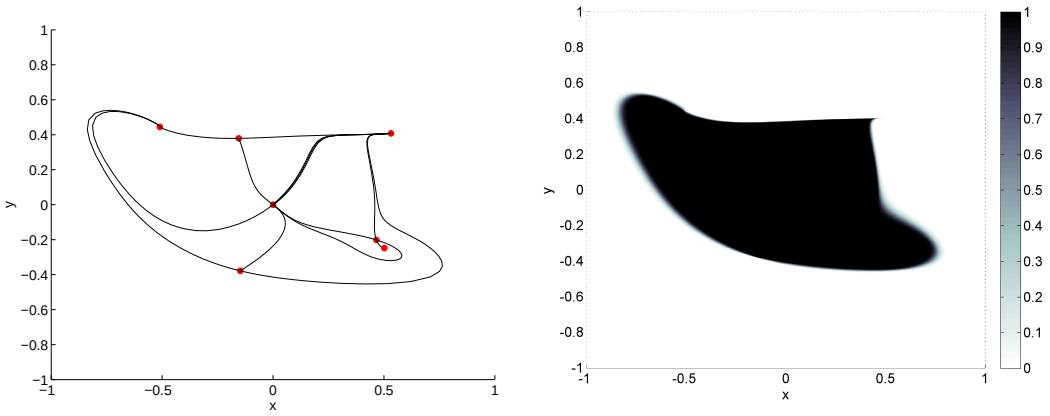


Figure 5.11: The approximate domain of attraction of the origin - fixed points and some connecting orbits (left); and the left eigenvector of the Ulam type approximate generator on a 1001×1001 box covering.

5.8 Conclusions and outlook

In this chapter we developed and extensively analyzed two numerical methods for the discretization of the infinitesimal generator of the semigroup $\mathcal{P}^t$ of Frobenius–Perron operators. The main benefit is that the expensive numerical integration, which is involved in any approximation of $\mathcal{P}^t$, can be avoided. Also, there is an "optimal"

exhaustion of the computed information, in the sense, that every evaluation of the vector field goes directly into the approximation. Meanwhile, e.g. the discretization of $\mathcal{P}^t$ by Ulam's method uses only the endpoints of the simulated trajectories, and does not consider the former points of the trajectory which were computed by the time integrator on the way getting to the endpoint.

The first method, the Ulam type approach for the infinitesimal generator, turned out to be the known upwind scheme from finite volume methods. An analysis by operator semigroup theory showed that it is an adequate approximation, even if the set of interesting dynamical behavior is a subset of $\mathbb{R}^d$ with complicated geometry. We believe, that the robustness of the method is strongly connected with the numerical diffusion arising by the discretization (just as numerical diffusion stabilizes the upwind scheme). However, the significance of this concept for our purposes is not perfectly understood yet. A drawback is, that we cannot "turn off" this diffusion; it is always present, we can only decrease it under a desired threshold by making the box sizes smaller. Nevertheless, the size of the numerical diffusion is the same as the magnitude of the phase space resolution (see Section 5.7.1), therefore, if one would like to resolve the spatial behavior further, one would have to increase the resolution anyway.
Convergence of the eigenfunctions (or at least of the invariant density) is still an open question (cf. Section 5.4.2). It would be desirable to understand as well, why the congruency of boxes is so important for the convergence of the generator (see Lemma 5.17 and the remark afterwards); and if there is an approximation which converges even for general box coverings?

The second method, the spectral method approach for the infinitesimal generator, can be proven to have the spectral convergence speed — at least for the Galerkin method. All our examples were computed to a sufficient accuracy by the collocation method, but there could be systems, in particularly in higher space dimensions, where the full occupancy of the matrix representation of the discretized operator sets computational limits. There, the Galerkin method should be applied.
We can exploit the full power of spectral methods only in spaces which are tensor products of intervals. On spaces with more complicated geometry so-called spectral elements (also called hp finite elements) could be used.

Note, that the discretization for both methods can be written as $\mathcal{A}_n^{(\varepsilon)} = \varepsilon\Delta_n + \mathcal{A}_n$, where $\mathcal{A}_n$ is the discretization with $\varepsilon = 0$. In order to study the properties of the

system for different values of ε, the discretized operator $\mathcal{A}_n$ has to be assembled only *once*. If we would discretize the transfer operator by Ulam's method, we would have to set up the transition matrix every time anew, since a different SDE (2.4) has to be integrated.

Chapter 6

Mean field approximation for marginals of invariant densities

6.1 Motivation

Every time we have to compute the macroscopic behavior of dynamical systems with high phase space dimension, by using transfer operator methods we run into difficulties. Unless we can exploit some dynamical structure to reduce the problem dimension (e.g. there are slow and fast variables [Pav08], or the attractor has a smaller fractal dimension [Del96, Del97]), or we can use adaptivity to find a partition we can still deal with, the curse of dimension puts these problems beyond the limits of current numerical methods. *General* approaches, like the one introduced in Chapter 4, allow us to access a few dimensions more, but the computational treatment of molecules with a few hundred atoms is still way out of reach for these.[1]

We abandon generality, and turn our attention to more specific systems. We assume that the dynamical system consists of subsystems, each acting on a low-dimensional space. Moreover, each subsystem interacts strongly only with a few other subsystems, and its interaction with the other ones is negligible or very weak; where it is always to specify what "weak" means. Furthermore, we will be only interested in the evolution (resp. long-term behavior) of some particular subsystems. Until now, one would have had to analyze the whole system to draw, in the end, the desired (*reduced* or *marginal*)

[1]Note, that in the context of conformation dynamics, *special* transfer operator based techniques have successfully been developed, see the references in Section 2.4.1.

information of the subsystem. Our aim in this chapter is to define proper reduced systems (on a low dimensional phase space) which give good approximations on the statistical behavior of the marginal system. Furthermore, we wish to use them for numerical computations, since these systems on low dimensional spaces are accessible via transfer operator methods.

To include the influence of interacting subsystems into the dynamics of the subsystem under consideration we use *mean field* theory. Here, one averages the action of the surrounding interacting subsystems w.r.t. appropriate distributions. The idea is not new, it has been successfully applied in many fields, e.g. in quantum chemistry in the Hartree–Fock theory of many-particle Schrödinger equations [Har28, Foc30].

Our guiding examples are coupled map lattices and molecular dynamical (MD) systems for chain molecules. First, the mean field theory for coupled maps is introduced in Section 6.2, where we concentrate on asymptotic results in dependence of the coupling strength. Second, we apply the methodology on MD systems in Section 6.3, and test it on the example of n-butane. While the results for the latter problem look promising, there are several important questions, to be discussed in the future:

- How to extend the method for larger molecules?
- Under which assumptions does the method work for large molecules?

They are topic of ongoing work, hence our answers can only be founded conjectures. The reader may also find, that this chapter is of highly experimental nature. Indeed, the behavior we analyzed elucidate only some aspects of mean field approximation of coupled dynamical systems. There are still many more interesting questions to ask.

6.2 Mean field for maps

6.2.1 Nondeterministic mean field

Let X and Y be compact spaces measurable with the Lebesgue measure m. Define the *full* system by $S : X \times Y \to X \times Y$, $S(x,y) = \big(S_1(x,y), S_2(x,y)\big)^\top$, where S is nonsingular and $S_i(\cdot, y)$ resp. $S_i(x,\cdot)$ are nonsingular[1] for $i = 1,2$ and for all $x \in X$, $y \in Y$. The transfer operator associated with S is denoted by $\mathcal{P}$. Although we restrict

[1] $T : X \to Y$ is nonsingular, if for all measurable $A \subset Y$ with $m(A) = 0$ we have $m(T^{-1}(A)) = 0$

our considerations on two subsystems, it is straightforward to generalize everything for an arbitrary number of subspaces.

Assume, that the full system has an invariant density. Let x be the variable of interest. We would like to characterize its long-term behavior, hence we search for the marginal of the invariant density w.r.t. x. How does x evolve, if the system is distributed according to its invariant density? Then, $\mathbf{y}$ is a random variable with a distribution depending on x itself, and x is mapped to a random variable $\mathbf{x} = S_1(x, \mathbf{y})$. Since we started with the invariant distribution, we expect (without justification, for now) $\mathbf{x}$ to be distributed nearly according to the x-marginal of the invariant density. As a further approximation step, we assume the subsystems being "sufficiently independent", such that the distribution of $\mathbf{y}$ can be well approximated by the density $u_2 \in L^1(Y)$, independent of x. Then we can look at u_2 as (an approximation) to the y-marginal of the invariant density. Now, we may define the approximate evolution of the x variable, given that the full system is in "equilibrium", i.e. it is distributed according to its invariant density. We call it the *mean field dynamics* of the x variable (or x-subsystem).

$$\mathbf{x}_{k+1} = S_1(\mathbf{x}_k, \mathbf{y}), \tag{6.1}$$

where $\mathbf{y}$ is distributed according to u_2. Let $p_{1,\mathrm{mf}}[u_2](\cdot,\cdot)$ be the transition function associated with this system, i.e.

$$p_{1,\mathrm{mf}}[u_2](x, A) = \int \chi_A\left(S_1(x, y)\right) u_2(y)\,\mathrm{d}y = \int_{\left\{y | S_1(x,y)\in A\right\}} u_2(y)\,\mathrm{d}y, \tag{6.2}$$

for all measurable $A \subset X$. By the non-singularity of $S_{1,x}$ and the Radon–Nikodym theorem, $p_{1,\mathrm{mf}}[u_2]$ has a transition density function as well; cf. Definition 2.1. In order to obtain it, we introduce a formal FPO $\mathcal{P}_{1,x} : L^1(Y) \to L^1(X)$ associated with the function $S_{1,x} := S_1(x,\cdot) : Y \to X$, by $\int_A \mathcal{P}_{1,x} f = \int_{S_{1,x}^{-1}(A)} f$[1]. The operator is well defined, since $S_{1,x}$ is nonsingular. We get

$$p_{1,\mathrm{mf}}[u_2](x, A) = \int_A \mathcal{P}_{1,x} u_2(z)\,\mathrm{d}z. \tag{6.3}$$

In other words, $q_{1,\mathrm{mf}}[u_2](x, z) = \mathcal{P}_{1,x} u_2(z)$ is the transition density function of the system (6.1).

[1]Note, that the first integral is over $A \subset X$, and the second one over $S_{1,x}^{-1}(A) \subset Y$.

The mean field system. One can, of course, do the same derivation, but with the aim to describe the evolution of the y variable. Then, one would fix a u_1 representing the distribution of the the random variable $\mathbf{x}$, and $S_2(\mathbf{x}, \cdot)$ defines the mean field dynamics of the y variable. So, even if the system is not in equilibrium, i.e. u_1 and u_2 do not necessary represent marginals of the invariant density, one can define a coupled system on X and Y — the mean field *system*, — by

$$\begin{array}{rcl} \mathbf{x}_{k+1} & = & S_1(\mathbf{x}_k, \mathbf{y}), \\ \mathbf{y}_{k+1} & = & S_2(\mathbf{x}, \mathbf{y}_k), \end{array} \tag{6.4}$$

where $\mathbf{x}$ (resp. $\mathbf{y}$) is a random variable independent of $\mathbf{x_k}$ and $\mathbf{y_k}$, having the same distribution as $\mathbf{x}_k$ (resp. $\mathbf{y}_k$).

The associated transfer operator. Let $\mathcal{P}_{1,\mathrm{mf}}[u_2]$ denote the FPO associated with $p_{1,\mathrm{mf}}[u_2](\cdot,\cdot)$. An explicit representation of $\mathcal{P}_{1,\mathrm{mf}}[u_2]$ can be given by (2.11), and the transition density above.

For $u_1 \in L^1(X)$ and an arbitrary measurable $A \subset X$ we have

$$\begin{array}{rcl} \displaystyle\int_A \mathcal{P}_{1,\mathrm{mf}}[u_2]u_1(x)\,\mathrm{d}x & = & \displaystyle\int_X u_1(x) p_{1,\mathrm{mf}}[u_2](x, A)\,\mathrm{d}x \\ & = & \displaystyle\int_X \int_{\{y|S_1(x,y)\in A\}} u_1(x)u_2(y)\,\mathrm{d}y\mathrm{d}x \\ & = & \displaystyle\iint_{\{(x,y)^\top|S_1(x,y)\in A\}} u_1(x)u_2(y)\,\mathrm{d}(x,y). \end{array} \tag{6.5}$$

Note, that the integration domain is actually $S^{-1}(A \times Y)$, but this depends only on S_1.

For comparison, we compute the marginal of a density $u \in L^1(X \times Y)$ iterated by $\mathcal{P}$ (integrated over $A \subset X$, just as above):

$$\begin{array}{rcl} \displaystyle\int_A \int_Y (\mathcal{P}u)(x,y)\,\mathrm{d}(x,y) & = & \displaystyle\iint_{S^{-1}(A\times Y)} u(x,y)\,\mathrm{d}(x,y) \\ & = & \displaystyle\iint_{S_1^{-1}(A)} u(x,y)\,\mathrm{d}(x,y), \end{array}$$

which is exactly (6.5) for all measurable sets $A \subset X$, if $u(x,y) = u_1(x)u_2(y)$. We have proven:

Proposition 6.1. *Let the full density $u \in L^1(X \times Y)$ be separable, i.e. $u = u_1 \otimes u_2$. Then the nondeterministic mean field system* (6.4) *describes the exact one-step evolution of the distributions of the subsystems, i.e.*

$$\mathcal{P}_{1,\mathrm{mf}}[u_2]u_1 = \int_Y \mathcal{P}(u_1 u_2) \quad \textit{and} \quad \mathcal{P}_{2,\mathrm{mf}}[u_1]u_2 = \int_X \mathcal{P}(u_1 u_2). \tag{6.6}$$

Moreover, if the invariant density of the full system is separable, the marginals are invariant under the respective mean field subsystem dynamics, i.e.

$$\mathcal{P}_{1,\mathrm{mf}}[u_2]u_1 = u_1 \quad \textit{and} \quad \mathcal{P}_{2,\mathrm{mf}}[u_1]u_2 = u_2.$$

The marginal of $\mathcal{P}u$. We derive here an expression for the marginal(s) of $\mathcal{P}u$, where $u \in L^1(X \times Y)$, not necessary separable. It will be useful in a later section, where we analyze the mean field model for weakly coupled systems. Also, we get a second explicit representation of the mean field transfer operator.

Lemma 6.2. *The marginal density of $\mathcal{P}u$ can be written as*

$$\int_Y \mathcal{P}u(x,y)\,\mathrm{d}y = \int_Y \left(\mathcal{P}_{1,y}u_y\right)(x)\,\mathrm{d}y, \tag{6.7}$$

where $u_y(x) = u(x,y)$, and $\mathcal{P}_{1,y}$ is the transfer operator associated with $S_1(\cdot,y)$. By (6.6) *we also get*

$$\mathcal{P}_{1,\mathrm{mf}}[u_2]u_1 = \int_Y \left(\mathcal{P}_{1,y}u_1\right)\ u_2(y)\,\mathrm{d}y. \tag{6.8}$$

One can derive analogously formulas for the y-marginal and the corresponding mean field transfer operator.

Proof. The idea for obtaining a representation formula is to split the integral below on integration over fibers:

$$\begin{aligned}
\int_A\int_Y \mathcal{P}u(x,y)\,\mathrm{d}(x,y) &= \int_{S_1^{-1}(A)} u(x,y)\,\mathrm{d}(x,y)\\
&= \int_Y \int_{S_{1,y}^{-1}(A)} u_y(x)\,\mathrm{d}x\mathrm{d}y\\
&= \int_Y \int_A \left(\mathcal{P}_{1,y}u_y\right)(x)\,\mathrm{d}x\mathrm{d}y\\
&\overset{\text{Fubini}}{=} \int_A \int_Y \left(\mathcal{P}_{1,y}u_y\right)(x)\,\mathrm{d}y\mathrm{d}x,
\end{aligned}$$

Since this holds for every measurable $A \subset X$, the proof is complete. □

6.2.2 Deterministic mean field

In cases, such as

- the y variable evolves much faster than the x variable (i.e. $\left|S_1(\cdot,y) - \mathrm{Id}_x\right|/\ell_x \ll \left|S_2(x,\cdot) - \mathrm{Id}_y\right|/\ell_y$ for all x, y, where ℓ_x and ℓ_y are typical length scales of the x and y variables, respectively), or

- the variance of $S_1(x, \mathbf{y})$ is small independently of the distribution of $\mathbf{y}$, and the variance of $S_2(\mathbf{x}, y)$ is small independently of the distribution of $\mathbf{x}$,

it is well-founded to approximate the non-deterministic mean field system with a deterministic one, just by setting the image of a point as the expectation value of the image random variable.[1]

Definition 6.3 (Deterministic mean field)**.** *The deterministic mean field*[2] *system is given by*[3]

$$\begin{array}{rclcrcl} x_{k+1} & = & S_{1,\mathrm{MF}}[u_{2,k}](x_k) & := & \mathbb{E}_{u_{2,k}}\left(S_1\left(x_k, \mathbf{y}\right)\right) & = & \int_Y S_1(x,y)u_{2,k}(y)\,\mathrm{d}y, \\ y_{k+1} & = & S_{2,\mathrm{MF}}[u_{1,k}](y_k) & := & \mathbb{E}_{u_{1,k}}\left(S_2\left(\mathbf{x}, y_k\right)\right) & = & \int_X S_2(x,y)u_{1,k}(x)\,\mathrm{d}x, \end{array} \tag{6.9}$$

where for $i = 1, 2$ the $u_{i,0}$ are given initial densities, $u_{i,k+1} = \mathcal{P}_{i,\mathrm{MF}}[u_{i^c,k}]u_{i,k}$, and $\mathcal{P}_{i,\mathrm{MF}}[u_{i^c,k}]$ is the FPO associated with $S_{i,\mathrm{MF}}[u_{i^c,k}]$ (i^c is the complement of i, i.e. $\{i, i^c\} = \{1, 2\}$).

6.2.3 Numerical computation with the mean field system

In order to be able to work with the mean field system, we introduce an Ulam type discretization, cf. Section 2.3. The densities $u_{i,k}$, $i = 1, 2$ and $k \in \mathbb{N}$, are approximated by the piecewise constant functions $u_{n,i,k} \in V_{n,i}$, $V_{n,i}$ being the approximation space associated with the partition of X (if $i = 1$), resp. of Y (if $i = 2$).

Iterating the mean field system. Following algorithms approximate the iterates of the mean field systems (6.4) and (6.9).

Algorithm 6.4 (Iterating the non-deterministic mean field system)**.** Let the initial densities $u_{n,1,0} \in V_{n,1}$ and $u_{n,2,0} \in V_{n,2}$ be given. For $k = 0, 1, \ldots$ we compute:
System samples. We sample the transition function $p_{1,\mathrm{mf}}[u_{n,2,k}](x, \cdot)$ for a given $x \in X$ by

1. drawing a random sample $y \in Y$ according to the distribution $u_{n,2,k}$, and

[1] The first case is a discrete-time analogon to "averaging", see [Pav08]. The x variable barely changes, meanwhile the y variable already samples its invariant density. In the second case, the dynamics resemble deterministic movement under a small random perturbation.

[2] To emphasize the difference between stochastic and deterministic mean field, we indicate the former with "mf", and latter with "MF".

[3] We denote the expectation value of the random variable $\mathbf{y}$ with density u by $\mathbb{E}_u(\mathbf{y})$.

2. then computing $S_1(x, y)$.

The transition function $p_{2,\mathrm{mf}}[u_{n,1,k}](y, \cdot)$ is sampled in the same fashion.
Discretized transfer operator. We set up the transition matrices $P_{n,i,\mathrm{mf}}[u_{n,i^c,k}]$ (matrix representations of the discretized transfer operators $\mathcal{P}_{n,i,\mathrm{mf}}[u_{n,i^c,k}]$) by (2.19). The images of the sample points are computed by the two-step system sampling from above.
Next iterates. Now we can sample $\mathbf{x}_{k+1}$ and $\mathbf{y}_{k+1}$, if we want to. Their distributions are approximated by $u_{n,i,k+1} := \mathcal{P}_{n,i,\mathrm{mf}}[u_{n,i^c,k}]u_{n,i,k}$.

Algorithm 6.5 (Iterating the deterministic mean field system)**.** Let the initial densities $u_{n,1,0} \in V_{n,1}$ and $u_{n,2,0} \in V_{n,2}$, as well as initial points $x_0 \in X$ and $y_0 \in Y$ be given. For $k = 0, 1, \ldots$ we compute:
Next iterates. The iterate

$$x_{k+1} = \int_Y S_1(x_k, y) u_{n,2,k}(y) \mathrm{d}y$$

is computed by numerical quadrature. If we expect the box resolution to be high enough, that the function $S_1(x, \cdot)$ does not vary strongly in a box, one map evaluation per box is sufficient. The iterate y_{k+1} is computed analogously.
Discretized transfer operator. We have just discussed, how the map $S_{i,\mathrm{MF}}[u_{n,i^c,k}]$ is evaluated. The corresponding transition matrix $P_{n,i,\mathrm{MF}}[u_{n,i^c,k}]$ is computed with (2.19). The new densities are obtained by $u_{n,i,k+1} := \mathcal{P}_{n,i,\mathrm{MF}}[u_{n,i^c,k}]u_{n,i,k}$.

Approximating marginals. If we expect the mean field system to approximate the dynamics of the subsystems qualitatively well, it is a natural choice to define the *mean field invariant marginal densities* as the pair (u_1, u_2) satisfying

$$\begin{array}{rcl} u_1 & = & \mathcal{P}_{1,\mathrm{mf}}[u_2]u_1, \\ u_2 & = & \mathcal{P}_{2,\mathrm{mf}}[u_1]u_2; \end{array} \tag{6.10}$$

analogously for the deterministic mean field system. It is a nonlinearly coupled eigenvalue problem. For its solution we propose to use a procedure which is inspired by the so-called Roothaan algorithm from quantum chemistry.

Algorithm 6.6 (Roothaan iteration)**.** Let $u^0_{n,1} \in V_{n,1}$ and $u^0_{n,2} \in V_{n,2}$ be initial (approximative) guesses for the invariant marginals.
By alternating i (or running through the subsystems cyclically, if there are more than two) we compute the density $u^{k+1}_{n,i}$ from

$$u^{k+1}_{n,i} = \mathcal{P}_{n,i,\mathrm{mf}}\left[u^{k^*}_{n,i^c}\right] u^{k+1}_{n,i},$$

where k^* is the largest index $u_{n,i^c}^{k^*}$ is already defined for.
End the iteration, if (6.10) is satisfied to a desired accuracy, or if no further improvement is observed.

Once the approximative invariant marginals $u_{n,1}$ and $u_{n,2}$ are obtained, we can use the operators $\mathcal{P}_{n,i,\mathrm{mf}}[u_{n,i^c}]$, $i = 1, 2$, to detect almost invariant structures in the subsystems. We simply compute their eigenmodes with eigenvalues near one, and proceed as in Section 2.2.2. By this, we reveal almost invariant structures under the assumption, that the surrounding subsystems are distributed according to their (marginal) invariant densities.

Complexity. For simplicity, assume that X and Y are full dimensional rectangular subsets of d_1 and d_2 dimensional spaces, respectively. Let them be partitioned by a uniform box covering, consisting of n boxes in each dimension. Hence, $\dim(V_{n,i}) = n^{d_i}$. To evaluate the deterministic mean field system, we have to compute transition matrices over a space of dimension d_i, which is done by Ulam's method in $\#\mathrm{flops}(S_{i,\mathrm{MF}}) \cdot \mathcal{O}\left(n^{d_i}\right)$ flops. However, one evaluation of the the mean field subsystem $S_{i,\mathrm{MF}}$ needs $\mathcal{O}\left(n^{d_{i^c}}\right)$ flops, because of the involved numerical quadrature. Overall, the $\mathcal{O}\left(n^{d_1+d_2}\right)$ costs are at the same order of magnitude as if we were applying Ulam's method for the full system with the tensor product partition, resulting in the approximation space $V_{n,1} \otimes V_{n,2}$. For the non-deterministic mean field system we may decide, how many sample points per box are needed. However, in order to get a good approximation on the distribution of $p_{1,\mathrm{mf}}[u_{n,2}](x, \cdot)$, we need to sample $u_{n,2}$ properly, i.e. the whole space Y has to be sampled. This results in an at least as large complexity, as before.

For completely coupled systems, until now the only gain of applying the mean field methods onto the system, is that the transfer operators involved are of smaller dimensions, since $n^{d_i} \ll n^{d_1+d_2}$. Their storage and any computation with them is of much less effort. Nevertheless, their assembly involves numerical costs of $\mathcal{O}\left(n^{d_1+d_2}\right)$.

We expect mean field to show a real advantage in the case where more subsystems are involved, but each one of them interacts strongly (directly) only with a few others. Then, weak interactions could be neglected, and computations on *one* subsystem are of complexity of computations made on a group of strongly interacting subsystems. Nonetheless, if we choose systems i and j, respectively j and k to be directly coupled in our model, then there is an indirect coupling between the systems i and k. In order

to include the effect of this indirect coupling in the computations, iterative algorithms have to be used, like the Roothaan iteration.

6.2.4 Numerical examples

Fast convergence of the approximative marginals. This example is inspired by coupled map lattices. We consider the approximation error of the mean field invariant density for a vanishing coupling strength. Let two maps on the unit interval be given by

$$S_1(x) = \begin{cases} \frac{2x}{1-x}, & \text{if } x < 1/3, \\ \frac{1-x}{2x}, & \text{otherwise,} \end{cases} \quad \text{and} \quad S_2(x) = \begin{cases} 2x, & \text{if } x < 1/2, \\ 2(1-x), & \text{otherwise,} \end{cases}$$

with invariant densities $u_1(x) = \frac{2}{(1+x)^2}$ and $u_2(x) = 1$, cf. [Din96]. They are assembled to define the two dimensional coupled map

$$S_\varepsilon(x,y) = \begin{pmatrix} (1-\varepsilon)S_1(x) + \varepsilon S_2(y) \\ \varepsilon S_1(x) + (1-\varepsilon)S_2(y) \end{pmatrix}, \tag{6.11}$$

with the coupling constant $\varepsilon > 0$.

Following computations are done for $\varepsilon = 2^{-1}, \ldots, 2^{-9}$. We use the uniform partition of $[0,1]$ into $n = 64$ boxes, which also yields a 64×64 box partition of $[0,1]^2$. On the latter, the approximate invariant density of S_ε is computed by Ulam's method. Then, the Roothaan iteration is done, to obtain the approximative (deterministic) mean field invariant marginals. For this, the Ulam approximations of the one dimensional invariant densities of S_1 and S_2 are chosen as initial vectors. The Roothaan iteration always converged after just several steps (~ 5). Figure 6.1 shows

- the L^1-difference of the two dimensional invariant densities of S_0 and S_ε (blue dots);
- the L^1-difference of the two dimensional invariant density of S_ε and $u^\varepsilon_{n,1} \otimes u^\varepsilon_{n,2}$, where $u^\varepsilon_{n,i}$ is the mean field invariant marginal of the ith subsystem computed by the Roothaan iteration (green squares);
- the L^1-difference of the one dimensional (x- resp. y-) marginals of the invariant density of S_ε, and $u^\varepsilon_{n,1}$ resp. $u^\varepsilon_{n,2}$ (red and cyan triangles).

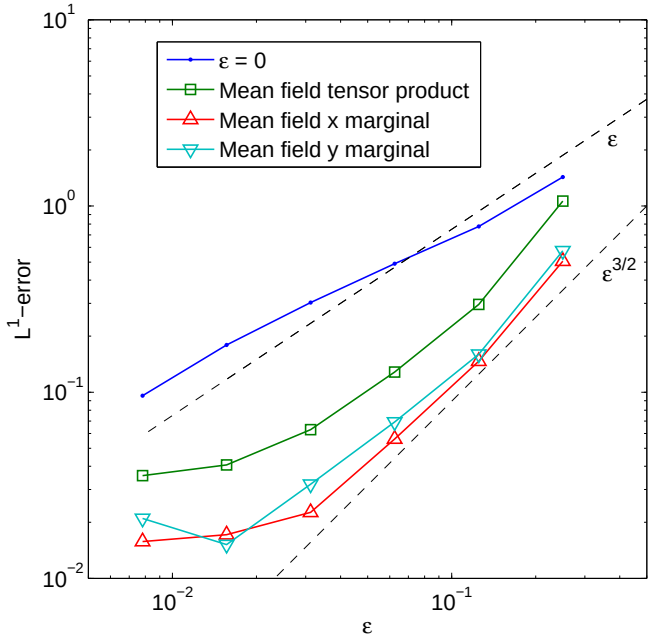


Figure 6.1: Error asymptotics in ε - The error of the mean field marginal invariant densities decay faster than linear in ε. The invariant density converges only at a linear rate to the invariant density of the decoupled system ($\varepsilon = 0$).

Where the best approximation error, $\mathcal{O}\left(n^{-1}\right)$, allowed by the approximation space, is reached, no further improvement is possible.

While the invariant density of the decoupled system seems to be only an order one approximation of the invariant density of S_ε, the mean field approximation shows better asymptotics. This observation lead to the error analysis in Section 6.2.5.

Connections with the tensor product approximability. The interplay between almost invariance and coupling yields an interesting behavior. Let us consider the parameter-dependent maps

$$S_{1,a}(x) = \left\{ \begin{array}{ll} 2x, & x < 1/4 \text{ or } x \geq 3/4 \\ 2(x-1/4)+a, & 1/4 \leq x < 3/4 \end{array} \right\} \mod 1,$$

and

$$S_{2,a}(x) = \left\{ \begin{array}{ll} 2x+a, & x < 1/4 \text{ or } x \geq 3/4 \\ 2(x-1/4), & 1/4 \leq x < 3/4 \end{array} \right\} \mod 1.$$

Both $S_{1,a}$ and $S_{2,a}$ have the almost invariant sets $[0,1/2]$ and $[1/2,1]$ with almost invariance ratio $1-a$. We define the coupled system $S_{\varepsilon,a}$ as in (6.11), with $S_{1,a}$ and $S_{2,a}$ replacing S_1 and S_2, respectively. Then, the Roothaan iteration is performed for all $a \in \left\{10^{-3}, 2{\cdot}10^{-3}, \ldots, 2{\cdot}10^{-2}\right\}$ and $\varepsilon \in \left\{10^{-3}, 2{\cdot}10^{-3}, \ldots, 2{\cdot}10^{-2}\right\}$, to obtain the (deterministic) mean field invariant marginals. The numerical computations are done by using a uniform partition of 128 boxes per dimension. As initial vectors we use here marginals of the two dimensional invariant densities computed with Ulam's method on a coarse partition ($n = 16$), embedded in the space of piecewise constant functions over the fine partition ($n = 128$). In the end, the L^1-errors of the mean field marginals to the marginals of the two dimensional invariant density are computed, cf. Figure 6.2. As we see, for some pairs (a, ε) both marginals are computed with a big error. The stochastic mean field approach gives qualitatively the same picture. It turns out, that these error plots are very similar to the ones obtained by plotting the error of the best approximation of the two dimensional invariant density by tensor product functions (i.e. functions u which can be represented as $u(x,y) = u_1(x)u_2(y)$); cf. [War10]. Observe also for the previous example, in Figure 6.1, that the good asymptotic behavior of the mean field marginals is accompanied by the good approximability of the two dimensional invariant density by tensor product functions. Since the Roothaan iteration seems to converge for all pairs (a, ε), we can draw the following conclusion:

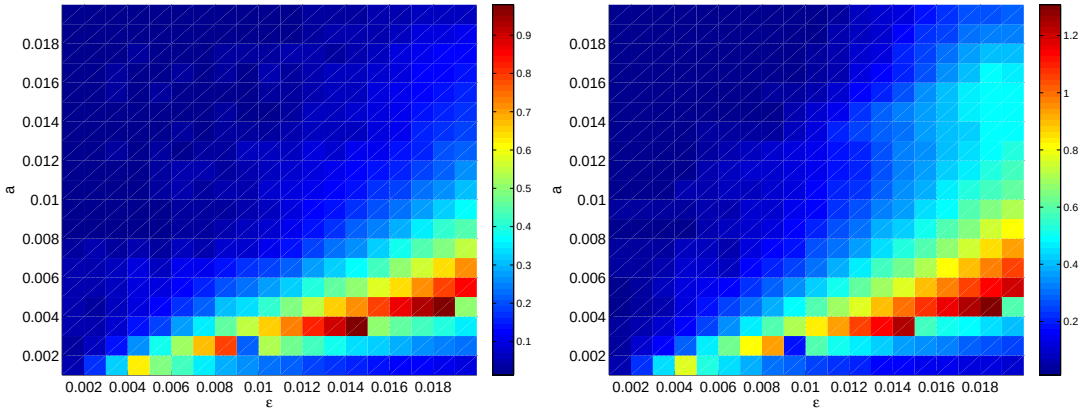


Figure 6.2: The errors of the mean field marginal invariant densities - L^1-error of the x-marginal (left) and L^1-error of the y-marginal (right), computed on a uniform partition of 128 boxes.

- the mean field invariant marginal densities are proper approximations of the invariant density marginals only if the multidimensional invariant density can be well approximated by a tensor product function; or
- the Roothaan iteration converges under some circumstances to the wrong fixed point.

In this work we do not analyze these problems any further.

6.2.5 Accuracy for weakly coupled systems

The first example in the previous section suggests, that the mean field system is capable of approximating the marginals at a higher order of accuracy. Here we prove a result on this.

Let $X \subset \mathbb{R}^d$, and let two maps $S : X \to X$ and $T : X \to \mathbb{R}^d$ be given. Define the perturbation of S, $S_\varepsilon : X \to \mathbb{R}^d$, by $S_\varepsilon = S + \varepsilon T$.

S_ε as a diffeomorphism. We restrict S_ε onto the reduced phase space $X_\varepsilon := \bigcap_{n\geq 0} S_\varepsilon^n(X)$, on which S_ε is surjective. Write $\mathrm{Id} = S^{-1} \circ S$ to obtain

$$\|x_1 - x_2\| \leq L_{S^{-1}} \|S(x_1) - S(x_2)\|,$$

where $L_{S^{-1}}$ is the Lipschitz constant of S^{-1}. A sufficient condition for S_ε to be one-to-one is

$$\|S(x_1) - S(x_2)\| > \varepsilon \|T(x_1) - T(x_2)\| \qquad \forall x_1, x_2,$$

since this implies $S_\varepsilon(x_1) \neq S_\varepsilon(x_2)$. We compute

$$\varepsilon \|T(x_1) - T(x_2)\| \leq \varepsilon L_T \|x_1 - x_2\| \leq \underbrace{\varepsilon L_{S^{-1}} L_T}_{:=\delta} \|S(x_1) - S(x_2)\|.$$

Hence we need $\delta < 1$ to get injectivity, i.e.

$$\varepsilon < \frac{1}{L_{S^{-1}} L_T}. \tag{6.12}$$

Further, we have for all x_1, x_2

$$\begin{aligned} \|S_\varepsilon(x_1) - S_\varepsilon(x_2)\| &= \|S(x_1) - S(x_2) + \varepsilon \left(T(x_1) - T(x_2)\right)\| \\ &\geq \|S(x_1) - S(x_2)\| - \underbrace{\varepsilon \|T(x_1) - T(x_2)\|}_{\leq \delta \|S(x_1) - S(x_2)\|} \\ &\geq (1-\delta) \|S(x_1) - S(x_2)\| \\ &\geq \frac{1-\delta}{L_{S^{-1}}} \|x_1 - x_2\|. \end{aligned}$$

This implies that DS_ε cannot have a singular value smaller than $(1-\delta)/L_{S^{-1}}$, and by the inverse function theorem we have that $S_\varepsilon : X_\varepsilon \to X_\varepsilon$ is a diffeomorphism, provided that T is continuously differentiable and the above bound on ε holds. Moreover, we may bound the Lipschitz constant of S_ε^{-1}:

$$L_{S_\varepsilon^{-1}} \leq \frac{L_{S^{-1}}}{1 - \varepsilon L_{S^{-1}} L_T}. \tag{6.13}$$

Expansion of S_ε^{-1} in ε. We would like to expand the inverse of S_ε up to $\mathcal{O}(\varepsilon)$ terms. Observe

$$S^{-1}\left(S_\varepsilon(x)\right) = x + \varepsilon DS^{-1}(S(x)) \cdot T(x) + \mathcal{O}\left(\varepsilon^2\right),$$

or, by setting $y = S_\varepsilon(x)$,

$$S_\varepsilon^{-1}(y) = S^{-1}(y) - \varepsilon DS^{-1}\left(S\left(S_\varepsilon^{-1}(y)\right)\right) \cdot T\left(S_\varepsilon^{-1}(y)\right) + \mathcal{O}\left(\varepsilon^2\right),$$

if S^{-1} is twice continuously differentiable. Since $S(x) = S_\varepsilon(x) + \mathcal{O}(\varepsilon)$, this inspires the approximation

$$S_\varepsilon^{-1} \approx \bar{S}_\varepsilon := S^{-1} - \varepsilon DS^{-1} \cdot T \circ S^{-1}.$$

The function $\bar{S}_\varepsilon$ is differentiable, if T is so. Then a first order Taylor expansion yields

$$\bar{S}_\varepsilon(\underbrace{S_\varepsilon(x)}_{=y}) = \underbrace{x}_{=S_\varepsilon^{-1}(y)} + \mathcal{O}\left(\varepsilon^2\right),$$

uniformly in x, thus uniformly in y (over compact sets), since S_ε is a diffeomorphism. This means that $\bar{S}_\varepsilon$ is the expansion we were searching for, and we need $S^{-1} \in C^2(X)$ and $T \in C^1(X)$. The following Lemma summarizes these results.

Lemma 6.7. *Let X be a compact set, and $S : X \to X$ a diffeomorphism with $S^{-1} \in C^2(X)$. Further let $T \in C^1(X)$. Then it holds for $S_\varepsilon = S + \varepsilon T$, ε small enough,*

$$S_\varepsilon^{-1}(x) = S^{-1}(x) - \varepsilon DS^{-1}(x) \cdot T \circ S^{-1}(x) + \mathcal{O}\left(\varepsilon^2\right) \tag{6.14}$$

uniformly in $x \in X_\varepsilon$; provided X_ε has nonempty interior.

Perturbation analysis of mean field. Let a dynamical system $S : X \times Y \to X \times Y$ be given with

$$S(x,y) = \begin{pmatrix} S_1(x) \\ S_2(y) \end{pmatrix},$$

where S_1 and S_2 are diffeomorphisms with twice continuously differentiable inverses. Define its perturbation

$$S_\varepsilon(x,y) = \begin{pmatrix} S_{1,\varepsilon}(x,y) \\ S_{2,\varepsilon}(x,y) \end{pmatrix} = \begin{pmatrix} S_1(x) + \varepsilon T_1(x,y) \\ S_2(y) + \varepsilon T_2(x,y) \end{pmatrix},$$

where we assume T_1 and T_2 to be differentiable. Further, we assume, that S_ε is surjective on $X \times Y$ for all ε in consideration. Let $\mathcal{P}_\varepsilon$ denote the FPO associated with S_ε. Given a separable density $u = u_1 \otimes u_2$ (u_1 and u_2 both twice continuously differentiable) we would like to compare the marginal of $\mathcal{P}_\varepsilon u$ with the *deterministic* mean field iterate of u_1. We have already seen in Proposition 6.1 that the stochastic mean field system gives the exact marginals.

Recall that for a diffeomorphism, the FPO can be written as

$$\mathcal{P}u = u \circ S^{-1} \cdot \left|DS^{-1}\right|, \tag{6.15}$$

where $\left|DS^{-1}\right| = \left|\det\left(DS^{-1}\right)\right|$. Since the determinant is continuous as a function of the matrix components, and DS is never singular, we may omit the absolute value brackets without loss. In the following $|A|$ denotes $\det(A)$.

We begin with the expansion of the determinant.

Lemma 6.8 (Perturbation expansion of the determinant)**.** *It holds*

$$\det(I + X) = 1 + \operatorname{tr}(X) + \mathcal{O}(X^2).$$

If A is a nonsingular matrix we also have

$$\det(A + \varepsilon B) = \det(A)\ \left(1 + \varepsilon\ \operatorname{tr}\left(A^{-1}B\right)\right) + \mathcal{O}\left(\varepsilon^2\right),$$

as $\varepsilon \to 0$.

First, we compute the perturbation expansion of the marginal density.

Lemma 6.9. *The expansion of the marginal density is*

$$\begin{aligned}\int_Y \mathcal{P}_\varepsilon u \,\mathrm{d}y &= u_1 \circ S_1^{-1} \left|DS_1^{-1}\right| - \varepsilon \left|DS_1^{-1}\right| \Bigg(\nabla u_1^\top \circ S_1^{-1} \cdot \int_Y T_{1,y}^{(-1)} u_2(y)\,\mathrm{d}y + \\ &\quad + u_1 \circ S_1^{-1} \int_Y \operatorname{tr}\left((DS_1^{-1})^{-1} DT_{1,y}^{(-1)} \right) u_2(y)\,\mathrm{d}y \Bigg) + \mathcal{O}\left(\varepsilon^2\right). \end{aligned} \tag{6.16}$$

Proof. By (6.14) we have, that the inverse of the first component map has following expansion:

$$S_{1,\varepsilon,y}^{-1} = S_1^{-1} - \varepsilon \underbrace{DS_1^{-1} \cdot T_{1,y} \circ S_1^{-1}}_{T_{1,y}^{(-1)}} + \mathcal{O}\left(\varepsilon^2\right). \tag{6.17}$$

Equation (6.7) gives

$$\int_Y \mathcal{P}_\varepsilon u \,\mathrm{d}y = \int_Y \mathcal{P}_{1,y,\varepsilon} u_1\ u_2(y)\,\mathrm{d}y.$$

Using the expression (6.15) for the FPO and using the expansions (6.17) and Lemma 6.8, we obtain

$$\begin{aligned}\int_Y \mathcal{P}_\varepsilon u \,\mathrm{d}y &= \int_Y u_1 \left(S_1^{-1} - \varepsilon T_{1,y}^{(-1)} + \mathcal{O}\left(\varepsilon^2\right) \right) \left| DS_1^{-1} - \varepsilon DT_{1,y}^{(-1)} \right| u_2(y)\,\mathrm{d}y \\ &= \int_Y \left(u_1 \circ S_1^{-1} - \varepsilon \nabla u_1^\top \circ S_1^{-1} \cdot T_{1,y}^{(-1)} + \mathcal{O}\left(\varepsilon^2\right) \right) \cdot \\ &\quad \cdot \left|DS_1^{-1}\right| \left(1 - \varepsilon\ \operatorname{tr}\left((DS_1^{-1})^{-1} DT_{1,y}^{(-1)} \right) + \mathcal{O}\left(\varepsilon^2\right) \right) u_2(y)\,\mathrm{d}y, \end{aligned}$$

where the second equation follows from Taylor expansions. Reordering the terms by their order of ε and pulling out the factors from the integral which depend solely on x, we get (6.16). □

Next we do the same for the deterministic mean field system

$$S_{1,\mathrm{MF},\varepsilon} = S_1 + \varepsilon \underbrace{\int_Y T_{1,y} u_2(y)\,\mathrm{d}y}_{T_{1,\mathrm{MF}}},$$

with associated transfer operator $\mathcal{P}_{1,\mathrm{MF},\varepsilon}$.

Lemma 6.10. *The expansion of the iterated density is*

$$\begin{aligned}\mathcal{P}_{1,\mathrm{MF},\varepsilon}u_1 &= u_1 \circ S_1^{-1}\left|DS_1^{-1}\right| - \varepsilon\left|DS_1^{-1}\right|\Big(\nabla u_1^{\top} \circ S_1^{-1} \cdot T_{1,\mathrm{MF}}^{(-1)} + \\ &\quad + u_1 \circ S_1^{-1}\mathrm{tr}\left((DS_1^{-1})^{-1}DT_{1,\mathrm{MF}}^{(-1)}\right)\Big) + \mathcal{O}\left(\varepsilon^2\right). \end{aligned} \tag{6.18}$$

Proof. The inverse of the mean field map has following expansion:

$$S_{1,\mathrm{MF},\varepsilon}^{-1} = S_1^{-1} - \varepsilon \underbrace{DS_1^{-1} \cdot T_{1,\mathrm{MF}} \circ S_1^{-1}}_{T_{1,\mathrm{MF}}^{(-1)}} + \mathcal{O}\left(\varepsilon^2\right). \tag{6.19}$$

By definition is $T_{1,\mathrm{MF}}$ continuously differentiable. The rest of the proof follows exactly the lines of the proof of Lemma 6.9 (except that we do not need (6.7)). □

Now we are ready to show, that the mean field iterate coincides with the marginal of the iterate of the full density up to first order. Naively, one would have expected only zeroth order match.

Theorem 6.11. *Under the assumptions made before it holds*

$$\mathcal{P}_{1,\mathrm{MF},\varepsilon}u_1(x) - \int_Y \mathcal{P}_{\varepsilon}u(x,y)\,\mathrm{d}y = \mathcal{O}\left(\varepsilon^2\right),$$

uniformly in x.

Proof. It is easy to see, that

$$T_{1,\mathrm{MF}}^{(-1)} = \int_Y T_{1,y}^{(-1)}\,\mathrm{d}y.$$

Comparing (6.16) and (6.18), we only need that the functional $A(y) \mapsto \int A(y)u(y)\,\mathrm{d}y$ and the trace function are interchangeable:

$$\int \mathrm{tr}A(y)u(y)\,\mathrm{d}y = \int \sum_i a_{ii}(y)u(y)\,\mathrm{d}y = \sum_i \int a_{ii}(y)u(y)\,\mathrm{d}y = \mathrm{tr}\int A(y)u(y)\,\mathrm{d}y.$$

Thus the proof is competed. □

Remark 6.12 (Deterministic mean field - general coupling). Theorem 6.11 holds for general couplings as well. Consider

$$S_{\varepsilon}(x,y) = \begin{pmatrix} S_{1,\varepsilon}(x,y) \\ S_{2,\varepsilon}(x,y) \end{pmatrix} = \begin{pmatrix} S_1(x) + \varepsilon T_{1,\varepsilon}(x,y) \\ S_2(y) + \varepsilon T_{2,\varepsilon}(x,y) \end{pmatrix},$$

i.e. the first order terms will depend on ε as well. We can omit higher order terms, since they can be included in the first order ones. This does not change the expansion of the inverse either, and allows an analogous derivation as above.

Remark 6.13. Comparing Figure 6.1 with Theorem 6.11 we observe a $\mathcal{O}\left(\varepsilon^{1/2}\right)$ loss of accuracy. A reason for this may be that the mapping is no diffeomorphism, merely piecewise differentiable and piecewise invertible, or that the reduced phase space X_{ε} is not the whole space, just a rhomboid with vertices $(0,0)$, $(1-\varepsilon,\varepsilon)$, $(1,1)$ and $(\varepsilon,1-\varepsilon)$.

6.3 Mean field for molecular dynamics

The idea of applying mean field theory for detecting dominant conformations of molecules goes back to Friesecke et al. [Fri09]. We give here a detailed explanation of the method presented in that publication. We expect the mean field description of particular classical MD systems to work well for reasons that follow.

The examples to the mean field theory for maps suggest, that our method works well in cases where the invariant density of the system is "as decoupled as possible"; i.e. a good tensor product approximability is available. Considering MD, the canonical density is decoupled (i.e. in tensor product form), if the Hamiltonian consists of independent summands, see Section 2.4.1. This can be partly achieved for chain molecules with the standard force field we are working with (the potential depending on bond lengths, bond angles and torsion angles), by using inner coordinates. Coupling occurs only in the kinetic energy term $\frac{1}{2}p^{\top}M(q)^{-1}p$, the potential part of the canonical density is decoupled. It turns out, that the coupling induced by $M(q)$ is not of negligible magnitude for "neighboring" degrees of freedom. Still, the exact details of *deterministic* momentum evolution do not seem to play a very inportant role in conformation dynamics. Conformational changes may be observed by modeling the system with the Langevin equation, where one applies perturbation of the momenta by suitably scaled white noise. Another successful approach (Schütte's spatial transfer operator, cf. Section 2.4.1) considers only fluctuations on the configuration space by building expectation values w.r.t. the distribution of momenta.

6.3.1 The continuous-time mean field system

We establish here our theory of mean field approximation for MD systems introduced in Section 2.4.1. Our starting point is a, for the moment arbitrary, partition of phase

space coordinates $z = (q, p)$ into subsystem coordinates:

$$z = (z_1, \dots, z_N) \in \Omega \times \mathbb{R}^d, \quad z_i = (q_i, p_i) \in \Omega_i \times \mathbb{R}^{d_i}, \quad \dim(\Omega_i) = d_i, \quad \sum_{i=1}^{N} d_i = d,$$

where p_i is the vector of momentum coordinates corresponding to the position coordinates q_i. Let $f_i = \left(\frac{\partial H}{\partial p_i}, -\frac{\partial H}{\partial q_i}\right)$. Then (2.26) can be rewritten as

$$\dot{z}_i = f_i(z), \quad i = 1, \dots, N. \tag{6.20}$$

We define the mean field system in an analogous manner, as for maps. Here we consider only the deterministic system, cf. Definition 6.3. Since we are dealing with time-continuous systems, it is natural to average the effect of the influencing subsystems on the right hand side. Let the $u_i(\cdot, t)$, $i = 1, \dots, N$, be probability density functions describing the distribution of the ith subsystem at time t. For notational convenience, let $\widehat{z_i}$ denote the coordinates $(z_j)_{j \neq i}$, $\widehat{u_i} = \prod_{j \neq i} u_j$, and $\widehat{\Omega_i} = \bigotimes_{j \neq i} \Omega_j$ the tensor product space. The mean field system is defined by the (time-depenent!) right hand sides

$$f_{i,\mathrm{MF}}[\widehat{u_i}](z_i, t) := \int_{\widehat{\Omega_i} \times \mathbb{R}^{d-d_i}} f_i(z) \widehat{u_i}(\widehat{z_i}, t) \, \mathrm{d}\widehat{z_i}, \tag{6.21}$$

where the evolution of subsystem densities is governed by

$$\partial_t u_i + \mathrm{div}_{z_i} \left(u_i f_{i,\mathrm{MF}}[\widehat{u_i}] \right), \quad i = 1, \dots, N. \tag{6.22}$$

We call the system of equations (6.22), $i = 1, \dots, N$, the *mean field approximation* to the Liouville equation. Note that it is a system of N coupled nonlinear partial integrodifferential equations on the lower-dimensional subsystem phase spaces $\mathbb{R}^{2d_i}$, whereas the original Liouville equation was a linear partial differential equation on $\Omega \times \mathbb{R}^d \subset \mathbb{R}^{2d}$, $d = \sum_i d_i$.

We record some basic properties of the mean field approximation. For more details, we refer to [Fri09].

1. The total densities $\int u_i(z_i, t) \, \mathrm{d}z_i$ are conserved.[1] This is immediate from the conservation law form (6.22). Thus we may continue to interpret the u_i as probability densities.

[1] Since the integration domains should be always clear, we omit indicating them from now on.

2. For noninteracting subsystems, i.e.,

$$H(z) = \sum_{i=1}^{N} \left(\frac{1}{2} p_i^\top M_i(q_i)^{-1} p_i + V_i(q_i) \right),$$

the mean field system is exact; that is, if the $u_i(z_i, t)$ evolve according to 6.22, then the product $u_1(z_1, t) \cdots u_N(z_N, t)$ solves the original Liouville equation (2.28).

3. For given u_j, $j \neq i$, the dynamics of the ith subsystem are governed by the time-dependent subsystem Hamiltonian

$$H_{i,\mathrm{MF}}[\widehat{u_i}](q_j, p_j, t) = \int H(q,p) \prod_{j \neq i} u_j(q_j, p_j, t) \, \mathrm{d}\widehat{z_i}; \tag{6.23}$$

so,

$$f_{i,\mathrm{MF}}[\widehat{u_i}](q_i, p_i, t) = \begin{pmatrix} \frac{\partial}{\partial p_i} H_{i,\mathrm{MF}}(p_i, q_i, t) \\ -\frac{\partial}{\partial q_i} H_{i,\mathrm{MF}}(p_i, q_i, t) \end{pmatrix}. \tag{6.24}$$

In particular, $f_{i,\mathrm{MF}}$ is divergence-free. Note that time-dependence of the effective subsystem Hamiltonian enters only through time-dependence of the u_j, $j \neq i$.

4. The total energy expectation

$$E(t) := \int H(z) \, u_1(z_1, t) \cdots u_N(z_N, t) \, \mathrm{d}z_1 \cdots \mathrm{d}z_N$$

is conserved.

Property 2 contains useful information regarding how the, up to now arbitrary, partitioning into subsystems should be chosen in practice. In order to maximize agreement with the full Liouville equation (2.27), the subsystems should be only weakly coupled. In the case of an N-atom chain, this suggests working with subsystems defined by inner, not cartesian, coordinates (as it has been done in the example of n-butane in Section 2.4.2). Namely, in inner coordinates, at least the *potential energy* decouples completely for standard potentials containing nearest neighbor bond terms, third neighbor angular terms, and fourth neighbor torsion terms: $V((r_{ij}), (\theta_{ijk})_{ijk}, (\phi_{ijk\ell})_{ijk\ell}) = \sum V_{ij}(r_{ij}) + \sum V_{ijk}(\theta_{ijk}) + \sum V_{ijk\ell}(\phi_{ijk\ell})$.

Remark 6.14. A deeper, and perhaps surprising, theoretical property of the mean field model which goes beyond property 2 concerns weakly coupled subsystems. Consider a Hamiltonian of the form $H(z) = H_0(z) + \varepsilon H_{\mathrm{int}}(z)$, where H_0 is a noninteracting Hamiltonian of the form given in 2 and ε is a coupling constant. We expect, in analogy

with Theorem 6.11, that in case of a tensor product initial density the exact marginal subsystem densities $\int u(\cdot,t)\,\mathrm{d}\widehat{z}_i$, obtained from (2.27), and the mean field densities obtained by solving (6.22) differ, up to any fixed time t, only by $O(\varepsilon^2)$, and not the naively expected $O(\epsilon)$. This means that the effect of coupling between subsystems is captured correctly to leading order (in the coupling constant) by the mean field approximation.

We do not prove this statement here, but leave it as a conjecture. In particular, note, that the coupling of the momenta, introduced by $M(q)$, is not of small magnitude. However, we assume mean field to work well here due to the reasons given above in the introduction.

The mean field transfer operator. The most natural way to define the mean field transfer operators would be as the evolution operator of the coupled system of mean field Liouville equations (6.22). This would be a nonlinear operator, since changing an initial subsystem density $u_i(\cdot,0)$ will affect all other mean field subsystems (which are coupled with the ith), which, in turn, influences the dynamics of the ith subsystem nonlinearly.

In order to obtain *linear* operators still appropriate for our purposes, let us recall what our aim is with the mean field approximation: we wish to characterize the long-term behavior of the subsystems by defining suitable dynamics, "averaged" w.r.t. the distribution of the other systems, on them. Assuming, that the full system is in equilibrium (i.e. distributed according to its invariant density), the subsystems are distributed according to the marginals of the invariant density. Therefore, we seek for subsystem densities u_i which are invariant under the mean field dynamics *induced by themselves*. Hence, we freeze time, and define time-independent right hand sides for the (time-independent) subsystem densities u_i, $i=1,\dots,N$,

$$f_{i,\mathrm{MF}}[\widehat{u}_i](z_i) := \int_{\mathbb{R}^{2(d-d_i)}} f_i(z)\prod_{j\neq i} u_j(z_j)\,\mathrm{d}\widehat{z}_i. \tag{6.25}$$

Thus, we have N autonomous systems with flow denoted by $\Phi^t_{i,\mathrm{MF}}$. We define the *mean field transfer operator* of the ith subsystem, $\mathcal{P}_{i,\mathrm{MF}}[\widehat{u}_i]$, by the transfer operator associated with $\Phi^t_{i,\mathrm{MF}}$.

Once we have the mean field approximations to the invariant marginals, i.e. u_i, $i=1,\dots,N$, with $\mathcal{P}_{i,\mathrm{MF}}[\widehat{u}_i]u_i = u_i$ for $i=1,\dots,N$, the mean field transfer operators

describe the density changes in equilibrium, or "averaged along a long iteration" of the system.[1] Hence, we expect eigenfunctions of $\mathcal{P}_{i,\mathrm{MF}}[\widehat{u_i}]$ at eigenvalues near one to give information about almost invariant behavior (or "rarely occurring transitions" in a long iteration — we think of conformation changes in MD) of the ith subsystem. Note, this operator is not suitable for describing the evolution of the mean field system in general, merely for characterizing evolution in equilibrium.

Recall, that $h(q,p)$ denotes the canonical density of the system, and the spatial transfer operator is given by

$$\mathcal{S}^t w = \int \mathcal{P}^t \left(w\bar{h}(\cdot,p)\right) \,\mathrm{d}p,$$

where $\bar{h}$ is the distribution of momenta for a given position q, i.e.

$$\bar{h}(q,p) = \frac{h(q,p)}{\int h(q,p)\,\mathrm{d}p}.$$

Now we define the spatial transfer operator corresponding to the mean field system. The (canonical) distribution of the ith subsystem is given by

$$h_i(q_i,p_i) = \int h(z)\,\mathrm{d}\widehat{z_i}.$$

The distribution of p_i for a given q_i is

$$\bar{h}_i(q_i,p_i) = \frac{h_i(q_i,p_i)}{\int h_i(q_i,p_i)\,\mathrm{d}p_i}.$$

We therefore define the *mean field spatial transfer operator* as

$$\mathcal{S}^t_{i,\mathrm{MF}}[\widehat{w_i}]w_i(q_i) = \int \mathcal{P}^t_{i,\mathrm{MF}}[\widehat{u_i}]u_i(q_i,p_i)\,\mathrm{d}p_i, \tag{6.26}$$

where $u_i := w_i\bar{h}_i$.

Mean field spatial eigenfunction approximation. We approximate the eigenfunctions of the spatial transfer operator the same way as indicated in the previous paragraph. In the first step, we search for the mean field invariant marginals $w_1,\ldots,w_N$. They satisfy $\mathcal{S}^t_{i,\mathrm{MF}}[\widehat{w_i}]w_i = w_i$, $i = 1,\ldots,N$. In the second step, dominant configurations are obtained as almost invariant sets in the configuration space of the subsystems, i.e. we search for eigenvalues near one of the operators $\mathcal{S}^t[\widehat{w_i}]$.

[1] Assuming ergodicity, states along a long trajectory will be distributed according to the invariant density of the system; see Section 2.2.1.

The computation of the first step is done by a Roothaan type iteration, cf. Algorithm 6.6. We fix initial values w_i^0 and solve the linear eigenvalue problems $\mathcal{S}^t[\widehat{w_i}]w_i^{\text{new}} = w_i^{\text{new}}$, by updating the w_i running cyclically over the subsystem index i. The iteration is terminated if no improvement is observable. Then, the second step is carried out by taking the final w_i, and computing eigenfunctions of the $\mathcal{S}^t[\widehat{w_i}]$ at eigenvalues near one. The computation of the numerical discretization of the $\mathcal{S}^t[\widehat{w_i}]$ is discussed in the next section.

6.3.2 Numerical realization

Equation (2.33) shows us a way to discretize the spatial transfer operator. However, in order to use it for the mean field spatial transfer operators, two questions have to be answered.

- How to sample $\bar{h}_i(q_i, \cdot)$, i.e. the distribution of the momenta p_i?
- Given the spatial distributions w_i, $i = 1, \ldots, N$, how to compute the flows $\Phi^t_{i,\text{MF}}$?

The computations here assume, that we use inner coordinates, where the potential is decoupled, i.e. $V(q) = \sum_{k=1}^N V_k(q_k)$. Recall the Hamiltonian

$$H(q,p) = \frac{1}{2}p^\top M(q)^{-1}p + V(q),$$

where $M(q)$ is symmetric positive definite for every q, and the canonical density

$$h(q,p) = C\exp(-\beta H(q,p)) = C\exp\left(-\frac{\beta}{2}p^\top M(q)^{-1}p\right)\ \prod_{k=1}^N \exp(-\beta V_k(q_k)).$$

Sampling of $\bar{h}_i(q_i, \cdot)$. First, we consider marginal canonical density h_i.

$$h_i(q_i,p_i) = C\int e^{-\beta V(q)} \underbrace{\int \exp\left(-\frac{\beta}{2}p^\top M(q)^{-1}p\right)\,\mathrm{d}\widehat{p}_i}_{\text{analytical solution?}}\,\mathrm{d}\widehat{q}_i. \tag{6.27}$$

A semi-analytical solution of the integral can be obtained as follows. Without loss, we may permute the subsystems such that $i = 1$. Decompose $M(q)^{-1}$ by

$$M(q)^{-1} = \begin{pmatrix} A & V^\top \\ V & \widehat{M} \end{pmatrix},$$

with $A \in \mathbb{R}^{d_i \times d_i}$, $V \in \mathbb{R}^{(d-d_i)\times d_i}$ and $\widehat{M} \in \mathbb{R}^{(d-d_i)\times(d-d_i)}$. The dependence on q is suppressed for notational simplicity. Just as $M(q)$, also A and $\widehat{M}$ are symmetric positive definite, and thus the latter can be diagonalized by the orthogonal matrix $\widehat{Q}$. Hence $\widehat{Q}^\top \widehat{M}\widehat{Q} = \widehat{D} = \operatorname{diag}(\hat{d})$. By coordinate transformation, exploiting $\int_{\mathbb{R}} e^{-\alpha x^2}\,\mathrm{d}x = \sqrt{\pi/\alpha}$ for $\alpha > 0$, and denoting the columns of the matrix $V^\top \widehat{Q}\widehat{D}^{-1}$ by $v_1, \dots, v_{d-d_i}$, we have

$$\begin{aligned}
\int \exp\left(-\frac{\beta}{2} p^\top M(q)^{-1} p\right) \mathrm{d}\widehat{p}_i = \\
&= \exp\left(-\frac{\beta}{2} p_i^\top A p_i\right) \int \exp\left(-\frac{\beta}{2}\left(2 p_i^\top V^\top \widehat{p}_i + \widehat{p}_i^{\,\top} \widehat{M}\widehat{p}_i\right)\right) \mathrm{d}\widehat{p}_i \\
&\overset{\widehat{p}_i = \widehat{Q}y}{=} \exp\left(-\frac{\beta}{2} p_i^\top A p_i\right) \int \exp\left(-\frac{\beta}{2}\left(2 p_i^\top V^\top \widehat{Q} y + y^\top \widehat{D} y\right)\right) \mathrm{d}y \\
&= \exp\left(-\frac{\beta}{2} p_i^\top A p_i\right) \int \exp\left(-\frac{\beta}{2}\left(\sum_{k=1}^{d-d_i} \hat{d}_k \left(y + p_i^\top v_k\right)^2 - \hat{d}_k (p_i^\top v_k)^2\right)\right) \mathrm{d}y \\
&= \exp\left(-\frac{\beta}{2} p_i^\top B p_i\right) \prod_{k=1}^{d-d_i} \sqrt{\frac{2\pi}{\beta \hat{d}_k}},
\end{aligned}$$

with $B = A - V^\top \widehat{Q}\widehat{D}^{-1}\widehat{Q}V = A - V^\top \widehat{M}^{-1} V$. Note, $B = B(q)$ is symmetric positive definite for all q. Numerical computations suggest, that $M(q)$ and $B(q)$ are smooth, thereby the integral w.r.t. $\widehat{q}_i$ in (6.27) can be approximated very well by a low order tensor product Gauss quadrature. Let q_i^ℓ, c_i^ℓ, $\ell = 1, \dots, L$, denote the quadrature nodes in $\widehat{\Omega}_i$ and weights, respectively. Then we have

$$h_i(q_i, p_i) \approx C \sum_{\ell=1}^{L} \underbrace{c_i^\ell \exp\left(-\beta V((q_i, q_i^\ell))\right) \prod_{k=1}^{d-d_i} \sqrt{\frac{2\pi}{\beta \hat{d}_k((q_i, q_i^\ell))}}}_{=: C_i^\ell(q_i)} \exp\left(-\frac{\beta}{2} p_i^\top B((q_i, q_i^\ell)) p_i\right) \tag{6.28}$$

For any fixed q_i, the density $\bar{h}_i(q_i, \cdot)$ is just the function $h_i(q_i, \cdot)$, normed to be a probability density. Hence, $\bar{h}(q_i, \cdot)$ can be approximated with a weighted sum of Gaussians, where the weights have the same pairwise ratios as the $C_i^\ell(q_i)$.

Note, that Gaussians can be sampled easily by suitably scaled normally distributed random variables.[1] The most programs used for numerical computations provide rou-

[1] A random variable $\mathbf{x}$ distributed according to a multivariate normal distribution with covariance matrix Σ can be sampled as follows. The symmetric positive definite covariance matrix has a Cholesky factorization $L^\top L = \Sigma$. The components of the random variable $\mathbf{y} = L^{-\top}\mathbf{x}$ are independent with variance 1. Hence, we can draw a sample y of $\mathbf{y}$, and set $L^\top y$ as a sample of $\mathbf{x}$.

tines for drawing random samples according to a normal distribution with variance one. Hence, we may sample $\bar{h}(q_i, \cdot)$ in two steps:

1. Choose $\tilde{\ell} \in \{1, \dots, L\}$ with probability $C_i^{\tilde{\ell}}(q_i) / \sum_{\ell=1}^{L} C_i^{\ell}(q_i)$.
2. Draw a random sample according to a normal distribution with covariance matrix $B(q^{\tilde{\ell}})$.

Regarding complexity, to set up the sampling of $\bar{h}(q_i, \cdot)$, an initial step is made, where L different $(d-d_i)$-by-$(d-d_i)$ matrices are diagonalized. If the latter step is performed by the QR algorithm, the complexity is $\mathcal{O}\left(L(d-d_i)^3\right)$. All other computational steps have costs of lower order, thus these are the leading order costs of sampling $\bar{h}(q_i, \cdot)$ for a fixed q_i.

The full representation of $\bar{h}_i$ will be needed in the following paragraph, so we give an explicit expression for $h_{i,q}(q_i) = \int h_i(q_i, p_i)\, \mathrm{d}p_i$ as well. Let $\sigma(B(q)) = \left\{b_1(q), \dots, b_{d_i}(q)\right\}$ denote the spectrum of $B(q)$. Then we have

$$\begin{aligned} h_{i,q}(q_i) &= C \int e^{-\beta V(q)} \prod_{k=1}^{d-d_i} \sqrt{\frac{2\pi}{\beta \hat{d}_k(q)}} \prod_{j=1}^{d_i} \sqrt{\frac{2\pi}{\beta b_j(q)}}\, \mathrm{d}\widehat{q}_i \\ &= C \left(\frac{2\pi}{\beta}\right)^{d/2} \int e^{-\beta V(q)} \prod_{k=1}^{d-d_i} \hat{d}_k(q)^{-1/2} \prod_{j=1}^{d_i} b_j(q)^{-1/2}\, \mathrm{d}\widehat{q}_i. \end{aligned} \tag{6.29}$$

Computing the flow $\Phi_{i,\mathrm{MF}}^t$. As discussed in Section 2.4.2, we apply small integration times t. Hence, some low order explicit integration schemes are suitable for the numerical approximation of the flow $\Phi_{i,\mathrm{MF}}^t$. Nevertheless, they all require some evaluations of the right hand side $f_{i,\mathrm{MF}}$, the computation of which is discussed in the following. Recall, that the subsystem distributions are given by $u_i = w_i \bar{h}_i$. This gives with (6.25) the differential equations describing the motion of the ith subsystem, $(\dot{q}_i, \dot{p}_i) = f_{i,\mathrm{MF}}(q_i, p_i)$ (remember, we use decoupled potentials), i.e.

$$\begin{aligned} \dot{q}_i &= \int \frac{\partial}{\partial p_i} \left(\tfrac{1}{2} p_i^\top M(q)^{-1} p_i\right) \prod_{k \neq i} w_k(q_k) \bar{h}_k(q_k, p_k)\, \mathrm{d}\widehat{z}_i, \\ \dot{p}_i &= \int \left(-\frac{\partial}{\partial q_i} \left(\tfrac{1}{2} p_i^\top M(q)^{-1} p_i\right) - \nabla_{q_i} V_i(q_i)\right) \prod_{k \neq i} w_k(q_k) \bar{h}_k(q_k, p_k)\, \mathrm{d}\widehat{z}_i. \end{aligned}$$

In the following, we assume all subsystems to be one dimensional. Hence, q_i and p_i can be viewed as the ith component of the vectors q and p, respectively. This will simplify

the derivation of the results below. Nevertheless, analogous results hold in the general case as well.

To $\dot{q}_i$. We have

$$\begin{aligned}\dot{q}_i &= \iint \left(M(q)^{-1}p\right)_i \prod_{k\neq i} w_k(q_k)\bar{h}_k(q_k,p_k)\,\mathrm{d}\widehat{p}_i\mathrm{d}\widehat{q}_i \\ &= \int \prod_{k\neq i} w_k(q_k) \int \sum_{\ell} \left(M(q)^{-1}\right)_{i\ell} p_\ell \prod_{j\neq i} \bar{h}_j(q_j,p_j)\,\mathrm{d}\widehat{p}_i\mathrm{d}\widehat{q}_i \\ &= \dots\end{aligned}$$

where $h_j(q_j,p_j)$ is an even function of p_j, thus $p_j h_j(q_j,p_j)$ is odd as a function of p_j, so its integral over the real line vanishes, and the above sum reduces to a single term:

$$\begin{aligned}\dots &= \int \prod_{k\neq i} w_k(q_k) \left(M(q)^{-1}\right)_{ii} p_i \int \prod_{j\neq i} \underbrace{\bar{h}_j(q_j,p_j)}_{\int\dots\mathrm{d}p_j=1}\,\mathrm{d}\widehat{p}_i\mathrm{d}\widehat{q}_i \\ &= p_i \int \left(M(q)^{-1}\right)_{ii} \widehat{w_i}(\widehat{q}_i)\,\mathrm{d}\widehat{q}_i.\end{aligned}$$

To $\dot{p}_i$. We deal with the two summands separately. It is an easy task to compute the mean field force contribution of the potentials, since the potential is decoupled, thus

$$\dot{p}_i^{\mathrm{II}} = \int -V_i'(q_i) \prod_{k\neq i} w_k(q_k)\bar{h}_k(q_k,p_k)\,\mathrm{d}\widehat{p}_i\mathrm{d}\widehat{q}_i = -V_i'(q_i).$$

Considering the first term, we have

$$\begin{aligned}\dot{p}_i^{\mathrm{I}} &= \iint -\tfrac{1}{2}p^\top \frac{\partial}{\partial q_i} M(q)^{-1}p \prod_{k\neq i} w_k(q_k)\bar{h}_k(q_k,p_k)\,\mathrm{d}\widehat{p}_i\mathrm{d}\widehat{q}_i \\ &= -\tfrac{1}{2}\int \prod_{k\neq i} w_k(q_k) \int p^\top \frac{\partial}{\partial q_i} M(q)^{-1}p \prod_{j\neq i} \bar{h}_j(q_j,p_j)\,\mathrm{d}\widehat{p}_i\mathrm{d}\widehat{q}_i \\ &= -\tfrac{1}{2}\int \prod_{k\neq i} w_k(q_k) \sum_{n,m=1}^{d} \frac{\partial}{\partial q_i}\left(M(q)^{-1}\right)_{nm} \left\{\int p_n p_m \prod_{j\neq i} \bar{h}_j(q_j,p_j)\,\mathrm{d}\widehat{p}_i\right\}\mathrm{d}\widehat{q}_i \\ &= \dots\end{aligned}$$

The integral in the brackets $\{\}$ vanishes every time $n \neq m$, because then either $n \neq i$ or $m \neq i$, and we integrate the function $p_j\bar{h}_j(q_j,p_j)$, which is odd in p_j, over the real line. We get

$$-\tfrac{1}{2}\int \prod_{k\neq i} w_k(q_k) \left\{ \sum_{n\neq i} \frac{\partial}{\partial q_i}\left(M(q)^{-1}\right)_{nn} \underbrace{\int p_n^2 \bar{h}_n(p_n,q_n)\,\mathrm{d}p_n}_{\text{analytical solution?}} + \frac{\partial}{\partial q_i}\left(M(q)^{-1}\right)_{ii} p_i^2 \right\}\mathrm{d}\widehat{q}_i.$$

Indeed, there can be given an expression for underbraced integral, by using the notation introduced in the previous paragraph. Since the subsystems are one dimensional, the matrix $B(q)$ is simply a scalar, denoted by $b_n(q)$, indicating the dependence on n.

$$\int p_n^2 \bar{h}_n(p_n, q_n)\,\mathrm{d}p_n = \frac{1}{\beta}\frac{\int b_n(q)^{-3/2}\prod_k \hat{d}_k(q)^{-1/2}\prod_{j\neq n} e^{-\beta V_j(q_j)}\,\mathrm{d}\widehat{q_n}}{\int b_n(q)^{-1/2}\prod_k \hat{d}_k(q)^{-1/2}\prod_{j\neq n} e^{-\beta V_j(q_j)}\,\mathrm{d}\widehat{q_n}}.$$

Note, that this expression does *not* depend on q_i. Hence, if we fix the quadrature nodes (see below) for the integral $\int \ldots \mathrm{d}\widehat{q_i}$ above, these values can be computed in advance and stored in a "lookup table".

While we managed to compute the integrals w.r.t. $\mathrm{d}\widehat{p_i}$ analytically, the integrals w.r.t. $\widehat{q_i}$ need numerical treatment. Since in our approximation the w_i are piecewise constant functions, these integrals are computed by evaluating the integrand at the center points of the boxes and summing it up with an appropriate scaling.

To conclude, we have seen, that the originally $2(d-d_i)$ dimensional integral which defines $f_{i,\mathrm{MF}}$ can be simplified analytically, such that for the numerical evaluation of the right hand side only $d-d_i$ dimensional numerical quadratures are required.

Complexity. Let us first investigate the costs of setting up the discretized transfer operator for (an arbitrary) subsystem i. Using Ulam's method as for (2.33), we need to perform the following steps for each partition element B_j:

- fix quadrature nodes $q^\ell \in B_j$ and corresponding weights;
- for each q^ℓ, sample several $p^{\ell,n}$ according to $\bar{h}(q^\ell, \cdot)$;
- integrate the mean field system for time t and initial data $(q^\ell, p^{\ell,n})$; and
- project the endpoint onto the configuration space and find the partition element B_k it is contained in.

Using the canonical density for the invariant density h, there is an explicit representation for the momentum distributions $\bar{h}_i(q^\ell, \cdot)$ which can be sufficiently well approximated by a linear combination of Gaussians. The numerical time integration of the initial points requires several evaluations of the mean field vector field (6.25). This, in turn, requires the numerical evaluation of a $2(d-d_i)$ dimensional integral. The integral with respect to the $\widehat{p_i}$ can be handled analytically by an a priori computation which is

independent of the w_i, p_i, and q_i. Naively, this leaves us with a $d - d_i$ dimensional integral. However, note, that in the case of noninteracting subsystems, i.e., $f_i(z) = f_i(z_i)$, f_i can be pulled out and the integral reduces to 1. For systems with small subsystems, i.e. $d_i \leq \bar{d}$ and $\bar{d}$ small, and in which only a fixed and small number of neighboring subsystems interact (strongly), the dimensionality of the integral is $\sum_{j\sim i} d_j = \mathcal{O}\left(\bar{d}\right)$, where $j \sim i$ means all subsystems j which interact with subsystem i.

Remark 6.15. Observe, that the eigenfunctions of the spatial transfer operator seem to be smooth. Hence, it could be advantageous to use sparse grid quadrature for computing the integrals w.r.t. $\widehat{q_i}$ in $f_{i,\mathrm{MF}}$. This variant of the method has not been implemented yet.

The solution of the resulting eigenvalue problems is simple compared with the assembling of the discretized mean field transfer operator, particularly since d_i is small, and we are interested only in the dominant part of the spectrum. Arnoldi-type iteration methods can be used.

On the observation with the interacting subsystems above relies the hope to be able to use our methods for larger chain molecules. We neglect the *direct* inclusion of weak interactions, reducing the dimension of the integration domain. Indirectly, these interactions enter through the solution of the coupled eigenvalue problem.

6.3.3 Example: n-butane

We analyze the n-butane molecule, cf. Section 2.4.2. We decompose the model into three subsystems; i.e., each configuration variable is treated separately. As discussed above, we perform the Roothaan iteration to compute fixed points of the mean field spatial transfer operators, and having these, we compute eigenfunctions at eigenvalues near one for $\mathcal{S}_{\phi,\mathrm{MF}}[\widehat{w_\phi}]$; i.e. for the mean field spatial transfer operator corresponding to the ϕ-subsystem.

The Roothaan iteration is initialized with $w_i^0(q_i) := C_i e^{-\beta V_i(q_i)}$, $i = 1, 2, 3$, where β is the inverse temperature corresponding to 1000 K and C_i is a corresponding normalizing factor. We partition the (one dimensional) subsystems into 32 subintervals each. The entries of the transition matrix are computed as discussed after Equation (2.33); where a one-node Gauss quadrature is used with 32 Monte Carlo sample points to approximate the integral w.r.t. momenta.

We denote the computed mean field invariant marginals by w_{θ_1}, w_{θ_2} and w_ϕ, and the (other) eigenfunctions of $\mathcal{S}_{\phi,\mathrm{MF}}[\widehat{w_\phi}]$ with $v_{\phi,j}$, $j = 2, 3, \ldots$ Then $w_{\theta_2} \otimes w_\phi$, $w_{\theta_2} \otimes v_{\phi,2}$ and $w_{\theta_2} \otimes v_{\phi,3}$, shown in Figure 6.3, approximate the θ_2-ϕ-marginals of the first three eigenfunctions of the full spatial transfer operator, shown in Figure 2.4. The "rough"

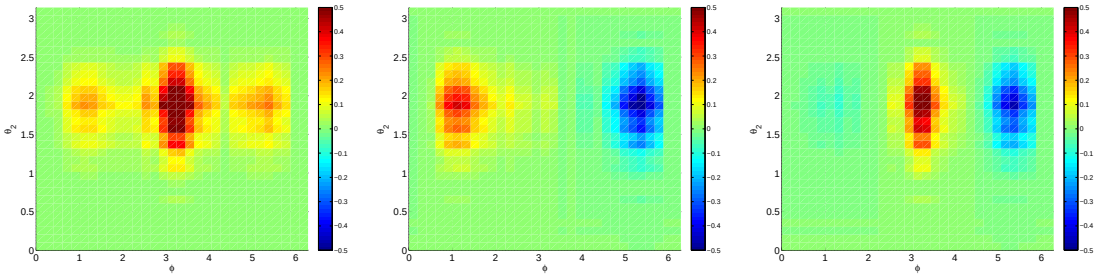


Figure 6.3: Mean field approximations to the marginals of the dominant eigenfunctions of the spatial transfer operator - the sign structure of the second and third eigenfunctions (middle and right, respectively) indicate the dominant configurations.

surface of the eigenfunctions indicate that the quadrature involved was not accurate enough. Because of the same reason, the second and third eigenfnuctions are swapped. Note, the results are qualitatively correct. This points out, that for a more efficient implementation of the mean field method other sampling strategies, or other discretization have to be applied. This will be the topic of future work.

6.4 Conclusions and outlook

We started our considerations with the aim of developing a method appropriate to describe the statistical evolution of subsystems of (large) coupled systems. We showed that, under certain regularity assumptions and weak coupling, the mean field model shows first order accuracy for the marginal densities; cf. Theorem 6.11. However, numerical experiments showed that if the full system invariant density can not be well approximated by tensor product functions (i.e. the long-term statistical behavior is not "decoupled"), then the mean field approximation of the marginal invariant densities is not adequate, or the Roothaan algorithm fails to converge to the right fixed point. To assess the real potential of the method, it would be desirable to show which of the above cases is responsible for the wrong results. Until then we have to take the worst

case into account, and conclude that the mean field approximation works well if the full system invariant density is having a "nearly tensor product structure".

The mean field description of classical MD systems in inner coordinates with standard force field shows very good qualitative results. A *quantitative* analysis, e.g. the comparison of the rates of conformation changes predicted by the mean field model with the rates computed by a suitable simulation, is the topic of future work. Further, one would like to have a theoretical explanation for the good performance, although the coupling (introduced by the momenta) between the subsystems is of order one.

Also, the extension of the method for larger chain molecules lies ahead. The interacting subsystems in the model have to be chosen such that we avoid the computation of high dimensional integrals. For long chain molecules, geometrical constraints have to be taken into account as well (the molecule may be folded, but two atoms are never allowed to come too close to each other). Hence, other potentials, like the Lennard–Jones potential, have to be included in the model.

References

[Agm65] S. Agmon. *Lectures on Elliptic Bonudary Value Problems.* Van Norstand Mathematical Studies 2, 1965.

[Ama83] H. Amann. "Dual Semigroups and Second Order Linear Elliptic Boundary Value Problems". *Israel Journal of Mathematics*, Vol. 45., No. 2–3, pp. 225–254, 1983.

[Arn65] V. Arnol'd. "Sur la topologie des écoulements stationnaires des fluides parfaits". *C. R. Acad. Sci. Paris*, Vol. 261, pp. 17–20, 1965.

[Aub82] T. Aubin. *Nonlinear Analysis on Manifolds. Monge–Ampère Equations.* Springer-Verl., 1982.

[Bab91] I. Babuška and J. Osborn. "Eigenvalue problems". In: *Handbook of Numerical Analysis, vol. 2*, pp. 641–787, Elsevier Science Publishers, North-Holland, 1991.

[Ben93] M. Benedicks and L.-S. Young. "Sinai–Bowen–Ruelle measures for certain Hénon maps". *Invent. Math.*, Vol. 112, pp. 541–576, 1993.

[Bir31] G. D. Birkhoff. "Proof of the ergodic theorem". *Proc. nat. Acad. Sci. U.S.A.*, Vol. 17, pp. 650–660, 1931.

[Bos01] C. J. Bose and R. Murray. "The exact rate of approximation in Ulam's method". *Disc. Cont. Dynam. Sys.*, Vol. 7, pp. 219–235, 2001.

[Boy01] J. P. Boyd. *Chebyshev and Fourier Spectral Methods.* Dover Publications, Inc., 2. Ed., 2001.

[Bro83] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". *Journal of Computational Chemistry*, Vol. 4, No. 2, pp. 187–217, FebruaryFebruary 1983.

[Bun04] H.-J. Bungartz and M. Griebel. "Sparse grids". *Acta Numerica*, Vol. 13, pp. 1–123, 2004.

[Can07] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zhang. *Spectral Methods in Fluid Dynamics.* Springer-Verl., 2007.

REFERENCES

[Del05] M. Dellnitz, O. Junge, W. S. Koon, F. Lekien, M. W. Lo, J. E. Marsden, K. Padberg, R. Preis, S. D. Ross, and B. Thiere. "Transport in Dynamical Astronomy and Multibody Problems". *J. of Bifurcation and Chaos*, Vol. 15, pp. 699–727, 2005.

[Del09] M. Dellnitz, G. Froyland, C. Horenkamp, K. Padberg-Gehle, and A. S. Gupta. "Seasonal variability of the subpolar gyres in the Southern Ocean: a numerical investigation based on transfer operators". *Nonlinear Processes in Geophysics*, Vol. 16, pp. 655–664, 2009.

[Del96] M. Dellnitz and A. Hohmann. "The computation of unstable manifolds using subdivision and continuation". In: H. W. Broer, S. A. van Gils, I. Hoveijn, and F. Takens, Eds., *Nonlinear Dynamical Systems and Chaos*, pp. 449–459, Birkhäuser, 1996.

[Del97] M. Dellnitz and A. Hohmann. "A subdivision algorithm for the computation of unstable manifolds and global attractors". *Numerische Mathematik*, Vol. 75, pp. 293–317, 1997.

[Del98] M. Dellnitz and O. Junge. "An adaptive subdivision technique for the approximation of attractors and invariant measures". *Comput. Visual. Sci.*, Vol. 1, pp. 63–68, 1998.

[Del99] M. Dellnitz and O. Junge. "On the approximation of complicated dynamical behavior". *SIAM J. Numer. Anal.*, Vol. 36, pp. 491–515, 1999.

[Deu01] P. Deuflhard, W. Huisinga, and C. Schütte. "Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems". In: *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pp. 191–223, Springer-Verl., 2001.

[Deu04a] P. Deuflhard and C. Schütte. "Molecular Conformation Dynamics and Computational Drug Design". In: *Applied Mathematics Entering the 21st Century*, pp. 91–119, SIAM, 2004.

[Deu04b] P. Deuflhard and M. Weber. "Robust Perron cluster analysis in conformation dynamics". *Linear Algebra Appl.*, Vol. 398, pp. 161–184, 2004. Special Issue on Matrices and Mathematical Biology.

[Deu96] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. "Computation of Essential Molecular Dynamics by Subdivision Techniques I: Basic Concept". In: *Computational Molecular Dynamics: Challenges, Methods, Ideas*, pp. 98–115, Springer-Verl., 1996.

[Din91] J. Ding and T.-Y. Li. "Markov finite approximation of the Frobenius-Perron operator". *Nonlin. Anal., Theory, Meth. & Appl.*, Vol. 17, pp. 759–772, 1991.

[Din93] J. Ding, Q. Du, and T.-Y. Li. "High Order Approoximation of the Frobenius–Perron Operator". *Applied Mathematics and Computation*, Vol. 53, pp. 151–171, 1993.

[Din96] J. Ding and A. Zhou. "Finite approximations of Frobenius-Perron operators. A solution of Ulam!s conjucture on multi-dimensional transformations". *Physica D*, Vol. 92, pp. 61–68, 1996.

[Dom86] T. Dombre, U. Frisch, M. Henon, J. M. Greene, and A. M. Soward. "Chaotic streamlines in the ABC flows". *J. of Fluid Mechanics*, Vol. 167, pp. 353–391, 1986.

[Doo60] J. L. Doob. *Stochastic Processes*. John Wiley, 1960.

[Eva98] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 1998.

[Fel71] W. Feller. *An introduction to probability theory and its applications*. Vol. 2., Wiley, 2. Ed., 1971.

[Foc30] V. A. Fock. "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems". *Zeitschrift für Physik*, Vol. 61, No. 1–2, pp. 126–148, 1930.

[Fri09] G. Friesecke, O. Junge, and P. Koltai. "Mean Field Approximation in Conformation Dynamics". *Multiscale Model. Simul.*, Vol. 8, pp. 254–268, 2009.

[Fro] G. Froyland, O. Junge, and P. Koltai. "Estimating long term behavior of flows without trajectory integration: the infinitesimal generator approach". in preparation.

[Fro03] G. Froyland and M. Dellnitz. "Detecting and locating near-optimal almost-invariant sets and cycles". *SIAM J. Sci. Comput.*, Vol. 24, No. 6, pp. 1839–1863, 2003.

[Fro05] G. Froyland. "Statistically optimal almost-invariant sets". *Physica D*, Vol. 200, pp. 205–219, 2005.

[Fro07] G. Froyland, K. Padberg, M. H. England, and A. M. Treguier. "Detection of Coherent Oceanic Structures via Transfer Operators". *Physical Review Letters*, Vol. 98, No. 22, 2007.

[Fro09] G. Froyland and K. Padberg. "Almost-invariant sets and invariant manifolds – connecting probabilistic and geometric descriptions of coherent structures in flows.". *Physica D*, Vol. 238, No. 16, pp. 1507–1523, 2009.

[Fro95] G. Froyland. "Finite Approximation of Sinai-Bowen-Ruelle Measures for Anosov Systems in Two Dimensions". *Random & Computational Dynamics*, Vol. 3, pp. 251–264, 1995.

[Fro96] G. Froyland. *Estimating Physical Invariant Measures and Space Averages of Dynamical Systems Indicators*. PhD thesis, University of Western Australia, 1996.

[Gav06] B. Gaveau and L. S. Schulman. "Multiple phases in stochastic dynamics: Geometry and probabilities". *Phys. Rev. E*, Vol. 73, No. 3, 2006.

[Gav98] B. Gaveau and L. S. Schulman. "Theory of nonequilibrium first-order phase transitions for stochastic dynamics". *J. Math. Phys.*, Vol. 39, No. 3, pp. 1517–1533, 1998.

[Giu84] E. Giusti. *Minimal Surfaces and Functions of Bounded Variation*. Vol. 80 of *Monographs in Mathematics*, Birkhäuser, 1984.

[Gol04] S. Goldschmidt, N. Neumann, and J. Wallaschek. "On the Application of Set-Oriented Numerical Methods in the Analysis of Railway Vehicle Dynamics". In: *ECCOMAS 2004*, 2004.

[Gor84] P. Góra. "On small stochastic perturbations of mappings of the unit interval". *Colloq. Math.*, Vol. 49, pp. 73–85, 1984.

[Gri07] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics*. Vol. 5 of *Texts in Computational Science and Engineering*, Springer, Berlin, Heidelberg, 2007.

[Gri99] M. Griebel, P. Oswald, and T. Schiekofer. "Sparse grids for boundary integral equations". *Numerische Mathematik*, Vol. 83, No. 2, pp. 279–312, 1999.

[Guc83] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verl., 1983.

[Gud97] R. Guder, M. Dellnitz, and E. Kreuzer. "An adaptive method for the approximation of the generalized cell mapping". *Chaos, Solitons and Fractals*, Vol. **8**, No. 4, pp. 525–534, 1997.

[Hai06] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*. Springer-Verl., 2 Ed., 2006.

[Hai96] E. Hairer and C. Lubich. "The Life-Span of Backward Error Analysis for Numerical Integrators". *Numer. Math*, Vol. 76, pp. 441–462, 1996.

[Har28] D. R. Hartree. "The wave mechanics of an atom with a non-Coulomb central field". *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 24, pp. 89–132, 1928.

[Hig02] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, USA, second Ed., 2002.

[Hsu87] C. S. Hsu. *Cell-to-Cell Mapping. A Method of Global Analysis for Nonlinear Systems*. Springer-Verl., 1987.

[Hui06] W. Huisinga and B. Schmidt. "Metastability and Dominant Eigenvalues of Transfer Operators". In: B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, and R. Skeel, Eds., *New Algorithms for Macromolecular Simulation*, pp. 167–182, Springer Berlin Heidelberg, 2006.

[Hun94] F. Y. Hunt. "A Monte Carlo approach to the approximation of invariant measures". *Random Comput. Dynam.*, Vol. 2, No. 1, pp. 111–133, 1994.

[Jun04] O. Junge, J. E. Marsden, and I. Mezic. "Uncertainty in the dynamics of conservative maps". In: 43*rd IEEE Conference on Decision and Control*, pp. 2225–2230, 2004.

[Jun09] O. Junge and P. Koltai. "Discretization of the Frobenius–Perron Operator Using a Sparse Haar Tensor Basis: The Sparse Ulam Method". *SIAM J. Numer. Anal.*, Vol. 47, pp. 3464–3485, 2009.

[Kat84] T. Kato. *Perturbation Theory for Linear Operators.* Springer-Verl., 2. Ed., 1984.

[Kha63] R. Z. Khas'minskii. "Principle of averaging for parabolic and elliptic differential equations and for Markov processes with small diffusion". *Theory of Probability and its Applications*, Vol. 8, pp. 1–21, 1963.

[Kif86] Y. Kifer. "General random perturbations of hyperbolic and expanding transformations". *Journal D'Analyse Mathématique*, Vol. 47, pp. 111–150, 1986.

[Kro97] D. Kröner. *Numerical Schemes for Conservation Laws.* Wiley & Teubner, 1997.

[Las94] A. Lasota and M. C. Mackey. *Chaos, Fractals, and Noise.* Springer-Verl., 2. Ed., 1994.

[LeV02] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University Press, 2002.

[Li76] T.-Y. Li. "Finite approximation for the Frobenius-Perron operator. A solution to Ulam's conjecture". *J. Approx. Theory*, Vol. 17, pp. 177–186, 1976.

[Lor63] E. N. Lorenz. "Deterministic Nonperiodic Flow". *J. Atmos. Sci.*, Vol. 20, pp. 130–141, 1963.

[Lun95] A. Lunardi. *Analytic Semigroups and Optimal Regularity in Parabolic Problems.* Birkhäuser, 1995.

[Mur97] R. Murray. *Discrete approximation of invariant densities.* PhD thesis, University of Cambridge, 1997.

[Nor97] J. R. Norris. *Markov Chains.* Cambridge Univ. Press, 1997.

[Osb75] J. E. Osborn. "Spectral approximation for compact operators". *Math. Comp.*, Vol. 29, pp. 712–725, 1975.

[Pav08] G. A. Pavliotis and A. M. Stuart. *Multiscale Methods.* Springer-Verl., 2008.

[Paz83] A. Pazy. *Semigroups of linear operators and applications to partial differential equations.* Springer-Verl., 1983.

[Qua00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerische Mathematik.* Vol. 1, Springer-Verl., 2000.

[Sch99] C. Schütte. "Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules". 1999. Habilitation Thesis, FU Berlin.

[Smo63] S. Smolyak. "Quadrature and interpolation formulas for tensor products of certain classes of functions". *Soviet Math. Dokl.*, Vol. 4, pp. 240–243, 1963.

REFERENCES

[Sta07] O. Stancevic. *Transfer operator methods in continuous time dynamical systems.* Honours thesis, University of New South Wales, 2007.

[Tre00] L. N. Trefethen. *Spectral Methods in MATLAB.* SIAM, 2000.

[Tre90] A. M. Treguier and J. C. McWilliams. "Topographic influences on wind-driven, stratified flow in a β-plane channel: An idealized model for the Antarctic Circumpolar Current". *J. Phys. Oceanogr.*, Vol. 20, No. 3, pp. 321–343, 1990.

[Tre94] A. M. Treguier and R. L. Panetta. "Multiple Zonal Jets in a Quasigeostrophic Model of the Antarctic Circumpolar Current". *J. Phys. Oceanogr.*, Vol. 24, No. 11, pp. 2263–2277, 1994.

[Tre97] L. N. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, 1997.

[Tuc99] W. Tucker. "The Lorenz attractor exists". *C. R. Acad. Sci. Paris*, Vol. 328, pp. 1197–1202, 1999.

[Ula60] S. M. Ulam. *A Collection of Mathematical Problems.* Interscience Publisher NY, 1960.

[War10] T. Wartewig. *Das Spektrum des Frobenius–Perron Operators im Fall schwach-gekoppelter Abbildungen.* Bachelor's Thesis, Technische Universität München, 2010.

[Web07] M. Weber, S. Kube, L. Walter, and P. Deuflhard. "Stable Computation of Probability Densities for Metastable Dynamical Systems". *Multiscale Model. Simul.*, Vol. 6, No. 2, pp. 396–416, 2007.

[You02] L.-S. Young. "What Are SRB Measures, and Which Dynamical Systems Have Them?". *Journal of Statistical Physics*, Vol. 108, pp. 733–754, 2002.

[Zee88] E. C. Zeeman. "Stability of dynamical systems". *Nonlinearity*, Vol. 1, pp. 115–155, 1988.

[Zen91] C. Zenger. "Sparse grids". In: *Parallel algorithms for partial differential equations (Kiel, 1990)*, pp. 241–251, Vieweg, Braunschweig, 1991.

[Zho98] H.-X. Zhou, S. T. Wlodek, and J. A. McCammon. "Conformation gating as a mechanism for enzyme specificity". *Proc. Natl. Acad. Sci. USA*, Vol. 95, pp. 9280–9283, 1998.