Computation Schemes for Transfer Operators

Reasons to believe in your computational results

Adam Nielsen

Dissertation

Eingereicht am Fachbereich Mathematik und Informatik der Freien Universität Berlin zur Erlangung des Grades eines Doktors der Naturwissenschaften.
Berlin, Oktober 2015

Adam Nielsen
Computation Schemes for Transfer Operators
Dissertation, 20 Oktober, 2015

Betreuer: PD Dr. Marcus Weber
1. Gutachter: PD Dr. Marcus Weber
2. Gutachter: Prof. Dr. Reinhold Schneider

Tag der Disputation: 31.03.2016

Acknowledgements

I would like to express my deepest gratitude to my advisor, PD. Dr. Marcus Weber, for his inspiring ideas, excellent guidance, caring, patience; and for introducing me to this exciting field, providing me with an outstanding atmosphere for doing research and giving me the opportunity to write this thesis. I am also deeply thankful to Prof. Dr. Christof Schütte for his support, his caring, insightful discussions, useful critiques and enthusiastic encouragement.

I would like to thank my office mate Jannes who was always open to a long discussion at the black board. I also thank the members of the research group *Computational Drug Design* from the Konrad Zuse Institute, the members of the *Berlin Mathematical School*, and the members of the *BioComputing Group* for this pleasant research environment. I thank Svenja for making this phase so much more enjoyable.

I would also like to express my gratitude to the *Sonderforschungsbereich 765* and the *Berlin Mathematical School* for their financial support.

Finally, I wish to thank my sister Lilli and my parents, Nandana and Karl, for everything. You were there for me when I needed you most. Thank you.

Contents

Introduction

> *Even if you fall on your face, you're still moving forward.*
>
> **— Victor Kiam**

The original motivation for this thesis was to address questions in the field of computational drug design. The main objective was how to evaluate whether a given ligand matches well to a malicious receptor molecule to inhibit the latter's biological activity. This required the development of methods for analyzing molecules. However, almost all the findings presented in this thesis are of such a general nature that they can be applied to any moving objects.

The topic concerned, the "analysis of moving objects", falls within the field of ergodic theory, which was originally developed in the year 1930 [6, 39]. At that time, the main concern was the long-term behavior of an object. Given data for a moving object, how could the probability of the object being found in a certain area be computed? Or, in other words, how could the stationary measure of the system be computed? In the field of drug design, a ligand is considered to be a good inhibitor if it binds to a receptor and remains there. This leads to the different question of how to find the so-called *metastable sets*, i.e. areas where there is a high probability that the object will remain. In order to clarify why this is actually a different question to that of computing the stationary measure, let us consider the following example. Suppose that our moving object is the German population, and consider as areas a supermarket and a jail. Since most people are usually more often in a supermarket then in jail, the stationary probability of being in a supermarket is much higher than that of being in jail. At the same time, the probability of remaining in prison once one has arrived there is much higher than the probability of staying in a supermarket.

For reversible processes, a machinery has been developed in recent decades [52] for extracting metastable sets from data using the clustering method *Robust Perron Cluster Analysis* (PCCA+) [14]. Reversible processes are described by the property that they keep the same probability law even if their movement is considered backwards in time. Many models for molecules turn out to be reversible. The machinery makes use of an old tool, known as the transfer operator. This continues to be the most modern and eloquent description of moving objects. Its universal ability to describe stochastic and deterministic processes for finite measures was developed in 1954 by Hopf [23]. Although this tool has now been known for quite some time, the author of this thesis is able to bring some

crucial properties of this operator to light. One of these is that transfer operators and Markov operators are equivalent. This is a slightly extension of a theorem proved by Foguel [20]. In addition, the author also characterizes the class of adjoint transfer operators with arbitrary measure. Finally, the author characterizes the class of adjoint transfer operators with invariant measures. The last characterization has not only remained undiscovered, but was actually denied by Brown in the year 1966 [10]. In short, the author gives a neat characterization for transfer and adjoint transfer operators in the first chapter.

After identifying what a transfer operator really is, we will be confronted by the problem of how to obtain a Galerkin projection of the transfer operator. A Galerkin projection is a matrix representation of a projection from the transfer operator to a finite dimensional space. The first steps in this direction were taken by Ulam in 1960 [57], and it is still the subject of ongoing research [53]. The entries of the Galerkin projection can be approximated by using Monte Carlo methods by a short-term trajectory and a long-term trajectory approach. For both concepts, the author is able to find an analytical expression of the error by the difference in the L^2 norm between the true Galerkin entry and the Monte Carlo approximation, depending on the number of random variables used. The analytical expression of the error can again be approximated to reveal the error of the entries from the Galerkin projection. As a byproduct, a universal property for reversible processes is revealed, showing that reversible processes are more likely to return to a set than to stay there. This characteristic had remained undiscovered from 1935, when Kolmogorov [16] first introduced reversible processes, until now. Further, the author provides a reweighting scheme, which was developed together with Christof Schütte and Marcus Weber, that offers a solution to one of the main problems in computational drug design, namely: How to test several similar ligands on a single, very large receptor molecule. Such computations are very cost intensive and it is unfortunate that for each new ligand a complete new Galerkin projection is necessary, even though the systems are very similar. In the proposed reweighting scheme, we exploit the fact that the systems are similar, and are able to compute a Galerkin projection of a system by reusing the trajectories from a different system. The author was able to extend this reweighting scheme to be usable even in cases where the ligands' dimensions differ.

Finally, we come to the main result of this thesis, which tackles one of the major problems of computing a Galerkin projection, namely the numerical error which occurs when a Galerkin projection is computed. Specifically, for a reversible process the exact Galerkin projection has a reversible property which guarantees that the matrix has real eigenvalues and eigenvectors. These real eigenvalues and eigenvectors are essential in order to identify the metastable sets with PCCA+. The computation of a Galerkin projection for a large molecule is extremely challenging, and the computed matrix all too easily loses its reversible property and returns complex eigenvalues and eigenvectors due to numerical errors. In such cases, identification of the metastable sets is not possible with PCCA+. The author shows that for each Galerkin projection we can find a closest matrix which maintains the reversible property and thus provides us with real eigenvalues and eigenvectors. The computation can be obtained by solving a convex quadratic minimization problem. Thus, regardless of how severely the computed Galerkin projection is influenced by numerical errors, it can be corrected to restore its reversible property and to enssure that it possesses real eigenvalues and eigenvectors, in order to be able to use PCCA+ for the identification of metastable sets. As a result, we succeed in devising a solid computation scheme for metastable sets for arbitrary processes.

Thesis Structure

Chapter 1 - Operators

We give a brief overview of relevant tools from measure theory, Markov chains on a measurable state space, and ergodic theory, in order to introduce rigorously the transfer operator. We show that the transfer operator propagates probability densities of a Markov chain on a measurable state space. Then we show that transfer operators are identical to Markov operators, and how they are related to sub Markov operators. Finally, we characterize adjoint transfer operators and relate them to Brown-Markov operators. This reveals a connection that has been denied for the past 50 years.

Chapter 2 - Computation Schemes

We start this chapter with a summary of the current methods for computing metastable sets and approximating the transfer operator. This leads to the task of computing certain entries of a Galerkin projection matrix. We describe how to approximate these entries and we give an exact formula for the error. As a byproduct, this reveals an interesting characteristic of reversible processes. Finally, we introduce a new reweighting scheme, showing how trajectories for a previous Galerkin projection can be used to receive a new one if the dynamical system is changed.

Chapter 3 - Make it Reversible

We show that every Galerkin projection of a reversible process has a certain property which is used in order to find the metastable sets. Unfortunately, this property is sometimes lost because of numerical estimation errors. In this chapter, we show that for any given stochastic matrix, and any norm induced by a scalar product, there exists a unique closest matrix that maintains this important property. We proof the theoretical existence and uniqueness, and we show how this recovered matrix can be found with the help of a convex optimization scheme. In addition, we introduce a specific norm such that the corresponding closets matrix will also preserver the spectrum.

This machinery enables use of the clustering method PCCA+ for arbitrary processes.

1 Operators

> *Poetry is the art of giving different names to the same thing - Mathematics is the art of giving the same name to different things.*
>
> **— Henri Poincaré**

In order to understand dynamical systems, it is fundamental to understand how probability densities evolve according to a dynamical system. The operator that transports associated probability densities is called the transfer operator. This operator was identified by Hopf in 1954, and has received particular attention in recent decades because of his application to analyze molecules [52]. The transfer operator will be comprehensively explicated in this chapter. Also, we will identify the relations between transfer, Markov, generalized Koopman and Brown-Markov operators.

1.1 Fundamentals

In order to introduce the main object of this thesis, namely the transfer operator, some mathematical tools are needed. Therefore, we start with a glimpse of one of the most beautiful areas of mathematics, which has its roots in the late 19th century.

Measure Theory

It all begins with the definition of a σ-algebra which was first introduced by E. Borel in 1933 [30, II. §2]. We denote with E an arbitrary set and with $\mathcal{P}(E) = \{A \mid A \subseteq E\}$ the *power set* of E.

Definition 1.1.1. A family of sets $\Sigma \subseteq \mathcal{P}(E)$ is called a *σ-algebra* of E if:

- $E \in \Sigma$ holds.
- For each $A \in \Sigma$ we have $E \backslash A \in \Sigma$.
- For any sequence $(A_k)_{k\in\mathbb{N}}$ with $A_k \in \Sigma$ we have $\bigcup_{k\in\mathbb{N}} A_k \in \Sigma$.

The beauty of the concept of an abstract integral is that it has a wide range of properties but only spartan requirements. We only need a σ-Algebra and a corresponding measure for its definition.

Definition 1.1.2. A map $\mu\colon \Sigma \to \mathbb{R}$ is called a *signed measure* on a σ-algebra Σ if

- $\mu(\emptyset) = 0$ and

- For any sequence $(A_k)_{k\in\mathbb{N}}$ with $A_k \in \Sigma$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, it follows that $\mu(\bigcup_{k\in\mathbb{N}} A_k) = \sum_{k\in\mathbb{N}} \mu(A_k)$ holds.

If μ is a non-negative function, then it is called a *measure*.

Since a σ-algebra together with a measure will build the fundament for the upcoming abstract integral, we will give it a name.

Definition 1.1.3. A touple (E, Σ) is called a *measurable space* if Σ is a *σ-algebra* of E. A triple (E, Σ, μ) is called a *measure space* if (E, Σ) is a measurable space and μ is a measure on Σ. A measure space is called a *σ -finite measure space* if there exists a sequence $(A_i)_{i\in\mathbb{N}}$, $A_i \in \Sigma$, satisfying

$$E = \bigcup_{i\in\mathbb{N}} A_i \text{ and } \mu(A_i) < \infty \text{ for all } i \in \mathbb{N}.$$

A measure space is called a *probability space* if $\mu(E) = 1$.

Throughout this thesis we will always consider σ-finite measure spaces. This stipulation may appear minor, but it is no exaggeration to state that the whole theory which will be developed in this chapter rest upon it. One may note that any finite measure is in particular σ-finite, and that $\mathbb{R}^d$ together with the Lebesgue Measure is σ-finite with $A_i = [-i, i]^d$. Finally, in order to define the abstract integral, a certain class of functions will play an essential role.

Definition 1.1.4. Denote with (E, Σ, μ) a measure space and with (M, Σ_M) a measurable space. A function $f\colon E \to M$ is called measurable if

$$A \in \Sigma_M \Rightarrow f^{-1}(A) \in \Sigma$$

holds. If (E, Σ, μ) is a probability space, then f is called random variable.

In the following, we denote with (E, Σ, μ) a σ-finite measure space and with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ either the real or complex numbers. We denote with $\mathcal{B}(\mathbb{K})$ the *Borel σ-algebra* on $\mathbb{K}$, i.e. the smallest σ-algebra of $\mathbb{K}$ which contains all open sets. For any measurable function $f\colon E \to \mathbb{K} \cup \{+\infty, -\infty\}$ one can define the Lebesgue integral [3, Definition 12.1]

$$\int_E |f(x)|\, \mu(dx) \tag{1.1}$$

which is a value in $\mathbb{R} \cup \{\infty\}$. When the integral (1.1) is finite, f is called μ-integrable and in this case the symbol $\int_E f(x)\, \mu(dx)$ can also be assigned to a value in $\mathbb{K}$. We will denote for a set $A \in \Sigma$ with

$$\mathbb{1}_A\colon E \to \{1, 0\}, \quad \mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else,} \end{cases}$$

the characteristic function of A. For a set $A \in \Sigma$, we define the integral over a set A for μ-integrable function f as $\int_A f(x)\, \mu(dx) := \int_E f(x)\, \mathbb{1}_A(x)\, \mu(dx)$. For a probability space (E, Σ, μ) we denote the expectation value for a random variable $Y\colon E \to \mathbb{R}$ as

$$\mathbb{E}_\mu[Y] = \int_E Y(x)\, \mu(dx) \tag{1.2}$$

if it exists. We will use the abbreviation

$$\mathbb{E} := \mathbb{E}_{\mathbb{P}}$$

whenever we want to denote the expectation according to the probability measure $\mathbb{P}$. We will often use that for a random variable $X\colon \Omega \to E$ on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which is distributed according to μ, i.e. $\mathbb{P}[X \in A] = \mu(A)$ for all $A \in \Sigma$, we have that for any measurable function $f\colon E \to [0,\infty]$ it holds that

$$\mathbb{E}[f(X)] = \int_E f(x)\,\mu(dx), \tag{1.3}$$

this follows from [3, Proposition 19.1]. Although it might be unclear how to imagine an abstract integral on some set E with an arbitrary measure, it still preserves many pleasant convergence properties which are true for the Lebesgue integral on $\mathbb{R}^d$, including some that turned out to be wrong for the ancient Riemann integral. Two measurable functions $f, g\colon E \to \mathbb{R}$ are said to be μ-almost surely equal if there exists a set $A \in \Sigma$ with $\mu(A) = 0$ and

$$f(x) = g(x) \quad \text{for all } x \in E\backslash A.$$

Proposition 1.1.5 ([34, Proposition 2.1.1])**.** *Given two measurable function $f, g\colon E \to \mathbb{R}$, then*

$$\int_A f(x)\,\mu(dx) = \int_A g(x)\,\mu(dx) \text{ for all } A \in \Sigma \quad \Leftrightarrow \quad f = g \quad \mu\text{-almost surely}.$$

This implies the following useful property.

Proposition 1.1.6. *For a measurable function $f\colon E \to \mathbb{R}$ we have*

$$\int_A f(x)\,\mu(dx) \geq 0 \quad \text{for all } A \in \Sigma \quad \Leftrightarrow \quad f \geq 0 \quad \mu\text{-almost surely}.$$

Proof. The implication from right to left follows from the monotonicity of the integral [3, Proposition 12.4, Property (12.3)]. For the other implication, it suffices to show that we have $\mu(A) = 0$ for $A := \{x \mid f(x) < 0\}$. From the precondition, we have $\int_B f(x)\,\mathbb{1}_A(x)\,\mu(dx) \geq 0$ for all $B \in \Sigma$ and from the monotonicity of the integral we get

$$\int_B f(x)\,\mathbb{1}_A(x)\,\mu(dx) = 0 = \int_B 0\,\mu(dx)$$

for all $B \in \Sigma$. By Proposition 1.1.5 we have $f(x) \cdot \mathbb{1}_A(x) = 0$ μ-almost surely and therefore $\mu(A) = 0$. □

Another often used property is the following.

Proposition 1.1.7 ([3, Proposition 13.2])**.** *For a measurable, non-negative function $f : E \to [0,\infty]$ it holds*

$$\int_E f(x)\,\mu(dx) = 0 \quad \Leftrightarrow \quad f = 0 \quad \mu\text{-almost surely}.$$

Since the integral is well-defined for any non-negative function, one can derive the following useful rule.

Proposition 1.1.8 (Monotone convergence theorem, [3, Proposition 11.4])**.** *For any monotonically increasing sequence of measurable, non-negative functions $(f_n)_{n\in\mathbb{N}}$ we have that $\lim_{n\to\infty} f_n(x)$ is μ-integrable and it holds*

$$\int_E \lim_{n\to\infty} f_n(x)\,\mu(dx) = \lim_{n\to\infty} \int_E f_n(x)\,\mu(dx).$$

We denote a sequence of functions (f_n) which converge monotonically increasing to a function f by

$$f_n \uparrow f.$$

A function $f\colon E \to \mathbb{R}$ is called *simple* if one can write it as

$$f(x) = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$$

for $\alpha_i \in \mathbb{R}$ and $A_i \in \Sigma$ for $i = 1, \dots, n$. The monotone convergence theorem is one of the most important theorems in measure theory, because for any non-negative and measurable function, one can find a sequence to which this theorem is applicable. As will be seen, this will become one of the major tools in the upcoming proofs.

Proposition 1.1.9 ([3, Proposition 11.6]). *For any non-negative and measurable function $f\colon E \to \mathbb{R}$ exists a sequence of simple functions $(f_n)_{n\in\mathbb{N}}$ such that*

$$f_n \uparrow f$$

holds.

If we look at a sequence of functions which are not necessarily non-negative and monotonic, then one cannot apply the monotone convergence theorem. However, the following theorem shows that one can obtain the same result by requiring instead an integrable and bounded sequence of functions.

Proposition 1.1.10 (Dominated convergence theorem, [3, Proposition 15.6]). *Given a convergent sequence of measurable functions $(f_n)_{n\in\mathbb{N}}$ which are bounded by an integrable function g, i.e. $|f_n| \leq g$ μ-almost surely, then $\lim_{n\to\infty} f_n$ is μ-integrable and*

$$\int_E \lim_{n\to\infty} f_n(x)\, \mu(dx) = \lim_{n\to\infty} \int_E f_n(x)\, \mu(dx)$$

holds.

The next proposition is the essential tool for all theorems in this chapter. This theorem is only valid because we required a σ-finite measure space (E, Σ, μ).

Proposition 1.1.11 (Radon-Nikodym theorem, [3, Proposition 17.10]). *Denote with $\nu\colon \Sigma \to [0, \infty]$ another measure. Then, there exists a measurable function $g\colon E \to [0, \infty]$ with*

$$\nu(A) = \int_A g(x)\, \mu(dx)$$

for every $A \in \Sigma$ if and only if $\mu(A) = 0$ implies $\nu(A) = 0$ for every $A \in \Sigma$.

The Radon-Nikodym theorem provides us with the following useful tool.

Definition 1.1.12. Denote with $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and with $\mathcal{A}' \subseteq \mathcal{A}$ another σ-algebra from Ω. A function $\phi\colon \Omega \to \mathbb{R}$ is called the *conditional expectation* of a random variable $Y\colon \Omega \to \mathbb{R}$ under the condition $\mathcal{A}'$ if

- ϕ is $\mathcal{A}'$ measurable
- for all $A' \in \mathcal{A}'$ we get $\int_{A'} Y(\omega)\mathbb{P}(d\omega) = \int_{A'} \phi(\omega)\mathbb{P}(d\omega)$.

We denote the conditional expectation then by $\mathbb{E}[Y \mid \mathcal{A}'] := \phi$.

It follows from the Radon-Nikodym theorem that such a function ϕ always exists and that it is unique [4, Chapter IV]. Thus the symbol $\mathbb{E}[Y \mid \mathcal{A}']$ is well defined.

For the coming Markov and transfer operators, we will denote for $1 \leq p < \infty$ with

$$L^p(E, \Sigma, \mu) := \{f \colon E \to \mathbb{K} \mid f \text{ is measurable, } ||f||_{L^p(\mu)} < \infty\}/N,$$
$$||f||_{L^p(\mu)} := \left(\int_E |f(x)|^p \mu(dx)\right)^{\frac{1}{p}}$$

and with

$$L^\infty(E, \Sigma, \mu) := \{f \colon E \to \mathbb{K} \mid f \text{ is measurable, } ||f||_\infty < \infty\}/N,$$
$$||f||_\infty := \inf_{A \in \Sigma, \mu(A)=0} \sup_{x \in E \setminus A} |f(x)|,$$

the Lebesgue spaces, where N is given by

$$N = \{f \colon E \to \mathbb{K} \mid \exists A \in \Sigma \text{ with } \mu(A) = 0 \text{ and } f|_{E \setminus A} = 0\}.$$

In the following we neglect E and Σ and write only $L^p(\mu)$ instead of $L^p(E, \Sigma, \mu)$. If it is clear from the context which measure μ is used, we use the abbreviation

$$||f||_p := ||f||_{L^p(\mu)}.$$

Due to Holder's inequality, the term

$$\langle f, g \rangle_\mu = \int_E f(x)\, \overline{g(x)}\, \mu(dx)$$

is well-defined for $f \in L^p(\mu)$ and $g \in L^q(\mu)$ with $\frac{1}{q} + \frac{1}{p} = 1$, for $1 \leq p, q \leq \infty$, where $\frac{1}{\infty} := 0$.

Finally, since this thesis is about the computation of a projected transfer operator, it seems wise to introduce the orthogonal projection. This map lives on a Hilbert space H with scalar product $\langle , \rangle$, i.e. H is a complete vector space according to the norm $|| \cdot ||$ induced by the scalar product by $||x|| = \sqrt{\langle x, x \rangle}$. In this thesis, we consider in most cases the Hilber space $L^2(\mu)$ together with the scalar product $\langle , \rangle_\mu$ where μ is a probability measure.

Definition 1.1.13. Given a Hilbert space H associated with the norm $|| \cdot ||$ induced by the scalar product. For a closed subspace $D \subseteq H$ we call a surjective map $Q \colon H \to D$ an *orthogonal projection* if $Q^2 = Q$ and $\sup_{x \in H, ||x||=1} ||Qx|| = 1$ holds.

For any closed subspace D, a unique orthogonal projection Q exists. Further, if $\{e_1, \ldots, e_n\}$ is a basis of D with

$$\langle e_i, e_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{else,} \end{cases}$$

then

$$Qx = \sum_{i=1}^{n} \langle x, e_i \rangle\, e_i \tag{1.4}$$

holds [61, Proposition V.3.4, V.4.8, and V.5.9].

Markov Chains

We denote with (E, Σ) a measurable space for a given set E and with $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space.

Markov chains describe a class of stochastic processes without memory. In order to introduce them, we need to begin with the definition of a stochastic process.

Definition 1.1.14. A *stochastic process* with index set I and state space E is a family of random variables

$$(X_i)_{i \in I}, \quad X_i \colon \Omega \to E$$

for all $i \in I$.

The definition of Markov chains has evolved over the past decades. In 1953 the term Markov chain was used for stochastic processes with discrete or continuous index set I living on a countable or finite state space E, see Doob or Chung [17, 13]. However, many models of practical interest live on the state space $E = \mathbb{R}^d$ and thus it is more natural to consider Markov chains on a measurable state space. We will follow the modern theory as it is found in [37, 46, 18] which defines a Markov chain as a stochastic process with a discrete time parameter but living on a measurable state space. It turns out that many results for the countable state space carry over virtually unchanged to the measurable state space with the proofs remaining beautifully clean and simple.

We are especially interested in time-homogeneous Markov chains, in which independent of the time the probability to move from one state to another will remain fixed. Thus, a Markov chain is defined by a law of transitions between states, which will be decoded in a transition kernel. For example, if we have n states, and between two states i, j the probability for a transition is p_{ij}, then the matrix $[p_{ij}] = P \in \mathbb{R}^{n \times n}$ represents a transition kernel. If E is not finite, then the transition kernel cannot be represented by a matrix. In this case, we need the following definition.

Definition 1.1.15. We call

$$p \colon E \times \Sigma \to [0, \infty]$$

a *kernel* iff

- $A \mapsto p(x, A)$ is a measure on Σ for all $x \in E$,
- $x \mapsto p(x, A)$ is measurable on E for all $A \in \Sigma$,

we call p a *sub transition kernel* if and only if $p(x, E) \leq 1$, and we call p a *transition kernel* if and only if $p(x, E) = 1$ holds.

A transition kernel is sometimes also called a Markov kernel [4, §36].

In order to introduce Markov chains on a measurable state space, it is helpful to mention that with the symbol used for the integral according to a measure μ as defined in (1.1), we already made an abuse of notation. Because on the one hand, a measure $\mu \colon \Sigma \to [0, \infty]$ is defined for measurable sets $A \in \Sigma$ and evaluated by writing $\mu(A)$ and on the other hand we write $\int_E f(x)\, \mu(dx)$ for every μ integrable function f. Note that $\mu(dx)$ is simply part of a symbol to denote the corresponding integral and should not be confused with evaluating

μ on dx, since dx is not a measurable set. Consider for some transition kernel p and fixed $x \in E$ the measure

$$\begin{aligned} \mu'_x \colon \Sigma &\to [0,1] \\ A &\mapsto p(x,A). \end{aligned}$$

Then, we can write:

$$\int_E f(y)\, p(x,dy) := \int_E f(y)\, \mu'_x(dy).$$

Equipped with this notation, we are now prepared to introduce Markov chains. There is a standard approach to construct Markov chains [37, 46, 18] that we will describe in the following. First, consider the product space

$$\prod_{i\in\mathbb{N}} A_i := \{(x_i)_{i\in\mathbb{N}} \mid x_i \in A_i \text{ for all } i \in \mathbb{N}\}.$$

For a general state space (E,Σ), one defines the probability space as $\Omega := \prod_{i\in\mathbb{N}} E$ together with a σ-algebra $\mathcal{F} := \sigma(M)$ where

$$M = \left\{ \prod_{i\in\mathbb{N}} E_i \mid E_i \in \Sigma, \text{ and there exists a } N \in \mathbb{N} \text{ with } E_i = E \text{ for all } i \geq N \right\}.$$

Then, we define for each $n \in \mathbb{N}$ a random variable X_n as follows. For $\omega = (x_i)_{i\in\mathbb{N}} \in \Omega$ we set

$$X_n(\omega) = x_n.$$

Now for each transition kernel p and probability distribution μ we can construct a probability measure $\mathbb{P}_\mu$ in the following sense.

Proposition 1.1.16 ([46, Theorem 2.8 and Proposition 2.10])**.** *For any transition kernel p and probability measure μ exists a probability measure $\mathbb{P}_\mu$ with*

$$\begin{aligned} &\mathbb{P}_\mu[X_0 \in A_0,\, X_1 \in A_1, \ldots, X_n \in A_n] \\ &\quad = \int_{A_0} \cdots \int_{A_{n-1}} p(y_{n-1}, A_n)\, p(y_{n-2}, dy_{n-1}) \ldots p(y_0, dy_1)\, \mu(dy_0) \end{aligned} \tag{1.5}$$

for any $n \in \mathbb{N}$, $A_0, \ldots, A_n \in \Sigma$.

Such a stochastic process will be called a time-homogeneous Markov chain, because the evolution from X_n to X_{n+1} always behaves in accordance with the probability measure $p(X_n, \cdot)$ for all $n \in \mathbb{N}$.

Definition 1.1.17. A stochastic process $(X_n)_{n\in\mathbb{N}}$ on $(\Omega, \mathcal{F}, \mathbb{P}_\mu)$ is called a time-homogeneous Markov chain with transition kernel p and initial distribution μ if (1.5) is satisfied. We denote this Markov chain by $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_\mu)$

We now extend our notation from (1.2) for the expectation value. For a Markov chain $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_\mu)$ we denote for a random variable $Z \colon \Omega \to \mathbb{R}$ the expectation value as

$$\mathbb{E}_{[\mu]}[Z] := \mathbb{E}_{\mathbb{P}_\mu}[Z] = \int_\Omega Z(\omega)\, \mathbb{P}_\mu(d\omega) \tag{1.6}$$

if the integral exists. If μ is a dirac delta measure in x, i.e.

$$\mu(A) = \begin{cases} 1 & \text{, if } x \in A \\ 0 & \text{, else} \end{cases}$$

we write $\mathbb{P}_x$ and $\mathbb{E}_x$ instead of $\mathbb{P}_\mu$ or $\mathbb{E}_{[\mu]}$. The expectation value and the transition kernel p have the following relation for a measurable function $f \geq 0$

$$\int_E f(y)\, p(x, dy) = \mathbb{E}_x[f(X_1)], \tag{1.7}$$

this follows from [46, Chapter 1, §2, (2.3)]. The beginning (X_0, X_1) of the Markov chain from the transition kernel p fulfills

$$\mathbb{P}_\mu[X_0 \in A, X_1 \in B] = \int_A p(x, B)\, \mu(dx) \tag{1.8}$$

for any $A, B \in \Sigma$. One can show that one can embed any two random variables $X, Y : \Omega \to E$ into the beginning of a Markov chain, i.e. there exists a Markov chain $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_\mu)$ with

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}_\mu[X_0 \in A, X_1 \in B]$$

for all $A, B \in \Sigma$, see [40, Theorem 2]. For a transition kernel p one can introduce a family of transition kernels $(p_n)_{n\in\mathbb{N}}$ by

$$p_{n+1}(x, A) := \int_E p_n(y, A)\, p(x, dy) \tag{1.9}$$

for $n \geq 1$ and $p_1 := p$. For the Markov chain $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_\mu)$ associated with the transition kernel p we have

$$\mathbb{P}_\mu[X_n \in B, X_0 \in A] = \int_A p_n(x, B)\, \mu(dx).$$

Similarly, one can define a Markov process for continuous time. In this situation, one requires a family of transition kernels $(p_t)_{t\in I}$ that fulfills

$$p_{s+t}(x, A) = \int_E p_t(x, A)\, p_s(x, dy)$$

and $p_0(x, A) := \mathbb{1}_A(x)$. We call such a family a *Markov family*. In addition, some requirements on (E, Σ) are needed. In particular, (E, Σ) must be a a *polish space*, i.e. (E, Σ) must be a separable, completely metrizable, topological space. Under these requirements, we can generalize the concept for Markov chains to any ordered index set I.

Proposition 1.1.18 ([4, Corollary 35.4 and Proposition 42.3])**.** *Let (E, Σ) be a polish space. For any Markov family $(p_t)_{t\geq 0}$ and probability measure μ exists a stochastic process $(X_t)_{t\geq 0}$ and probability measure $\mathbb{P}_\mu$ with*

$$\begin{aligned} &\mathbb{P}_\mu[X_{t_1} \in A_1,\, X_{t_2} \in A_2, \ldots, X_{t_n} \in A_{t_n}] \\ &\quad = \int_E \int_{A_1} \cdots \int_{A_{n-1}} p_{s_n}(y_{n-1}, A_n)\, p_{s_{n-1}}(y_{n-2}, dy_{n-1}) \ldots p_{s_2}(y_1, dy_2)\, p_{t_1}(y_0, dy_1)\, \mu(dy_0) \end{aligned} \tag{1.10}$$

with $s_i = t_i - t_{i-1}$.

We can now define a Markov process in a similar way.

Definition 1.1.19. A stochastic process $(X_t)_{t\geq 0}$ on $(\Omega, \mathcal{F}, \mathbb{P}_\mu)$ is called a time-homogeneous Markov process with transition kernel p and initial distribution μ if (1.10) is satisfied.

In particular, if $(X_t)_{t\geq 0}$ is a Markov process, then for any lag time τ, the subfamily $(X_{n\tau})_{n\in\mathbb{N}}$ is a Markov chain with transition kernel $p(x, A) := p_\tau(x, A)$.

Ergodic Theory

Ergodic theory is a branch of mathematics that studies the long-term behaviour of objects that move according to a given law. Probably the first result in this field was that of Henri Poincaré [44], which can be stated in the following measure theoretic form.

> Denote with (E, Σ, μ) a finite measure space and with $S\colon E \to E$ a map that satisfies $\mu(A) = \mu(S^{-1}(A))$ for all $A \in \Sigma$. Then for any $A \in \Sigma$ with $\mu(A) > 0$ it follows that μ-almost surely all points $x \in A$ have the property that $S^{n_k}(x) \in A$ where $(n_k) \subset \mathbb{N}$ is a monotonic sequence.

The condition $\mu(A) = \mu(S^{-1}(A))$ implies that μ is an invariant measure. This shows that regardless of how small the set A is, provided it has a positive invariant measure, almost all trajectories that start in A will return infinitely often.

The word *ergodic* is a combination of two Greek words: ergon (work) and odos (path), and was coined by Ludwig Boltzmann [7] at the time he stated his hypothesis:

> *"For large systems of interacting particles in equilibrium, the time average along a single trajectory equals the space average."* [1]

In the early days of ergodic theory, a trajectory was thought as the sequence of an iteration of a map $S\colon E \to E$ on a measurable space (E, Σ, μ), then, the time average for a trajectory starting in $x \in E$ was defined as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(S^k x\right)$$

and the space average could be defined if $0 < \mu(E) < \infty$ as

$$\frac{1}{\mu(E)} \int_E f(x)\, \mu(dx).$$

Thus, the hypothesis could be be transformed in the measure theoretic formulation

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(S^k x\right) = \frac{1}{\mu(E)} \int_E f(x)\, \mu(dx) \tag{1.11}$$

which, in general, is not correct. The first discovery of conditions under which these two quantites proved equal was developed forty years later, in 1931, by von Neumann and

[1] During the 1870s and 1880s, various forms of the ergodic hypothesis were used by Boltzmann. The here stated advanced formulation is inspired from [56].

Birkhoff. Their respective versions of the ergodic theorem are often credited as the birth of ergodic theory. Birkhoff's version [6] can be stated as follows.

> Given a probability space (E, Σ, μ) and a map $S\colon E \to E$ which satisfies $\mu(A) = \mu(S^{-1}(A))$ for all $A \in \Sigma$ and for each invariant set $A \in \Sigma$, i.e. $S^{-1}(A) = A$, follows $\mu(A) = 1$ or $\mu(A) = 0$, then the Hypothesis (1.11) holds for any μ-integrable function f.

The results were followed by a long list of publications in which authors tried to generalize the concept, dropping constraints on S and on μ. A well-written general overview of the development of ergodic theory between 1931 and 1948 is given by Halmos [22].

While Birkhoff's formulation was only concerned with what happens when a single trajectory converges, von Neumann's formulation focused on the behavior of the whole ensemble. For the moment, let us consider an operator U on a Hilbert space H with

$$\sup_{f \in H, ||f||=1} ||Uf|| \leq 1.$$

It will become clear in the next section why operators that propagate probability distributions fulfill this condition. Consider the space $D = \{h \in H \mid Uh = h\}$ with the orthogonal projection Q onto D from Definition 1.1.13. Von Neumanns's mean ergoric theorem [33, Theorem 1.4] states that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} U^n f = Qf \tag{1.12}$$

holds for all $f \in H$. Initially, von Neumann's theorem was only applied to operators associated with a deterministic system. Twenty years later, Hopf extended this framework to operators associated with Markov chains. This led to the rise of the transfer operator, which will be described in the next section.

1.2 Transfer Operators

We denote with (E, Σ, μ) a σ-additive measure space and with $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space.

We start with the motivation of the transfer operator for a deterministic system. In this case, Ulam [57] proposed calling this operator Frobenius-Perron operator, because he claimed a conjecture[2] about this operator that was motivated by the Frobenius-Perron theorem for matrices. The Frobenius-Perron operator is usually introduced by an integral equation, which can be motivated as follows. A deterministic system is given by a map $S\colon E \to E$. One can think of S as the deterministic movement from a point x to $S(x)$. Consider the random variable $X\colon \Omega \to E$ with probability density $f \in L^1(\mu)$, i.e $\mathbb{P}[X \in A] = \int_A f(x)\,\mu(dx)$ for $A \in \Sigma$. Then, the Frobenius-Perron operator $\mathcal{T}$ propagates the probability distribution in the following sense:

$$\mathbb{P}[S(X) \in A] = \int_A (\mathcal{T}f)(x)\,\mu(dx).$$

[2] See Chapter 2.

Also, we get

$$\mathbb{P}[S(X) \in A] = \mathbb{P}[X \in S^{-1}(A)] = \int_{S^{-1}(A)} f(x)\,\mu(dx).$$

This motivates the original definition of the Frobenius-Perron operator, which is given by the solution of the integral equation

$$\int_A (\mathcal{T}f)(x)\,\mu(dx) = \int_{S^{-1}(A)} f(x)\,\mu(dx) \tag{1.13}$$

for all $A \in \Sigma$ and all μ-integrable functions f. It follows from the Radon-Nykodým Theorem and Proposition 1.1.5 that this integral equation possess a unique solution

$$\mathcal{T}\colon L^1(\mu) \to L^1(\mu),$$

if S^{-1} is μ-non singular, i.e. for any $A \in \Sigma$ with $\mu(A) = 0$ it follows $\mu(S^{-1}(A)) = 0$.

For a long time, ergodic theory was restricted to deterministic systems, i.e. determined by a map S as above, and separated from the theory of Markov chains. In order to bring ergodic theory together with Markov chains, a generalized operator was necessary which was introduced by Hopf in 1954. Hopf's operator propagates probability densities of a Markov chain on a general state space with a finite measure and includes the Frobenius-Perron operator. We will name the operator introduced by Hopf in [23] *transfer operator*. Hopf's article provides a short overview of the historical background, the origin and the development of the transfer operator. A very short but nonetheless comprehensive book about the ergodic theory of Markov chains on a measurable state space which is based on the transfer operator introduced by Hopf was written by Foguel in 1969 [20].

We will now introduce the transfer operator and show its correspondence with general Markov chains. In order to do this, one needs to replace the deterministic map S with a transition kernel p and Equation (1.13) will change to Equation (1.14).

Definition 1.2.1. We call a linear operator

$$\mathcal{T}\colon L^1(\mu) \to L^1(\mu)$$

a *transfer operator* if there exists a transition kernel p such that

$$\int_E p(x, A)\, f(x)\,\mu(dx) = \int_A (\mathcal{T}f)(x)\,\mu(dx) \tag{1.14}$$

holds for all $A \in \Sigma$ and all $f \in L^1(\mu)$.

We now show that the transfer operator propagates probability densities. Recall that the transition kernel p induces a family of transitions kernels $(p_n)_{n\in\mathbb{N}}$ by (1.9). For each *probability density* $f \in L^1(\mu)$, i.e. $f \geq 0$ and $\int_E f(x)\,\mu(dx) = 1$, Proposition 1.1.16 guarantees a Markov chain $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_f)$ with

$$\mathbb{P}_f[X_n \in B, X_0 \in A] = \int_A p_n(x, B)\, f(x)\,\mu(dx)$$

for any $A, B \in \Sigma$. Thus, for any probability density f, the transfer operator $\mathcal{T}_n$ according to transition kernel p_n fulfills by definition

$$\int_A \mathcal{T}_n f(x)\, \mu(dx) = \mathbb{P}_f[X_n \in A] \tag{1.15}$$

this means $\mathcal{T}_n f$ is the probability density of X_n according to μ. One can show that the family $(\mathcal{T}_n)$ is a semigroup, i.e.

$$\mathcal{T}_n \mathcal{T}_m = \mathcal{T}_{n+m}$$

for $n, m \in \mathbb{N}$, see [40, Proposition 2.1.9].

To clarify under which relation between a measure μ and a transition kernel p we can assure the existence of a transfer operator, we introduce the following concept[3].

Definition 1.2.2 (μ-compatible)**.** A transition kernel p is called *μ-compatible* if and only if for each $A \in \Sigma$ with $\mu(A) = 0$ we find a set $B \in \Sigma$ with $\mu(B) = 0$ and

$$p(x, A) = 0 \text{ for all } x \in E \backslash B. \tag{1.16}$$

In other words, p is μ-compatible if for any $A \in \Sigma$ with $\mu(A) = 0$ it holds that $p(\cdot, A) = 0$ μ-almost surely. One may notice that μ-compatibility is a weaker demand than absolute continuity. To see that, consider the Lebesgue measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \lambda)$ and a transition kernel

$$p(x, A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases}$$

Then, for each $x \in E$ the measure $p(x, \cdot)$ is a Dirac delta measure, and, therefore, not absolutely continuous. However, p is λ-compatible, since for each set $A \in \mathcal{B}(\mathbb{R}^n)$ with $\lambda(A) = 0$ the set $B := A$ meets (1.16). If the transition kernel is absolutely continuous for every $x \in E$, then it is in particular μ-compatible, since Equation (1.16) is fulfilled for all μ null sets with $B = \emptyset$.

We will show in the following that μ-compatibility is the necessary and sufficient condition between μ and p to guarantee the existence of the transfer operator. In order to show the existence, consider the following operator

$$\begin{aligned} \mathcal{U} \colon L^\infty(\mu) &\to L^\infty(\mu) \\ f &\mapsto \left(x \mapsto \int_E f(y)\, p(x, dy) \right). \end{aligned} \tag{1.17}$$

Without any relation between the transition kernel p and the measure μ the operator $\mathcal{U}$ is not necessarily well-defined.

Proposition 1.2.3. *The operator from Equation (1.17) is well-defined if and only if the associated transition kernel p is μ-compatible.*

Proof. If p is μ-compatible, then consider two representatives $f, g \in L^\infty(\mu)$ of the same equivalence class $[f] = [g]$, i.e.

$$f(x) = g(x) \quad \text{for all } x \in E \backslash A$$

[3] This definition can also be found in [33] under the name *null preserving* kernel.

for a μ null set A. From the μ-compatibility, we obtain a μ null set B with

$$p(x, A) = 0 \quad \text{for all } x \in E \backslash B.$$

For $x \in E \backslash B$, we obtain

$$\begin{aligned}\int_E g(y)\, p(x, dy) &= \int_{E\backslash A} g(y)\, p(x, dy) \\ &= \int_{E\backslash A} f(y)\, p(x, dy) \\ &= \int_E f(y)\, p(x, dy).\end{aligned}$$

In other words:

$$\left[\int_E g(y)\, p(\cdot, dy)\right] = \left[\int_E f(y)\, p(\cdot, dy)\right].$$

It is straightforward to show that $\mathcal{U}f \in L^\infty(\mu)$ holds.

We show now the other direction. If p is not μ-compatible, we find a set $A \in \Sigma$ with $\mu(A) = 0$ and

$$p(\cdot, A) \neq 0$$

μ-almost surely. It holds

$$\mathbb{1}_A = 0$$

μ-almost surely. Since

$$\mathcal{U}0 = \int_E 0\, p(x, dy) = 0$$

and

$$\mathcal{U}\mathbb{1}_A = \int_E \mathbb{1}_A\, p(x, dy) = p(x, A)$$

we get that if $\mathcal{U}$ is well-defined, then

$$p(\cdot, A) = 0$$

μ-almost surely. This is a contradiction. □

The following proposition states the existence of an operator $\mathcal{T}$ that satisfies

$$\langle \mathcal{U}g, f\rangle_\mu = \langle g, \mathcal{T}f\rangle_\mu \tag{1.18}$$

for all $g \in L^\infty$, $f \in L^1(\mu)$. This proves the existence of the transfer operator for μ-compatible transition kernels, because replacing g by the indicator function $\mathbb{1}_B$ leads to Equation (1.14). In general, for an arbitrary operator $\mathcal{U}\colon L^\infty(\mu) \to L^\infty(\mu)$ an operator $\mathcal{T}$ that satisfies Equation (1.18) does not exist. However, for an operator of the special form as in (1.17) it does exist.

Proposition 1.2.4 ([40, Theorem 1])**.** *Let p be a μ-compatible transition kernel and $\mathcal{U}$ be defined as in Equation (1.17). Then there exists a unique operator $\mathcal{T}\colon L^1(\mu) \to L^1(\mu)$ with*

$$\langle \mathcal{U}g, f\rangle_\mu = \langle g, \mathcal{T}f\rangle_\mu$$

for all $g \in L^\infty$, $f \in L^1(\mu)$.

The following result shows the importance of μ-compatibility.

Theorem 1.2.5. *If and only if the transition kernel p is μ-compatible, there exists a transfer operator*

$$\mathcal{T}\colon L^1(\mu) \to L^1(\mu)$$

associated with the transition kernel p as stated in Equation (1.14).

Proof. If p is μ-compatible, then the existence of such an operator $\mathcal{T}$ follows from Proposition 1.2.4, because $\mathcal{T}$ fulfills in particular Equation (1.14).

We now show that if such an operator $\mathcal{T}$ exists, then p is μ-compatible. Let us take a given set $A \in \Sigma$ with $\mu(A) = 0$. It follows then that

$$\int_E p(x, A)\, f(x)\, \mu(dx) = \int_A (\mathcal{T} f)(x)\, \mu(dx) = 0 \tag{1.19}$$

for any $f \in L^1(\mu)$, where the last equality follows from Proposition 1.1.7. Since (E, Σ, μ) is σ-finite, we find a sequence $(A_i)_{i\in\mathbb{N}}$ with $\mu(A_i) < \infty$ and $\bigcup_{i\in\mathbb{N}} A_i = E$. Replacing f by $\mathbb{1}_{A_i}$ in Equation (1.19) reveals a zero set $B_i \in \Sigma$ with

$$p(x, A) = 0 \quad \text{for all } x \in A_i \backslash B_i$$

for all $i \in \mathbb{N}$, see again Proposition 1.1.7. Defining $B := \bigcup_{i\in\mathbb{N}} B_i$ gives together with $\mu(B) = 0$ and

$$p(x, A) = 0 \quad \text{for all } x \in E \backslash B$$

that $p(\cdot, A)$ is μ-almost surely zero. □

One may also note that Definition 1.2.1 includes the definition of the Perron-Frobenius operator[4]. For a μ non-singular map S one can define the μ-compatible transition kernel

$$p(x, A) := \mathbb{1}_{S^{-1}(A)}(x).$$

In fact, it can easily be shown that p is μ-compatible if and only if S is μ non-singular. The associated transfer operator coincides with the Froebnius-Perron operator.

Stationary Measure

In most applications, stochastic systems often inherit a stationary measure. This is an essential part of the clustering theory that will be presented shortly.

Definition 1.2.6. A probability measure $\mu\colon \Sigma \to [0, 1]$ is called a *stationary measure* of p if

$$\int_A \mu(dx) = \int_E p(x, A)\, \mu(dx) \tag{1.20}$$

holds for all $A \in \Sigma$.

[4]This will follow immediately from Theorem 1.3.5, but one can also check it directly by hand.

Firstly, one may note that Equation (1.20) implies for the associated family of transition kernels $(p_n)_{n\in\mathbb{N}}$ from (1.9) that

$$\begin{aligned}\int_E p_{n+1}(x,A)\,\mu(dx) &= \int_E \int_E p_n(y,A)\,p(x,dy)\,\mu(dx)\\ &\overset{(1.20)}{=} \int_E p_n(x,A)\,\mu(dx)\\ &= \mu(A)\end{aligned}$$

holds. Further, if $((X_n)_{n\in\mathbb{N}},\mathbb{P}_\mu)$ is the Markov chain from Proposition 1.1.16 that evolves according to p with stationary measure μ, then

$$\mathbb{P}_\mu[X_n\in A] = \int_E p_n(x,A)\,\mu(dx) = \int_A \mu(dx) = \mathbb{P}_\mu[X_0\in A]$$

holds, which shows that all random variables X_n are equally distributed according to $\mathbb{P}_\mu$, and thus legitimizes the name of the above definition.

If μ is a stationary measure of p, then p is μ-compatible. To show this, consider a set $A\in\Sigma$ with $\mu(A)=0$, we then obtain

$$0 = \int_A \mu(dx) = \int_E p(x,A)\,\mu(dx).$$

Since p is non-negative, it follows that $p(\cdot,A)$ is μ-almost surely zero , see Proposition 1.1.7. Therefore, p meets (1.16).

As long as μ is a stationary measure of p, one can show that the transfer operator defined here meets

$$\mathcal{T}_n(L^p(\mu))\subseteq L^p(\mu)$$

for $p>1$ and is a contraction, this follows from the Jensen inequality, see [5]. It is unknown if this still holds if p is only μ-compatible. It follows from Hölder's inequality that whenever μ is a finite measure, i.e. $\mu(E)<\infty$, the $L^p(\mu)\subseteq L^1(\mu)$ for $p\geq 1$. Therefore, when μ is a stationary measure, we can consider $\mathcal{T}_n$ as an operator acting on

$$\mathcal{T}_n\colon L^p(\mu)\to L^p(\mu)$$

for $p\geq 1$. In particular for $p=2$ we can apply von Neumann's mean ergodic theorem since $L^2(\mu)$ is a Hilbert space and $\mathcal{T}$ is a contraction. If we further assume that $\mathcal{T}$ is ergodic, i.e. that $\mathcal{T}f=f$ implies that f is constant, then one obtains a generalization of Equation (1.11) for stochastic processes as follows

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}(\mathcal{T}^i f)(x) \overset{(1.12)}{=} Qf(x) \overset{(1.4)}{=} \langle \mathbb{1}, f\rangle_\mu\,\mathbb{1}(x) = \int_E f(x)\,\mu(dx).$$

Reversibility

Reversible processes were introduced in 1935 by Kolmogorov [31, 16]. We will see later that the transfer operator inherits very advantageous properties in the reversible case, which we will exploit to introduce a machinery to analyze molecules.

Definition 1.2.7. A transition kernel p is called *reversible* if for each $A, B \in \Sigma$ it holds

$$\int_A p(x, B)\, \mu(dx) = \int_B p(x, A)\, \mu(dx) \tag{1.21}$$

for a probability measure μ.

Replacing $B = E$ in (1.21) shows that if a transition kernel p is reversible according to a probability measure μ, then μ must be a stationary measure. The transfer operator $\mathcal{T}$ on $L^2(\mu)$ is called *self-adjoint* if and only if

$$\langle \mathcal{T} f, g \rangle_\mu = \langle f, \mathcal{T} g \rangle_\mu$$

for all $f, g \in L^2(\mu)$. The transfer operator $\mathcal{T}$ is self-adjoint if and only if its associated transition kernel p is reversible [26, Proposition 1.1]. Thus, if the transition kernel p is reversible, we have

$$\mathcal{T} f(x) = \mathcal{U} f(x) \overset{(1.7)}{=} \mathbb{E}_x[f(X_1)]. \tag{1.22}$$

Examples

Deterministic Process

Let us consider the Lebesgue measure space $([0,1], \mathcal{B}([0,1]), \lambda)$ and the map $S(x) = x^2$. If one wants to think of S as the deterministic process that maps $x \to x^2$, the transition kernel has to be defined as

$$p(x, A) = \mathbb{1}_{S^{-1}(A)}(x),$$

in other words, the probability of moving from x to A is 1 if $x \in S^{-1}(A)$ and 0 otherwise. In this case, Equation (1.14) reduces to

$$\int_A \mathcal{T} f(x)\, dx = \int_{S^{-1}(A)} f(x)\, dx,$$

where we have replaced $\lambda(dx)$ by dx. For a set $[0, x] \in \mathcal{B}(X)$ we have $S^{-1}([0, x]) = [0, \sqrt{x}]$. With the fundamental theorem of calculus and the substitution rule it is possible to determine the transfer operator analytically.

$$\begin{aligned}
(\mathcal{T} f)(x) &= \frac{\partial}{\partial x} \int_0^x (\mathcal{T} f)(y)\, dy \\
&= \frac{\partial}{\partial x} \int_0^{\sqrt{x}} f(y)\, dy \\
&= \frac{\partial}{\partial x} \int_0^x f(\sqrt{y}) \frac{1}{2\sqrt{y}}\, dy \\
&= f(\sqrt{x}) \frac{1}{2\sqrt{x}}.
\end{aligned}$$

The propagation of $f(x) = 1$ is visualized in Figure 1.1.

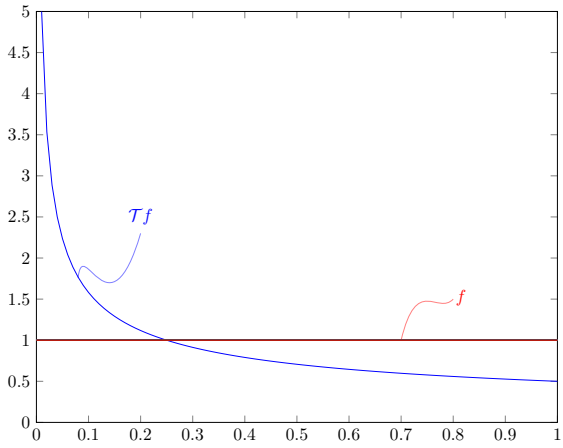

Figure 1.1: Propagation of equal distribution $f(x) = 1$.

Markov Chain

Let us consider the case where $E = \{1, \dots, n\}$ is finite. A homogeneous Markov chain $(X_n)_{n\in\mathbb{N}}$ is then described by a transition matrix $P \in \mathbb{R}^{n\times n}$, i.e.

$$P(i,j) \geq 0 \quad \text{and} \quad \sum_{j=1}^{n} P(i,j) = 1$$

for all $i, j = 1, \dots, n$. The transition kernel of this Markov chain is then defined on the measure space (E, Σ) with $\Sigma := \mathcal{P}(\{1, \dots, n\})$ by

$$p(i, A) := \sum_{j\in A} p_{ij}$$

for $A \in \Sigma$ and $i = 1, \dots, n$. One can identify a non-negative vector $x \in \mathbb{R}^n$ with a measure ν_x acting on $\Sigma := \mathcal{P}(\{1, \dots, n\})$ through

$$\nu_x(A) = \sum_{i\in A} x_i.$$

The transfer operator according to the measure $\nu_{\mathbb{1}}$ for $\mathbb{1} = (1, \dots, 1)^T$ is given by

$$\mathcal{T}_n^{\mathbb{1}} x = (P^n)^T x.$$

A vector $\pi \in \mathbb{R}^n$ is stationary according to equation (1.20) if

$$\pi^T P = \pi^T$$

holds. If we denote

$$D := \begin{pmatrix} \pi_1 & & \\ & \ddots & \\ & & \pi_n \end{pmatrix}$$

then the reversibility condition of Equation (1.21) according to measure ν_π is fulfilled if

$$DP = P^T D$$

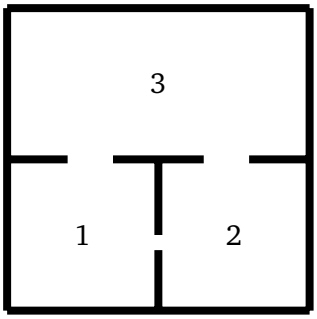

Figure 1.2: Three rooms.

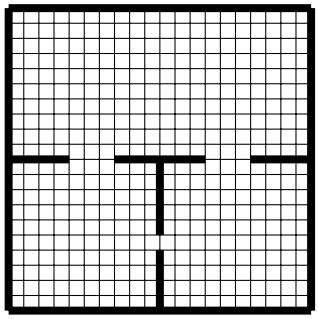

Figure 1.3: Three rooms with grid.

holds. The transfer operator $\mathcal{T}_n^\pi$ according to measure ν_π is then given by

$$\mathcal{T}_n^\pi x = P^n x.$$

Note that $\mathcal{T}_n^\pi$ and $\mathcal{T}_n^{\mathbb{1}}$ both still propagate probability densities, but according to different measures. If $x \in \mathbb{R}^n$ is a probability density, i.e $\sum_{i=1}^n x_i = 1$ and $x_i \geq 0$, then one can either compute the propagated density with $\mathcal{T}_n^{\mathbb{1}} x$ or equivalently, but more complexly, as $D\mathcal{T}_n^\pi D^{-1} x$. This propagation with $\mathcal{T}^\pi$ looks complicated, because it requires converting the probability density x to a probability density $\tilde{x} := D^{-1}x$ associated with π, this means $\sum_{i=1}^n \tilde{x}_i \pi_i = 1$ and $\tilde{x}_i \geq 0$, and then after the propagation, it must be converted back to a normal probability density. For now it may just seem confusing to consider the operator $\mathcal{T}_n^\pi$, but later we will see that we are primarily interested in the eigenvalues and eigenfunctions of $\mathcal{T}_n^\pi$ in order to identify metastable sets. For a reversible Markov chain on a finite dimensional state space, these are simply the right eigenvectors of the transition matrix P.

As an example of a Markov Chain, consider a three-room flat as in Figure 1.2. We discretize the flat by a 20×20 grid into 400 cells, as visualized in Figure 1.3. We now define the probability of moving from one such cell to another as follows. Any one cell can only move to a neighboring cell with which it shares an edge. The probability of moving to any neighbored cell is equally distributed. This gives rise to a transition matrix

$$P \in \mathbb{R}^{400 \times 400}$$

which is sparse, because in each row there are only a maximum of four non-zero entries. Furthermore, let us denote with $x \in \mathbb{R}^{400}$ a probability vector, i.e

$$\sum_{i=1}^{400} x(i) = 1 \quad \text{and} \quad x(i) \geq 0$$

for all $i = 1, \ldots, 400$, which is mainly distributed in the room in the bottom right corner. This probability vector will then be propagated by the transfer operator and spreads through all the rooms as shown in Figure 1.4.

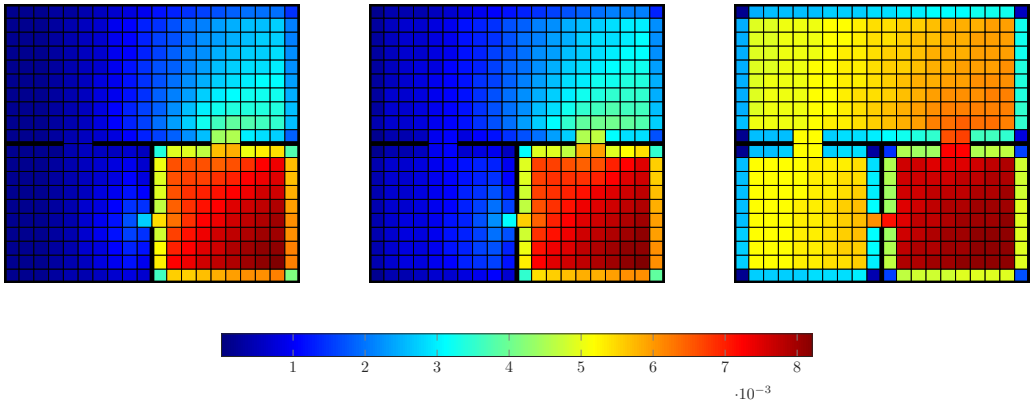

Figure 1.4: Propagation from vector x (left) to $\mathcal{T}^{\mathbb{1}}_{100}x$ (middle) and to $\mathcal{T}^{\mathbb{1}}_{700}x$ (right).

1.3 Related Operators

In this section, we will discover characterizations of the transfer operator and adjoint transfer operator that are independent of a transition kernel. Transfer operators will be identified by Markov operators, adjoint transfer operators will be identified by Koopman operators, and Brown-Markov operators will be identified to be the class of adjoint transfer operators with an invariant measure.

We denote again with (E, Σ, μ) a σ-finite measure space on any given set E.

Markov Operators

Note that each transfer operator $\mathcal{T}$ uniquely determines its transition kernel p μ-almost surely and, conversely, a μ-compatible transition kernel p uniquely determines the transfer operator[5] by the integral Equation (1.14). We will show in the following that there is a completely different characterization of transfer operators.

Definition 1.3.1. A linear operator $P\colon L^1(\mu) \to L^1(\mu)$ satisfying

(i) $Pf \geq 0$ for all $f \geq 0$, $f \in L^1(\mu)$

(ii) $||Pf||_1 \leq ||f||_1$ for all $f \in L^1(\mu)$

is called a *sub Markov operator*. If in addition

(iii) $||Pf||_1 = ||f||_1$ for all $f \geq 0$, $f \in L^1(\mu)$

holds, then P is called a *Markov Operator*.

A Markov Operator is already defined by the properties (i) and (iii), since they imply Property (ii) [34, Proposition 3.1.1]. The following property of Markov operators will be of great help in the proofs that follow.

Proposition 1.3.2. *For a sub Markov operator P and for $f, f_n \in L^1(\mu)$ with $f_n \uparrow f$ and $f_n \geq 0$ it follows*

$$\int_A \lim_{n\to\infty} Pf_n(x)\,\mu(dx) = \lim_{n\to\infty} \int_A Pf_n(x)\,\mu(dx) = \int_A Pf(x)\,\mu(dx)$$

[5] Both directions follow from Proposition 1.1.5. The first direction additionally requires a σ-additive argument.

for any $A \in \Sigma$.

Proof. We get for $A \in \Sigma$

$$\begin{aligned} 0 &\leq \int_A (P(f - f_n))(x)\, \mu(dx) \\ &\overset{(i)}{\leq} \int_E (P(f - f_n))(x)\, \mu(dx) \\ &\overset{(ii)}{\leq} \int_E (f - f_n)(x)\, \mu(dx) \to 0 \end{aligned}$$

by the monotonic convergence theorem. This shows

$$\int_A Pf(x)\, \mu(dx) = \lim_{n\to\infty} \int_A Pf_n(x)\, \mu(dx).$$

Again, since $(Pf_n)_{n\in\mathbb{N}}$ is monotonically increasing, applying the monotonic convergence theorem reveals

$$\lim_{n\to\infty} \int_A Pf_n(x)\, \mu(dx) = \int_A \lim_{n\to\infty} Pf_n(x)\, \mu(dx).$$

□

Further, sub Markov operators can be characterized by

Proposition 1.3.3 ([20, Chapter 1])**.** *For any sub Markov Operator* $P\colon L^1(\mu) \to L^1(\mu)$ *exists a sub transition kernel* p *with*

$$\int_E p(x, A)\, f(x)\, \mu(dx) = \int_A Pf(x)\mu(dx)$$

for all $f \in L^1(\mu)$.

From the last two propositions we get, with the help of the monotonic convergence theorem, the following useful characterization.

Corollary 1.3.4. *For* $f, f_n \in L^1(\mu)$ *with* $f_n \uparrow f$ *and* $f_n \geq 0$ *it follows*

$$\lim_{n\to\infty} \int_E Pf_n(x)\, \mu(dx) = \int_E p(x, E)\, f(x)\, \mu(dx).$$

This enables the following characterization between Markov and transfer operators. For finite measures, this has been already shown in [23, Theorem 2.1].

Theorem 1.3.5. *A operator* $P\colon L^1(\mu) \to L^1(\mu)$ *is a transfer operator if and only if* P *is a Markov operator.*

Proof. Let P be a transfer operator and $f \in L^1(\mu)$ with $f \geq 0$. Then, $\tilde{f} := \frac{f}{\|f\|_1}$ is a probability density. Thus, we have from (1.15)

$$0 \leq \mathbb{P}_{\tilde{f}}[X_1 \in A] = \int_A (P\tilde{f})(x)\, \mu(dx)$$

for any $A \in \Sigma$, and thus $Pf \geq 0$ μ-almost surely by Proposition 1.1.6. Further, we have

$$1 = \mathbb{P}_{\tilde{f}}[X_1 \in E] = \int_E (P\tilde{f})(x)\,\mu(dx)$$

and therefore

$$\int_E f(x)\,\mu(dx) = \|f\|_1 = \int_E (Pf)(x)\,\mu(dx).$$

Thus, P is indeed a Markov operator.

If P is now given as a Markov operator, we are given a sub transition kernel p by Proposition 1.3.3 with

$$\int_E p(x,A)\,f(x)\,\mu(dx) = \int_A Pf(x)\mu(dx)$$

for all $f \in L^1(\mu)$. Since (E,Σ,μ) is σ-additive, we find a family of sets $(C_i)_{i\in\mathbb{N}}$ with

$$\mu(C_i) < \infty \text{ and } \bigcup_{i\in\mathbb{N}} C_i = E.$$

We then get for $A \in \Sigma$ and $D_n := A \cap \bigcup_{i=1}^n C_i$

$$\begin{aligned}
\int_A p(x,E)\,\mu(dx) &\overset{(*)}{=} \lim_{n\to\infty} \int_E \left(P\mathbb{1}_{D_n}\right)(x)\,\mu(dx) \\
&\overset{(iii)}{=} \lim_{n\to\infty} \int_E \mathbb{1}_{D_n}(x)\,\mu(dx) \\
&= \int_E \mathbb{1}_A(x)\,\mu(dx) \\
&= \int_A \mathbb{1}_E(x)\,\mu(dx),
\end{aligned}$$

where we have used Corollary (1.3.4) in $(*)$. Thus, by Proposition 1.1.5 we obtain that p is a transition kernel. □

A transfer operator propagates probability densities of a process. In contrast, sub Markov operators map probability densities only to integrable functions where the integral over the state space E might be less then 1. However, any sub Markov operator can be associated with a transfer operator in the following way.

Corollary 1.3.6. *For any sub Markov operator P with*

$$||Pf||_1 \geq \gamma||f||_1$$

for all $f \in L^1(\mu)$ and a fixed $\gamma \in (0,1]$ exists a transfer operator $\mathcal{T}$ and a function $g\colon E \to (0,1]$ with

$$Pf = \mathcal{T}(f \cdot g)$$

for all $f \in L^1(\mu)$.

Proof. If P is a sub Markov operator, then we find with the help of Proposition 1.3.3 a sub transition kernel p with

$$\int_E p(x,A)\,f(x)\,\mu(dx) = \int_A Pf(x)\mu(dx).$$

Then, for $A \in \Sigma$ we obtain

$$\begin{aligned}\int_A p(x,E)\,\mu(dx) &\overset{(*)}{=} \lim_{n\to\infty} \int_E \left(P\mathbb{1}_{D_n}\right)(x)\,\mu(dx)\\ &\geq \lim_{n\to\infty} \gamma \int_E \mathbb{1}_{D_n}(x)\,\mu(dx)\\ &= \gamma \int_E \mathbb{1}_A(x)\,\mu(dx)\\ &= \gamma \int_A \mathbb{1}_E(x)\,\mu(dx),\end{aligned}$$

where $(*)$ follows from Corollary 1.3.4 and thus $p(x,E) \geq \gamma$ μ-almost surely by Proposition 1.1.6. Define $\alpha(x) := \frac{1}{p(x,E)} \leq \frac{1}{\gamma}$. Then, for $f \in L^1(\mu)$ we have $f \cdot \alpha \in L^1(\mu)$ and we can define for $f \in L^1(\mu)$ the operator

$$\mathcal{T} f := P(f \cdot \alpha),$$

which satisfies

$$\int_E \tilde{p}(x,A)\, f(x)\,\mu(dx) = \int_A \mathcal{T} f(x)\,\mu(dx)$$

for the transition kernel $\tilde{p}(x,A) = \frac{p(x,A)}{p(x,E)}$. Thus, $\mathcal{T}$ is a transfer operator that shows together with $g(x) = p(x,E)$ the stated property. □

Koopman Operators

In the following, we give a characterization of adjoint transfer operators, which is again independent of a transition kernel.

If we denote with

$$M = \{f\colon E \to \mathbb{R}_+ \cup \{\infty\} \mid f \text{ measurable}\}$$

the set of non-negative measurable functions and denote by p a kernel, one can define

$$Vf(x) := \int_E f(y)\, p(x,dy) \tag{1.23}$$

for $f \in M$. For the operator in Equation (1.17) we needed μ-compatibility to guarantee existence. In this situation, we do not need any relation between p and μ, because μ does not appear anywhere, and functions in M are not equivalence classes[6]. There is a very neat characterization for operators in this general form[7].

Proposition 1.3.7 ([46, Proposition 1.3])**.** *For a linear map $V\colon M \to M$ exists a kernel p such that Equation (1.23) holds if and only if for every increasing sequence (f_n) of functions in M one has*

$$V(\lim_{n\to\infty} f_n) = \lim_{n\to\infty} V f_n. \tag{1.24}$$

[6] For the operator in Equation (1.17) the measure μ is hidden in the space $L^\infty(\mu)$, on which the operator is defined.

[7] In [46] the reader is left to find the proof of Proposition 1.3.7. We provide it here for completeness.

Proof. If V is defined as in Equation (1.23), then the monotone convergence theorem implies Equation (1.24).

If, on the other hand, V is an operator witch fulfills Equation (1.24), then one can define $p(x, A) := V\mathbb{1}_A(x)$. Then from the linearity we have $p(x, \emptyset) = V0 = 0$ and for a sequence of pairwise disjoint sets $A_i \in \Sigma$ we have

$$p(x, \bigcup_{i\in\mathbb{N}} A_i) = V \sum_{i\in\mathbb{N}} \mathbb{1}_{A_i} \overset{(1.24)}{=} \sum_{i\in\mathbb{N}} V\mathbb{1}_{A_i} = \sum_{i\in\mathbb{N}} p(x, A_i),$$

thus p is a measure, and if $f \in M$ we get from Proposition 1.1.9 a sequence (f_n) of simple functions with $f_n \uparrow f$, and obtain

$$Vf(x) = \lim_{n\to\infty} Vf_n(x) = \lim_{n\to\infty} \int f_n(y)\, p(x, dy) = \int f(y)\, p(x, dy),$$

where we have used the monotonic convergence theorem in the last step. □

The above proposition gives rise to the question whether a similar characterization is possible, when M is replaced by $L^\infty(\mu)$. It turns out that such a characterization exists:

Definition 1.3.8. We call a linear operator

$$\mathcal{U} : L^\infty(\mu) \to L^\infty(\mu)$$

a *generalized Koopman operator* if it fulfills

$(K1)$ $\mathcal{U}f \geq 0$ for $f \geq 0$, $f \in L^\infty(\mu)$

$(K2)$ $\mathcal{U}\mathbb{1} = \mathbb{1}$

and Property $(K3)$ which is that for any sequence of disjoint sets $(A_i)_{i\in\mathbb{N}} \subset \Sigma$ we get for $B_n := \bigcup_{i=1}^n A_i$ and $B := \bigcup_{i\in\mathbb{N}} A_i$ that

$$\mathcal{U}\mathbb{1}_{B_n} \uparrow \mathcal{U}\mathbb{1}_B$$

holds.

The name generalized Koopman operator is chosen, because the definition generalizes the Koopman operator definied in [34]. We can now identify the class of adjoint transfer operators.

Theorem 1.3.9. *A operator $\mathcal{U} : L^\infty(\mu) \to L^\infty(\mu)$ is a generalized Koopman operator if and only if $\mathcal{U}$ is the adjoint operator from a transfer operator.*

Proof. If $\mathcal{U}$ is the adjoint operator from a transfer operator, then the operator is of the form as in Equation (1.17) for some transition kernel p. It follows directly that $\mathcal{U}$ fulfills condition $(K1)$ and $(K2)$. Property $(K3)$ follows from the monotonic convergence theorem. Thus $\mathcal{U}$ is a generalized Koopman operator.

We now show that any generalized Koopman operator is the adjoint operator from a transfer operator. Thus, let us given a generalized Koopman operator $\mathcal{U}$. By [40, Theorem 1

on page 29], it follows that $\mathcal{U}$ has an adjoint operator $\mathcal{T}$ that satisfies Equation (1.18). It remains to show that $\mathcal{T}$ is a transfer operator. First, for $f \geq 0$ we have for any $A \in \Sigma$

$$\int_A \mathcal{T}f(x)\,\mu(dx) = \langle \mathbb{1}_A, \mathcal{T}f\rangle_\mu = \langle \mathcal{U}\mathbb{1}_A, f\rangle_\mu \geq 0$$

and thus $\mathcal{T}f \geq 0$ by Proposition 1.1.6. In addition, we have

$$\int_E \mathcal{T}f(x)\,\mu(dx) = \langle \mathbb{1}, \mathcal{T}f\rangle_\mu = \langle \mathcal{U}\mathbb{1}, f\rangle_\mu = \langle \mathbb{1}, f\rangle_\mu = \int_E f(x)\,\mu(dx).$$

Thus, $\mathcal{T}$ is a Markov operator and therefore by Theorem 1.3.5 a transfer operator. □

This theorem shows that one can replace Property $(K3)$ by the property that for $f_n, f \in L^\infty(\mu)$, with $f_n, f \geq 0$ and $f_n \to f$ it always follows $\lim_{n\to\infty} \mathcal{U}f_n = \mathcal{U}f$.

Brown-Markov Operators

In this section, we will show that Brown-Markov operators characterize the class of operators that are adjoint to a transfer operator with an invariant measure. Although the Definition 1.3.1 of Markov operators is used frequently [48, 34, 1, 2, 32], it is not used consistently in the literature. One definition of the Markov operator can be found in [10], introduced in 1966 by James Russel Brown as follows.

Definition 1.3.10. A linear operator $P\colon L^\infty(\mu) \to L^\infty(\mu)$ satisfying

(I) $Pf \geq 0$ for all $f \geq 0,\ f \in L^\infty(\mu)$

(II) $||Pf||_1 = ||f||_1$ for all $f \geq 0,\ f \in L^\infty(\mu) \cap L^1(\mu)$

(III) $P\mathbb{1} = \mathbb{1}$

is called a *Brown-Markov operator*.

The name Brown-Markov operator is chosen, because the definition has been introduced by Brown and at the first glance, this operator looks like a good mix between a Markov operator and an adjoint Markov operator. Its true nature will be revealed at the end of this section.

To characterize Brown-Markov operators, we need the following definition.

Definition 1.3.11. A measure μ is called *invariant measure* according to a kernel p if and only if

$$\int_E p(x, B)\,\mu(dx) = \mu(B)$$

holds for all $B \in \Sigma$.

Thus, if a invariant measure is in addition a probability measure, then it is also a stationary measure. In [10], Brown introduces for a given transition kernel p and associated *invariant measure* the operator already considered in Equation (1.17)

$$\begin{aligned}\mathcal{U}&\colon L^\infty(\mu) \to L^\infty(\mu)\\ (\mathcal{U}f)(x) &= \int_E f(y)\,p(x, dy).\end{aligned}$$

Since μ is a invariant measure, this operator is well-defined. This follows from Proposition 1.2.3, because the invariance property implies that p is μ-compatible. One can verify that the operator from Equation (1.17) is a Brown-Markov operator if μ is a invariant measure. This raises the question of whether the reverse is also true, i.e. whether any Brown-Markov operator can be obtained as in Equation (1.17) where μ is invariant according to p. Brown suggested the following answer to this question [10, Page 15]:

> *"In general, a [Brown-]Markov operator can not be defined in terms of a stochastic transition function [as in Equation (1.17)]."*

The note is stated without any evidence. The reason for that is that the statement is simply not correct. This is shown by the following proposition.

Proposition 1.3.12. *A Brown-Markov operator is a generalized Koopman operator.*

Proof. Let us given a Brown-Markov operator P. It remains to show that for a sequence of disjoints sets $(A_i)_{i\in\mathbb{N}} \subset \Sigma$ with $B_n = \bigcup_{i=1}^{n} A_i$ and $B = \bigcup_{i\in\mathbb{N}} A_i$ we have

$$P\mathbb{1}_{B_n} \uparrow P\mathbb{1}_B.$$

We have for any $A \in \Sigma$

$$\begin{aligned} 0 &\leq \int_A (P(\mathbb{1}_B - \mathbb{1}_{B_n}))(x)\,\mu(dx) \\ &\overset{(I)}{\leq} \int_E (P(\mathbb{1}_B - \mathbb{1}_{B_n}))(x)\,\mu(dx) \\ &\overset{(II)}{=} \int_E \mathbb{1}_B(x) - \mathbb{1}_{B_n}(x)\,\mu(dx) \to 0 \end{aligned}$$

which reveals

$$\lim_{n\to\infty} \int_A (P\mathbb{1}_{B_n})(x)\,\mu(dx) = \int_A (P\mathbb{1}_B)(x)\,\mu(dx).$$

Since $(P\mathbb{1}_{B_n})_{n\in\mathbb{N}}$ is monotonically increasing, applying the monotonic convergence theorem gives

$$\int_A \lim_{n\to\infty} (P\mathbb{1}_{B_n})(x)\,\mu(dx) = \int_A (P\mathbb{1}_B)(x)\,\mu(dx).$$

Thus by Proposition 1.1.5, we have

$$P\mathbb{1}_{B_n} \uparrow P\mathbb{1}_B.$$

□

Thus by Theorem 1.3.9, any Brown-Markov operator can be defined in terms of a stochastic transition function as in Equation (1.17).

In the following, we will show that the class of Brown-Markov operators is identical to the class of operators that are adjoint to transfer operators with invariant measure. Property (II) enables us to consider any Brown-Markov operator as a Markov operator in the following sense.

Proposition 1.3.13. *For any Brown-Markov operator P exists a Markov operator $\mathcal{T}$ with*

$$Pf = \mathcal{T}f$$

for all $f \in L^1(\mu) \cap L^\infty(\mu)$.

Proof. For $f \in L^1(\mu)$, $f \geq 0$ define

$$g_n(x) := f(x) \cdot \mathbb{1}_{\{y\,:\, f(y) \leq n\}}(x).$$

Then, Pg_n is monotonously increasing and we can define

$$(\mathcal{T}f)(x) := \lim_{n\to\infty} (Pg_n)(x).$$

In addition, we have

$$\begin{aligned}
\int_E (\mathcal{T}f)(x)\,\mu(dx) &= \int_E \lim_{n\to\infty} (Pg_n)(x)\,\mu(dx) \\
&\overset{(*)}{=} \lim_{n\to\infty} \int_E (Pg_n)(x)\,\mu(dx) \\
&\overset{(II)}{=} \lim_{n\to\infty} \int_E g_n(x)\,\mu(dx) \\
&\overset{(*)}{=} \int_E f(x)\,\mu(dx) < \infty
\end{aligned}$$

and thus $\mathcal{T}f \in L^1(\mu)$, where we have used the monotonic convergence theorem in $(*)$. Thus for $f \in L^1$ with $f = f^+ - f^-$ it is well-defined to set

$$\mathcal{T}f := \mathcal{T}f^+ - \mathcal{T}f^-$$

and $\mathcal{T}$ is a Markov operator.

For $f \in L^1(\mu) \cap L^\infty(\mu)$ we have

$$\lim_{n\to\infty} \int_E (Pf)(x) - (Pg_n)(x)\,\mu(dx) = \lim_{n\to\infty} \int_E (f - g_n)(x)\,\mu(dx) = 0$$

and thus for an appropriate subsequence[8] we get

$$(Pf)(x) = \lim_{k\to\infty} (Pg_{n_k})(x) = (\mathcal{T}f)(x).$$

□

In the case where μ is invariant, it is possible to find a reversed process in the following sense. This result has been known for finite measures [29, 23] and is now extended to σ-finite measures.

Proposition 1.3.14. *For any transition kernel p and associated invariant measure μ exists a μ-compatible transition kernel $\tilde{p}$ with*

$$\int_A p(x, B)\,\mu(dx) = \int_B \tilde{p}(x, A)\,\mu(dx) \tag{1.25}$$

for all $A, B \in \Sigma$. We call p the reverse *of $\tilde{p}$ and vice versa.*

[8] Note that in general from $||Pf - Pf_n||_1 \to 0$ it does not follow that $Pf \to Pf_n$ converges pointwise. This can only be shown for a subsequent, see [3, Proposition 15.7].

Proof. Define the operator

$$Pf(x) := \int_E f(y)\, p(x, dy) \tag{1.26}$$

for $f \in L^\infty(\mu)$ which is well-defined according to Proposition 1.2.3. Denote with $\mathcal{T}\colon L^1(\mu) \to L^1(\mu)$ the associated Markov operator from Proposition 1.3.13. According to Theorem 1.3.5 we can assure the existence of a μ-compatible transition kernel $\tilde{p}$ with

$$\int_A (\mathcal{T}f)(x)\, \mu(dx) = \int_E f(x)\, \tilde{p}(x, A)\, \mu(dx) \tag{1.27}$$

for $f \in L^1(\mu)$. For a set $B \in \Sigma$ define[9] $D_n^B := B \cap \bigcup_{k=1}^n C_k$. Then, one obtains

$$\begin{aligned}\lim_{n\to\infty} \int_A (P\mathbb{1}_{D_n^B})(x)\, \mu(dx) &\overset{(1.26)}{=} \lim_{n\to\infty} \int_A \int_E \mathbb{1}_{D_n^B}(x)\, p(x, dy)\, \mu(dx) \\ &\overset{(*)}{=} \int_A p(x, B)\, \mu(dx)\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty} \int_A (P\mathbb{1}_{D_n^B})(x)\, \mu(dx) &= \lim_{n\to\infty} \int_A (\mathcal{T}\mathbb{1}_{D_n^B})(x)\, \mu(dx) \\ &\overset{(1.27)}{=} \lim_{n\to\infty} \int_E \mathbb{1}_{D_n^B}(x)\, \tilde{p}(x, A)\, \mu(dx) \\ &\overset{(*)}{=} \int_B \tilde{p}(x, A)\, \mu(dx)\end{aligned}$$

where we have used the monotone convergence theorem in $(*)$. □

One may note that if μ is not invariant according to p, then p cannot possess a reverse transition kernel, because replacing A by E in Equation (1.25) yields

$$\int_E p(x, B)\, \mu(dx) = \int_B \tilde{p}(x, E)\, \mu(dx) = \mu(B).$$

The following theorem together with Equation (1.7) shows that propagated probability densities can be evaluated pointwise by an expectation value.

Proposition 1.3.15. *For any transfer operator $\mathcal{T}$ with transition kernel p and associated invariant measure μ exists a transition kernel $\tilde{p}$ with*

$$(\mathcal{T}f)(x) = \int_E f(y)\, \tilde{p}(x, dy)$$

for any $f \in L^1(\mu)$, where $\tilde{p}$ is the reverse of p.

Proof. From the pre-condition we have

$$\int_A (\mathcal{T}f)(x)\, \mu(dx) = \int_E p(x, A)\, f(x)\, \mu(dx),$$

[9] With C_i defined as in Theorem 1.3.5.

and from Proposition 1.3.14 we know the existence of a transition kernel $\tilde{p}$ which is the reverse of p. Note that this implies

$$\int_E \mathbb{1}_B(x)\, p(x, A)\, \mu(dx) = \int_A \int_E \mathbb{1}_B(y)\, \tilde{p}(x, dy)\, \mu(dx)$$

for all $A, B \in \Sigma$ and with the help of the monotone convergence theorem we get

$$\int_E f(x)\, p(x, A)\, \mu(dx) = \int_A \int_E f(y)\, \tilde{p}(x, dy)\, \mu(dx)$$

for all measurable functions $f \geq 0$. For any $A \in \Sigma$, we get

$$\begin{aligned}\int_A (\mathcal{T}f)(x)\, \mu(dx) &= \int_E f(x)\, p(x, A)\, \mu(dx) \\ &= \int_A \int_E f(y)\, \tilde{p}(x, dy)\, \mu(dx)\end{aligned}$$

for $f \in L^1(\mu)$ with $f \geq 0$ and by Proposition 1.1.5 we have

$$(\mathcal{T}f)(x) = \int_E f(y)\, \tilde{p}(x, dy)$$

for $f \in L^1(\mu)$ with $f \geq 0$. Since $\mathcal{T}f \in L^1(\mu)$ for $f \in L^1(\mu)$, we can split $f = f^+ - f^-$ use the linearity of $\mathcal{T}$ and obtain

$$(\mathcal{T}f)(x) = \int_E f(y)\, \tilde{p}(x, dy)$$

for all $f \in L^1(\mu)$. □

In the case of the Frobenius-Perron operator $\mathcal{T}$, one can state under certain conditions on S that

$$(\mathcal{T}f)(x) = f(S^{-1}(x)) \tag{1.28}$$

holds for all $x \in E$ μ-almost surely [34, Corollary 3.2.1]. This can be read as follows: Assuming all points are distributed according to f and propagated, then the probability to be in state x is given by $(\mathcal{T}f)(x)$, which is identical to the probability of being in $S^{-1}(x)$ before, which is given by $f(S^{-1}(x))$. Denote with $((\widetilde{X}_n)_{n\in\mathbb{N}}, \widetilde{\mathbb{P}}_x)$ the associated Markov chain from the reversed transition kernel $\widetilde{p}$ and with $\mathcal{T}$ the associated transfer operator according to the transition kernel p, then we obtain from the previous theorem and Equation (1.7)

$$(\mathcal{T}f)(x) = \widetilde{\mathbb{E}}_x[f(\widetilde{X}_1)].$$

This is the extension of Equation (1.28) to stochastic processes. Also it explains the nature of the adjoint operator of a transfer operator, as long as μ is invariant: The adjoint transfer operator can be viewed as the transfer operator of the reversed process.

Ultimately, we arrived at the last theorem of this chapter: The characterization of Brown-Markov operators.

Theorem 1.3.16. *A operator P is a Brown-Markov operator if and only if P is an adjoint operator from a transfer operator with invariant measure.*

Proof. If P is the adjoint operator from a transfer operator with invariant measure μ, then it only remains to validate (II), because the remaining properties follow from Theorem 1.3.9. Since P is the adjoint of a transfer operator, we find a transition kernel p with

$$Pf(x) = \int_E f(y)\, p(x, dy)$$

for all $f \in L^\infty(\mu) \cap L^1(\mu)$. This implies

$$\int_E Pf(x)\, \mu(dx) = \int_E \left[\int_E f(y)\, p(x, dy) \right] \mu(dx) \overset{(*)}{=} \int_E f(x)\, \mu(dx)$$

for all $f \in L^\infty(\mu) \cap L^1(\mu)$, where it was used in $(*)$ that μ is invariant according to p. This shows that P is indeed a Brown-Markov operator.

Let us now given a Brown-Markov operator P. It follows from Proposition 1.3.12 that P is a generalized Koopman operator, and thus from Theorem 1.3.9 that P is the adjoint of a transfer operator $\mathcal{T}$ with transition kernel p. It remains to show that μ is invariant according to p. Denote with $\widetilde{\mathcal{T}}$ the associated Markov operator according to Proposition 1.3.13. Denote with $\widetilde{p}$ the μ-compatible transition kernel associated with $\widetilde{\mathcal{T}}$ from Theorem 1.3.5, i.e.

$$\int_A (\widetilde{\mathcal{T}} f)(x)\, \mu(dx) = \int_E f(x)\, \widetilde{p}(x, A)\, \mu(dx)$$

for all $f \in L^1(\mu)$, $A \in \Sigma$. Since μ is σ-additive, we are given a sequence of disjoint sets $(C_i)_{i\in\mathbb{N}}$ with $C_i \in \Sigma$, $\bigcup_{i\in\mathbb{N}} C_i = E$ and $\mu(C_i) < \infty$. Set $D_n := \bigcup_{i=1}^n C_i$, For $A \in \Sigma$ we get from the monotonic convergence theorem and Property $(K3)$

$$\int_A P\mathbb{1}(x)\, \mu(dx) = \lim_{n\to\infty} \int_A P\mathbb{1}_{D_n}(x)\, \mu(dx). \tag{1.29}$$

Further, we get

$$\begin{aligned}
\mu(A) &= \int_A \mathbb{1}(x)\, \mu(dx) \\
&\overset{(III)}{=} \int_A (P\mathbb{1})(x)\, \mu(dx) \\
&\overset{(1.29)}{=} \lim_{n\to\infty} \int_A (P\mathbb{1}_{D_n})(x)\, \mu(dx) \\
&= \lim_{n\to\infty} \int_A (\widetilde{\mathcal{T}}\mathbb{1}_{D_n})(x)\, \mu(dx) \\
&= \lim_{n\to\infty} \int_E \widetilde{p}(x, A)\, \mathbb{1}_{D_n}(x)\, \mu(dx) \\
&= \int_E \widetilde{p}(x, A)\, \mu(dx)
\end{aligned}$$

which shows that μ is invariant according to $\tilde{p}$. Thus, by Proposition 1.3.14 the transition kernel $\tilde{p}$ possesses a reverse transition kernel which we denote by $\hat{p}$. It will turn out that $\hat{p} = p$ and then

$$\int_E p(x, B)\, \mu(dx) = \int_B \widetilde{p}(x, E)\, \mu(dx) = \mu(B)$$

proves the claim.

To see that $\hat{p} = p$ holds, we define for any $A \in \Sigma$ the set $D_n^A = A \cap \bigcup_{i=1}^n C_i$. Then, we have for any $A, B \in \Sigma$

$$\begin{aligned}
\int_B \hat{p}(x, A)\, \mu(dx) &= \int_A \widetilde{p}(x, B)\, \mu(dx) \\
&= \lim_{n\to\infty} \int_E \widetilde{p}(x, B)\, \mathbb{1}_{D_n^A}(x)\, \mu(dx) \\
&= \lim_{n\to\infty} \int_B (\widetilde{\mathcal{T}} \mathbb{1}_{D_n^A})(x)\, \mu(dx) \\
&= \lim_{n\to\infty} \int_B (P \mathbb{1}_{D_n^A})(x)\, \mu(dx) \\
&= \lim_{n\to\infty} \int_B p(x, D_n^A)\, \mu(dx) \\
&= \int_B p(x, A)\, \mu(dx),
\end{aligned}$$

and by Proposition 1.1.5 we have $\hat{p} = p$. □

Computation schemes

2

> *Truth* [...] *is much too complicated to allow anything but approximations.*
>
> **— John von Neumann**

So far, we have revealed the essence of a transfer operator. In this chapter, we explain how to obtain and how to extract knowledge from a Galerkin projection of the transfer operator. We will then discuss conceptional convergence results, which have been of interest since Ulam posed a conjecture about transfer operators in 1960. We continue by explaining how to compute a single entry of the Galerkin projection. We offer a numerical scheme for estimating the exact error of the computation from such an entry. We are also able to deduce a stochastic interpretation for the exact numerical error. Finally, we introduce a Girsanov reweighting scheme, which shows how one can make use of pre-existing trajectories in order to receive a new Galerkin projection for a different system.

We denote with (E, Σ, μ) a probability space on any given set E, and with $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space.

2.1 State of the Art

In this section, we give a brief overview of why we are interested in the eigenvalues and eigenfunctions of the transfer operator, and how one can reduce the transfer operator from its infinite dimensional state space to a matrix representation. Also, we will briefly discuss some function spaces that are often used to reduce the original state space. Finally, we describe how we model the molecule and present some of the benefits of this method.

The Clustering Method PCCA+

We call a set $A \subset E$ a *metastable* set for a stochastic process $(X_t)_{t \in I}$ if

$$\mathbb{P}[X_\tau \in A \mid X_0 \in A] \approx 1,$$

where close to one and the time step τ has to be specified for each model individually. In fact, the value depends on the eigenvalues of the transfer operator. In order to apply the theorem which gives rise to the clustering methods considered here, certain conditions are required on the transfer operator $\mathcal{T}$. The definition of the transfer operator includes a transition kernel p. We denote with $((X_n)_{n \in \mathbb{N}}, \mathbb{P}_\mu)$ the associated Markov chain and with μ the stationary measure of p.

Definition 2.1.1. We say that a self-adjoint transfer operator $\mathcal{T}\colon L^2(\mu) \to L^2(\mu)$ *fulfills Assumption S* if:

- It exhibits n eigenvalues
$$\lambda_n \leq \cdots \leq \lambda_2 < \lambda_1 = 1$$
counted according to their multiplicity. The corresponding set of μ-orthonormal eigenvectors will be denoted by $\{f_n, \ldots, f_1\}$.
- The spectrum $\sigma(\mathcal{T})$ of $\mathcal{T}$ satisfies
$$\sigma(\mathcal{T}) \subset [a, b] \cup \{\lambda_n, \ldots, \lambda_2, 1\}$$
for some constants $a, b \in (-1, +1)$ satisfying $-1 < a \leq b < \lambda_n$.

If a transfer operator fulfills assumption S, one can show the following connection between metastable sets and eigenvalues.

Theorem 2.1.2 ([27, Theorem 1 and Theorem 2])**.** *Consider a transfer operator $\mathcal{T}\colon L^2(\mu) \to L^2(\mu)$ according to a transition kernel p with stationary measure μ satisfying Assumption S. Denote with $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_\mu)$ the associated Markov chain. For an arbitrary decomposition of E into sets $A_1, \ldots, A_n$ holds:*

$$\sum_{i=1}^{n} p_i \lambda_i + c \leq \sum_{i=1}^{n} \mathbb{P}_\mu(X_1 \in A_i \mid X_0 \in A_i) \leq \sum_{i=1}^{n} \lambda_i$$

with

$$p_i = ||Qf_i||^2 = 1 - ||Qf_i - f_i||^2$$

and

$$c = m(\mathcal{T}) \left(\sum_{i=1}^{n} ||Qf_i - f_i||^2 \right)$$

where

$$m(\mathcal{T}) := \inf \{\langle \mathcal{T}, x\rangle \mid ||x|| = 1\} \in (-1, 1)$$

and Q denotes the orthogonal projection onto $\{\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_n}\}$.

This theorem reveals the connection between metastable sets and the eigenfunctions and eigenvalues of the transfer operator. When the eigenfunctions $f_1, \ldots, f_n$ of the transfer operator are given, then one may extract the metastable sets from the eigenfunctions with the clustering method Robust Perron Cluster Analysis (PCCA+) invented by Marcus Weber and Peter Deuflhard [14]. The idea is described as follows. If the eigenvectors are well approximated by the indicator functions $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_n}$, then the above theorem shows that $A_1, \ldots, A_n$ must be metastable sets. In this case, the indicator functions should be well approximated through a linear combination of the eigenfunctions $f_1, \ldots, f_n$. Since it is our aim to find the metastable sets $A_1, \ldots, A_n$, this leads to the question of whether it is possible to find functions $\chi_1, \ldots, \chi_n$ in $\operatorname{span}\{f_1, \ldots, f_n\}$ with

$$\chi_i \approx \mathbb{1}_{A_i} \tag{2.1}$$

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
1	0.9969	0.9966	0.9912	0.9731	0.9725

Table 2.1: The largest 6 eigenvalues.

for $i = 1, \ldots, n$. This can be achieved by questioning if it is possible to find functions $\chi_1, \ldots, \chi_n$ in $\operatorname{span}\{f_1, \ldots, f_n\}$ with

$$\sum_{i=1}^{n} \chi_i(x) = 1 \quad \text{and} \quad \chi_i \geq 0 \tag{2.2}$$

where χ_i is a linear combination of the eigenfunctions, i.e.

$$\chi_i = \sum_{j=1}^{n} \alpha_{ij} f_j$$

or denoted in matrix notation

$$\begin{pmatrix} | & & | \\ f_1 & \cdots & f_n \\ | & & | \end{pmatrix} \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} = \begin{pmatrix} | & & | \\ \chi_1 & \cdots & \chi_n \\ | & & | \end{pmatrix}.$$

The set of possible matricies $\mathcal{A}(i,j) := \alpha_{i,j}$ that lead to a basis $\chi_1, \ldots, \chi_n$ with the properties (2.2) span a convex polytop in the space $\mathbb{R}^{n \times n}$. The algorithm PCCA+ solves to a given convex function $I \colon \mathbb{R}^{n \times n} \to \mathbb{R}$ an optimization problem on the polytope and returns the matrix $\mathcal{A}$ which maximizes the function I. In general, the solution is not unique. In addition, there are multiple choices for a reasonable function I. The condition (2.1) has to be decoded in the function I. Possible choices for I can be found in [59, 47].

Example

Let us return to the example of the three rooms from Figure 1.2. Theorem 2.1.2 shows that the eigenvalues and eigenvectors of $\mathcal{T}_1^\pi$ are of interest, in order to find the metastable sets. The largest 6 eigenvalues of the transfer operator $\mathcal{T}_1^\pi$ are given in Table 2.1. There is a spectral gap between the third and fourth eigenvalue, which indicates that we have three metastable sets. The emergence of so many eigenvalues close to one occurs because we only consider the transfer operator according to one timestep, thus many possible sets have a high metastability. If we were to increase the timestep, the eigenvalues would dissociate more clearly from another. However, the eigenfunctions do fully decode the metastable sets. Unlike the eigenvalues, the eigenfunctions do not change for different timesteps. Thus, approximating the eigenfunctions for a small timestep is sufficient for finding the metastable sets for larger timesteps, but it is difficult to identify the spectral gap. The first dominant vector is constant 1, the other non-trivial dominant eigenvectors are shown in Figure 2.1. Figure 2.2 shows the three linear combinations χ_1, χ_2 and χ_3 of the dominant eigenvectors, computed with PCCA+.

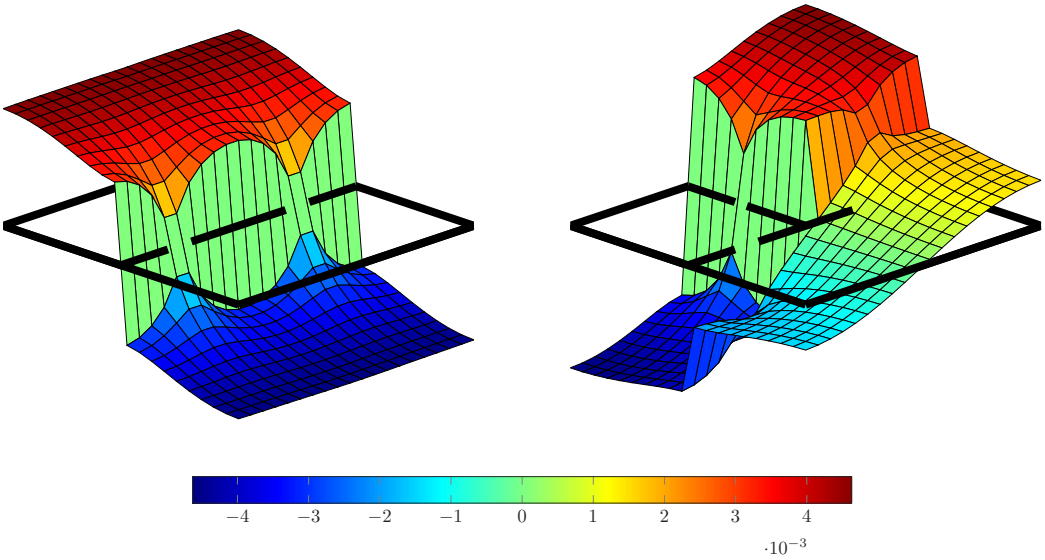

Figure 2.1: Second (left) and third (right) dominant eigenvector of $\mathcal{T}_1^{\pi}$.

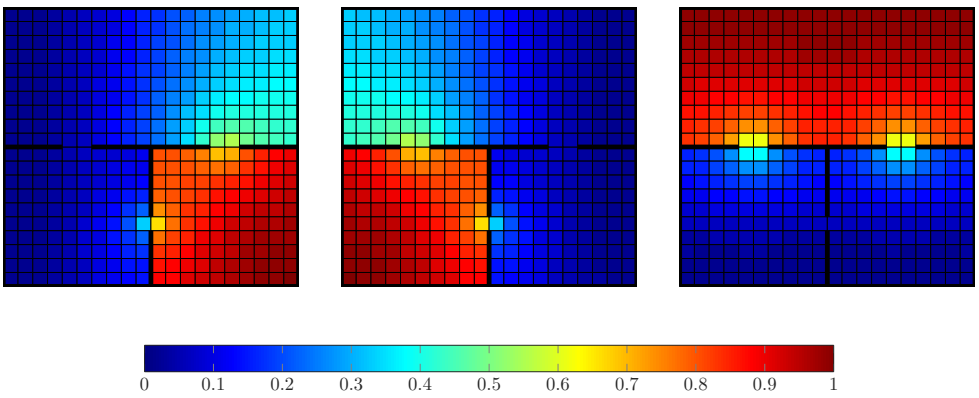

Figure 2.2: Cluster χ_1, χ_2 and χ_3.

Galerkin Method

Galerkin methods are used to convert a continuous operator problem to a discrete problem. In our case we want to discretize the transfer operator

$$\mathcal{T}\colon L^1(\mu) \to L^1(\mu).$$

One way of getting numerical access to the operator is to project it on a finite space. We choose μ as the stationary measure of the process, because then we have for every $1 \leq r \leq \infty$ that

$$\mathcal{T}(L^r(\mu)) \subseteq L^r(\mu)$$

holds [5]. Thus, we can project $\mathcal{T}$ on the Hilbert space $L^2(\mu)$. For functions $\phi_1, \ldots, \phi_n \in L^2(\mu)$ and the unique orthogonal projection $Q\colon L^2(\mu) \to D$ from Definition 1.1.13 where $D = span\{\phi_1, \ldots, \phi_n\}$ one can define the projection $Q\mathcal{T}Q$. The projected transfer operator $Q\mathcal{T}Q$ lives on an n-dimensional space and thus, there exists a matrix representation $\mathcal{M} \in \mathbb{R}^{n\times n}$ of $Q\mathcal{T}Q$. We call $\mathcal{M}$ *left matrix representation* according to the basis $\{\phi_1, \ldots, \phi_n\}$ if for any function $f \in D$ with

$$f = \sum_{i=1}^{n} \alpha_i \psi_i \quad \text{and} \quad \hat{f} = (\alpha_1, \ldots, \alpha_n)$$

it holds

$$(Q\mathcal{T}Q)f = \sum_{i=1}^{n} \beta_i \psi_i \quad \text{with} \quad \hat{f}\mathcal{M} = (\beta_1, \ldots, \beta_n)$$

and we call it *right matrix representation* if we multiply $\hat{f}$ from the right of $\mathcal{M}$ to obtain the corresponding β_i. In both cases, $\mathcal{M}$ is called the *Galerkin projection* of the transfer operator $\mathcal{T}$.

Marco Sarich showed how the left matrix representation appears for the transfer operator.

Theorem 2.1.3 ([49, Theorem 1])**.** *Let $D = span\{\phi_1, \ldots, \phi_n\} \subset L^2(\mu)$ be an n-dimesional space and μ denote any measure and $Q\colon L^2(\mu) \to D$ the ortoghonal projection. And let $\mathcal{T}\colon L^2(\mu) \to L^2(\mu)$ denote a linear operator. If $\langle \phi_i, \mathbb{1}\rangle_\mu > 0$ for $i = 1, \ldots, n$ then*

$$\mathcal{M} = TS^{-1} \quad T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} \quad S_{ij} = \frac{\langle \phi_i, \phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu}$$

is a left matrix representation of $Q\mathcal{T}Q$ to basis $A = \{f_1, \ldots, f_n\}$ with $f_i = \phi_i \frac{1}{\langle \phi_i, \mathbb{1}\rangle_\mu}$.

In cases where the operator is self-adjoint, it is possible to give a right matrix representation of $Q\mathcal{T}Q$ to an unweighted basis, which is useful for numerical application since the weights are known to be ill-conditioned [60].

Theorem 2.1.4. *Let $\{\phi_1, \ldots, \phi_n\} \subset L^2(\mu)$ be a basis with $\langle \phi_i, \mathbb{1}\rangle_\mu > 0$ of a subspace D, and $Q\colon L^2(\mu) \to D$ the orthogonal projection onto D. For any self-adjoint continuous operator $\mathcal{T}\colon L^2(\mu) \to L^2(\mu)$ we have*

$$\mathcal{M} = S^{-1}T, \quad T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu}, \quad S_{ij} = \frac{\langle \phi_i, \phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu}$$

is a right matrix representation of $Q\mathcal{T}Q$ according to the basis $A = \{\phi_1, \dots, \phi_n\}$, i.e. for any

$$f = \sum_{i=1}^{n} \alpha_i \phi_i, \quad Q\mathcal{T}Qf = \sum_{i=1}^{n} \beta_i \phi_i$$

it holds

$$\mathcal{M}(\alpha_1, \dots, \alpha_n)^T = (\beta_1, \dots, \beta_n)^T.$$

Proof. Consider the Gram matrix of $\{\phi_1, \dots, \phi_n\}$

$$\hat{S}_{ij} = \langle \phi_i, \phi_j \rangle_\mu .$$

This matrix is invertible since $\{\phi_1, \dots, \phi_n\}$ is a basis and the orthogonal projection Q can be represented as

$$Qv = \sum_{i,j=1}^{n} \hat{S}_{ij}^{-1} \langle v, \phi_i \rangle_\mu \phi_j.$$

This can be verified by checking $\langle Qv - v, g \rangle_\mu = 0$ for all $g \in D$, $v \in L^2(\mu)$. From

$$S = D^{-1}\hat{S} \quad \text{with} \quad D = diag\left(\langle \phi_1, \mathbb{1} \rangle_\mu , \dots, \langle \phi_n, \mathbb{1} \rangle_\mu \right)$$

we obtain

$$S^{-1} = \hat{S}^{-1} D \quad \text{and, therefore,} \quad \hat{S}_{ij}^{-1} = S_{ij}^{-1} \frac{1}{\langle \phi_j, \mathbb{1} \rangle_\mu} = S_{ji}^{-1} \frac{1}{\langle \phi_i, \mathbb{1} \rangle_\mu},$$

in the last step it was used that $\hat{S}^{-1}$ is symmetric since $\hat{S}$ is symmetric. This implies

$$Qv = \sum_{i,j=1}^{n} S_{ji}^{-1} \langle v, \phi_i \rangle_\mu \frac{\phi_j}{\langle \phi_i, 1 \rangle_\mu}.$$

Therefore,

$$\begin{aligned} Q\mathcal{T}Q\phi_k &= Q\mathcal{T}\phi_k \\ &= \sum_{i,j=1}^{n} S_{ji}^{-1} \langle \mathcal{T}\phi_k, \phi_i \rangle_\mu \frac{\phi_j}{\langle \phi_i, 1 \rangle_\mu} \\ &= \sum_{i,j=1}^{n} S_{ji}^{-1} \frac{\langle \phi_k, \mathcal{T}\phi_i \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} \phi_j \\ &= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} S_{ji}^{-1} T_{ik} \right) \phi_j \\ &= \sum_{j=1}^{n} \left(S^{-1} T \right)_{kj} \phi_j. \end{aligned}$$

□

Choosing the Function Space

In order to obtain meaningful results with PCCA+, the functions ϕ_i should be chosen such that they can be correlated to a set of conformations. This is fulfilled if

$$\phi_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \phi_i(x) = 1. \tag{2.3}$$

To see this, let us recall how PCCA+ computes the cluster. Assume we have a state space discretization on the finite-dimensional space spanned by the functions $\{\phi_1, \ldots, \phi_n\}$. Then, we can compute the matrix representation $\mathcal{M}$. The eigenfunctions $f_1, \ldots, f_s$ of the first s dominant eigenvalues near 1 from the projected transfer operator correspond to eigenvectors $\beta^i = (\beta_1^i, \ldots, \beta_n^i)$ of $\mathcal{M}$ in the sense that

$$\sum_{k=1}^{n} \beta_k^i \, \phi_k = f_i$$

holds for $i = 1, \ldots, s$. The clustering method PCCA+ works on the representation vectors β^i of the eigenfunctions f_i, i.e. it computes a matrix $\mathcal{A} = [\alpha_{ij}]$ and vectors $\chi^1, \ldots, \chi^s$ where $\chi^i = (\chi_1^i, \ldots, \chi_n^i)$ with

$$\sum_{i=1}^{s} \chi_l^i = 1 \quad \text{and} \quad \chi_l^i \geq 0$$

for all $i, l \geq 0$ and

$$\chi^i = \sum_{l=1}^{s} \alpha_{li} \, \beta^l.$$

The vector $\chi^i = (\chi_1^i, \ldots, \chi_n^i)$ is associated with the function

$$\chi_\phi^i := \sum_{k=1}^{n} \chi_k^i \, \phi_k.$$

We have $\chi_\phi^i(x) \geq 0$ and $\sum_{i=1}^{s} \chi_\phi^i(x) = 1$ from (2.3), the latter can be seen by

$$\begin{aligned}
\sum_{i=1}^{s} \chi_\phi^i(x) &= \sum_{i=1}^{s} \sum_{k=1}^{n} \chi_k^i \, \phi_k(x) \\
&= \sum_{k=1}^{n} \phi_k(x) \left(\sum_{i=1}^{s} \chi_k^i \right) \\
&= \sum_{k=1}^{n} \phi_k(x) \\
&= 1.
\end{aligned}$$

Thus, after projecting the transfer operator on $\phi_1, \ldots, \phi_n$, the metastable sets A_i can be identified by $A_i = \{x \in E \mid \chi_\phi^i(x) \approx 1\}$ for $i = 1, \ldots, s$.

Another direct consequence of (2.3) is that the matricies T and S are both transition matrices, i.e. $\sum_j T_{ij} = \sum_j S_{i,j} = 1$ and $T_{ij}, S_{ij} \geq 0$. Thus, they both give rise to two different Markov chains on a finite state space. However the Galerkin projection $\mathcal{M}$ is generally not a transition matrix.

In the following, three particularly interesting choices of functions that fulfill these conditions are considered in more detail.

Indicator Functions

Let $(A_i)_{i=1,\dots,n}$ be a partition of E, i.e.

$$\bigcup_{i=1}^{n} A_i = E \quad \text{and} \quad A_i \cap A_j = \emptyset$$

then consider the family of *indicator functions*

$$\phi_i(x) = \mathbb{1}_{A_i}(x).$$

In this case, S is the identity matrix and we have $\mathcal{M} = T$, thus the Galerkin projection of the transfer operator is a transition matrix of a Markov chain and therefore often referred to as the Markov State Model [8]. In this case, the entries of the Galerkin projection have a stochastic interpretation. This can be derived from

$$\left\langle \mathcal{T}\mathbb{1}_{A_i}, \mathbb{1}_{A_j} \right\rangle_\mu = \int_{A_j} (\mathcal{T}\mathbb{1}_{A_i})(x)\, \mu(dx) \overset{(1.14)}{=} \int_{A_i} p(x, A_j)\, \mu(dx) \overset{(1.8)}{=} \mathbb{P}_\mu[X_1 \in A_j,\, X_0 \in A_i]$$

and

$$\langle \mathbb{1}_{A_i}, \mathbb{1} \rangle_\mu = \int_E \mathbb{1}_{A_i}(x)\, \mu(dx) = \mu(A_i) = \mathbb{P}_\mu[X_0 \in A_i]$$

where $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_\mu)$ is the associated Markov chain of the transfer operator. Combined, the two equations produce

$$T_{ij} = \frac{\left\langle \mathcal{T}\mathbb{1}_{A_i}, \mathbb{1}_{A_j} \right\rangle_\mu}{\langle \mathbb{1}_{A_i}, \mathbb{1} \rangle_\mu} = \mathbb{P}_\mu[X_1 \in A_j \mid X_0 \in A_i]. \tag{2.4}$$

If the dimension d of the state space E grows, the size N of an equidistant partition of sets will explode with d and leads to the curse of dimension. A possible way to slow down the curse of dimension is to use sparse grids [62, 21] as used by Junge and Koltai [28] to discretize the transfer operator for a deterministic system.

Radial Basis Functions

In [58, 59, 47, 11] a meshless discretization of the transfer operator is presented. It uses the functions

$$\phi_i(x) = \frac{1}{Z(x)} \exp(-\alpha d^2(x, x_i))$$

with

$$Z(x) = \sum_j \exp(-\alpha d^2(x, x_j)),$$

and the function d is a function that measures a distance between points; this distance could either be certain internal chemical coordinates of the molecule or the euclidean distance in $\mathbb{R}^d$. While this does not lead to a curse of dimension, it does leads to the problem of how to place the points (x_i) properly.

Committor Functions

In [54] another meshless discretization of the transfer operator is presented. One starts with some disjoint sets $(C_i)_{i=1,\dots,n}$. Unlike the indicator functions, the sets $(C_i)_{i=1,\dots,n}$ do not form a partition. These sets are called *core sets* and each core set gives rise to a committor function $(q_i)_{i=1,\dots,n}$. The committor function $q_i(x)$ at point $x \in E$ is defined as the probability that a process starting in x will visit C_i before it visits any other core set. It is possible to compute the Galerkin projection without explicitly computing the committor functions. In addition, it has been shown that eigenfunctions and eigenvalues of the transfer operator are extremely well approximated by the committor functions if the core sets are placed close to metastable sets.

Similarly to the problem for radial basis functions, this approach has the advantage that it does not lead to a curse of dimension, but it is unclear where to place the core sets. As a method for identifying where to place the core sets, in [49] a high temperature sampling procedure is proposed to give a first indication for the metastable sets, since the core sets should be placed near to the metastable sets.

If one is interested in an approximation of the eigenfunctions of the transfer operator, then the committor functions ϕ_i must be computed explicitly. This is because the eigenvector of the Galerkin approximation only reveals β^j and not the eigenfunction $f_j = \sum_{k=1}^n \beta_k^i \phi_k$ of the projected transfer operator. Also, if one wants to get from χ^i to $\chi_\phi^i = \sum_{k=1}^n \chi_k^i \phi_k$, then also the computation of the committor functions is necessary. Unfortunately, efficient explicit computation of the committor function is not feasible in high dimensions. This is a disadvantage for computing metastable sets with PCCA+ by committor functions, because in contrast to indicator functions or radial basis functions, it is a barrier to move from χ^i to $\chi_\phi^i = \sum_{k=1}^n \chi_k^i \phi_k$.

Molecule Model

We model our molecule in state space with a stochastic differential equation. Possible state spaces are either the configuration space $\mathbb{R}^{3d}$ for molecules with d atoms, or $[0, 2\pi]^d$ for molecules with d torsion angles. We assume that a potential V is given which describes the energy landscape of the state space. We obtain trajectories of the molecule's state space from a stochastic differential equation[1]

$$dX_t = -\nabla V(X_t)\, dt + \sigma\, dW_t, \quad X_0 \sim \mu \tag{2.5}$$

where $\sigma = \sqrt{2\beta^{-1}}$ with temperature $\beta^{-1} = k_B T$, where k_B denotes the Boltzmann's constant and T the absolute temperature.

For each probability measure μ as an initial condition, the solution of the stochastic differential equation is a Markov process $(X_t)_{t\geq 0}$, see [4, Proposition 42.7] and [42, Theorem 7.1.2]. Further, for any lag time τ we get a discretized Markov chain $(X_{n\tau})_{n\in\mathbb{N}}$ with a transition kernel

$$p(x, A) := p_\tau(x, A).$$

[1] A good introduction to the theory of stochastic differential equations can be found in [42].

There is a fundamental advantage to model the molecule as a Markov process[2]. We are interested in the clusters of the molecule and the necessary information is decoded in the eigenfunctions of the transfer operator $\mathcal{T}_{\tau n}$. But any eigenfunction of $\mathcal{T}_\tau$ is also an eigenfunction of $\mathcal{T}_{\tau n} = \mathcal{T}_\tau^n$; thus we only need to compute trajectories for a very short lag time and approximate eigenfunctions of $\mathcal{T}_\tau$. Another advantage is that the process obtained through Equation (2.5) is always reversible [43, Proposition 4.5]. This has a direct consequence for the Galerkin projection. As seen in the last chapter, reversibility of the process is equivalent to the self-adjointness of the transfer operator $\mathcal{T}$. Since the orthogonal projection Q is also self-adjoint, the immediate consequence is that the Galerkin projection itself is also self-adjoint

$$(Q\mathcal{T}Q)^* = Q^*\mathcal{T}^*Q^* = Q\mathcal{T}Q.$$

Thus, the eigenvalues of the Galerkin projection are all real-valued and contained within $[-1, 1]$ since

$$||Q\mathcal{T}Q||_2 \leq 1$$

and a basis of orthogonal eigenvectors exists. However, it might be worth noting that the Galerkin projection P_τ for some fixed τ on finite sets $A_1, \ldots, A_n$ of X_τ does not provide the Galerkin projection $P_{2\tau}$ of $X_{2\tau}$. This is because

$$P_\tau^2 \neq P_{2\tau}$$

in general. A counterexample can be found in [40, Example 3.2.2.].

If V is smooth, and fulfills

$$\lim_{|x|\to\infty} V(x) = \infty$$

and

$$e^{-\beta V(x)} \in L^1(\mathbb{R}^d)$$

for all $\beta > 0$, then the process is also ergodic and the unique invariant distribution is the Gibbs distribution which is given by

$$\mu(x) = \frac{1}{Z} e^{-\beta V(x)}$$

where the normalization factor Z is the partition function

$$Z = \int_{\mathbb{R}^d} e^{-\beta V(x)}\, dx,$$

see [43, Proposition 4,2]. It is often useful to know the (unnormalized) Gibbs distribution analytically, in order to use Monte Carlo methods to compute stationary distributed points. In cases where the process is ergodic, a very recent result [38] shows that the associated restricted transfer operator $\mathcal{T}\colon L^2(\mu) \to L^2(\mu)$ posses a spectral gap if there exists a $p > 2$ with

$$\sup_{||f||_2=1} ||\mathcal{T}f||_p < \infty.$$

[2]There are other reasonable models for a molecule which are not Markov, see [40, Section 2.3.2].

2.2 Basic Computation

We have seen that in order to find metastable sets of a dynamical system, we need to compute the eigenfunctions of the transfer operator. This raises the question whether a Galerkin projection of the transfer operator is a good source for the eigenvectors. This question was first posed in 1960 and is known as *Ulam's Conjecture*. We will present two standard methods for computing the Galerkin projection, and we will show that one can actually find a stochastic interpretation for the exact error of our Galerkin projection. Further, we will reveal some unexpected properties of reversible processes.

In the following we denote with

$$\mathcal{T}\colon L^1(\mu) \to L^1(\mu)$$

a self-adjoint transfer operator, i.e. it is associated to a reversible transition kernel p, and with μ a stationary measure of p. We denote with $((X_n)_{n\in\mathbb{N}}, \mathbb{P}_x)$ the associated Markov chain and we denote with lag time τ the time which is required to move from X_0 to X_1.

Ulam's Conjecture

Rechard [45] drew attention to the operator for the deterministic case in 1956, by following the introduction of the generalized transfer operator in 1954 by Hopf [23]. This work was acknowledged by Ulam in 1960, when he dedicated three pages of his book "A collection of mathematical problems" [57, page 73-75] to the transfer operator. It was in this book that he posed the following conjecture, which is still not completely resolved.

Consider the Lebesuge measure space $([0,1], \mathcal{B}([0,1]), \lambda)$ together with a transfer operator $\mathcal{H}\colon L^1(\lambda) \to L^1(\lambda)$. Ulam then considered an equidistant partition of $[0,1]$ into n sets $A_1^n, \ldots, A_{k(n)}^n$, and defined for each n the matrix $A(n) = [a_{ij}^n]$ by

$$a_{ij}^n = \frac{\left\langle \mathcal{H}\mathbb{1}_{A_i^n}, \mathbb{1}_{A_j^n} \right\rangle_\lambda}{\left\langle \mathbb{1}_{A_i^n}, \mathbb{1}_E \right\rangle_\lambda}.$$

The matrix $A(n)$ is a Markov chain because by the Markov operator property we have

$$\left\langle \mathcal{H}\mathbb{1}_{A_i^n}, \mathbb{1}_E \right\rangle_\lambda = ||\mathcal{H}\mathbb{1}_{A_i^n}||_1 = ||\mathbb{1}_{A_i^n}||_1 = \left\langle \mathbb{1}_{A_i^n}, \mathbb{1}_E \right\rangle_\lambda,$$

and hence

$$\sum_{j=1} a_{ij}^n = \frac{\left\langle \mathcal{H}\mathbb{1}_{A_i^n}, \mathbb{1}_E \right\rangle_\lambda}{\left\langle \mathbb{1}_{A_i^n}, \mathbb{1}_E \right\rangle_\lambda} = 1.$$

Thus, the matrix inherits at least one left normalized invariant vector, which we denote by $\pi^n = (\pi_1^n, \ldots, \pi_{k(n)}^n)$. We are now ready to pose Ulam's conjecture:

Assume $\mathcal{H}$ has a non-negative, unique invariant function $\mu \in L^1([0,1])$, does

$$\sum_{i=1}^{k(n)} \pi_i^n \mathbb{1}_{A_i^n} \to \mu$$

then converge in $L^1(\lambda)$ for $n \to \infty$?

As soon as 16 years later, in 1976, the author Li [15] was able to show that for the deterministic case where the dynamical system is described by a map S, Ulam's conjecture is fulfilled if S is piecewise twice continuously differentiable and $\inf |S'(x)| > 2$.

For our purpose, we are interested in a different matrix $T(n) = [T_{ij}^n]$ with

$$T_{ij}^n = \frac{\left\langle \mathcal{T}\mathbb{1}_{A_i^n}, \mathbb{1}_{A_j^n} \right\rangle_\mu}{\left\langle \mathbb{1}_{A_i^n}, \mathbb{1}_E \right\rangle_\mu},$$

where μ is the unique Gibbs distribution. This matrix coincides with our Galerkin projection on indicator functions. As Theorem 2.1.2 showed, we are interested in the eigenvalues and eigenfuctions of the μ-weighted transfer operator in order to identify metastable sets. Fortunately, there are a variety of advantages to consider the μ weighted Galerkin projection. First, one only needs points distributed according to μ in order to compute the Galerkin projection. Computing equal distributed points according to the Lebesgue measure λ would lead directly to curse of dimension and could not be applied for molecules with many atoms. Secondly, the weighted Galerkin projection inherits beneficial properties of the self-adjoint transfer operator, like a real valued spectrum and real eigenvectors. Thirdly, the eigenvectors and eigenvalues of $T(n)$ converge to the eigenfunctions and eigenvalues of $\mathcal{T}$ in the $||\cdot||_2$ norm if the corresponding eigenvalues are isolated [50, Corollary 5.4].

Computing the Galerkin Projection

When we have a Galerkin projection on a finite dimensional space spanned by a basis $\{\phi_1, \ldots, \phi_n\}$ with the property stated in (2.3), we have to compute estimates of the matrices $T = [T_{ij}]$ and $S = [S_{ij}]$ with

$$T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}, \quad S_{ij} = \frac{\langle \phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}$$

where each entry represents a high-dimensional integral that needs to be solved. It is known that standard approximation techniques such as the trapezian rule fail to compute such an integral in high dimensions, because an accuracy of ε would require $O(\frac{1}{\varepsilon^d})$ computations and grows exponentially with dimension d. However, it is possible to break the curse of dimension in the following sense. To explain the procedure in simplified notation, we consider the general problem of approximating an integral of the form

$$I := \int_E \phi(x)\, \mu(dx), \tag{2.6}$$

where μ is a probability distribution. The term T_{ij} can be written in this form by replacing ϕ with

$$\phi_{ij}(x) = \mathcal{T}\phi_i(x) \cdot \phi_j(x) \cdot \frac{1}{\langle \phi_i, \mathbb{1} \rangle_\mu}.$$

If we have $Y, Y_1, \ldots, Y_N \colon \Omega \to E$ independent random variables which are distributed according to μ, then one can consider the random variable

$$\hat{I} := \frac{1}{N} \sum_{i=1}^{N} \phi(Y_i).$$

Because of

$$\mathbb{E}[\hat{I}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\phi(Y_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\phi(Y_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\phi(Y)\right] = \mathbb{E}\left[\phi(Y)\right] = I,$$

we obtain

$$\mathrm{VAR}[\hat{I}] = ||I - \hat{I}||^2_{L^2(\mathbb{P})}.$$

Since the Y_i are independent, one can compute the variance as

$$\mathrm{VAR}[\hat{I}] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{VAR}[\phi(Y_i)] = \frac{1}{N}\mathrm{VAR}[\phi(Y)]$$

and thus

$$||I - \hat{I}||_{L^2(\mathbb{P})} = \frac{\sqrt{\mathrm{VAR}[\phi(Y)]}}{\sqrt{N}}.$$

Therefore, $\hat{I}$ converges to I in $O(\frac{1}{\sqrt{N}})$ in the $||\cdot||_{L^2(\mathbb{P})}$ norm and is independent of the dimension. However, one should be aware that for higher dimensions the preceding factor $\mathrm{VAR}[\phi(Y)]$ may increase, and obtaining independent samples from a given distribution is generally a difficult task.

We will now start discussing two major approaches for obtaining a Markov State Model. The first approach is to compute a single long-term trajectory. The idea is that we first approximate the term

$$C_{ij} = \langle \mathcal{T}\phi_i, \phi_j \rangle_\mu = \langle \phi_i, \mathcal{U}\phi_j \rangle_\mu \overset{(1.7)}{=} \int_E \phi_i(x)\, \mathbb{E}_x[\phi_j(X_1)]\, \mu(dx).$$

Thus we need $Y_1, \ldots, Y_N$ μ-distributed random variables and can estimate C_{ij} by

$$\frac{1}{N}\sum_{k=1}^{N}\phi_i(Y_k)\,\mathbb{E}_{Y_k}[\phi_j(X_1)].$$

Then, for each k we have to compute m_k trajectories starting at Y_k of length τ and ending in $Y_1^k, \ldots Y_{m_k}^k$; the final approximation is then given by

$$\widetilde{C}_{ij} = \frac{1}{N}\sum_{k=1}^{N}\phi_i(Y_k)\left(\frac{1}{m_k}\sum_{l=1}^{m_k}\phi_j(Y_l^k)\right).$$

The computation of

$$\frac{\langle \mathcal{T}\mathbb{1}_{A_i}, \mathbb{1}_{A_j} \rangle_\mu}{\langle \mathbb{1}_{A_i}, \mathbb{1}_E \rangle_\mu}$$

can be obtained after normalizing the matrix $\widetilde{C}$ row-wise

$$\widetilde{T}^L_{ij} := \frac{\widetilde{C}_{ij}}{\sum_{j=1}^N \widetilde{C}_{ij}}. \tag{2.7}$$

Analogously we can compute $\widetilde{S}^L_{ij}$. Setting $m_k = 1$ for each k allows us to approximate the terms of one single long-term trajectory. Summarizing, we end up with the following scheme.

Long-Term Trajectories To gain an approximation of $\widetilde{T}^L$, we compute a long trajectory $(y_i)_{i=0,\dots,r-1}$ for the dynamics (2.5) by performing r timesteps of size dt using the Euler-Maruyama discretization

$$y_{i+1} = y_i - \nabla V(y_i)\, dt + \sigma\sqrt{dt}\, \eta_i$$

of (2.5), where $\eta_i = (\eta_i^1, \dots, \eta_i^d)$ are independent d-dimensional random variables distributed according to the standard normal distribution. This trajectory is divided into pieces of length l yielding M subtrajectories $(y_i^k)_{i=1,\dots,l} := (y_{lk}, \dots, y_{l(k+1)-1})$ for $k = 0, \dots, M-1$. If the trajectory is long enough, it can be assumed that the points $y_1^0, \dots, y_1^{M-1}$ are distributed according to μ and we can estimate $\widetilde{T}^L$ by

$$\hat{C}_{ij} \quad = \quad \sum_{k=0}^{M-1} \phi_i(y_1^k)\, \phi_j(y_l^k)$$

and

$$\widetilde{T}^L_{ij} \approx \frac{\hat{C}_{ij}}{\sum_{j=1}^n \hat{C}_{ij}}.$$

To compute an estimation of $\widetilde{S}$ we only need $y_1^0, \dots, y_1^{M-1}$. It is good practice to take only every m-th point of the trajectory, to make them at least seem to be independent. Then we can estimate $\widetilde{S}$ by

$$\widetilde{D}_{ij} \quad = \quad \sum_{k=0}^{M-1} \phi_i(y_1^k)\, \phi_j(y_1^k)$$

and

$$\widetilde{S}^L_{ij} \approx \frac{\widetilde{D}_{ij}}{\sum_{j=1}^n \widetilde{D}_{ij}}.$$

In accordance with the explanation above, this will converge in order $O(\frac{1}{\sqrt{M}})$ in the $||\cdot||_{L^2(\mathbb{P})}$ norm to the exact value. However, long-term trajectories are unfeasible in high dimensions. Thus, a different method is needed. We follow the idea of vertical and horizontal sampling which was initially developed for a Galerkin projection onto radial basis functions [58, 47, 59] but can analogously be used for arbitrary functions. The idea is to rewrite

$$\frac{\langle \mathcal{T}\phi_i, \phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} = \frac{\langle \phi_i, \mathcal{T}\phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} = \langle h_i, \mathcal{T}\phi_j\rangle_\mu$$

with

$$h_i(x) := \frac{\phi_i(x)}{\langle \phi_i, \mathbb{1} \rangle_\mu}.$$

If we define the measure $\mu_i(A) := \int_A h_i(x)\, \mu(dx)$, then we are faced with the task of computing

$$T_{ij} = \int_E \mathbb{E}_x[\phi_j(X_1)]\, \mu_i(dx).$$

Thus, we once again need random variables $Y_1, \ldots, Y_N$, but this time distributed according to μ_i and we can estimate T_{ij} by

$$\widetilde{T}^S_{ij} := \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{Y_k}[\phi_j(X_1)]. \tag{2.8}$$

Then, for each k we have to compute m_k trajectories starting at Y_k of length τ and ending in $Y_1^k, \ldots Y_{m_k}^k$, the final approximation is then given by

$$\frac{1}{N} \sum_{k=1}^{N} \frac{1}{m_k} \sum_{l=1}^{m_k} \phi_j(Y_l^k).$$

In this case, a normalization is not needed. The computation of μ_i distributed points can be obtained via the Metropolis Monte Carlo Method, which is also feasible in high dimensions. A good overview of Monte Carlo methods, including the Metropolis method, can be found in [35]. We use the Metropolis methods with Gaussian centered proposal density on the current point, which we will explain in the following. We assume that a probability measure ν with a Lebesgue density f_ν that is known up to a multiplicative constant, i.e.

$$\nu(A) = \int_A \frac{f_\nu(x)}{Z}\, dx$$

with $A \in \Sigma$ and $Z = \int_E f_\nu(x), dx$.

Compute ν distributed points First choose any random x_1^i. We describe now how to pick x_{m+1}^i if x_m^i is given. Consider a proposal state x' which is selected according to the normal distribution $\mathcal{N}(x_m^i, r^2)$ centered on x_m^i with standard derivation r which is chosen appropriate. To decide whether to accept the new state, we compute the quantity

$$a = \frac{f_\nu(x')}{f_\nu(x_m^i)}.$$

If $a \geq 1$, then the new state is accepted, i.e. $x_{m+1}^i := x'$. If $a < 1$, then the new state is accepted with probability a, and otherwise we have $x_{m+1}^i := x_m^i$.

This tool leads to the following approximation scheme for short trajectories.

Short-Term Trajectories To gain an approximation of $\widetilde{T}^S$, we compute for each set A_i points $x_1^i, \ldots, x_N^i$ that are distributed according to μ_i, where x_0^i is taken randomly from A_i. This can be obtained for our molecule model using the scheme explained above with

$$f_\nu(x) = \phi_i(x)\, e^{-\beta V(x)}.$$

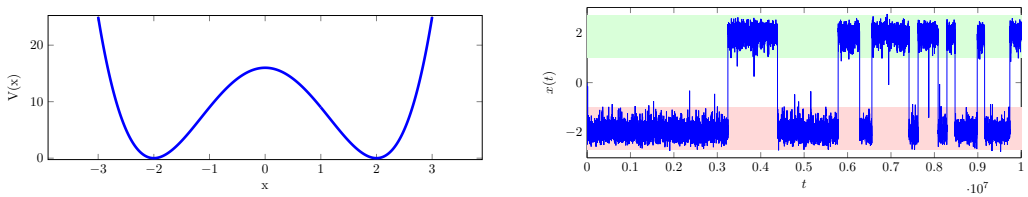

Figure 2.3: Potential (left) and trajectory (right).

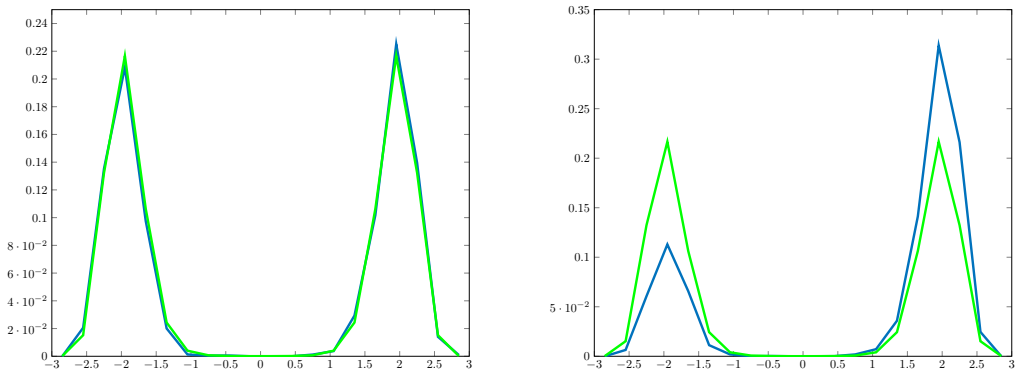

Figure 2.4: Vector π is green and vector v is blue. The error left is $e \approx 0.0195$ and right $e \approx 0.1889$.

For each point x_l^i we compute m_i trajectories again, using a Euler-Maruyama discretization and saving the endpoints $y_1^{i,l}, \dots, y_{m_i}^{i,l}$. The entry $\widetilde{T}_{ij}^S$ is then estimated as

$$\widetilde{T}_{ij}^S \approx \frac{1}{N} \sum_{l=1}^{N} \frac{1}{m_i} \sum_{k=1}^{m_i} \phi_j(y_k^{i,l}).$$

Analogous, the entry $\widetilde{S}_{ij}^S$ can be estimated as

$$\widetilde{S}_{ij}^S \approx \frac{1}{N} \sum_{k=1}^{N} \phi_j(x_k^i)$$

We will now discuss both computation schemes on an elementary example. Consider the double well potential $V(x) = (x-2)^2(x+2)^2$ as shown in Figure 2.3 with $\beta = 0.5$ according to our model given by Equation (2.5). We consider in the following the in magnitude closest three eigenvalues to 1, i.e $|\lambda_3| < |\lambda_2| < \lambda_1 = 1$. We compute a Galerkin projection of our model using trajectories of length $\tau = 200 \cdot dt$, with timestep $dt = 0.001$, and with 20 equidistant sets $(A_i)_{i=1,\dots,20}$ partitioning the interval $[-3, 3]$. The vector $\pi = (\mu(A_1), \dots, \mu(A_{20}))$ can be computed using the trapeze rule. Since the vector π should be equal to the normalized left eigenvalue v to eigenvalue 1, we will compare the error $e = ||\pi - v||$ in the standard euclidean norm. Figure 2.4 shows two examples of the impact of the size of e. We conducted an experiment by computing 100 Galerkin projections for long and short-term trajectories. The results are shown in Table 2.2 and Table 2.3 respectively. For each Galerkin projection, the values λ_2, λ_3 and e have been evaluated. The terms $\mathbb{E}[\lambda_2], \mathbb{E}[\lambda_3]$ and $\mathbb{E}[e]$ describe the mean value of the 100 computations, and $\sigma(\lambda_2), \sigma(\lambda_3)$ and $\sigma(e)$ describe the standard variance of the 100 computations. We conducted the experiment with $5,000$ trajectories and repeated the experiment with an increasing

trajectories	$\mathbb{E}[\lambda_2]$	$\sigma(\lambda_2)$	$\mathbb{E}[\lambda_3]$	$\sigma(\lambda_3)$	$\mathbb{E}[e]$	$\sigma(e)$
$50,000$	0.9995	$1,3 \cdot 10^{-4}$	$0.0632 + 0.0065i$	0.907	0.0931	0.0681
$500,000$	0.9995	$3.7 \cdot 10^{-5}$	0.0598	0.012	0.0269	0.02

Table 2.2: Long-term trajectory computation.

trajectories	$\mathbb{E}[\lambda_2]$	$\sigma(\lambda_2)$	$\mathbb{E}[\lambda_3]$	$\sigma(\lambda_3)$	$\mathbb{E}[e]$	$\sigma(e)$
$5,000$	0.9996	$4,0 \cdot 10^{-4}$	$0.0398 + 0.0046i$	0.0594	0.2383	0.0935
$50,000$	0.9995	$1,3 \cdot 10^{-4}$	0.0593	0.0026	0.1542	0.0364
$500,000$	0.9995	$5,6 \cdot 10^{-5}$	0.0592	$7,3 \cdot 10^{4}$	0.1216	0.0096
$5,000,000$	0.9995	$1,5 \cdot 10^{-5}$	0.0591	$2,210^{4}$	0.1152	0.0012

Table 2.3: Short-term trajectories computation.

number of trajectories up to $5,000,000$, each of length $\tau = 200 \cdot dt$ with timestep $dt = 0.001$. Thus, the number of total steps is between $200 \cdot 5,000 = 10^7$ and $200 \cdot 5,000,000 = 10^9$. We used a standard derivation $r = 0.0345$ for the Metropolis method. A trajectory computed with the Euler-Maruyama discretization of this model can be found in Figure 2.3. In Figure 2.5, the distribution of $M = 500,000$ starting points is shown for the long-term trajectory, i.e the points $(y_1^k)_{k=0,\dots,M-1}$, and for the short-term trajectories, i.e. the points $(x_j^i)_{i=1,\dots,n,j=1,\dots,N}$ with $n \cdot N = M$. The variance of the experiment is already very small, which indicates convergence. Nonetheless, 10^7 steps is a lot of computation for a one-dimensional problem. Each step is saved as a double and therefore consumes 64 bit. Thus, 10^7 steps need 80 MB storage space, which is certainly far too much data for a simple one dimensional problem. However, we cannot compute a Markov State Model of a long-term trajectory with less then 10^7 steps, because in general that would result in no jumps between the two wells. For the calculation of 10^9 steps, already 8 GB storage space is needed. For the short-term trajectory estimation, we can get better approximations of the eigenvalues, using less data as shown in Table 2.3. On the other hand, the computation of π seems to get better if the sampling is concentrated on the metastable sets. However, the computation of the weights is known to be ill-conditioned [36, 12] and there are better methods available to compute the weights [60].

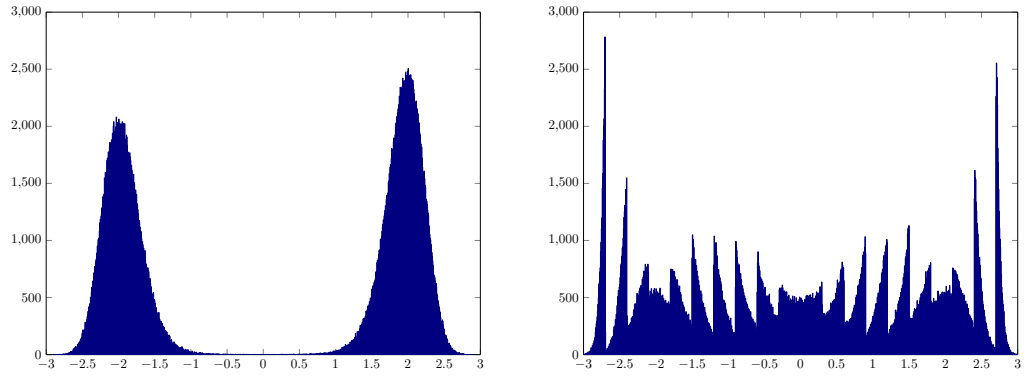

Figure 2.5: Histogram of starting points of long-term (left) and short-term (right) scheme.

Computing the Error

We are again provided with a basis $\{\phi_1, \dots, \phi_n\}$ of measurable functions with the property stated in (2.3), and with the goal to compute estimates of the matrices $T = [T_{ij}]$ and $S = [S_{ij}]$ with

$$T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}, \quad S_{ij} = \frac{\langle \phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}.$$

In the previous section, we have introduced two major concepts of computing the Galerkin projection: The long-term trajectory approach by approximating $\widetilde{T}^L$ from Equation (2.7) and the short-term trajectory approach by approximating $\widetilde{T}^S$ from Equation (2.8). We will now derive for both quantities the exact error in dependency of the number N of trajectories used.

Error for Long-Term Trajectories

We start with the long-term trajectory approach. First of all, the term $\widetilde{T}^L$ can be rewritten as

$$\widetilde{T}^L_{ij} = \frac{1}{N} \sum_{k=1}^{N} \phi_{ij}(Y_k)$$

with $Y_1, \dots, Y_N$ distributed according to μ and

$$\phi_{ij}(x) = \mathcal{T}\phi_i(x) \cdot \phi_j(x) \cdot \frac{1}{\langle \phi_i, \mathbb{1} \rangle_\mu}.$$

The error between the entries $\widetilde{T}^L_{ij}$ and T_{ij} is exactly given by

$$||T_{ij} - \widetilde{T}^L_{ij}||_{L^2(\mathbb{P})} = \frac{\sqrt{\mathrm{VAR}(\phi_{ij}(Y))}}{\sqrt{N}},$$

where Y is distributed according to μ, this follows from the last section, as explained for the approximation of the term in (2.6). Thus, if we could compute $\sqrt{\mathrm{VAR}(\phi_{ij}(Y))}$, we could exactly compute the error of the Galerkin approximation, depending on the number of points N. Surprisingly, we can actually compute this term analytically. To do so, we first

limit ourselves to the simple case of a Galerkin projection onto indicator functions according to a partition $(A_i)_{i=1,\dots,n}$ of E. The term ϕ_{ij} is then given as

$$\phi_{ij}(x) = \mathcal{T}\mathbb{1}_{A_i}(x) \cdot \mathbb{1}_{A_j}(x) \cdot \frac{1}{\mu(A_i)}.$$

We have

$$\mathrm{VAR}[\phi_{ij}(Y)] = \mathbb{E}[\phi_{ij}^2(Y)] - \mathbb{E}[\phi_{ij}(Y)]^2$$

and

$$\mathbb{E}[\phi_{ij}(Y)] \overset{(*)}{=} \int_E \phi_{ij}(x)\,\mu(dx) \overset{(**)}{=} \mathbb{P}_\mu[X_1 \in A_j \mid X_0 \in A_i]$$

where $(*)$ follows from (1.3) and $(**)$ has been shown in (2.4). The other term can be computed as follows

$$\begin{aligned}
\mathbb{E}[\phi_{ij}^2(Y)] &\overset{(*)}{=} \frac{1}{\mu(A_i)^2} \int \mathcal{T}\mathbb{1}_{A_i}(x) \cdot \left(\mathcal{T}\mathbb{1}_{A_i} \cdot \mathbb{1}_{A_j}\right)(x)\,\mu(dx) \\
&\overset{(**)}{=} \frac{1}{\mu(A_i)^2} \int \mathbb{1}_{A_i}(x) \cdot \left(\mathcal{U}\left(\mathcal{T}\mathbb{1}_{A_i} \cdot \mathbb{1}_{A_j}\right)\right)(x)\,\mu(dx) \\
&= \frac{1}{\mu(A_i)^2} \int_{A_i} \int_{A_j} \mathcal{T}\mathbb{1}_{A_i}(y)\,p(x,dy)\,\mu(dx),
\end{aligned}$$

where $(*)$ follows from Equation (1.3) and $(**)$ follows from Euqation (1.18). Since $\mathcal{T}$ is associated with a reversible process, we have $\mathcal{T}\mathbb{1}_{A_j}(y) \overset{(1.22)}{=} \mathcal{U}\mathbb{1}_{A_j}(y) = p(y, A_j)$ and the variance simplifies to

$$\begin{aligned}
\mathbb{E}[\phi_{ij}^2(Y)] &= \frac{1}{\mu(A_i)^2} \int_{A_i} \int_{A_j} p(y, A_i)\,p(x,dy)\,\mu(dx) \\
&\overset{(*)}{=} \frac{1}{\mu(A_i)^2} \mathbb{P}_\mu[X_2 \in A_i,\, X_1 \in A_j,\, X_0 \in A_i] \\
&= \frac{1}{\mu(A_i)} \mathbb{P}_\mu[X_2 \in A_i,\, X_1 \in A_j \mid X_0 \in A_i]
\end{aligned}$$

where $(*)$ follows from Definition 1.1.17. Thus, the variance for reversible processes is exactly given by

$$\mathrm{VAR}[\phi_{ij}(Y)] = \frac{\mathbb{P}_\mu[X_2 \in A_i,\, X_1 \in A_j \mid X_0 \in A_i]}{\mathbb{P}_\mu[X_0 \in A_i]} - \mathbb{P}_\mu[X_1 \in A_j \mid X_0 \in A_i]^2.$$

Computation of the variance is similar to the previous approximations. First of all, note that the entries $\mathbb{P}[X_1 \in A_j \mid X_0 \in A_i]$ are given by the Galerkin projection itself and $\mathbb{P}[X_0 \in A_i]$ can be computed as the left eigenvector of the Galerkin projection, or with the better conditioned method presented in [60]. Thus it remains to compute the term

$$\mathbb{P}_\mu[X_2 \in A_i,\, X_1 \in A_j \mid X_0 \in A_i].$$

To do so, we write

$$\mathbb{P}_\mu[X_2 \in A_i,\, X_1 \in A_j \mid X_0 \in A_i] = \frac{1}{\mu(A_i)} \int_{A_i} \int_{A_j} p(y, A_i)\,p(x,dy)\,\mu(dx) = \int_E f(x)\,\mu_i(dx)$$

with $\mu_i(A) = \frac{1}{\mu(A_i)} \int_A \mathbb{1}_{A_i}(x)\, \mu(dx)$ and $f(x) = \int_{A_j} p(y, A_i)\, p(x, dy)$. As a first step, we need to compute points $Y_1, \dots, Y_{N_1}$ according to μ_i. Then the term is approximated by

$$\frac{1}{N_1} \sum_{l=1}^{N_1} f(Y_l).$$

To compute the term $f(Y_l)$, we need points $Y_1^l, \dots, Y_{N_2}^l$ distributed according to $p(Y_l, \cdot)$. This can be obtained by starting N_2 trajectories of length τ from Y_l; the endpoints are the corresponding points. Then

$$f(Y_l) \approx \frac{1}{N_2} \sum_{k=1}^{N_2} \mathbb{1}_{A_j}(Y_k^l)\, p(Y_k^l, A_i).$$

Finally, for each k we need to approximate the term $p(Y_k^l, A_i) = \int_E \mathbb{1}_{A_i}(x)\, p(Y_k^l, dx)$. For this we require N_3 points $Z_1^k, \dots, Z_{N_3}^k$ distributed according to $p(Y_k^l, \cdot)$, which can be obtained by starting N_3 trajectories from Y_k^l of length τ and taking the endpoints. Then, the term is approximated by

$$p(Y_k^l, A_i) \approx= \frac{1}{N_3} \sum_{r=1}^{N_3} \mathbb{1}_{A_i}(Z_r^k).$$

This leads to the following approximation of the variance.

Variance Computation In order to approximate the term $\mathrm{VAR}[\phi_{ij}(Y)]$ we assume that the matrix $\widetilde{T}^L$ has been computed together with an approximation $\widetilde{\pi}$ of π. As a first step, we need to compute points $y_1, \dots, y_{N_1}$ distributed according to μ_i. For each point y_l we compute N_2 trajectories with start point y_l of length τ and denote the end points with $y_1^l, \dots, y_{N_2}^l$. Then, for each point y_k^l we compute N_3 trajectories with the start point y_k^l of length τ and denote the end points with $z_1^k, \dots, z_{N_3}^k$. The variance is then approximated by

$$\mathrm{VAR}[\phi_{ij}(Y)] \approx \frac{1}{\widetilde{\pi}_i \cdot N_1 \cdot N_2 \cdot N_3} \Big(\sum_{l=1}^{N_1} \sum_{k=1}^{N_2} \sum_{r=1}^{N_3} \mathbb{1}_{A_j}(y_k^l)\, \mathbb{1}_{A_i}(z_r^k) \Big) - (\widetilde{T}_{ij}^L)^2.$$

We now look at the general case where the transfer operator is projected to a subspace D with basis of measurable functions $\phi_1, \dots, \phi_n \geq 0$ on E. In this scenario, we must compute the terms

$$T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} \quad \text{and} \quad S_{ij} = \frac{\langle \phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}.$$

To compute the variance associated to the error of $\widetilde{S}_{ij}^L$, we need to compute $\mathrm{VAR}[\hat{\phi}_{ij}(Y)]$ with

$$\hat{\phi}_{ij}(x) = \phi_i(x)\, \phi_j(x) \frac{1}{\langle \phi_i, \mathbb{1} \rangle_\mu}$$

where Y is still a random variable distributed according to μ. Considering the measure

$$\mu_i(A) := \frac{1}{\langle \phi_i, \mathbb{1} \rangle_\mu} \int_A \phi_i(x)\, \mu(dx),$$

one directly obtains

$$\begin{aligned}\mathrm{VAR}[\hat{\phi}_{ij}(Y)] &= \mathbb{E}[\hat{\phi}_{ij}^2(Y)] - \mathbb{E}[\hat{\phi}_{ij}(Y)]^2 \\ &= \mathbb{E}_\mu[\hat{\phi}_{ij}^2] - \mathbb{E}_\mu[\hat{\phi}_{ij}]^2 \\ &= \mathbb{E}_{\mu_i}[\phi_i\,\phi_j^2]\,\mathbb{E}_\mu[\phi_i]^{-1} - \mathbb{E}_{\mu_i}[\phi_j]^2,\end{aligned}$$

where we have used Equation (1.3). In order to compute the variance associated to the error of $\widetilde{T}_{ij}^L$, we need the following observation. For sets $A, B \in \Sigma$, we have[3]

$$\mathbb{E}_{[\mu]}[\mathbb{1}_A(X_0)\,\mathbb{1}_B(X_1)] = \mathbb{P}_\mu[X_0 \in A, X_1 \in B] \overset{(*)}{=} \int_E \mathbb{1}_A(x) \int_E \mathbb{1}_B(y)\,p(x,dy)\,\mu(dx)$$

where $(*)$ follows from Definition 1.1.17. For measurable functions $f, g \geq 0$, we then obtain

$$\mathbb{E}_{[\mu]}[f(X_0)\,g(X_1)] = \int_E f(x) \int_E g(y)\,p(x,dy)\,\mu(dx). \tag{2.9}$$

Now, for $\phi_{ij}(x) = \mathcal{T}\phi_i(x)\,\phi_j(x)\frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu}$ we first compute

$$\begin{aligned}\mathbb{E}[\phi_{ij}^2(Y)] &\overset{(1.3)}{=} \frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu^2} \int_E \left(\mathcal{T}\phi_i(x)\,\phi_j(x)\right)^2 \mu(dx) \\ &= \frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu^2} \int_E \mathcal{T}\phi_i(x) \left(\phi_j^2 \cdot \mathcal{T}\phi_i\right)(x)\,\mu(dx) \\ &\overset{(1.18)}{=} \frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu^2} \int_E \phi_i(x) \left(\mathcal{U}\left(\phi_j^2 \cdot \mathcal{T}\phi_i\right)(x)\right)\mu(dx) \\ &= \frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu^2} \int_E \phi_i(x) \left(\int_E \phi_j^2(y)\,\mathcal{T}\phi_i(y)\,p(x,dy)\right)\mu(dx).\end{aligned}$$

Knowing that $\mathcal{T}$ is reversible, one obtains from Equation (1.22)

$$\mathcal{T}\phi_i(y) = \mathbb{E}_y[\phi_i(X_1)],$$

leading to

$$\mathbb{E}[\phi_{ij}^2(Y)] = \frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu^2} \int_E \phi_i(x) \int_E \phi_j^2(y)\,\mathbb{E}_y[\phi_i(X_1)]\,p(x,dy)\,\mu(dx).$$

From Equation (2.9) we obtain

$$\mathbb{E}[\phi_{ij}^2(Y)] = \frac{1}{\langle\phi_i,\mathbb{1}\rangle_\mu^2}\mathbb{E}_{[\mu]}\Big[\phi_i(X_0)\,\phi_j^2(X_1)\,\mathbb{E}_{X_1}[\phi_i(X_1)]\Big].$$

Further, according to [37, Equation (3.28) in Chapter 3], we have

$$\mathbb{E}_{X_1}[\phi_i(X_1)] = \mathbb{E}_{[\mu]}[\phi_i(X_2) \mid \mathcal{F}_1],$$

where $\mathcal{F}_1 = \sigma(X_0, X_1)$. Thus, we obtain

$$\mathbb{E}_{[\mu]}[\phi_i(X_0)\phi_j^2(X_1)\mathbb{E}_{X_1}[\phi_i(X_1)] = \mathbb{E}_{[\mu]}\Big[\phi_i(X_0)\phi_j^2(X_1)\mathbb{E}_{[\mu]}[\phi_i(X_2) \mid \mathcal{F}_1]\Big]$$

[3] One may want to recall the notation made in (1.6).

Also note that we have

$$\mathbb{E}_{[\mu]}[\mathbb{1}_A \, \mathbb{E}_{[\mu]}[\phi_i(X_2) \mid \mathcal{F}_1]] = \mathbb{E}_{[\mu]}[\mathbb{1}_A \, \phi_i(X_2)]$$

for any $A \in \mathcal{F}_1$; this follows directly from Definition 1.1.12. Since $\phi_i(X_0)\,\phi_j^2(X_1)$ is measurable according to $\mathcal{F}_1$, we obtain

$$\mathbb{E}_{[\mu]}\big[\phi_i(X_0)\,\phi_j^2(X_1)\,\mathbb{E}_{[\mu]}[\phi_i(X_2) \mid \mathcal{F}_1]\big] = \mathbb{E}_{[\mu]}\big[\phi_i(X_0)\,\phi_j^2(X_1)\,\phi_i(X_2)\big].$$

Thus we finally get

$$\mathbb{E}[\phi_{ij}^2(Y)] = \mathbb{E}_{\nu_i}[\phi_j^2(X_1)\,\phi_i(X_2)]\,\mathbb{E}_{\mu}[\phi_i]^{-1}$$

with

$$\nu_i(A) = \int_A \frac{\phi_i(X_0(\omega))}{\int_\Omega \phi_i(X_0(\tilde{\omega}))\,\mathbb{P}_{[\mu]}(d\tilde{\omega})}\,\mathbb{P}_{[\mu]}(d\omega)$$

for all measurable sets A. Similarly, but more simply, one obtains

$$\mathbb{E}[\phi_{ij}(Y)] = \mathbb{E}_{\nu_i}[\phi_j(X_1)].$$

This gives us the exact formula for the variance

Theorem 2.2.1. *It holds*

$$\mathrm{VAR}[\phi_{ij}(Y)] = \frac{\mathbb{E}_{\nu_i}[\phi_j^2(X_1)\,\phi_i(X_2)]}{\mathbb{E}_{[\mu]}[\phi_i(X_0)]} - \mathbb{E}_{\nu_i}[\phi_j(X_1)]^2.$$

For committor functions, the term $\mathbb{E}_{\nu_i}[\phi_j^2(X_1)\,\phi_i(X_2)]$ is visualized in Figure 2.6, and can be interpreted as the conditioned probability that one came last from core set A_i, moves for lag time τ, goes next to core set A_j, then comes back from core set A_j, moving for lag time τ again, and goes next to core set A_i.

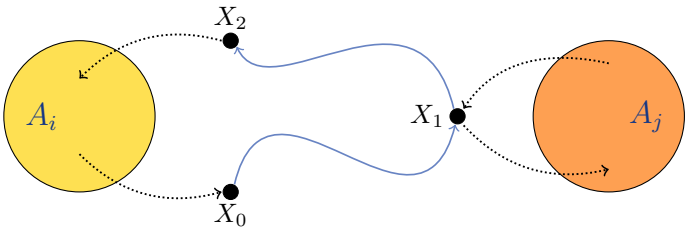

Figure 2.6: Visualization of partial variance term.

Error for Short-Term Trajectories

If we take a close look at the variance formula for the long-term trajectory scheme in the set-based case, it is obvious that the variance is quite high for transient sets. This makes sense, because if we sample points globally by a long-term trajectory, then we need a lot of

points in order to keep the estimation error low for transition regions, since they will hardly be visited by the long-term trajectory. This fundamental error can be overcome by using the short-term trajectory approach. We will see that the variance will become dramatically smaller, even in transition regions.

We look again at the general case where the transfer operator is projected to a subspace D with basis of measurable functions $\phi_1, \dots, \phi_n \geq 0$ on E. In this scenario, we must compute the terms

$$T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} \quad \text{and} \quad S_{ij} = \frac{\langle \phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}.$$

To compute the variance associated to the error of $\widetilde{S}_{ij}^S$, we need to compute $\mathrm{VAR}[\hat{\phi}_j(Y)]$ where Y distributed according to

$$\mu_i(A) = \frac{1}{\langle \phi_i, \mathbb{1} \rangle_\mu} \int_A \phi_i(x)\, \mu(dx).$$

The variance is simply given as

$$\begin{aligned} \mathrm{VAR}[\phi_j(Y)] &= \mathbb{E}[\phi_j^2(Y)] - \mathbb{E}[\phi_j(Y)]^2 \\ &= \mathbb{E}_{\mu_i}[\phi_j^2] - \mathbb{E}_{\mu_i}[\phi_j]^2. \end{aligned}$$

In order to compute the variance associated to the error of $\widetilde{T}_{ij}^S$, note first that we can rewrite this term by

$$\widetilde{T}_{ij}^S = \frac{1}{N} \sum_{k=1}^{N} \mathcal{T}\phi_j(Y_k),$$

where $Y_1, \dots, Y_N$ are distributed according to μ_i The error between the entries $\widetilde{T}_{ij}^S$ and T_{ij} is exactly given by

$$||T_{ij} - \widetilde{T}_{ij}^S||_{L^2(\mathbb{P})} = \frac{\sqrt{\mathrm{VAR}(\mathcal{T}\phi_j(Y))}}{\sqrt{N}},$$

where Y is distributed according to μ_i, this follows from the last section, as explained for the approximation of the term in (2.6). Again, we split the term into

$$\mathrm{VAR}[\mathcal{T}\phi_j(Y)] = \mathbb{E}[(\mathcal{T}\phi_j)^2(Y)] - \mathbb{E}[\mathcal{T}\phi_j(Y)]^2.$$

The last term can be rewritten to

$$\mathbb{E}[\mathcal{T}\phi_j(Y)]^2 \overset{(*)}{=} \left(\int_E \mathcal{T}\phi_j(x)\, \mu_i(dx) \right)^2 = \mathbb{E}_{\mu_i}[\mathbb{E}_x[\mathcal{T}\phi_j(X_1)]]^2,$$

and the first term simplifies to

$$\mathbb{E}[(\mathcal{T}\phi_j)^2(Y)] \overset{(*)}{=} \int_E (\mathcal{T}\phi_j)^2(x)\, \mu_i(dx) = \mathbb{E}_{\mu_i}[\mathbb{E}_x[\phi_j(X_1)]^2],$$

where we have again used Equation (1.3) in $(*)$. Thus, for local sampling the variance is given by

$$\mathrm{VAR}[\mathcal{T}\phi_j(Y)] = \mathbb{E}_{\mu_i}[\mathbb{E}_x[\phi_j(X_1)]^2] - \mathbb{E}_{\mu_i}[\mathbb{E}_x[\phi_j(X_1)]]^2.$$

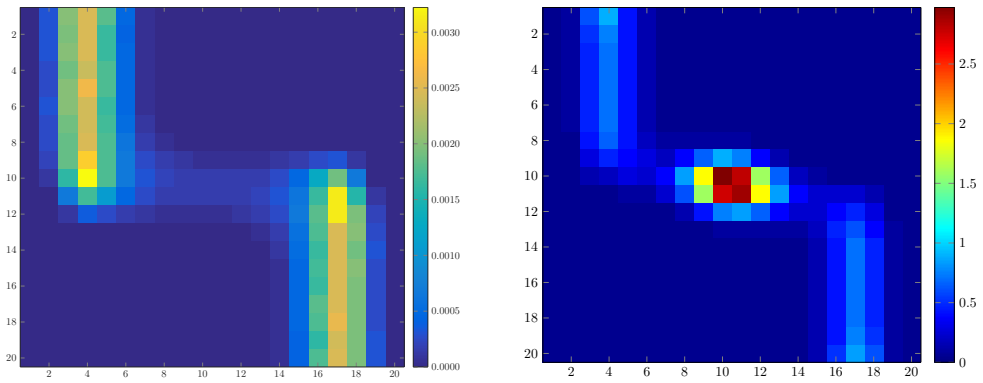

Figure 2.7: Variance matrix of short-term trajectory approach (left) and long-term trajectory approach (right) of V.

Whenever $\phi_j \leq 1$ one has $\mathrm{VAR}[\mathcal{T}\phi_j(Y)] \leq 1$ and thus

$$||T_{ij} - \widetilde{T}_{ij}|| \leq \frac{1}{\sqrt{N}}.$$

In the set based case where $\phi_i(x) = \mathbb{1}_{A_i}(x)$ and $\phi_j(x) = \mathbb{1}_{A_j}(x)$ the variance reduces to

$$\mathrm{VAR}[\mathcal{T}\phi_j(Y)] = \mathbb{E}_{\mu_i}[\mathbb{P}_x[X_1 \in A_j]^2] - \mathbb{P}[X_1 \in A_j \mid X_0 \in A_i]^2.$$

Another advantage of the local sampling method is that it can be computed without the need to double the trajectory length.

Smart Starting Points

So far it has become clear that the long-term trajectory approach is unfeasible in high dimensions at that the short-term trajectory has advantageous error bounds and is feasible in high dimensions. We will now discuss if we can improve the short-term trajectory approach further by adjusting the number of starting points in correlation to a precomputed variance matrix, as opposed to using the same number of starting points for all sets.

Going back to the example from Figure 2.3 in which we considered a partition of 20 equidistant sets, we obtain a short-term variance matrix $M \in \mathbb{R}^{20\times 20}$ which is visualized in Figure 2.7. To introduce the following scheme, consider the vector $\hat{c} \in \mathbb{R}^{20}$ with $\hat{c}_i = \max\{M_{i,1}, \ldots, M_{i,20}\}$ and denote with c the normalization of $\hat{c}$, i.e.

$$c_i = \frac{\hat{c}_i}{\sum_{j=1}^{20} \hat{c}_j}.$$

For a given number of starting points N, we use approximately $N \cdot c_i$ starting points in set A_i. The question arises of whether this distribution of starting points performs better then the equal distribution of starting points between the 20 sets. The result is given in Table 2.4 and shows that whether the starting points are chosen in accordance with the variance, or are equally distributed, does not appear to make a significant difference.

trajectories	$\mathbb{E}[\lambda_1]$	$\sigma(\lambda_1)$	$\mathbb{E}[\lambda_2]$	$\sigma(\lambda_2)$	$\mathbb{E}[e]$	$\sigma(e)$
$5,000$	0.9995	$4,6 \cdot 10^{-4}$	$0.0331 - 0.0052i$	0.0621	0.2605	0.1019
$50,000$	0.9995	$1,4 \cdot 10^{-4}$	0.0589	0.16	0.16	0.0447
$500,000$	0.9995	$4,7 \cdot 10^{-5}$	0.0591	$7.1 \cdot 10^{-4}$	0.1225	0.0092
$5,000,000$	0.9995	$1,5 \cdot 10^{-5}$	0.0590	$2.1 \cdot 10^{-4}$	0.1150	$9.7 \cdot 10^{-4}$

Table 2.4: Computation of short-term trajectories with special start points of V.

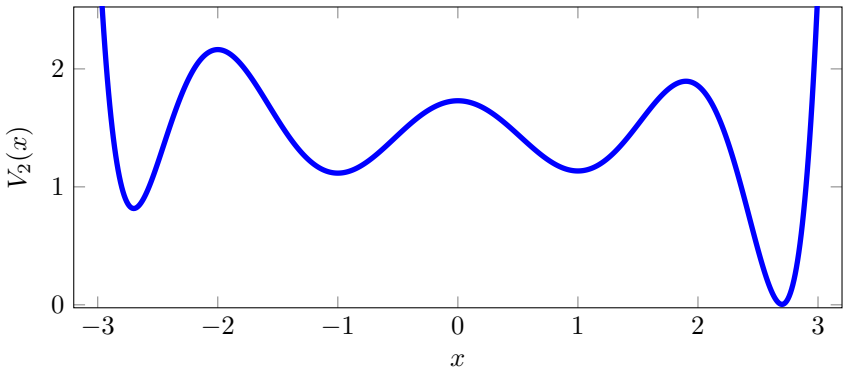

Figure 2.8: Intricate potential V_2.

Since the outcome of the example for the double potential was not insightful, we investigate this further for a more complex potential V_2 given in Figure 2.8 with the gradient

$$\nabla V_2(x) = 0.3(x+2.7)(x+2)(x+1)x(x-1)x(-1.9)(x-2.7).$$

Although we change the potential, we still keep the step size $dt = 0.001$ and the trajectory length $\tau = 200 \cdot dt$, and consider the Galerkin projection onto the partition of 20 equidistant sets from $[-3, 3]$. The associated variance matrix can be seen in Figure 2.9. We conducted an experiment by computing 100 Galerkin projections for short-term trajectories with equally distributed starting points. For each Galerkin projection, the eigenvalues λ_2, λ_3 were computed. The term $\mathbb{E}[\lambda_i]$ describes the mean value of the i-th eigenvalue from the 100 computations, and $\sigma(\lambda_i)$ describes the associated standard variance of the 100 computations. We conducted the same experiment again, but this time with starting points distributed according to the vector c as explained above. We denote the eigenvalues of this Galerkin projection with $\hat{\lambda}_i$, the mean value of the experiment with $\mathbb{E}[\hat{\lambda}_i]$ and the standard variance by $\sigma(\hat{\lambda}_i)$. The results can be found in Table 2.5 and Table 2.6. Even for the more complicated potential V_2, the proposed smart strategy seems not to make a notable difference for the Galerkin projection.

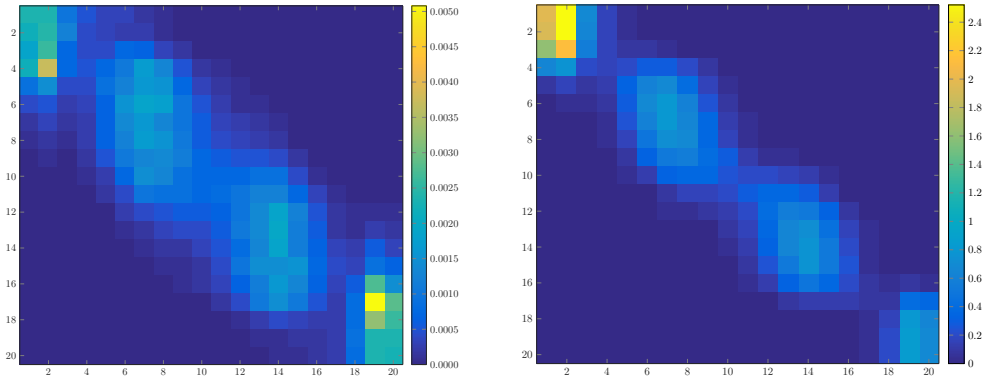

Figure 2.9: Variance matrix of short-term trajectory approach (left) and long-term trajectory approach (right) for V_2

trajectories	$\mathbb{E}[\lambda_1]$	$\sigma(\lambda_1)$	$\mathbb{E}[\lambda_2]$	$\sigma(\lambda_2)$
500	0.9440	0.0209	0.79	0.0463
5,000	0.9394	0.0149	0.78	0.027
50,000	0.9384	0.0209	0.793	0.0216

Table 2.5: Computation of short-term trajectories with special starting points of V_2

trajectories	$\mathbb{E}[\lambda_1]$	$\sigma(\lambda_1)$	$\mathbb{E}[\lambda_2]$	$\sigma(\lambda_2)$
500	0.9438	0.0139	0.7832	0.00441
5,000	0.9382	0.0215	0.7936	0.0225
50,000	0.9397	0.0015	0.7869	0.0261

Table 2.6: Computation of short-term trajectories with equidistant starting points of V_2

Non Reversible Case for Long-Term Trajectories

If $\mathcal{T}$ is not reversible, we can use Proposition 1.3.15 which shows the existence of a process $(\widetilde{X}_n)$ with $\mathcal{T}\mathbb{1}_A(x) = \mathbb{E}_x[\mathbb{1}_A(\widetilde{X}_1)]$ and one could rewrite the variation as

$$\mathbb{E}[\phi_{ij}^2(Y)] = \frac{1}{\mu(A_i)^2}\mathbb{E}_\mu\left[\mathbb{1}_{A_i}(x)\mathbb{E}_x\left[\mathbb{1}_{A_j}(X_1)\,\mathbb{E}_{X_1}\left[\mathbb{1}_{A_i}(\tilde{X}_1)\right]\right]\right],$$

or more shortly and only in dependency of the reversed transition kernel as

$$\mathbb{E}[\phi_{ij}^2(Y)] = \frac{1}{\mu(A_i)^2}\int_{A_j}\left(\tilde{p}(x, A_i)\right)^2\mu(dx),$$

which shows

$$\mathbb{E}[\phi_{ij}^2(Y)] \leq \frac{1}{\mu(A_i)^2}\int_{A_j}\tilde{p}(x, A_i)\,\mu(dx) = \frac{\mathbb{P}[\tilde{X}_0 \in A_j \mid \tilde{X}_1 \in A_i]}{\mathbb{P}[\tilde{X}_1 \in A_i]}.$$

The same upper bound can be derived more shortly as follows. From $\mathbb{1}_A \leq \mathbb{1}$ we certainly have $\mathcal{T}\mathbb{1}_A \leq \mathcal{T}\mathbb{1} = \mathbb{1}$ and thus we get

$$\begin{aligned}\mathbb{E}[\phi_{ij}^2(Y)] &\leq \frac{1}{\mu(A_i)^2}\int_{A_i}\mathbb{E}_x[\mathbb{1}_{A_j}(X_1)]\,\mu(dx)\\ &= \frac{1}{\mu(A_i)^2}\int_{A_i}p(x, A_j)\,\mu(dx)\\ &= \frac{\mathbb{P}_\mu[X_1 \in A_j \mid X_0 \in A_i]}{\mathbb{P}_\mu[X_0 \in A_i]}.\end{aligned}$$

In short, for any process whether reversible or not, one can estimate the variance corresponding to long term-trajectories as:

$$\mathrm{VAR}[\phi_{ij}(Y)] \leq \mathbb{P}_\mu[X_1 \in A_j \mid X_0 \in A_i]\cdot\left(\frac{1}{\mathbb{P}_\mu[X_0 \in A_i]} - \mathbb{P}_\mu[X_1 \in A_j \mid X_0 \in A_i]\right).$$

Non Reversible Case for Short-Term Trajectories

If $\mathcal{T}$ is not reversible, we can use again Proposition 1.3.15 which shows the existence of a process $(\widetilde{X}_n)$ with $\mathcal{T}\phi(x) = \mathbb{E}_x[\phi(\widetilde{X}_1)]$. Then, the variance for local sampling in the non-reversible case is exactly given by

$$\mathrm{VAR}[\mathcal{T}\phi_j(Y)] = \mathbb{E}_{\mu_i}[\mathbb{E}_x[\phi_j(X_1)]^2] - \mathbb{E}_{\mu_i}[\mathbb{E}_x[\phi_j(\widetilde{X}_1)]]^2.$$

The Jensen Inequality for Reversible Markov Chains

If we combine the formula for the variance and the fact that the variance is always non negative[4], we obtain a surprising relation which proves true for all reversible Markov chains:

$$\mathbb{P}[X_{2\tau} \in A,\, X_\tau \in B, X_0 \in A] \geq \mathbb{P}[X_\tau \in B, X_0 \in A]^2$$

[4]This is also a special case of the Jensen Inequality.

for all $A, B \in \Sigma$. Setting $B = E$ in particular produces the even more peculiar trueness

$$\mathbb{P}[X_{2\tau} \in A \mid X_0 \in A] \geq \mathbb{P}[X_0 \in A] = \mu(A)$$

for all reversible Markov chains. Specifically, if $(X_t)_{t\geq 0}$ is the solution of Equation (2.5), then we have

$$\mathbb{P}[X_t \in A \mid X_0 \in A] \geq \mathbb{P}[X_0 \in A]$$

for any $t \geq 0$, $A \in \Sigma$. In other words, it is always more likely for a reversible process to return to a set A than to be in it. In a case where the reversible Markov chain only has a finite state space and is decoded by a transition matrix P with stationary π, the equation states

$$\sum_{i,j\in A} P^2(i,j) \geq \sum_{i\in A} \pi_i,$$

and in particular for $A = \{i\}$, we obtain

$$P^2(i,i) \geq \pi_i \,. \tag{2.10}$$

One may note that if $P \in \mathbb{R}^{n\times n}$ is a reversible Markov chain according to a stationary distribution π, and if

$$\sum_{i=1}^{n} P^2(i,i) = 1$$

holds, then this implies that $P^2(i,i) = \pi_i$ for $i = 1, \ldots, n$, which also shows that π is actually the unique invariant measure to which P is reversible[5]. In particular, for the case $n = 2$ this can be characterized by

$$\begin{aligned} 1 &= \sum_{i=1}^{2} P^2(i,i) \\ &= (p_{11})^2 + p_{12}p_{21} + p_{21}p_{12} + (p_{22})^2 \\ &= (p_{11})^2 + 2(1-p_{11})(1-p_{22}) + (p_{22})^2 \end{aligned}$$

which is equivalent to

$$(p_{11} + p_{22} - 1)^2 = 0.$$

Notably, the Inequality (2.10) is sharp and we obtain equality for 2 by 2 matrices if and only if $p_{11} + p_{22} = 1$. Equality (2.10) is sharp if the function

$$e(p_{11}, p_{22}) = p_{11}^2 + 2(1-p_{11})(1-p_{22}) + p_{22}^2$$

is equal to 1. The function is plotted in Figure 2.10.

From the variance formula for measurable functions one obtains by setting $\phi_j = \mathbb{1}$ the following trueness for reversible Markov chains:

$$\mathbb{E}_{\nu_i}[\phi_i(X_2)] \geq \mathbb{E}_{[\mu]}[\phi_i(X_0)]$$

[5]Not every reversible transition matrix possesses a unique invariant measure for which it is reversible, consider for example $P = \mathrm{Id}$.

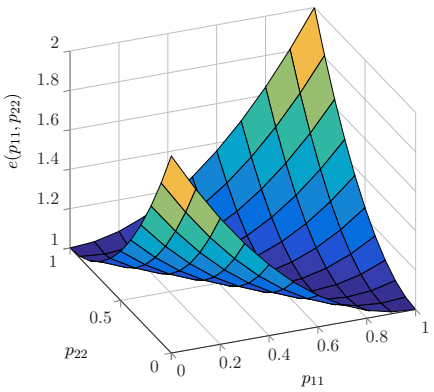

Figure 2.10: Sharpness of Equality (2.10).

for any measurable function ϕ_i. In the case where ϕ_i represents a committor function of a core set C_i, this shows that the conditioned probability that one came last from core set C_i, moves for two time steps, and goes next to core set C_i is always greater or equal then the probability that one came last from core set C_i.

2.3 The Girsanov Reweighting Scheme

We now return to the special problem in computational drug design, of how to find a perfect matching ligand molecule that binds to a malicious receptor molecule in order to inhibit the latter's biological activity. In application, one is often interested in testing multiple slightly different ligands on the same large receptor in order to identify the best fitting ligand. A special receptor-ligand pair is visualized in Figure 2.11. Although the total system is barely altered when the small ligand is exchanged, a complete new Galerkin projection has to be computed each time, involving the dismissal of all previously computed trajectories and calculation of new trajectories in the slightly changed system. In this section we will present a method by which trajectories from the preceding system can be used to help to compute the Galerkin projection after modifying the ligand, saving us from having to compute all the trajectories again. Instead, we only have to compute weights.

These results have previously been published in [51].

Girsanov Transformation

The reweighting scheme is based on the Girsanov transformation. Therefore, we need a short introduction of it.

To this end, let $Y_t^x = Y_t^x(\omega)$ and $X_t^x = X_t^x(\omega)$ be the solutions on a probability space $(\Omega, \Sigma, \mathbb{P})$ of the stochastic differential equations

$$dY_t^x = -\nabla V(Y_t^x)dt + \sigma dB_t \tag{2.11a}$$
$$dX_t^x = -(\nabla V(X_t^x) + \nabla U(X_t^x))dt + \sigma dB_t \tag{2.11b}$$

and deterministic initial conditions

$$Y_0^x(\omega) = X_0^x(\omega) = x \quad \text{almost surely.}$$

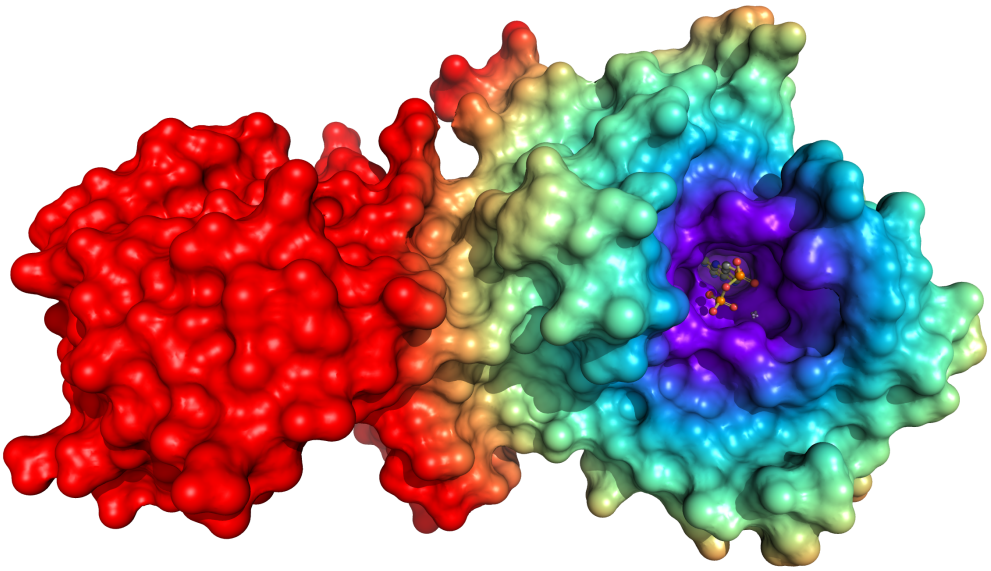

Figure 2.11: Crystal structure of the Phosphoinositol(3,4)-Bisphosphate

For both stochastic differential equations the associated Gibbs distributions are

$$\mu(x) = \frac{1}{Z} \exp(-\beta V(x))$$

and

$$\mu_R(x) = \frac{1}{Z_R} \exp(-\beta(V(x) + U(x)))$$

with associated normalization constant

$$Z = \int \exp(-\beta V(x))\, dx \quad \text{and} \quad Z_R = \int \exp(-\beta(V(x) + U(x)))\, dx.$$

Define $\xi_t \in \mathbb{R}^n$ by

$$\xi_t = \sigma^{-1} \nabla U(Y_t^x) = \sqrt{\frac{\beta}{2}} \cdot \nabla U(Y_t^x).$$

It follows from the Girsanov theorem [42, Thm. 8.6.8], also known as the *Cameron-Martin-Girsanov theorem* [55] that for

$$\mathbb{Q}[A] := \int_A M_t(\omega) \mathbb{P}(d\omega)$$

with

$$M_t := \exp\left(- \int_0^t \xi_s \cdot dB_s - \frac{1}{2} \int_0^t |\xi_s|^2 ds \right),$$

for any measurable set A we get

$$\mathbb{P}[X_t^x \in A] = \mathbb{Q}[Y_t^x \in A],$$

which can also be written as

$$\int \mathbb{1}_A(X_t^x(\omega))\mathbb{P}(d\omega) = \int \mathbb{1}_A(Y_t^x(\omega))\mathbb{Q}(d\omega).$$

In particular, we obtain

$$\begin{aligned}\mathbb{E}[\mathbb{1}_A(X_t^x)] &= \int \mathbb{1}_A(X_t^x(\omega))\mathbb{P}(d\omega) \\ &= \int \mathbb{1}_A(Y_t^x(\omega))\mathbb{Q}(d\omega) \\ &= \int \mathbb{1}_A(Y_t^x(\omega))M_t(\omega)\mathbb{P}(d\omega) \\ &= \mathbb{E}[M_t\mathbb{1}_A(Y_t^x)]\end{aligned}$$

for any measurable set A.

Weighting Scheme

This weighting scheme was developed by the author in collaboration with Christof Schütte and Marcus Weber [51].

Let us take a partition $A_1, \dots, A_n$ of E and denote with T the Galerkin projection from the stochastic differential equation (2.11a) and with T^R the Galerkin projection from the stochastic differential equation (2.11b), in which both Galerkin projections are onto the associated indicator functions. Then, our result yields that

$$\begin{aligned}T_{ij}^R &= \frac{1}{\mu_R(A_i)} \left\langle \mathcal{T}\mathbb{1}_{A_i}, \mathbb{1}_{A_j} \right\rangle_{\mu_R} \\ &= \frac{1}{\mu_R(A_i)} \left\langle \mathbb{1}_{A_i}, \mathcal{U}\mathbb{1}_{A_j} \right\rangle_{\mu_R} \\ &= \frac{1}{\mu_R(A_i)} \int \mathbb{1}_{A_i}(x) \cdot \mathbb{E}\left[\mathbb{1}_{A_j}(X_t^x)\right] \mu_R(x)dx \\ &= \frac{1}{\mu_R(A_i)} \int \mathbb{1}_{A_i}(x) \cdot \mathbb{E}\left[\mathbb{1}_{A_j}(Y_t^x) \exp\left(-\int_0^t \xi_s \cdot dB_s - \frac{1}{2}\int_0^t |\xi_s|^2 ds\right)\right] \mu_R(x)dx.\end{aligned}$$

We have

$$\mu_R(x) = \frac{1}{Z_R} \exp\Big(-\beta(V(x)+U(x))\Big) = \frac{Z}{Z_R}\mu(x)\exp(-\beta U(x)),$$

integrating on both sides and multiplying with Z_R reveals

$$\int \exp\Big(-\beta(V(x)+U(x))\Big)dx = Z \cdot \int_E \mu(x)\, \exp(-\beta U(x))\, dx = Z \cdot \mathbb{E}_\mu[e^{-\beta U}],$$

which reveals

$$Z_R = \int \exp\Big(-\beta(V(x)+U(x))\Big)dx = Z \cdot \mathbb{E}_\mu[e^{-\beta U}],$$

and thus

$$\mu_R(x) = \frac{\mu(x)\cdot\exp(-\beta U(x))}{\mathbb{E}_\mu[e^{-\beta U}]}.$$

All in all, we have

$$\begin{aligned} T_{ij}^R &= \frac{1}{\mu_R(A_i)} \int_{A_i} w_j(t,x)\, g(x)\, \mu(x) dx, \\ w_j(t,x) &= \mathbb{E}\left[\mathbb{1}_{A_j}(Y_t^x) \exp\left(-\int_0^t \xi_s \cdot dB_s - \frac{1}{2}\int_0^t |\xi_s|^2 ds \right) \right], \\ \xi_s &= \sqrt{\frac{\beta}{2}} \cdot \nabla U(Y_s^x), \\ g(x) &= \frac{e^{-\beta U(x)}}{\mathbb{E}_\mu(e^{-\beta U})}. \end{aligned}$$

Consequently, based on the trajectory information that was gained to compute T_{ij}, in principle we can also compute T_{ij}^R.

Note that for

$$C_{ij} = \int_{A_i} w_j(t,x)\, \tilde{g}(x)\, \mu(x) dx$$

with

$$\tilde{g}(x) = e^{-\beta U(x)},$$

we obtain

$$\sum_{j=1}^m C_{ij} = \int_{A_i} \tilde{g}(x)\, \mu(x) dx,$$

and, therefore,

$$\frac{C_{ij}}{\sum_{j=1}^m C_{ij}} = \frac{C_{ij}}{\sum_{j=1}^m C_{ij}} \frac{\frac{1}{c}}{\frac{1}{c}} = T_{ij}^R$$

for $c = \mathbb{E}_\mu(e^{-\beta U})$

The content which follows is once again the independent work of the author.

Algorithmic Realization

In the following, we compare different approximations of the transition matrix T^R related to (2.11b). One form of approximation of T^R is through direct computation, i.e., using the trajectories from (2.11b). We will denote this approximation by $\widetilde{T}^{R,\mathrm{dir}}$. The other one results from the reweighting scheme based on trajectories of (2.11a), denoted simply by $\widetilde{T}^R$. The computation of $\widetilde{T}^{R,\mathrm{dir}}$ is gained by the long-term trajectory approach as explained in the preceding section. We now explain the implementation of the reweighing scheme in detail.

Reweighted computation To gain $\widetilde{T}^R$, we compute a long trajectory $(Y_i)_{i=0,\dots,n-1}$ for the unperturbed dynamics (2.11a) by performing n timesteps of size dt using the Euler-Maruyama discretization

$$Y_{i+1} = Y_i - (\nabla V(Y_i))\, dt + \sigma\sqrt{dt}\, \eta_i$$

of (2.11a), where $\eta_i = (\eta_i^1, \dots, \eta_i^d)$ are independent d-dimensional random variables distributed according to the standard normal distribution. We divide this trajectory into pieces of length l yielding M subtrajectories $(Y_i^k)_{i=1,\dots,l} := (Y_{lk}, \dots, Y_{l(k+1)-1})$ for $k = 0, \dots, M-1$. For a long enough trajectory, it can be assumed that the points

$Y_1^0, \dots, Y_1^{M-1}$ are distributed according to μ. Now we have to approximate for each subtrajectory $(Y_i^k)_{i=1,\dots,l}$ the term $M_t(Y_l^k)$ with $t = l \cdot dt$. To do so, we note that for

$$\mathcal{R} = \int_0^t \xi_s \, dB_s + \frac{1}{2}\int_0^t |\xi_s|^2 ds = \sum_{i=1}^d \left(\int_0^t \xi_s(i) \, dB_s^i \right) + \frac{1}{2}\int_0^t |\xi_s|^2 ds$$

we have $M_t = \exp(-\mathcal{R})$, where $B_s = \left(B_s^1, \dots, B_s^d\right)$ denotes the d-dimensional Brownian Motion with independent components. Thus we need to approximate $\mathcal{R}$. Now each component $\int_0^t \xi_s(i) \, dB_s^i$ can be computed with the Euler-Maruyama discretization by

$$\int_0^t \xi_s(i) \, dB_s^i \approx r_l^i - r_0^i,$$

where

$$\begin{aligned} r_0^i &= x_1^k \\ r_{j+1}^i &= r_j^i + [\xi(r_j^i)(i)] \, \eta_{(kl+j)}^i \sqrt{dt} \end{aligned}$$

and

$$\xi(r) = \sigma^{-1} \nabla U(r).$$

Therefore, for each trajectory $(x_i^k)_{i=1,\dots,l}$ we calculate the weight w_k by

$$\begin{aligned} r_k &= \sum_{i=1}^d (r_l^i - r_0^i) + \frac{1}{2}\sum_{i=1}^l |\xi(Y_i^k)|^2 \, dt \\ w_k &= \exp(-r_k). \end{aligned}$$

Finally, we can compute $\widetilde{T}^R$ by

$$\widetilde{C}_{ij} = \sum_{k=0}^{M-1} \mathbb{1}_{A_i}(Y_1^k) \, \mathbb{1}_{A_j}(Y_l^k) \, w_k \, \tilde{g}(x_1^k)$$

and

$$\widetilde{T}_{ij}^R = \frac{\widetilde{C}_{ij}}{\sum_{i=1}^m \widetilde{C}_{ij}}.$$

We conclude this section with some remarks.

- The division of the long trajectory into subtrajectories can be be optimized by sampling μ distributed points according to the Metropolis Monte Carlo method and sampling short trajectories.
- It is essential that the the random vector $\eta_{(kl+i)}$ which was used to compute the term Y_{i+1}^k is also used to compute the corresponding term r_{kl+i}.

From Butane to Pentane

We now use the following example to show how Girsanov reweighting can be implemented when the ligand's dimension is changed. Consider the one-dimensional, 2π-periodic, artificial potential $V_B : \mathbb{R} \to \mathbb{R}$ of the dihedral angles for butane given by

$$V_B(x) = a + b\,\cos(x) + c\,\cos^2(x) + d\,\cos^3(x)$$

with a=2.0567, b=-4.0567, c=0.3133, d=6.4267 and the two-dimensional, 2π-periodic, artificial potential $V_P : \mathbb{R}^2 \to \mathbb{R}^2$ of the dihedral angles for pentane given by

$$V_P(x, y) = V_B(x) + V_B(y).$$

We will compute the Galerkin projection from pentane by only using trajectories from

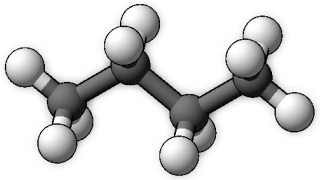

Figure 2.12: Butan

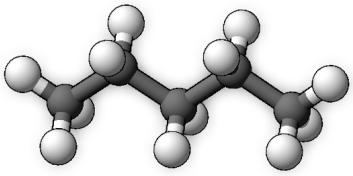

Figure 2.13: Pentane

butane together with the computed weights. Thus, we use the Girsanov reweighting scheme with $V(x, y) = V_B(x)$ and $U(x, y) = V_P(x, y) - V_B(x)$.

We keep the notation $\widetilde{T}^R$ for the Galerkin projection associated with (X_t^x) from (2.11b), which depends on V and U. If we choose $\sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_1 \end{pmatrix}$ with $\sigma_1^2 = 2\beta^{-1}$ and $\beta = 0.5$, then, when replacing $Y_t^x = \begin{pmatrix} y_t \\ z_t \end{pmatrix}$, the equation (2.11a) changes to

$$\begin{aligned} dy_t &= \frac{\partial V_B(y_t)}{\partial x} + \sigma_1 dB_t^1 \\ dz_t &= \sigma_1 dB_t^2. \end{aligned}$$

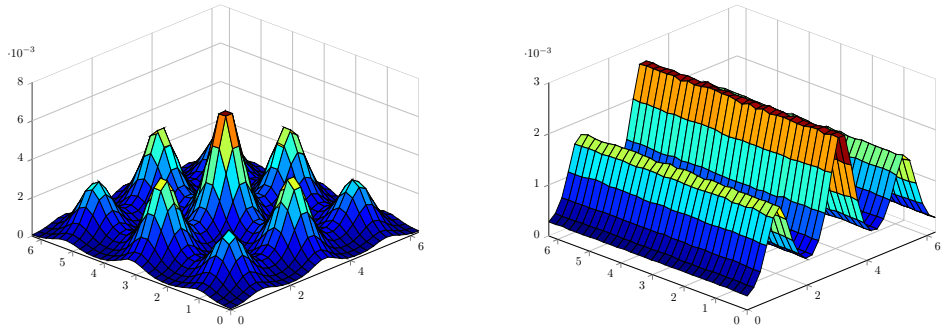

Figure 2.14: Stationary distribution of of Markov State Model from $\widetilde{T}^R$ (left) and Stationary distribution of $\widetilde{T}$ (right).

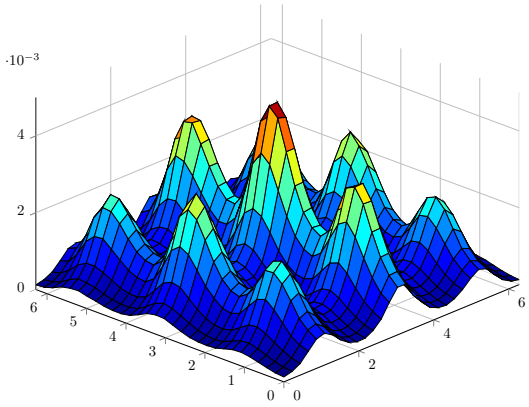

Figure 2.15: Stationary distribution of $\widetilde{T}^{R,\text{dir}}$.

Eigenvalues	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
$\widetilde{T}^{R}$	0.952	0.947	0.946	0.941	0.902	0.896	0.895	0.890	0.648
$\widetilde{T}^{R,\text{dir}}$	0.952	0.952	0.946	0.946	0.906	0.901	0.900	0.895	0.643

Table 2.7: From butane to pentane: First dominating eigenvalues of $\widetilde{T}^{R}$ and $\widetilde{T}^{R,\text{dir}}$.

Note that both terms y_t and z_t can be solved independently. The term y_t represents a trajectory of butan and the term z_t represents the Brownian motion.

We now compute both trajectories independently by performing $n = 4 \cdot 10^8$ timesteps of size $dt = 0.001$ using the Euler-Maryama discretization

$$y_{i+1} = y_i - \nabla V(y_i)\, dt + \sigma_1 \sqrt{dt}\, \eta_i^1,$$
$$z_{i+1} = z_i + \sigma_1 \sqrt{dt}\, \eta_i^2.$$

This yields two statistically independent discrete trajectories y_i, z_i, $i = 0, \ldots, 4 \cdot 10^8 - 1$.

We cut the long trajectory into pieces of length $l = 400$, yielding $M = 10.000.000$ subtrajectories $(y_i^k)_{i=1,\ldots,l}$, $(z_i^k)_{i=1,\ldots,l}$, $k = 0, \ldots, M-1$.

We divide $[0, 2\pi)$ into 30 sets $A_i = [x_i, x_i + \Delta x)$, $i = 1, \ldots, 30$ with $x_i = (i-1)\Delta x$ and $\Delta x = 2\pi/30$. Then, we partition $[0, 2\pi)^2$ into 900 sets B_{ij} with $B_{ij} = A_i \times A_j$ for $i, j = 1, \ldots, 30$ and use the above scheme to construct $\widetilde{T}^R$. In addition we compute an analogous trajectory for pentane and construct $\widetilde{T}^{R,\text{dir}}$ based on the same complete partition (B_{ij}) to compare our approximation.

The eigenvalues given in Table 2.7 and the eigenvector for eigenvalue $\lambda = 1$ given in Figure 2.14 and Figure 2.15 show that the weighted transition matrix $\widetilde{T}^R$ is a good approximation of $\widetilde{T}^{R,\text{dir}}$.

From Butane to Pentane in Atomic Resolution

We have seen that we can compute a Galerkin projection from pentane simply by simulating a trajectory of butan, a Brownian motion on the interval $[0, 2\pi]$, and the Girsanov weights. However, for trajectories in full atomic resolution instead of the torsian angle potential, a Brownian motion on $\mathbb{R}^{3n}$ will not be helpful and one can has to proceed in a different way. First, one needs to split the molecule of interest into two parts[6] that coincide in one single atom, as shown in Figure 2.16. Then, for each part one has to simulate an independent trajectory in full atomic resolution, but one needs to fix the point of coincide[7] in both simulations. Then one can merge both trajectories together, and the resulting trajectory can be seen as the outcome of the original molecule - without taking into account the interactions between the two parts that were once separated. To be more accurate, let us label the position in state space of the atoms from pentane at timestep i by $x_i^1, \dots, x_i^{17} \in \mathbb{R}^3$. We will now fix the yellow atom 13 shown in Figure 2.16; for instance we could choose $x_i^{13} = (0, 0, 0)$ for any $i \in \mathbb{N}$. We then consider the separated parts independently. The lower half of pentane including the yellow atom consists now consists of 13 atoms, but only 12 of them can have a freedom degree. Thus, if we denote the corresponding force field with V_1 it only takes into account the interactions of the first 12 atoms, since the 13th is fixed. A trajectory of the lower part can then be computed by

$$\begin{pmatrix} x_{i+1}^1 \\ \vdots \\ x_{i+1}^{12} \end{pmatrix} = -\nabla V_1 \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^{12} \end{pmatrix} dt + \sigma \begin{pmatrix} dB_i^1 \\ \vdots \\ dB_i^{12} \end{pmatrix}.$$

Analogously we can evaluate the upper half with four degrees of freedom by a force field V_2

$$\begin{pmatrix} x_{i+1}^{14} \\ \vdots \\ x_{i+1}^{17} \end{pmatrix} = -\nabla V_2 \begin{pmatrix} x_i^{14} \\ \vdots \\ x_i^{17} \end{pmatrix} dt + \sigma \begin{pmatrix} dB_i^{14} \\ \vdots \\ dB_i^{17} \end{pmatrix}.$$

Thus, if we join the two independent trajectories together at atom x_i^{13} for all $i \in \mathbb{N}$, then the resulting artificial trajectory z can be viewed as the outcome of the solution of the stochastic differential equation

$$dz_t = -\nabla V(z_t)\, dt + \sigma dB_t. \tag{2.12}$$

where

$$z_i := \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^{12} \\ x_i^{14} \\ \vdots \\ x_i^{17} \end{pmatrix} \quad \text{and} \quad V(z_i) := V_1 \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^{12} \end{pmatrix} + V_2 \begin{pmatrix} x_i^{14} \\ \vdots \\ x_i^{17} \end{pmatrix}.$$

[6] Note that these parts are artificial constructions and no physically meaningful molecules.
[7] The yellow point in Figure 2.16.

If one now denotes with $U(z)$ the potential that only takes into account the interactions between $(x^1, \dots, x^{12})$ and $(x^{14}, \dots, x^{17})$, we can reformulate the problem of how to construct the Galerkin projection of the solution of

$$dz_t = -\nabla(V(z_t) + U(z_t))\, dt + \sigma dB_t$$

when only trajectories of (2.12) are available. With this approach, one can compute new ligands which are created by joining together preexisting ligands. Calculating the weights is more efficent then computing new trajectories, because one only needs to consider the interactions between the separated parts.

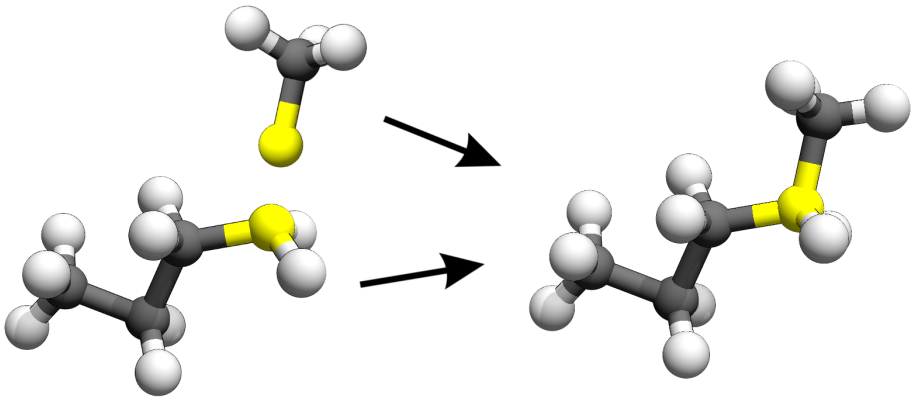

Figure 2.16: Left: Splitting pentane in two parts. Right: Merged parts.

The Advantage for Applications

In this section we presented a method by which trajectories from the preceding system can be used to help to compute the Galerkin projection after modifying the ligand, saving us from having to compute all the trajectories again. Instead, we only have to compute weights. Unfortunately, it turned out that in order to compute the weights, we have to compute as many trajectories as we would have when computing new trajectories for the preceding system from scratch. However, it also turned out that these trajectories can be computed for a much simpler potential. In practice, this means that for a molecule like that shown in Figure 2.11, computation of the new trajectories for the weights need not include all the interactions between all the atoms of the receptor, but only the interaction between the atoms of the receptor and the ligand.

As show in the last section, one may benifit from trajectories of some artificial constructed objects as shown on the left in Figure 2.16 that are not physically meaningful molecules, but which help to create the Galerkin projection for physically meaningful molecules.

Make it Reversible

3

> *If you make a mistake and do not correct it, this is called a mistake.*
>
> **— Confucius**

In the preceding chapters, we have discussed the nature of the transfer operator, and how to obtain a Galerkin projection. In this chapter, we will reveal that the Galerkin projection has an advantageous property that unfortunately is sometimes lost due to numerical errors in the computation schemes. We will present a method for restoring the property after the computation of the Galerkin projection has been obtained. These results have previously been published in [41].

We denote with (E, Σ, μ) a probability space.

3.1 The Reversible Property

We return to the general situation in which we are concerned with a Galerkin projection of a self-adjoint transfer operator $\mathcal{T}\colon L^2(\mu) \to L^2(\mu)$ onto a finite subspace D with basis $\{\phi_1, \ldots, \phi_n\}$ which fulfills

$$\sum_{i=1}^{n} \phi_i(x) = 1$$

for all $x \in E$ and $\phi_i(x) \geq 0$ for all $x \in E$. The Galerkin projection can be computed by approximating the matrices $T = [T_{ij}]$ and $S = [S_{ij}]$ where

$$T_{ij} = \frac{\langle \mathcal{T}\phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}$$

and

$$S_{ij} = \frac{\langle \phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu}$$

by the standard methods outlined in Chapter 2. We will now derive some simple properties of T. Since S is a special case of T where $\mathcal{T}$ is the identity map, i.e. $\mathcal{T}f = f$ for all $f \in L^2(\mu)$, all properties of T also apply to S.

The first property is that T is a transition matrix. This is because $T_{ij} \geq 0$ and

$$\sum_{j=1}^{n} T_{ij} = \sum_{j=1}^{n} \frac{\langle \mathcal{T}\phi_i, \phi_j \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} = \frac{\langle \mathcal{T}\phi_i, \mathbb{1} \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} = \frac{\langle \phi_i, \mathcal{U}\mathbb{1} \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} = \frac{\langle \phi_i, \mathbb{1} \rangle_\mu}{\langle \phi_i, \mathbb{1} \rangle_\mu} = 1.$$

Secondly, the vector $\pi = (\langle \phi_1, \mathbb{1} \rangle_\mu, \ldots, \langle \phi_n, \mathbb{1} \rangle_\mu)$ is a stationary vector of the matrix T, i.e. $\pi T = \pi$, because

$$\pi T(j) = \sum_{k=1}^{n} \pi_k T_{kj} = \sum_{k=1}^{n} \langle \mathcal{T} \phi_k, \phi_j \rangle_\mu = \langle \mathbb{1}, \phi_j \rangle_\mu = \pi_j.$$

Thirdly, because we are modeling our molecule using the stochastic differential equation (2.5), we know that the transfer operator $\mathcal{T}$ is self-adjoint. This implies that the Markov chain T is reversible according to π, i.e.

$$\pi_i T_{ij} = \pi_j T_{ji}$$

for $i, j = 1, \ldots, n$, this follows from

$$\pi_i T_{ij} = \langle \mathcal{T} \phi_i, \phi_j \rangle_\mu = \langle \phi_i, \mathcal{T} \phi_j \rangle_\mu = \pi_j T_{ji}.$$

Summarizing, T fulfills the following four conditions

(i) $\pi T = \pi$,

(ii) $DT = T^T D$,

(iii) $\sum_{j=1}^{n} T_{ij} = 1$ for all $i = 1, \ldots, n$,

(iv) $T_{ij} \geq 0$,

where $D = diag(\pi(1), \ldots, \pi(n))$ denotes the diagonal matrix of π. A matrix fulfilling these 4 conditions is called a π-reversible Markov chain.

If we use the computation schemes from Chapter 2, we end up with a matrix $\widetilde{T}$ which might not fulfill these conditions because of the numerical estimation errors. Thus, our problem can now be formulated as the following:

> For a given transition matrix $\widetilde{T}$ together with a norm, does a π-reversible matrix T^* exists that is closest to $\widetilde{T}$?

This question will be answered in the next section.

3.2 Finding the Closest Reversible Matrix

Given a transition matrix $\widetilde{T} \in \mathbb{R}^{n \times n}$ and a probability vector $\hat{\pi} \in \mathbb{R}^n$, it will be proven in the following that for every norm $|| \cdot ||$ which is induced by a scalar product, there exists a unique $\hat{\pi}$-reversible matrix T^* which minimizes $||T^* - \widetilde{T}||$. Note that the probability vector $\hat{\pi}$ can be chosen arbitrarily. As a little teaser, let us consider the Markov chain

$$T_{Ex} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

represented in Figure 3.1. Its stationary vector is $\pi_1 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and the closest reversible Markov chain of T_{Ex} according to the Frobeniusnorm and π_1 can be found in Figure 3.2.

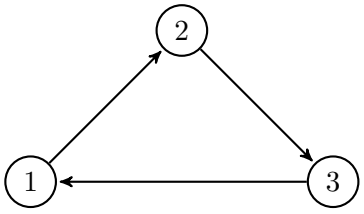

Figure 3.1: Simple Markov chain.

For the arbitary probabilty vector $\pi_2 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$, the closest reversible Markov chain of T_{Ex} according to the Frobeniusnorm with π_2 as stationary measure is highly non-trivial and can also be found approximately in Figure 3.2. At the end of this section, it will become clear how these matricies can be computed.

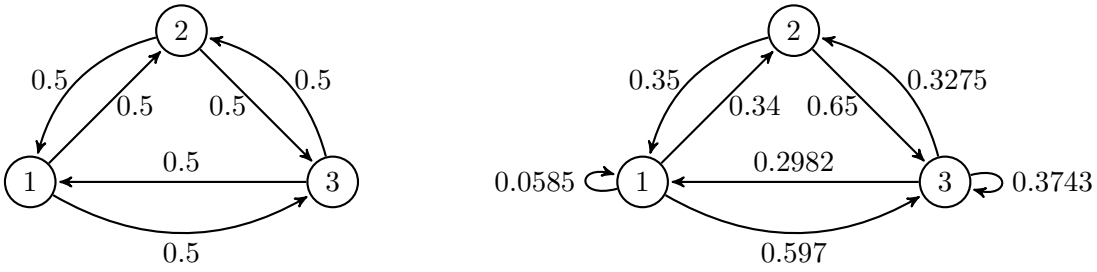

Figure 3.2: Corrected reversible Markov chain according to π_1 (left) and π_2 (right).

We now give a brief outline of the proof. For the moment, let us assume that $n = 2$, then any 2×2 transition matrix P can be presented as a point $(a, b) \in [0, 1]^2$ in the unit square as

$$P = \begin{pmatrix} a & 1-a \\ b & 1-b \end{pmatrix}, \quad \text{with } a, b \in [0, 1].$$

We then consider the set X of all $\hat{\pi}$-reversible matrices as a subset of the unit square, which is represented as the blue line in Figure 3.3. We show then that the map

$$\begin{aligned} X &\to [0, \infty) \\ A &\mapsto ||A - \widetilde{T}|| \end{aligned}$$

is strongly convex and thus possesses a unique minimum on the convex set X. The map is represented in Figure 3.3 as the red graph above X. We will show how to pose this problem as a quadratic convex optimization problem which can be used to compute the closest $\hat{\pi}$-reversible Markov chain T^*. As Stephen P. Boyd and Lieven Vandenberghe [9] have stated,

> *"With only a bit of exaggeration, we can say that, if you formulate a practical problem as a convex optimization problem, then you have solved the original problem."*

We will see later how elastic the term exaggeration actually is.

We will now give the proof. As a first step, we need to find out how to describe a matrix in X. We do so by considering the smallest subspace U that contains X, and then we construct

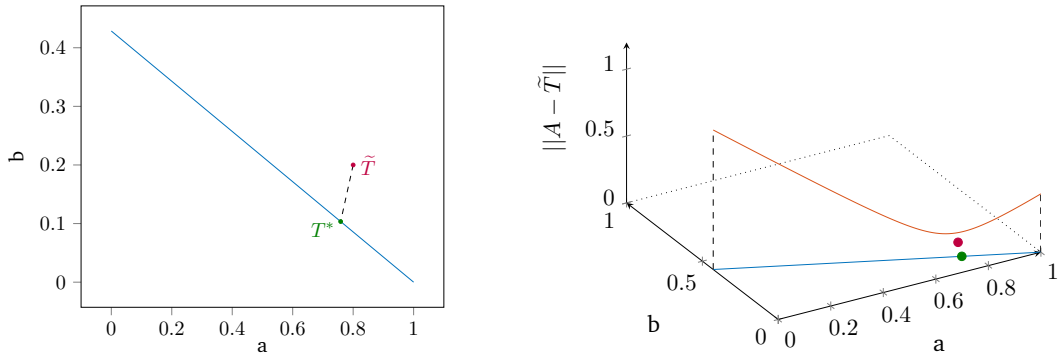

Figure 3.3: Convex set X (blue) and convex map $A \to ||A - \widetilde{T}||$ (red).

a basis of U. Note that X itself is not a subspace, because for $A \in X$ we have $\alpha A \notin X$ for $\alpha \neq 1$, $\alpha \in \mathbb{R}$, since αA is not stochastic anymore. Therefore, we start with the subspace

$$U = \{A \in \mathbb{R}^{n \times n} \mid DA = A^T D \text{ and } \exists k \in \mathbb{R} \text{ with } \sum_{j=1}^{n} A_{ij} = k \text{ for } i = 1, \ldots, n\}$$

where $D = \operatorname{diag}(\hat{\pi}_1, \ldots, \hat{\pi}_n)$ denotes the diagonal matrix with the values $\hat{\pi}_i$ on the diagonal. Notice that U is actually a subspace of $\mathbb{R}^{n \times n}$ because for $A, B \in U$ with $\sum_{j=1}^{n} a_{1j} = k_1$ and $\sum_{i=1}^{n} b_{1j} = k_2$ we get

$$D(\alpha A + \beta B) = \alpha DA + \beta DB = \alpha A^T D + \beta B^T D = (\alpha A + \beta B)^T D$$

for any $\alpha, \beta \in \mathbb{R}$, and

$$\sum_{j=1}^{n} \alpha a_{ij} + \beta b_{ij} = \alpha k_1 + \beta k_2 \quad \text{for all } i = 1, \ldots, n$$

holds. The subspace U contains X. A $\hat{\pi}$-reversible Markov chain always has $\hat{\pi}$ as a stationary distribution. For a matrix in U we get the following generalization of this property.

Lemma 3.2.1. *For $A \in U$ with $\sum_{j=1}^{n} A_{ij} = k$ it holds $\hat{\pi}^T A = k\hat{\pi}^T$.*

Proof. Since $A \in U$, we have $\hat{\pi}_i A_{ij} = \hat{\pi}_j A_{ji}$. Therefore, we obtain

$$(\hat{\pi}^T A)_i = \sum_{l=1}^{n} \hat{\pi}_l A_{li} = \hat{\pi}_i \sum_{l=1}^{n} A_{il} = \hat{\pi}_i k.$$

□

To understand the space U, the following matrices are of interest:

$$A_{ij}^{[r,s]} = \begin{cases} \hat{\pi}_s & \text{if } i = r \text{ and } j = s, \\ \hat{\pi}_r & \text{if } i = s \text{ and } j = r, \\ 1 - \hat{\pi}_s & \text{if } i = j = r, \\ 1 - \hat{\pi}_r & \text{if } i = j = s, \\ 1 & \text{if } i = j \text{ and } s \neq i \neq r, \\ 0 & \text{else,} \end{cases}$$

which is given in matrix notation as

$$A^{[r,s]} = \begin{array}{c} \\ r \\ \\ \\ \\ s \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} r & & & & & & s & & \end{array} \\ \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 1-\hat{\pi}_s & & & \hat{\pi}_s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \hat{\pi}_r & & & 1-\hat{\pi}_r & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix} \end{array}$$

and

$$\delta_{ij}^{[r,s]} = \begin{cases} 1 & \text{if } i = j \text{ and } i \neq r, \\ 1 & \text{if } i = r \text{ and } j = s, \\ 0 & \text{else,} \end{cases}$$

which is given in matrix notation as

$$\delta^{[r,s]} = \begin{array}{c} \\ r \\ \\ \\ \\ s \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} & & & r & & & s & & \end{array} \\ \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & & 1 & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix} \end{array}.$$

Proposition 3.2.2. *The family*

$$((A^{[r,s]})_{(r,s)\in I_A}, (\delta^{[r,s]}, \delta^{[s,r]})_{(r,s)\in I_B}, \mathrm{Id})$$

is a basis of U, where Id *is the identity matrix and*

$$\begin{aligned} I &= \{(r,s) \mid 1 \le r < s \le n\}, \\ A &= \{i \colon \hat{\pi}_i \neq 0\}, \\ B &= \{i \colon \hat{\pi}_i = 0\}, \\ I_A &= \{(r,s) \in I \mid r \in A \text{ or } s \in A\}, \\ I_B &= \{(r,s) \in I \mid r,s \in B\}. \end{aligned}$$

The dimension of U is given by

$$dim(U) = \binom{n}{2} + 1 + \binom{|B|}{2}.$$

Proof. First of all, we need to show that the claimed basis actually belongs to U. Further, we will see that this basis also belongs to X. First, we observe that the row sum of Id, $A^{[r,s]}$ and $\delta^{[r,s]}$ is always 1, and that

$$DA^{[r,s]} = \left(A^{[r,s]}\right)^T D$$

holds. The latter follows from

$$\begin{aligned} \left(DA^{[r,s]}\right)_{ij} &= \sum_{k=1}^{n} D_{ik} A^{[r,s]}_{kj} \\ &= \hat{\pi}_i A^{[r,s]}_{ij} \\ &\overset{(*)}{=} \hat{\pi}_j A^{[r,s]}_{ji} \\ &= \sum_{k=1}^{n} A^{[r,s]}_{ki} D_{kj} \\ &= \left((A^{[r,s]})^T D\right)_{ij}, \end{aligned}$$

where $(*)$ holds because for $i = r$ and $j = s$ we have

$$\hat{\pi}_i A^{[r,s]}_{ij} = \hat{\pi}_r \hat{\pi}_s = \hat{\pi}_s \hat{\pi}_r = \hat{\pi}_j A^{[r,s]}_{ji},$$

the same holds for $i = s$ and $j = r$. If $i = j$ holds, then the equation in $(*)$ is trivial and in all other cases we have $A^{[r,s]}_{ij} = A^{[r,s]}_{ji} = 0$, and therefore $A^{[r,s]} \in U$. Also, we have

$$D\delta^{[r,s]} = {\delta^{[r,s]}}^T D$$

for all $(r,s) \in I_B$. This is because $\hat{\pi}_r = 0$ and, therefore,

$$D\delta^{[r,s]}_{i,j} = \begin{cases} \hat{\pi}_i & \text{if } i = j, \\ 0 & \text{else.} \end{cases}$$

Since $D\delta^{[r,s]}$ is just a diagonal matrix, it is symmetric, and we obtain

$$D\delta^{[r,s]} = (D\delta^{[r,s]})^T = {\delta^{[r,s]}}^T D;$$

the argument for $\delta^{[s,r]}$ is analogous . Therefore the family is actually contained in X. In order to prove that these matrices are indeed a basis of U, it remains to show that they are linearly independent and that they span the subspace U.

We start by showing that the family is linearly independent. To do so, let us take an arbitrary linear combination of zero:

$$\sum_{(r,s)\in I_A} \alpha_{r,s} A^{[r,s]} + \sum_{(r,s)\in I_B} \alpha_{r,s}\delta^{[r,s]} + \beta_{r,s}\delta^{[s,r]} + \alpha I = 0.$$

For $(r,s) \in I_A$ the matrix $A^{[r,s]}$ is the only matrix in the above linear combination that could have a non-zero entry in row r and column s and in row s and column r. Therefore, we obtain

$$0 = \alpha_{r,s}\hat{\pi}_s \quad \text{and} \quad 0 = \alpha_{r,s}\hat{\pi}_r.$$

Thus, $\hat{\pi}_s \neq 0$ or $\hat{\pi}_r \neq 0$ provides $\alpha_{r,s} = 0$. For $(r,s) \in I_B$, we obtain analogously $\alpha_{r,s} = 0 \quad \text{and} \quad \beta_{r,s} = 0$. The linear combination reduces to $\alpha \operatorname{Id} = 0$ which, finally, leads us to $\alpha = 0$.

It remains to show that the given matrices span the subspace U. Consider a matrix $C \in U$ with $\sum_{j=1}^{n} C_{ij} = k$ for some $k \in \mathbb{R}$. For $(r,s) \in I_A$ define $\alpha_{r,s} := \frac{C_{sr}}{\hat{\pi}_r}$ if $\hat{\pi}_r \neq 0$ and otherwise $\alpha_{r,s} := \frac{C_{rs}}{\hat{\pi}_s}$. From $\hat{\pi}_r C_{rs} = \hat{\pi}_s C_{sr}$ we get

$$\alpha_{r,s} A^{[r,s]}_{rs} = C_{rs} \quad \text{and} \quad \alpha_{r,s} A^{[r,s]}_{sr} = C_{sr}.$$

For $(r,s) \in I_B$ choose

$$\alpha_{r,s} = C_{r,s} \quad \text{and} \quad \beta_{r,s} = C_{s,r}.$$

Since each off-diagonal element appears in exactly one matrix, C differs from

$$\widetilde{C} := \sum_{(r,s)\in I_A} \alpha_{r,s} A^{[r,s]} + \sum_{(r,s)\in I_B} \alpha_{r,s}\delta^{[r,s]} + \beta_{r,s}\delta^{[s,r]}$$

only in the diagonal. Because $\widetilde{C}$ is a linear combination of the elements in U, we know that $\widetilde{C} \in U$ and thus that a $l \in \mathbb{R}$ exists with $\sum_{j=1}^{n} \widetilde{C}_{ij} = l$ for $i = 1, \ldots, n$. Therefore, the matrix $\hat{C} := \widetilde{C} + (k-l)\operatorname{Id}$ has row-sum k and

$$\hat{C}_{ii} = k - \sum_{j=1, j\neq i}^{n} \hat{C}_{ij} = k - \sum_{j=1, j\neq i}^{n} C_{ij} = C_{ii},$$

holds. Therefore, we have $C = \hat{C}$ which shows that the given matrices span U. Furthermore, counting all the family parts together leads to

$$\dim U = |I_A| + |I_B| + |I_B| + 1 = |I| + |I_B| + 1.$$

The statement follows from

$$|I| = \binom{n}{2}$$

and

$$|I_B| = \binom{|B|}{2}.$$

□

We have now identified a basis of U. Next, we will use this basis to find a neat characterization of the set X. This characterization of the set

$$X = \{A \in U \mid A_{ij} \geq 0 \text{ for } i,j = 1,\dots,n \text{ and } \sum_{j=1}^{n} a_{1j} = 1\}$$

will lead us to the direct path to the closest $\hat{\pi}$-reversible matrix. To simplify notation, let us denote the basis of Proposition 3.2.2 by $(v_i)_{i=1,\dots,m}$ with $m = \dim U$ and $v_m = \mathrm{Id}$. For any matrix $A \in U$ a unique coefficient vector $\mathbf{x} \in R^m$ exists with $A = \sum_{i=1}^{m} x_i v_i$. We will now find the conditions on the coefficient vector $\mathbf{x} \in R^m$ that are sufficient and necessary for assuring $A \in X$. First, the row-sum of a matrix and the coefficient vector are related in the following way

$$\begin{aligned}\sum_{j=1}^{n} a_{ij} &= \sum_{j=1}^{n}\left(\sum_{l=1}^{m} x_l v_l(i,j)\right)\\ &= \sum_{l=1}^{m} x_l \left(\sum_{j=1}^{n} v_l(i,j)\right)\\ &= \sum_{l=1}^{m} x_l.\end{aligned}$$

Thus, we have that a matrix $A \in U$ has row-sum one, if and only if $\mathbb{1}^T\mathbf{x} = 1$ where $\mathbb{1} \in \mathbb{R}^m$ is the constant vector $\mathbb{1}_i = 1$ for $i = 1,\dots,m$. However, this condition is not enough to assure that $A \in X$. For that we also need to assure $a_{ij} \geq 0$ for all i,j. In the case of $i \neq j$ we get $a_{ij} \geq 0$ if and only if $x_l \geq 0$ for all $l = 1,\dots,m-1$, since each off-diagonal element appears in exactly one of the matrices $v_1,\dots,v_{m-1}$. This can be rewritten as $-\mathbf{x}e_i \leq 0$ for $i = 1,\dots,m-1$. However, the diagonal entries of A can be non-negative even if x_m is negative. Thus, to assure the non-negative diagonal entries as well requires a little effort. To see how to find the associated condition, let A be given as

$$A = \sum_{(r,s)\in I_A} \alpha_{r,s} A^{[r,s]} + \sum_{(r,s)\in I_B} \alpha_{r,s}\delta^{[r,s]} + \beta_{r,s}\delta^{[s,r]} + \alpha Id,$$

which reveals

$$a_{ii} = \sum_{(r,s)\in I_A} \alpha_{r,s} A_{ii}^{[r,s]} + \sum_{(r,s)\in I_B} \alpha_{r,s}\delta_{ii}^{[r,s]} + \beta_{r,s}\delta_{ii}^{[s,r]} + \alpha I_{ii}.$$

This leads to

$$a_{ii} = \sum_{\substack{(r,s)\in I_A\\ r=i}} \alpha_{r,s}(1-\hat{\pi}_s) + \sum_{\substack{(r,s)\in I_A\\ s=i}} \alpha_{r,s}(1-\hat{\pi}_r) + \sum_{\substack{(r,s)\in I_A\\ r\neq i\neq s}} \alpha_{r,s} + \sum_{\substack{(r,s)\in I_B\\ r\neq i}} \alpha_{r,s} + \sum_{\substack{(r,s)\in I_B\\ s\neq i}} \beta_{r,s} + \alpha.$$

The condition $a_{ii} \geq 0$ is thus equivalent to $-\mathbf{x}g_i \leq 0$ where

$$g_i(j) = \begin{cases} 1-\hat{\pi}_s & \text{if } v_j = A^{[i,s]} \text{ for some } s > i,\\ 1-\hat{\pi}_r & \text{if } v_j = A^{[r,i]} \text{ for some } r < i,\\ 0 & \text{if } v_j = \delta^{[i,s]} \text{ for some } s,\\ 1 & \text{else,}\end{cases}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. All in all, given the matrix

$$C = -1 \cdot \begin{bmatrix} — & e_1^T & — \\ & \vdots & \\ — & e_{m-1}^T & — \\ — & g_1^T & — \\ — & \vdots & — \\ — & g_n^T & — \end{bmatrix} \in \mathbb{R}^{(n+m-1)\times m},$$

the condition that $A = \sum_{i=1}^m x_i v_i$ is in set X is equivalent to

$$Cx \leq 0 \quad \text{and} \quad \mathbb{1}^T x = 1.$$

Our original intention was to show that for any transition matrix $\widetilde{T}$ we find a $\hat{\pi}$-reversible Markov chain $T^* \in X$ with

$$||T^* - \widetilde{T}|| \leq ||A - \widetilde{T}|| \quad \text{for all } A \in X.$$

Recall that the norm $||\cdot||$ is induced by a scalar product $\langle \cdot, \cdot \rangle$, i.e.

$$||A|| = \sqrt{\langle A, A \rangle}$$

for any matrix $A \in \mathbb{R}^{n\times n}$. Therefore we can rewrite the term to

$$\begin{aligned} \left\| \sum_{i=1}^m x_i v_i - T \right\|^2 &= \sum_{i,j=1}^m x_i x_j \langle v_i, v_j \rangle - 2 \sum_{i=1}^m x_i \left\langle v_i, \widetilde{T} \right\rangle + \left\langle \widetilde{T}, \widetilde{T} \right\rangle \\ &= \frac{1}{2} x^T Q x + x^T f + c \end{aligned}$$

where

$$Q(i,j) := 2 \langle v_i, v_j \rangle, \quad f(i) = -2 \langle v_i, T \rangle$$

and

$$c = \langle T, T \rangle .$$

Thus, we want to minimize the function

$$x \mapsto \frac{1}{2} x^T Q x + x^T f + c \tag{3.1}$$

with the restraints

$$Cx \leq 0 \quad \text{and} \quad \mathbb{1}^T x = 1.$$

Since Q is a Gram matrix of linear independent vectors, it is positive definite. This follows because for any $x \in \mathbb{R}^n$

$$x^T Q x = \sum_{i,j} x_i Q_{ij} x_j = \left\langle \sum_{i=1}^n x_i v_i, \sum_{i=1}^n x_i v_i \right\rangle$$

holds. Since $v_1, \ldots, v_n$ is a basis, we have that $\sum_{i=1}^n x_i v_i \neq 0$ for $x \neq 0$ and, in consequence, $x^T Q x > 0$ for $x \neq 0$. Since Q is positive definite, the quadratic function (3.1) is strongly

convex. Therefore, we have formulated the problem into a strongly convex quadratic programming problem that attains its global minimum [25, Theorem 1.15], since the quadratic function is coercive, continuous and the set X is non-empty because of $\mathrm{Id} \in X$. Also, the global minimum is unique because the quadratic function is strongly convex. This section can be summarized as follows.

Theorem 3.2.3. *For any transition matrix $\widetilde{T}$, any stochastic vector $\hat{\pi}$ and any norm $||\cdot||$ induced by a scalar-product, there exists a unique $\hat{\pi}$-reversible Markov chain $T^* \in X$ with*

$$||T^* - \widetilde{T}|| \leq ||A - \widetilde{T}|| \quad \textit{for all } A \in X.$$

3.3 Complexity and Eigenvalues

To avoid technical difficulties, we will assume in this chapter that $\hat{\pi}_i > 0$ for $i = 1, \ldots, n$. For $A \in \mathbb{R}^{n\times n}$ the trace $tr(A)$ is given by the sum of the diagonal elements

$$tr(A) = \sum_{i=1}^{n} A_{ii}.$$

The following definition

$$\langle A, B\rangle_F := tr(A^T B)$$

is a scalar product on $\mathbb{R}^{n\times n}$ for $A, B \in \mathbb{R}^{n\times n}$. This scalar product induces the Frobenius norm

$$||A||_F = \sqrt{\langle A, A\rangle_F} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2}.$$

The following complexity analysis will be given according to the Frobenius norm.

Complexity

Unfortunately, the matrix Q is quite large. This is because we are optimizing the entries of a matrix in $\mathbb{R}^{n\times n}$ and thus we have $O(n^2)$ unknowns. Specifically, we have $Q \in \mathbb{R}^{m\times m}$ where

$$m = \dim(U) = 1 + \frac{n^2}{2}.$$

Each entry of Q is given by a trace of two sparse matrices which can be computed using the following formula

$$\left\langle A^{[r,s]}, A^{[r',s']} \right\rangle_F = \begin{cases} n - \hat{\pi}_r - \hat{\pi}_{r'} - \hat{\pi}_s - \hat{\pi}_{s'} & \text{if } r, r', s, s' \text{ are distinct,} \\ n - 1 - 2\hat{\pi}_s + (1-\hat{\pi}_r)(1-\hat{\pi}_{r'}) & \text{if } r \neq r', s = s', \\ n - 1 - 2\hat{\pi}_r + (1-\hat{\pi}_s)(1-\hat{\pi}_{s'}) & \text{if } r = r', s \neq s, \\ n - 1 + (1-\hat{\pi}_s)(1-\hat{\pi}_{s'}) + (1-\hat{\pi}_r)(1-\hat{\pi}_{r'}) & \text{if } r = s' \text{ or } r' = s, \\ n - 2 + (1-\hat{\pi}_r)^2 + \hat{\pi}_r^2 + (1-\hat{\pi}_s)^2 + \hat{\pi}_s^2 & \text{if } r = r' \text{ and } s = s'. \end{cases}$$

It might be worth noting that because of $r < s$ and $r' < s'$ other cases are not possible. Since the proof of the formula is very technical, we only give a proof for the case $r \neq r'$ and $s = s'$; all other cases can be obtained analogously. First, if $A_{ij}^{[r,s]} \neq 0$ for $i \neq j$, it follows that either $i = r, j = s$ or $i = s, j = r$ holds. Since $r \neq r', s = s'$ we have $A_{ij}^{[r',s']} = 0$, therefore,

$$\begin{aligned}\left\langle A^{[r,s]}, A^{[r',s']} \right\rangle_F &= \sum_{i=1}^{n} A_{ii}^{[r,s]} A_{ii}^{[r',s']} \\ &= n - 3 + A_{rr}^{[r,s]} A_{rr}^{[r',s']} + A_{r'r'}^{[r,s]} A_{r'r'}^{[r',s']} + A_{ss}^{[r,s]} A_{ss}^{[r',s']} \\ &= n - 3 + 1 - \hat{\pi}_s + 1 - \hat{\pi}_{s'} + (1 - \hat{\pi}_r)(1 - \hat{\pi}_{r'}) \\ &= n - 1 - 2\hat{\pi}_s + (1 - \hat{\pi}_r)(1 - \hat{\pi}_{r'}).\end{aligned}$$

Hence, the computation required for a single entry of Q does not increase with n and the effort to compute and store Q is $O(n^4)$.

Furthermore, the matrix Q is well-conditioned. For

$$k := \min_{r,r',s,s'} \left\langle A^{[r,s]}, A^{[r',s']} \right\rangle_F,$$

we obtain

$$k \operatorname{Id} \leq Q \leq n \operatorname{Id}$$

where the inequality has to be read component-wise. The upper bound follows from

$$\langle v_i, v_j \rangle_F \leq \sqrt{\langle v_i, v_i \rangle_F} \sqrt{\langle v_j, v_j \rangle_F} \leq \max_{l=1,\ldots,m} \{\langle v_l, v_l \rangle_F\}$$

and the fact that

$$\left\langle A^{[r,s]}, A^{[r,s]} \right\rangle_F = n - 2 + (1 - \hat{\pi}_r)^2 + \hat{\pi}_r^2 + (1 - \hat{\pi}_s)^2 + \hat{\pi}_s^2 \leq n$$

and

$$\langle \operatorname{Id}, \operatorname{Id} \rangle_F = n$$

holds. Since Q is symmetric, its condition number according to the spectral norm is given by

$$\kappa(Q) = \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{n}{k}.$$

Since $k = n - c$ for a $0 < c < 4$, we have

$$\kappa(Q) \leq \frac{1}{1 - c/n} \to 1$$

for $n \to \infty$, which validates the claim.

The convex minimization problem can be solved using a barrier method, e.g. the interior point method. This method consists of N Newton iterations. The number N of Newton iterations to find a strictly feasible point is bounded by

$$N \leq \sqrt{n + n^2/2 - 1} \log\left(\frac{n + n^2/2 - 1}{\varepsilon}\right) \gamma$$

where $\varepsilon > 0$ is the demanded accuracy and γ is a constant depending on the choice of backtracking parameter, see [9, Section 11.5.5]. Therefore, the number of Newton iterations

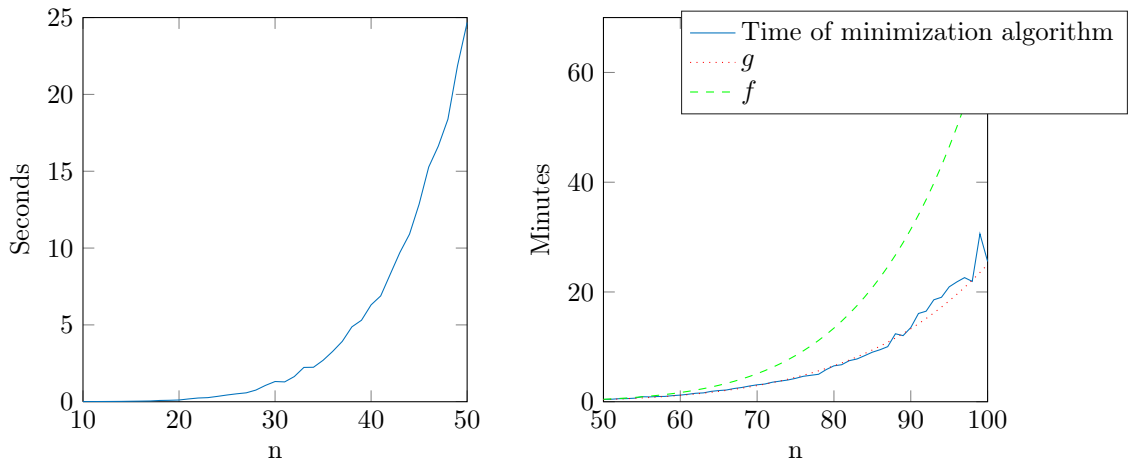

Figure 3.4: Duration of convex minimization problem to find closest reversible Markov chain.

is bounded by $O(n \log n)$. Unfortunately, each Newton iteration has to solve a linear equation system and cost $O((n^2)^3)$ since $Q \in \mathbb{R}^{(\frac{n^2}{2}+1)\times(\frac{n^2}{2}+1)}$. Therefore, the total cost for the optimization problem is bounded by $O(n^7 \log n)$. In the field of convex optimization, is is known that the upper bound for the number of Newton iterations is a large overestimation [9]. If we assume this number to be independent from n, then we can expect the time to find a solution using the convex optimization problem to be given by

$$g(n) = \alpha n^6$$

where n is the size of the matrix $\widetilde{T} \in \mathbb{R}^{n\times n}$. If we include the bad estimation for the Newton iteration, then the time should be represented by

$$f(n) = \beta n^7 \log(n).$$

In order to explore the computation time of the convex optimization problem, for each $n = 10, 11 \ldots, 100$ we generated a stochastic matrix $A \in \mathbb{R}^{n\times n}$ and a random probability vector $\hat{\pi} \in \mathbb{R}^n$, in which each entry was drawn from the standard uniform distribution on the open interval $(0, 1)$, and then we normalized A and $\hat{\pi}$. We used Matlab R2012b on a 3 GHz computer with 8 GB RAM. We solved the convex optimization problem using the *interior-point-convex* algorithm of the provided Matlab method *quadprog* with default options, i.e. relative dual feasibility $=2.31e-15$ with TolFun$=1e-0.8$, complementarity measure $=1.68e-10$ with TolFun$=1e-0.8$ and relative max constraint violoation $= 0$ with TolCon$=1e-0.8$. For $n = 45$ the execution time was about 12.46 seconds. The scalars α, β where chosen such that $f(45) = g(45) = \frac{12.46}{60}$ holds. In Figure 3.4 one can see that g seems to be a reasonable approximation of the execution time. Also one can see that the execution is a matter of only seconds for matrices in $\mathbb{R}^{50\times 50}$, but takes more then 20 minutes for matrices in $\mathbb{R}^{100\times 100}$.

Eigenvalue Analysis

The reason for wanting to approximate a Galerkin projection, was to obtain the corresponding eigenvalues and eigenvectors. It is beneficial to correct the error in the Galerkin projection by finding the nearest $\hat{\pi}$-reversible matrix, because this results in real eigenvalues and real eigenvectors. However, we have not discussed which matrix norm would be the best choice in order to maintain the spectrum of the unperturbed Galerkin discretization. At first glance the Frobenius norm seems to be a reasonable choice, because it would return the $\hat{\pi}$-reversible matrix which is closest according to the euclidean distance. To find out whether the Frobenius norm is a good choice, we will conduct a small experiment. This time, we consider the $2\hat{\pi}$-periodic function $V_B \colon \mathbb{R} \to \mathbb{R}$

$$V_B(x) = a + b\,\cos(x) + c\,\cos^2(x) + d\,\cos^3(x)$$

with a=2.0567, b=-4.0567, c=0.3133 and d=6.4267. This can be seen as an approximation of butane's potential energy function, see Figure 2.12, where x is the central dihedral angle. We are going to realize a trajectory with $5 \cdot 10^8$ timesteps of size $dt = 0.001$ of the dihedral angles

$$X_t = \widetilde{X}_t \mod 2\hat{\pi}$$

of butane from the stochastic differential equation

$$d\widetilde{X}_t = -\nabla V_B(\widetilde{X}_t)dt + \sigma dB_t$$

with perturbation $\sigma = \sqrt{\frac{2}{2.5}}$. We divide $[0, 2\pi)$ into 31 equidistant sets $A_i = [\frac{(i-1)2\pi}{31}, \frac{i2\pi}{31})$ for $i = 1, \ldots, 31$. We then construct a Markov State Model from the long-term trajectory as explained in the previous chapter, based upon the long trajectory but considering only between 10^8 and up to $5 \cdot 10^8$ timesteps. We approximate the vector π using the trapezian rule. For high-dimensional problems, the vector π can be approximated as explained in [60] and should not be computed as the left eigenvector of $\widetilde{T}$, since the problem is ill-conditioned [12, 36]. For each Markov State Model $\widetilde{T}$, we compute the closest $\hat{\pi}$-reversible Matrix T^* according to the Frobenius norm. The Markov State Models created in this way only becomes reasonable when considered for a trajectory longer than 10^6 timesteps, due to rare transition events.

We will now compare the eigenvalues of the Markov State Model $\widetilde{T}$ to the eigenvalues of the corrected estimation T^*. Since the eigenvalues of $\widetilde{T}$ sometimes turned out to be complex , we simply set the imaginary part to zero, in order to compare the eigenvalues with T^*. Butane's potential energy function has three metastable sets, and thus in addition to eigenvalue 1 it has two dominant eigenvalues close two 1. In Figure 3.5 one can observe that the eigenvalues of the correction T^* are having difficulties staying close to the spectrum of the standard estimation $\widetilde{T}$, even if we consider $2 \cdot 10^8$ timesteps from our long-term trajectory. For the approximation of the non-dominant eigenvalues, the estimation of the eigenvalues becomes even worse, resulting in a fixed gap between the eigenvalues of $\widetilde{T}$ and T^* as shown in Figure 3.6. Thus, the question arises whether the Frobenius norm really is the best choice. Recall that there exists an exact π-reversible transition matrix T and we only have access to an approximation $\widetilde{T}$ of T which is not necessarily π-reversible due to numerical errors. We can then construct a closest $\hat{\pi}$-reversible matrix T^* that is in general different from T, but we assume that T^* is at least close to T. Thus, to assure that the eigenvalues

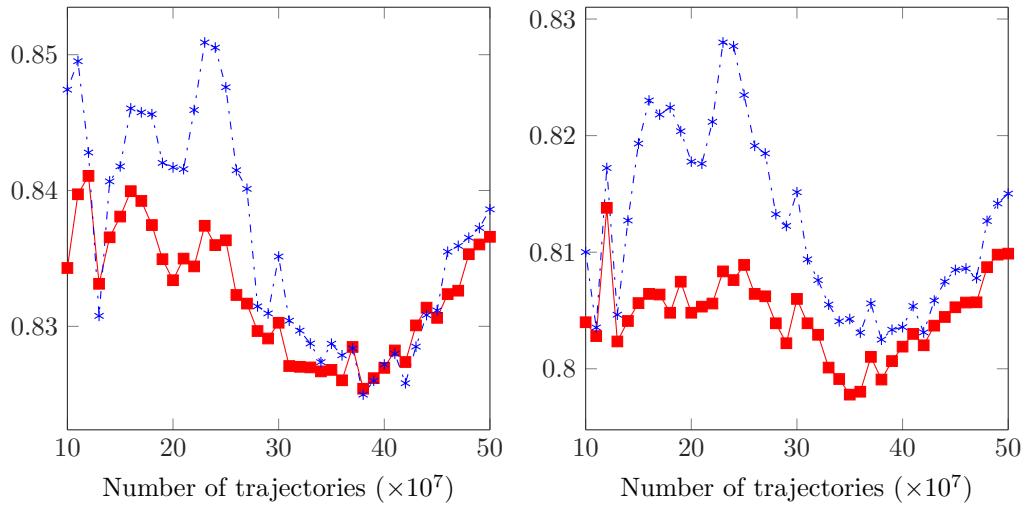

Figure 3.5: Approximation of second-largest eigenvalue (left) and the thrid-largest eigenvalue (right) for different Markov State Models. Red is the eigenvalue from the standard approximation T, blue is the eigenvalue from the corrected approximation T^* according to Frobenius norm.

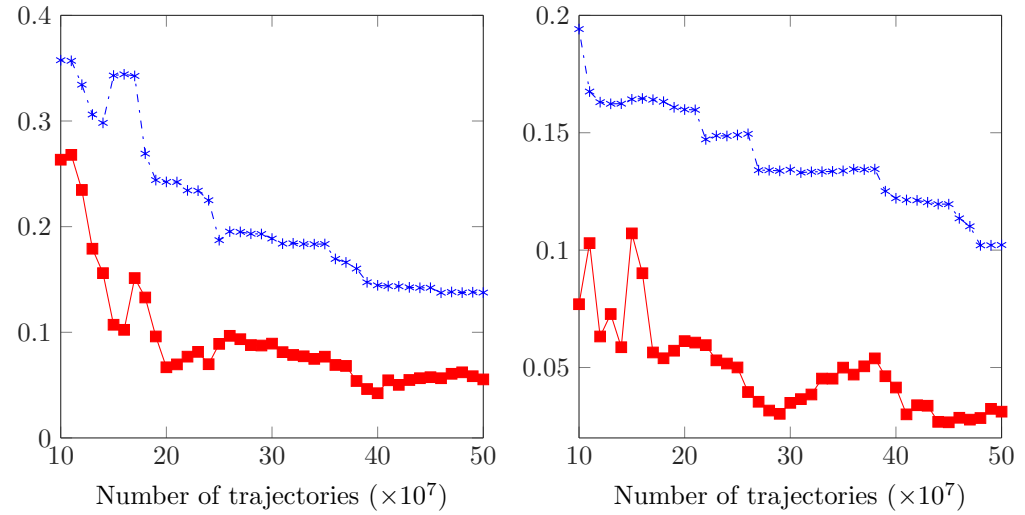

Figure 3.6: Approximation of fourth-largest eigenvalue (left) and the fifth-largest eigenvalue (right) for different Markov State Models. Red is the eigenvalue from the standard approximation $\widetilde{T}$, blue is the eigenvalue from the corrected approximation T^* according to Frobenius norm.

of T^* are close to the eigenvalues of T, it would be useful to find a relation between the eigenvalue approximation and the distance between T^* and T. Such a relation can be found by considering the following weighted Frobenius norm

$$||A||_{\widetilde{F}} := ||D^{\frac{1}{2}} A D^{-\frac{1}{2}}||_F$$

with

$$D = diag(\hat{\pi}_1, \ldots, \hat{\pi}_n), \quad D^{\frac{1}{2}} D^{\frac{1}{2}} = D$$

and

$$D^{-\frac{1}{2}} := (D^{\frac{1}{2}})^{-1}.$$

Note that the weighted Frobenius norm is given by

$$||A||^2_{\widetilde{F}} = \sum_{i,j=1}^{n} a^2_{i,j} \frac{\hat{\pi}_i}{\hat{\pi}_j}.$$

From this weighted Frobenius norm we gain the following relation between the distance and the eigenvalues. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of T, and $\hat{\lambda_1}, \ldots, \hat{\lambda_n}$ are the eigenvalues of T^*, then a permutation σ exists such that

$$\sum_{i=1}^{n} |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2 \leq ||T - T^*||^2_{\widetilde{F}}$$

holds. This is shown in the following theorem.

Theorem 3.3.1. *Let $A, B \in X$, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A and $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ be the eigenvalues of B. There then exists a permutation σ of the integers $1, 2, \ldots, n$ such that*

$$\sum_{i=1}^{n} |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2 \leq ||A - B||^2_{\widetilde{F}}.$$

Proof. The matrices $A, B \in X$ are self-adjoint according to the scalar product $\langle x, y \rangle_\pi := x^T D y$. Let us denote by $\{w_1, \ldots, w_n\}$ a $\langle , \rangle_\pi$-orthonormal basis, and denote with W the matrix with columns containing the vectors w_i, i.e.

$$W = \begin{pmatrix} | & | & & | \\ w_1 & w_2 & \ldots & w_n \\ | & | & & | \end{pmatrix}.$$

Then

$$A' := W^{-1} A W \quad \text{and } B' = W^{-1} B W$$

are symmetric, see [19, Chapter 5.6.1]. By the Hoffman and Wielandt Theorem [24, Theorem 6.3.5] we obtain

$$\sum_{i=1}^{n} |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2 \leq ||A' - B'||^2_F$$

for a permutation σ, because similar matrices have the same eigenvalues. It remains to show

$$||W^{-1} C W||^2_F = ||C||^2_{\widetilde{F}}$$

or equivalently

$$||C||_F^2 = ||WCW^{-1}||_{\widetilde{F}}^2$$

for any matrix $C \in \mathbb{R}^n$. By construction of W we have

$$W^T D W = I \quad \text{and} \quad W^{-1} D^{-1} (W^T)^{-1} = I. \tag{3.2}$$

Therefore,

$$\begin{aligned} ||WCW^{-1}||_{\widetilde{F}}^2 &= ||D^{\frac{1}{2}} WCW^{-1} D^{-\frac{1}{2}}||_F^2 \\ &= \operatorname{tr}(D^{-\frac{1}{2}} (W^{-1})^T C^T W^T D^{\frac{1}{2}} D^{\frac{1}{2}} WCW^{-1} D^{-\frac{1}{2}}) \\ &\overset{(*)}{=} \operatorname{tr}(W^{-1} D^{-1} (W^{-1})^T C^T W^T DWC) \\ &\overset{(3.2)}{=} \operatorname{tr}(C^T C) \\ &= ||C||_F^2, \end{aligned}$$

where in $(*)$ it is used that the trace is invariant under cyclic permutations .

□

This shows that in order to guarantee good approximations for the eigenvalues, one has to assure a good approximation of a_{ij} for those i, j in which $\hat{\pi}_j << \hat{\pi}_i$. These transitions are also known as *rare events* and often difficult to compute. Approximating the closest $\hat{\pi}$-reversible matrix according to this weighted norm actually improves the eigenvalue estimation dramatically. This is shown in the Figures 3.7-3.10.

So far, we know how to regain reversibility for the approximation $\widetilde{S}$ of S and the approximation $\widetilde{T}$ of T separately. The question arises of whether the reversibility of $\widetilde{S}$ and $\widetilde{T}$ implies that the final result $(\widetilde{S}^*)^{-1}\widetilde{T}^*$ also inherits a real spectrum and real eigenvectors. To address this, we firstly rewrite

$$S = D^{-1}\hat{S}, \quad T = D^{-1}\hat{T}$$

with $\hat{S}_{ij} = \langle \phi_i, \phi_j \rangle_\mu$ and $\hat{T}_{ij} = \langle \mathcal{T}\phi_i, \phi_j \rangle_\mu$, hence

$$S^{-1}T = \hat{S}^{-1} D D^{-1} \hat{T} = \hat{S}^{-1}\hat{T}.$$

Analogously to what has been shown for Q, $\hat{S}$ and $\hat{S}^{-1}$ are also symmetric positive definite matrices. Now, since $\hat{S}^{-1}$ is positive definite, we know that a symmetric square matrix A exists such that $A^2 = \hat{S}^{-1}$. Thus, $A^{-1}\hat{S}^{-1}\hat{T}A = A\hat{T}A$. Consequently, $\hat{S}^{-1}\hat{T}$ is similar to a symmetric matrix and hence diagonalizable. This shows that the spectrum of $\hat{S}^{-1}\hat{T}$ is real and that we know the existence of a basis of eigenvectors of $S^{-1}T$. Thus we can also assume that after correcting $\widetilde{S}$ and $\widetilde{T}$ separately, the resulting matrix $(\widetilde{S}^*)^{-1}\widetilde{T}^*$ has the desired property to assure the applicability of the method PCCA+.

Finally, we can extend the method PCCA+ for arbitrary processes with a stationary measure as follows. First, compute the matrices $\widetilde{S}$ and $\widetilde{T}$ associated to the arbitrary process. Secondly, compute $(\widetilde{S}^*)^{-1}\widetilde{T}^*$ which is applicable for PCCA+. This correction can be seen as a small perturbation of the system. If one assumes that the perturbation of the system does not fundamentally change the metastable sets, then the metastable sets of this adjusted Galerkin Projection can be assumed to be metastable sets of the arbitrary process.

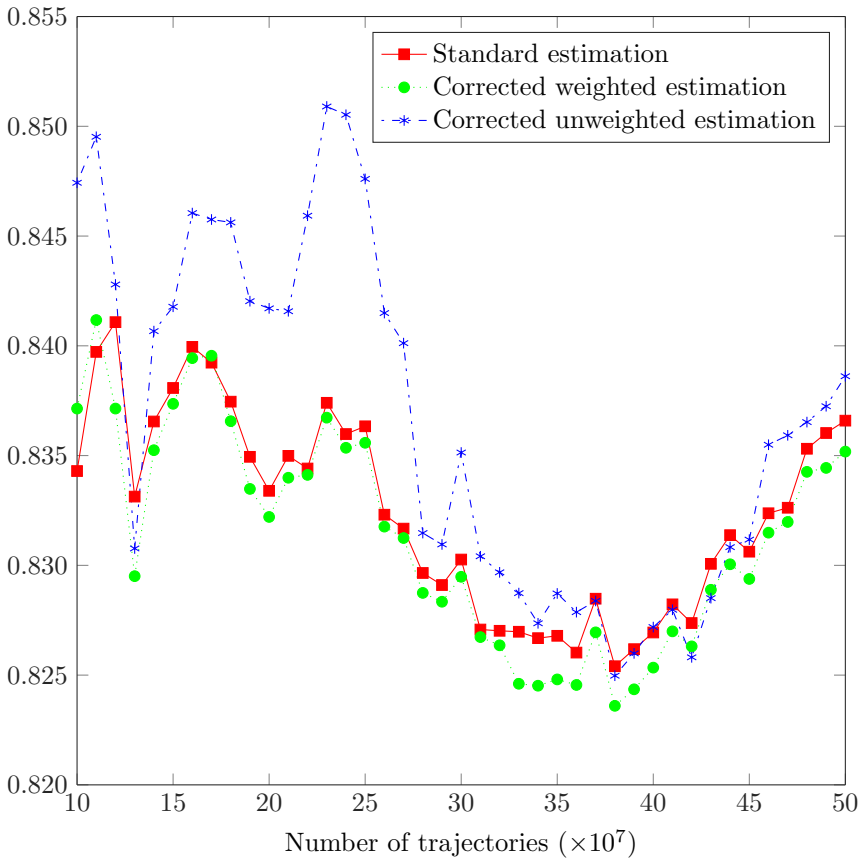

Figure 3.7: Second-largest eigenvalue.

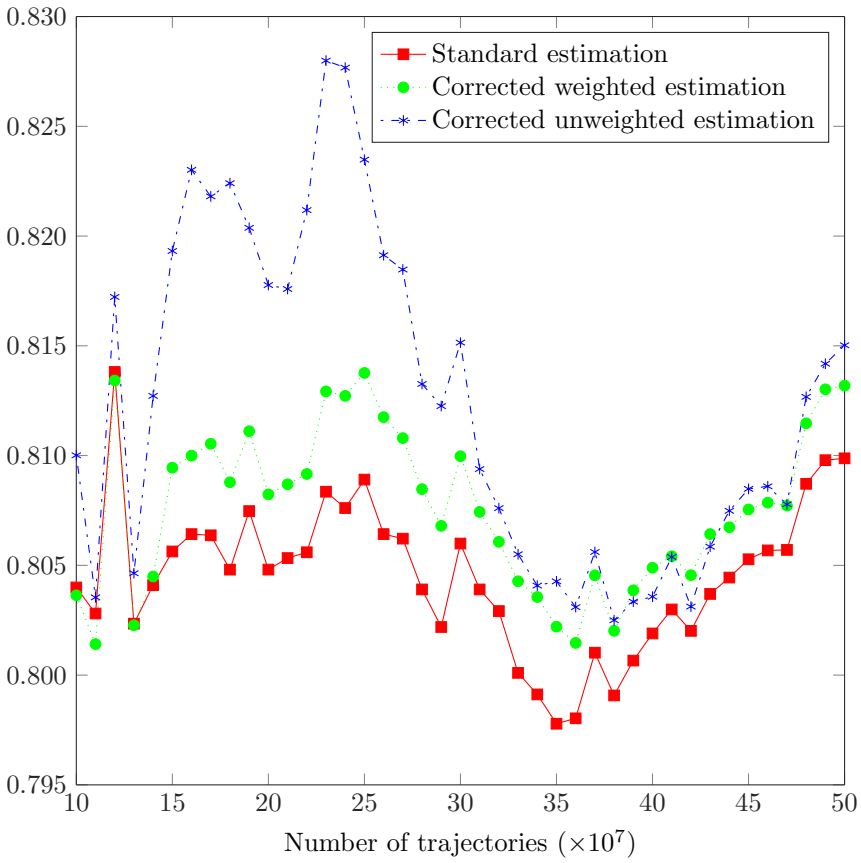

Figure 3.8: Third-largest eigenvalue.

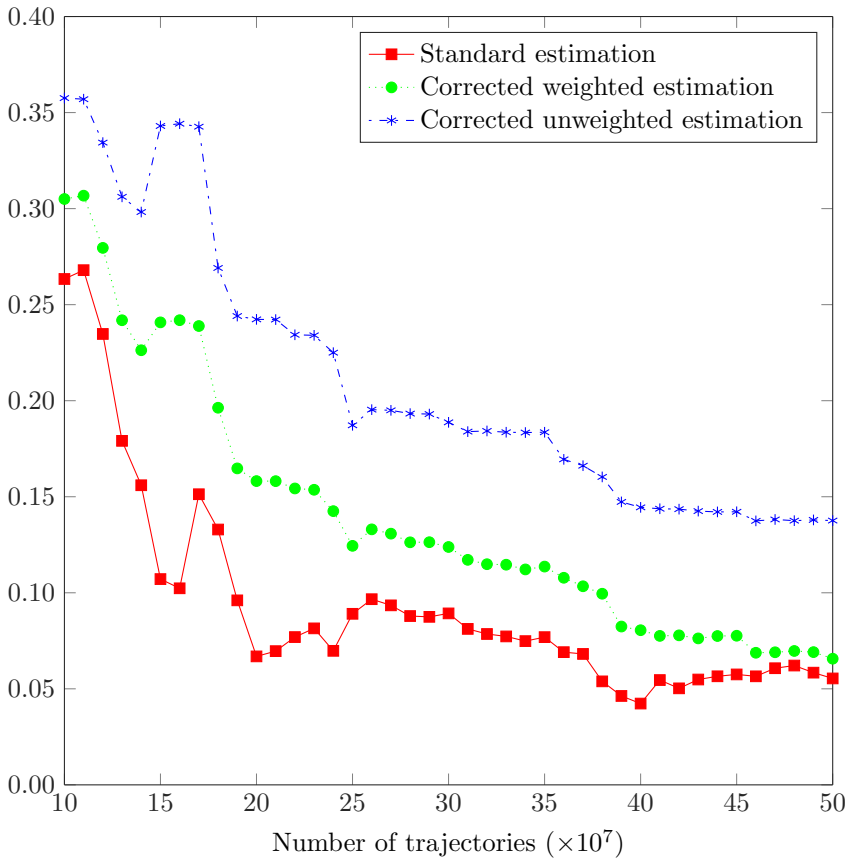

Figure 3.9: Fourth-largest eigenvalue.

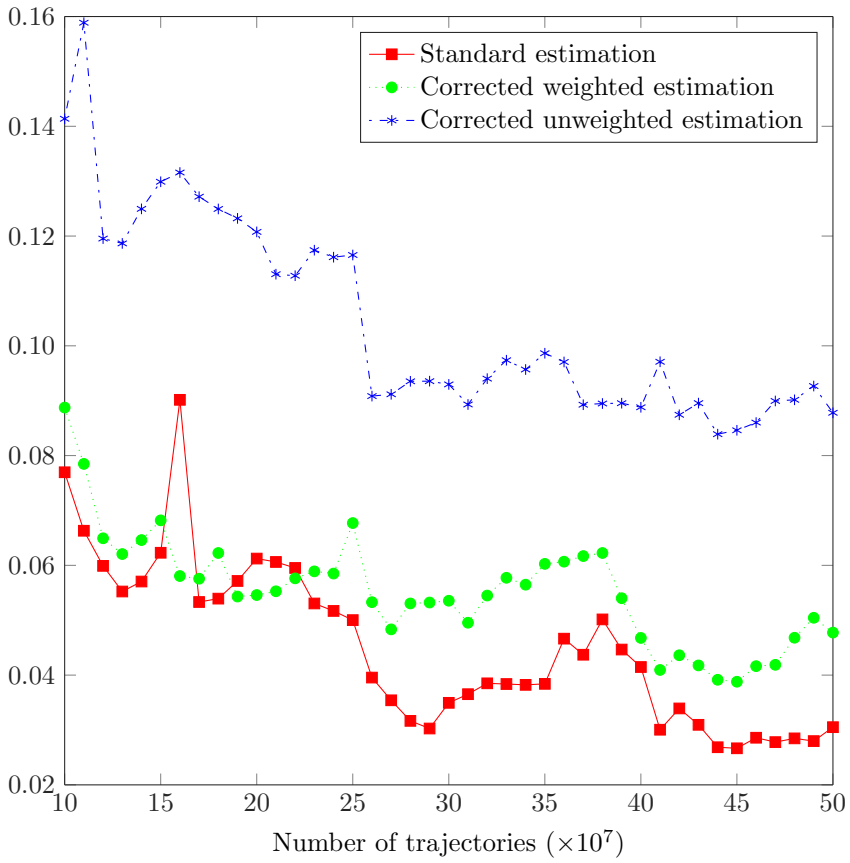

Figure 3.10: Sixth-largest eigenvalue.

Summary

The focus of this doctoral thesis is the transfer operator, a tool that describes the propagation of probability densities of an arbitrary dynamical system. This tool is usable for any moving object that one wants to analyze, and thus has applications in subjects like population statistics, the prediction of stock prices, and computational drug design.

The first part of this doctoral thesis is a purely theoretical investigation of the transfer operator. Characterizations of transfer operators and adjoint transfer operators are revealed. It is shown that Markov operators and transfer operators are equivalent. Further it is shown that an adjoint operator of a transfer operator is equivalent to a generalized Koopman operator, and that an adjoint operator of a transfer operator with an invariant measure is equivalent to a Brown-Markov operator. All three characterizations are independent of a transition kernel. The last characterization is disproving a claim made in 1966.

Diverse applications require a Galerkin projection of the transfer operator. Therefore, the second part of this thesis reveals possible ways of improving the computation of a Galerkin projection on an arbitrary function space. An exact formula of the error by the difference in the L^2 norm between the Galerkin entry and its approximation through a Monte Carlo method is deduced for long and short-term trajectory approaches. The formula enables us to approximate the Galerkin error itself by trajectories. It is shown that the error of the Galerkin projection is dramatically reduced when using short-term trajectories instead of a single long-term trajectory. Further, a characteristic of reversible processes is discovered, which shows that reversible processes are more likely to return to set than to be there. Next, a reweighting scheme is introduced that improves available techniques for obtaining a Galerkin projection for a typical scenario that often appears in computational drug design. It is shown that the Galerkin projections for multiple, similar ligands that bind to one receptor can be computed using trajectories of just one single ligand and the corresponding weights. Computation of the weights proves more advantageous than computing the trajectories separately for each ligand.

The final result presented in this thesis shows how to correct the numerical error of a Galerkin projection. This is useful for cases in which the numerical error might render the frequently employed clustering method PCCA+ to be inapplicable. It is shown that one can restore a particular property of a Galerkin projection that assures applicability of the method PCCA+. More precisely, for almost any norm and any transition matrix a closest reversible matrix exists, which can be computed by solving a strongly convex quadratic problem. Further, a norm is introduced which heavily weights transition probabilities of rare events. This norm has the property that the closest reversible matrix will preserve the spectrum. Application of the method PCCA+ was until now restricted to reversible processes. However, the correction scheme for the Galerkin projection opens the door to use of the method PCCA+ for arbitrary systems.

In summary, this thesis reveals theoretical aspects of the transfer operator that are then used to derive methods to optimize and correct the computation of the Galerkin projection.

Zusammenfassung

Der Fokus dieser Dissertation ist der Transferoperator. Dies ist ein Werkzeug, um die Ausbreitung von Wahrscheinlichkeitsdichten eines beliebigen dynamischen Systems zu beschreiben. Dieses Werkzeug ist benutzbar für jedes bewegende Objekt, welches man analysieren möchte und hat deswegen auch Anwendungen in Bereichen wie Bevölkerungsstatistik, die Vorhersage von Aktienkursen und Wirkstoffdesign.

Im ersten Teil dieser Arbeit wird gezeigt, dass Markov Operatoren und Transfer Operatoren identisch sind. Außerdem wird die Klasse der adjungierten Transfer Operatoren durch generalisierte Koopman Operatoren charakterisiert. Schließlich wird die Klasse der adjungierten Transfer Operatoren bezüglich eines invarianten Maßes durch Brown-Markov Operatoren charakterisiert. Dies widerlegt eine Behauptung aus dem Jahr 1966.

Im zweiten Teil dieser Arbeit wird gezeigt, wie man die Berechnung der Galerkin Projektion optimieren kann. Eine exakte Formel für den Galerkin Fehler für kurz und lang-zeit Trajektorien wird hergeleitet. Der Galerkin Fehler selbst kann durch die Formel wiederum mit Trajektorien approximiert werden. Es wurde gezeigt, dass der Fehler durch kurz-zeit Trajektorien dramatisch reduziert wird gegenüber lang-zeit Trajektorien. Eine Charakteristik von reversiblen Prozessen ist entdeckt worden, welche zeigt, dass reversible Prozesse lieber in eine Menge zurückkehren, anstatt sich in der Menge zu befinden. Abschließend wird eine Umgewichtungsstrategie eingeführt, welche die Berechnung der Galerkin Projektion optimiert. Es wird gezeigt, dass Galerkin Projektionen für mehrere, ähnliche Systeme berechnet werden können, nur durch Trajektorien von einem System und Gewichten. Es stellt sich heraus, dass die Gewichte einfacher zu berechnen sind als die Berechnung neuer Trajektorien für jedes System.

Das abschließende Resultat dieser Arbeit zeigt, wie man den numerischen Fehler von einer Galerkin Projektion korrigieren kann. Dies ist sinnvoll in Fällen, in denen der numerische Fehler die häufig benutze Methode PCCA+ unanwendbar macht. Es wird gezeigt, dass man eine bestimmte Eigenschaft der Galerkin Projektion wieder herstellen kann, welche garantiert, dass die Methode PCCA+ anwendbar ist. Genauer: Es wird gezeigt, dass für jede stochastische Matrix eine am nächsten gelegene reversible stochastische Matrix existiert, welche durch ein stark konvexes Optimierungsproblem berechnet werden kann. Die Methode PCCA+ war bisher auf reversible Prozesse eingeschränkt. Mit der oben erklärten Korrekturmethode, ist PCCA+ nun für beliebige Systeme anwendbar.

Insgesamt hat diese Arbeit theoretische Aspekte vom Transferoperator aufgezeigt, durch deren Kenntnis Methoden entwickelt wurden, um die Berechnung der Galerkin Projektion vom Transferoperator zu optimieren und zu korrigieren.

Bibliography

[1]Karol Baron and Andrzej Lasota. „Asymptotic properties of Markov operators defined by Volterra type integrals“. In: *Annales Polonici Mathematici* 58 (1993) (cit. on p. 28).

[2]Wojciech Bartoszek and Tom Brown. „On Frobenius-Perron operators which overlap supports“. In: *Polish Academy of Sciences and Mathematics* 45 () (cit. on p. 28).

[3]Heinz Bauer. *Maß und Integrationstheorie*. 2nd. Berlin; New York: Walter de Gruyter, 1992 (cit. on pp. 6–8, 30).

[4]Heinz Bauer. *Wahrscheinlichkeitstheorie*. 5nd. Berlin; New York: Walter de Gruyter, 2002 (cit. on pp. 9, 10, 12, 43).

[5]John R. Baxter and Jeffrey S. Rosenthal. „Rates of convergence for everywhere-positive Markov chains“. In: *Statistics & Probability Letters* 22 (1995), pp. 333–338 (cit. on pp. 19, 39).

[6]George D. Birkhoff. „Proof of the ergoic theorem“. In: *Proc Natl Acad Sci USA* 17.12 (1931), pp. 656–660 (cit. on pp. 1, 14).

[7]Ludwig Boltzmann. *Vorlesung über Gastheorie*. J.A.Barth, 1898 (cit. on p. 13).

[8]Gregory R. Bowman, Vijay S. Pande, and Frank Noé. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. 1th. Springer, 2011 (cit. on p. 42).

[9]Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004 (cit. on pp. 75, 83, 84).

[10]James R. Brown. „Approximation theorems for Markov operators.“ In: *Pacific J. Math.* 16.1 (1966), pp. 13–23 (cit. on pp. 2, 28, 29).

[11]Alexander Bujotzek, Ole Schütt, Adam Nielsen, Konstantin Fackeldey, and Marcus Weber. „ZIBgridfree: Efficient Conformational Analysis by Partition-of-Unity Coupling.“ In: *Journal of Mathematical Chemistry* 52.3 (2014), pp. 781–804 (cit. on p. 42).

[12]Grace E. Cho and Carl D. Meyer. „Comparison of perturbation bounds for the stationary distribution of a Markov chain“. In: *Linear Algebra and its Applications* 335.1–3 (2001), pp. 137 –150 (cit. on pp. 51, 85).

[13]Kai L. Chung. *Markov Chains with Stationary Transition Probabilities*. second edition. Berlin: Springer-Verlag, 1974 (cit. on p. 10).

[14]Peter Deuflhard and Marcus Weber. „Robust Perron Cluster Analysis in Conformation Dynamics“. In: *In Linear Algebra and Its Applications - Special Issue on Matrices and Mathematical Biology,volume 398C* (2005), pp. 161–184 (cit. on pp. 1, 36).

[15]Jiu Ding and Tien Yien Li. „Markov finite approximation of Frobenius-Perron operator“. In: *Nonlinear Analysis: Theory, Methods & Applications* 17.8 (1991), pp. 759 –772 (cit. on p. 46).

[16]Roland L. Dobrushin, Yuri M. Sukhov, and József Fritz. „A. N. Kolmogorov - the founder of the theory of reversible Markov processes“. In: *Russian Mathematical Surveys* 43.6 (1988), p. 157 (cit. on pp. 2, 19).

[17]Joeph L. Doob. *Stochastic Processes*. New York: John Wiley & Sons, 1953 (cit. on p. 10).

[18]Rick Durrett. *Probability: Theory and Examples*. fourth edition. Cambridge University Press, 2005 (cit. on pp. 10, 11).

[19]Gerd Fischer. *Lineare Algebra*. 15th. Vieweg, 2005 (cit. on p. 87).

[20]Shaul R. Foguel. *The Ergodic Theory of Markov Processes*. Van Nostrand Reinhold Company, 1969 (cit. on pp. 2, 15, 24).

[21]Thomas Gerstner and Michael Griebel. „Sparse Grids“. In: *Encyclopedia of Quantitative Finance*. Ed. by R. Cont. John Wiley and Sons, Feb. 2010 (cit. on p. 42).

[22]Paul R. Halmos. „Measurable transformations“. In: *Bull. Amer. Math. Soc.* 55.11 (Nov. 1949), pp. 1015–1034 (cit. on p. 14).

[23]Eberhard Hopf. *The general temporally discrete Markoff process*. Vol. 3. 1954, pp. 13–45 (cit. on pp. 1, 15, 24, 30, 45).

[24]Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985 (cit. on p. 87).

[25]Reiner Horst, Panos M. Pardalos, and Nguyen V. Thoai. *Introduction to Global Optimization*. Kluwer Academic Publishers, 1995 (cit. on p. 82).

[26]Wilhelm Huisinga. „Metastability of Markovian systems“. PhD thesis. Freie Universität Berlin, 2001 (cit. on p. 20).

[27]Wilhelm Huisinga and Bernd Schmidt. „Metastability and Dominan Eigenvaues of Transfer Operators“. In: *Lecture Notes in Computational Science and Engineering* 49 (2006), pp. 167–182 (cit. on p. 36).

[28]Oliver Junge and Péter Koltai. „Discretization of the Frobenius–Perron Operator Using a Sparse Haar Tensor Basis: The Sparse Ulam Method“. In: *SIAM Journal on Numerical Analysis* 47.5 (2009), pp. 3464–3485 (cit. on p. 42).

[29]Shizuo Kakutani. „Random Ergodic Theorems and Markoff Processes with a Stable Distribution“. In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, Calif.: University of California Press, 1951, pp. 247–261 (cit. on p. 30).

[30]Andrei N. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933 (cit. on p. 5).

[31]Andrei N. Kolmogoroff. „Zur Umkehrbarkeit der statistischen Naturgesetze“. German. In: *Mathematische Annalen* 113.1 (1937), pp. 766–772 (cit. on p. 19).

[32]Tomasz Komorowski and Joanna Tyrcha. „Asymptotic Properties of Some Markov Operators“. In: *Polish Academy of Sciences and Mathematics* 37 (1989) (cit. on p. 28).

[33]Ulrich Krengel. *Ergodic theorems*. English. W. de Gruyter Berlin ; New York, 1985, vii, 357 p. ; (cit. on pp. 14, 16).

[34]Andrzej Lasota and Michael C. Mackey. *Chaos, Fractals, and Noise*. Springer, 1994 (cit. on pp. 7, 23, 27, 28, 32).

[35]David J. C. MacKay. „Introduction to Monte Carlo Methods“. In: *Learning in Graphical Models*. Ed. by M. I. Jordan. NATO Science Series. Kluwer Academic Press, 1998, pp. 175–204 (cit. on p. 49).

[36]Carl D. Meyer. „Sensitivity Of The Stationary Distribution Of A Markov Chain“. In: *SIAM Journal on Matrix Analysis and Applications* 15 (1994), pp. 715–728 (cit. on pp. 51, 85).

[37]Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. second edition. New York: Cambridge University Press, 2009 (cit. on pp. 10, 11, 55).

[38]Laurent Miclo. „On hyperboundedness and spectrum of Markov operators“. English. In: *Inventiones mathematicae* 200.1 (2015), pp. 311–343 (cit. on p. 44).

[39]John von Neumann. „Proof of the Quasi-ergodic Hypothesis“. In: *Proc Natl Acad Sci USA* 18 (1932), pp. 70–82 (cit. on p. 1).

[40]Adam Nielsen. „Von Femtosekunden zu Minuten“. Updated version from September, 2015. MA thesis. Freie Universität Berlin, 2012 (cit. on pp. 12, 16, 17, 27, 44).

[41]Adam Nielsen and Marcus Weber. „Computing the nearest reversible Markov chain“. In: *Numerical Linear Algebra with Applications* 22.3 (2015), pp. 483–499 (cit. on p. 73).

[42]Bernt K. Øksendal. *Stochastic differential equations: an introduction with applications*. 6th ed. Berlin, Heidelberg, New York: Springer Verlag, 2003 (cit. on pp. 43, 64).

[43]Grigorios A. Pavliotis. *Stochastic Processes and Applications*. Springer, 2014 (cit. on p. 44).

[44]Henri Poincaré. „Sur le problème des trois corps et les équations de la dynamique“. In: *Acta Math.* 13 (1890), pp. 1–270 (cit. on p. 13).

[45]Ottis W. Rechard. „Invariant measures for many-one transformations“. In: *Duke Math. J.* 23.3 (Sept. 1956), pp. 477–488 (cit. on p. 45).

[46]Daniel Revuz. *Markov Chains*. 2nd. North-Holland Mathematical Library, 1984 (cit. on pp. 10–12, 26).

[47]Susanna Röblitz. „Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics“. PhD thesis. Freie Universität Berlin, 2008 (cit. on pp. 37, 42, 48).

[48]Ryszard Rudnicki. „Markov operators: applications to diffusion processes and population dynamics“. eng. In: *Applicationes Mathematicae* 27.1 (2000), pp. 67–79 (cit. on p. 28).

[49] Marco Sarich. „Projected Transfer Operators“. PhD thesis. Freie Universität Berlin, 2011 (cit. on pp. 39, 43).

[50] Christof Schütte. „Conformational Dynamics: Modelling, Theorey, Algorithm, and Applcation to Biomolecules.“ Habilitation. Freie Universität Berlin, 1999 (cit. on p. 46).

[51] Christof Schütte, Adam Nielsen, and Marcus Weber. „Markov state models and molecular alchemy“. In: *Molecular Physics,* 113.1 (2014) (cit. on pp. 63, 65).

[52] Christof Schütte, Wilhelm Huisinga, and Peter Deuflhard. „Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems“. English. In: (2001). Ed. by Bernold Fiedler, pp. 191–223 (cit. on pp. 1, 5).

[53] Christof Schütte and Marco Sarich. „A critical appraisal of Markov state models“. English. In: *The European Physical Journal Special Topics* (2015), pp. 1–18 (cit. on p. 2).

[54] Christof Schütte and Marco Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. A co-publication of the AMS and the Courant Institute of Mathematical Sciences at New York University. 2014 (cit. on p. 43).

[55] Daniel W. Stroock and Sathamangalam R.S. Varadhan. *Multidimensional Diffusion processes*. Berlin, Heidelberg: Springer, 2006 (cit. on p. 64).

[56] Domokos Szász. „Hard Ball Systems and the Lorentz Gas“. In: ed. by Domokos Szász. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. Chap. Boltzmann's Ergodic Hypothesis, a Conjecture for Centuries?, pp. 421–446 (cit. on p. 13).

[57] Stanisław M. Ulam. *A Collection of Mathematical Problems*. 1st ed. Interscience, 1960 (cit. on pp. 2, 14, 45).

[58] Marcus Weber. „A Subspace Approach to Molecular Markov State Models via a New Infinitesimal Generator“. Habilitation. Freie Universität Berlin, 2012 (cit. on pp. 42, 48).

[59] Marcus Weber. „Meshless Methods in Conformation Dynamics“. PhD thesis. Freie Universität Berlin, 2006 (cit. on pp. 37, 42, 48).

[60] Marcus Weber, Susanna Kube, Lionek Walter, and Peter Deuflhard. „Stable Computation of Probability Densities for Metastable Dynamical Systems“. In: *SIAM J. Multisc. Mod. Sim.* 6.2 (2007) (cit. on pp. 39, 51, 53, 85).

[61] Dirk Werner. *Funktionalanalysis*. 7th. Springer, 2011 (cit. on p. 9).

[62] Christoph Zenger. „Sparse grids“. In: *Parallel algorithms for partial differential equations* (1991). Ed. by W.Hackbusch (cit. on p. 42).