PROJECTED TRANSFER OPERATORS

Discretization of Markov Processes in High-Dimensional State Spaces

PhD Dissertation
Fachbereich Mathematik und Informatik
Freie Universität Berlin

Marco Sarich
June, 2011

Betreuer: Prof. Dr. Christof Schütte
Freie Universität Berlin
Fachbereich Mathematik und Informatik
Arnimallee 2-6
14195 Berlin

Gutachter: Prof. Dr. Christof Schütte
Prof. Dr. Michael Dellnitz (Universität Paderborn)

Tag der Disputation: 01. November 2011

Meinem Vater gewidmet.

Contents

Introduction

Models for the dynamics of complex systems are used in a wide range of scientific fields. Particularly in physical and chemical applications, nowadays many systems of interest are of so high complexity that one tries to avoid to model every single interaction within the system explicitly. Therefore, a multitude of these models make use of stochastic descriptions. Then, the evolution of the system is rather described by a stochastic process than by a deterministic dynamical system. Without doubt time-continuous Markov processes are most prominent under the stochastic processes considered not only in a physical or chemical context, but also in economic sciences, biology, meteorology, and other applications. On the other hand, Markov processes resulting from models of complex systems are usually too complicated for a direct analysis. In the last years, the size and complexity of the systems has been growing tremendously, which has led to high-dimensional state spaces for the associated Markov processes. Many systems exhibit multiscale dynamics in addition.

There have been various approaches to dimension reduction and simplification of Markov processes. One very successful class of such methods is given by so called Markov State Models (MSM). For 15 years, Markov State Models have been used as low-dimensional models for processes on very large, mostly continuous state spaces exhibiting metastable dynamics. This means that one can subdivide state space into metastable sets in which the system remains for long periods of time before it exits quickly to another metastable set. Here the words "long" and "quickly" mainly state that the typical residence time has to be much longer than the typical transition time so that the jump process between the metastable sets is approximately Markovian. Then, the goal of Markov State Modeling is the approximation of the original Markov process by a Markov chain on a small finite state space. For this purpose, the transition probabilities of the Markov chain are calculated from transition probabilities of the original process between the subsets of state space. In the first section of Chapter 2 the construction of Markov State Models is explained in detail.

Particularly in molecular dynamics, MSM have become popular as approximations of the conformational dynamics [72, 73, 84] of large biomolecules, which exhibits various timescales ranging from protein folding to fast vibrations and oscillations within a molecular conformation. There, the subdivision of the conformational state space has been usually achieved by partitioning [56, 11, 15, 35, 39, 60, 64, 52, 71, 80, 81]. Since such Markov State Models are defined by transition probabilities between sets, they can be estimated from trajectories. That is, one can sample the probabilities by many short and independent realizations of the Markov process. Usually, the required length of these trajectories is not comparable to the timescales of interest [58, 77]. This property attracted also the attention of computer sciences and even large projects concentrating on the construction of MSM

from heavy parallelized simulation have been established [5].

Until 2005, the construction of MSM has been always based on so called full partitions, i.e. sets that cover the whole state space. We will also refer to this type of MSM as classical or Standard Markov State Models. Then, in [19] a variant was introduced that used fuzzy affiliation functions instead of sets in state space. Two years ago it was proposed in [11] to use a set based approach again, but to construct a fuzzy MSM variant by defining small disjoint sets in the most dominant metastable regions. We will also call these sets *core sets*. Another approach that relies on milestoning [28] and on core sets has been recently discussed in [70].

In this thesis, we will develop a mathematical theory for a general class of approximations that will contain all these former MSM approaches. For this purpose, we will use stochastic and functional analytic concepts, which will require a higher level of abstraction. Since we know that Markov State Models are used very interdisciplinarily, we will carefully introduce in Chapter 1 the mathematical background that is needed to follow the argumentation for the rest of the thesis.

Chapter 2 will introduce the framework of projected transfer operators. That is, we will consider operators of the form QTQ, where T is the transfer operator of the original Markov process and Q is an orthogonal projection onto an appropriate space D. In Section 2.3 it will become clear how classical MSM and Markov State Models based on fuzzy affiliations are directly related to this approach. Especially, the projection onto the space spanned by the so called committors [24, 25] will turn out to be valuable. We will show that the associated projected operator has a stochastic interpretation. We will interpret the action of the operator QTQ to a density, explain under which conditions it preserves probability and how one can estimate a matrix representation from trajectories in terms of stopping times.

In Chapter 3 we will finally answer questions about the approximation quality of Markov State Models by analyzing the associated projected transfer operators QTQ. In Section 3.1 we focus on the issue that switching processes between sets as they appear in Markov State Modeling are not Markovian. Nevertheless, one uses a Markov chain on a small state space to approximate the dynamics. We will analyze under which conditions this approximation is reasonable. Then, Section 3.2 is about eigenvalues of the transfer operator, i.e. the timescales of the Markov process (see Sec. 1.3), and how they are reflected in the timescales of the MSM. We will find out that a small projection error of the dominant eigenvectors to the subspace D is an important criterion to guarantee a good quality in the sense of approximation of the switching processes and reflection of the longest timescales in the system. All these results will also be valid for the classical Markov State Models or other methods that can be connected to projected operators. In [63], these results are used to provide a guideline for the generation and validation of Standard Markov State Models. On the other hand, we

will find that one does not necessarily have to approximate the dominant timescales. We will show that it is theoretically also possible to pick certain interesting eigenvalues of T and ensure that the projected operator will describe the dynamics on these timescales. This will also lead to a discussion of a multilevel discretization of the Markov process. That is, one chooses several different numbers of core sets to achieve better approximations of different timescales. Section 3.3 will then discuss the consequences of the abstract results for the developed core set approach. Here, the fundamental question is how to choose the sets, i.e. the discretization, in order to build a good Markov State Model. Throughout the analysis, projection errors of eigenvectors with respect to the subspace D will turn out to be important. On the other hand, these eigenvectors are usually unknown. Moreover, often one cannot compute the projection onto the subspace spanned by the committors as well. So, we will have to gain insight into this projection error and discuss under which conditions it will be small. We will prove that we can estimate the projection error for every timescale from properties of the sets, which measure the flow from and into the core sets compared to the timescales of interest. From this we deduce how to define candidates for core sets aiming at the approximation of a specific part of the spectrum of the transfer operator. Finally, we construct an algorithm for the identification of appropriate core sets and the estimation of the corresponding Markov State Model from trajectories. We test the algorithm by approximating two diffusions in a one dimensional and in a two dimensional potential, respectively. The resulting Markov State Models that will be estimated completely simulation based will turn out to match the theoretical results. This implies that they will be superior to full partition methods for more complicated processes. Moreover, we will show how the new Markov State Modeling technique corresponds to fuzzy cluster methods in Section 3.5. It will turn out that for finite Markov jump processes we can avoid the sampling problem in the algorithm that we developed in Sec. 3.4. This will provide a new approach to fuzzy clustering, which has many advantages compared to other methods. We will demonstrate its behaviour for a network example.

This will point out the broad impact and applicability of the developed mathematical framework. We will start with Markov State Models, i.e. discretizations for Markov processes on continuous state spaces, and using the results from this analysis we will end with a proposal for fuzzy clustering of networks.

Acknowledgments

Many people supported me so much in the last years of research. First of all, I want to thank Christof Schütte for giving me the possibility to write this thesis, for the steady support, and that he introduced me to this vibrant field of research. I want to mention the whole biocomputing group, not only to

avoid forgetting somebody, but because the group in total provides this very special, amicable and motivating working environment. A special thanks goes to Carsten Hartmann, Frank Noé, and Illia Horenko for explanations and general, also not research connected advice. Natasa Djurdjevac for walking all the way together and making research become much more fun. My office mates Tomaso and Steffi, who have always been there for a talk, scientifically or not. Eric Vanden-Eijnden and John Chodera, who gave important input and ideas during several discussions. Finally, I want to thank Sandi. She has always been the foundation for everything.

This work was funded and supported by the DFG Research Center MATHEON "Mathematics for Key Technologies" and the Berlin Mathematical School.

1 Fundamentals of Markov Theory

The goal of this chapter is to present the mathematical concepts which are needed to follow the argumentation of the next chapters. Of course, there are a lot of textbooks about elementary stochastics, stochastic processes, and more specifically about Markov processes [10, 65, 27], but I wanted to provide an introduction that exactly fits to this thesis. The intention is to give understandable interpretations and motivations of the definitions and formulas, from the very basic ones to more complicated concepts such that one can quickly get an overview of the framework.

1.1 Probabilistic Framework

Probabilities

The very first definition in an elementary textbook about stochastics will be most likely the probability space as a triple $(\Omega, \mathcal{A}, P)$. When I say "most likely", I mean that if you took all elementary textbooks dealing with probabilities, the number of textbooks which start with the probability space would be much larger than the number of textbooks which define something else first. Actually, this is already our *objective view* onto probability. When we look at some experiment, e.g. take one of these probabilistic textbooks, we want to talk about the probability that some event will happen, e.g. that the triple $(\Omega, \mathcal{A}, P)$ will be the first definition in the book. The objective interpretation of probability asks for the ratio how often this event would happen if we ran this experiment more and more frequently, compared to the total number of tries. In life there is also a *subjective interpretation* of probability because sometimes one does not have the possibility to retry an experiment. Even in situations like this one tries to estimate how likely some event is although one can hardly validate this quantification. Mathematically we try to cover both of these understandings by following the dominant number of textbooks and making the first definition.

Definition 1 (Probability Space) *We call a triple $(\Omega, \mathcal{A}, P)$ probability space if $\mathcal{A}$ is a sigma field over the arbitrary set Ω, that is, $\mathcal{A}$ is a collection of subsets of Ω with*

1. $\emptyset, \Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$
3. $A_1, A_2, ... \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

and P is a probability measure, i.e. it is a map $P : \mathcal{A} \to [0, 1]$ fulfilling

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$, *for* $A_1, A_2, ... \in \mathcal{A}$, $A_i \cap A_j = \emptyset, i \neq j$.

So looking back at the motivation, what we try to model is the following. We have some experiment or a situation with a random outcome, so maybe we don't have the tools, or the insight, or it is simply not possible to forecast the result. The set Ω is the set of all elementary results that theoretically can happen. One simple standard example is the dice with six faces. Here, we could set $\Omega = \{1, 2, 3, 4, 5, 6\}$ which corresponds to the possible results of one dice roll. Then we want to ask for probabilities, e.g. what is the probability that the dice will show an even number? Here, the sigma field $\mathcal{A}$ comes into play. The question we just asked can be reformulated in the following way. Let us call the number the dice will show $\omega \in \Omega$, then what is the probability for $\omega \in \{2, 4, 6\} =: A$? So, asking questions about properties of the result correspond one-to-one to questions of the form, is $\omega \in A$ for subsets $A \subset \Omega$. The sigma field $\mathcal{A}$ is nothing else than the collection of these questions that are allowed to be asked or that we want to calculate probabilities for, and the probability measure P assigns a number between 0 and 1 to every of these questions. For an experiment that we can repeat under the same conditions several times, this number will represent the limit of the ratio

$$\frac{N(A)}{n} \to P(A), \quad n \to \infty,$$

where $N(A)$ is the number of runs of the experiment where the result was in set A, and n is the total number of tries. This corresponds to the objective interpretation of probability, which is absolutely suitable for this thesis because in our applications we concentrate on repeatable experiments.

Random Variables

We will see that the triple $(\Omega, \mathcal{A}, P)$ is very difficult to get access to directly, in general. Often it is even anything but trivial to construct the probability

space that should be a model for some application. So what we need is a tool that can be used to filter information about the probability space, that can be used for further analysis, or even for the construction of probability spaces. This tool is given by

Definition 2 (Random Variable) *A random variable on* $(\Omega, \mathcal{A}, P)$ *with state space* $(E, \mathcal{E})$*, where* $\mathcal{E}$ *is a sigma field on* E*, is a map* $X : \Omega \to E$*, such that*

$$\{\omega \in \Omega | X(\omega) \in A\} = X^{-1}(A) \in \mathcal{A}, \quad \forall A \in \mathcal{E}. \tag{1.1}$$

In short, a random variable does nothing else than constructing a new probability space, namely $(E, \mathcal{E}, P_X)$ where the probability measure P_X is given by

$$P_X(A) = P(X^{-1}(A)) \quad \forall A \in \mathcal{E}.$$

Therefore, we need the condition (1.1) because P is only defined on $\mathcal{A}$, so $X^{-1}(A)$ has to be an element of it. Usually the state space E will be much simpler than the probability space Ω itself. Particularly two classes of random variables will be important for the rest of the thesis.

1. Discrete random variable: $E = \{1, 2, ...\}$, where E is finite or denumerable.
2. Continuous random variable: $E = \mathbb{R}^n$ for some $n \geq 1$.

In both cases we consider the canonical Borel sigma field on E.

Conditional Probability

Information can change probabilities. This might be the common slogan for conditional probabilities. Take the following example. We throw the dice such that we do not see the result, and we still ask the question: "What is the probability for an even number?" Using $A = \{2, 4, 6\}$ we can write and easily calculate $P(A) = 1/2$ but things change if we have additional information. Assume that somebody who knows the result tells us that it is a number larger than or equal to 4. So, we have the information $\omega \in \{4, 5, 6\} =: B$. Because two out of these three numbers are even, the probability for an even number changed to $P(A|B) = 2/3$. We write $P(A|B)$ for the probability that the result is in A under the condition that it is in B. In general, one can calculate

Definition 3 (Conditional Probability) *The conditional probability of set* A *given set* B *is defined by*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

for $P(B) > 0$*.*

In general, the conditional probability is undefined for events B with $P(B) = 0$. At first sight, this does not seem to be an obstacle because if B is an event with probability $P(B) = 0$, why should we explicitly think about a probability for an event A in the case that event B happens? The problem is that in applications, especially when dealing with stochastic processes on continuous state space, the best information one will have access to will be events B with probability $P(B) = 0$. One example are two random variables X_0, X_1 on $\mathbb{R}$. Assume that $X_1 = 2X_0$. If we are interested in probabilities for X_1, the best information we can get is the explicit value of X_0 because it should hold

$$\mathbb{P}[X_1 = 2x | X_0 = x] = 1 \qquad \mathbb{P}[X_1 = y | X_0 = x] = 0, y \neq 2x.$$

On the other hand, for a continuous random variable, i.e. a random variable X_0 on $\mathbb{R}$, the event $B = \{X_0 = x\}$ will have zero probability $P(B) = 0$. In the next Section 1.2 we will find a different way to define a conditional probability that will overcome this problem, and that will fit better into the theory of stochastic processes.
Of course, information does not always change probabilities.

Definition 4 (Independence) *We call two events A and B independent of each other if*

$$P(A \cap B) = P(A)P(B),$$

which implies

$$P(A|B) = P(A) \qquad P(B|A) = P(B).$$

Moreover, we call two random variables X and Y independent if all pairs of events $X^{-1}(A), Y^{-1}(B)$ are independent, where A and B are elements of the corresponding sigma fields of state space.

So, knowledge about A or B does not influence the probability for the other event. For random variables it means that never any kind of knowledge about one random variable can change the probability for events of the other random variable.

1.2 Stochastic Processes

Let us straightforwardly define stochastic processes in the abstract way first and afterwards discuss the definition.

Definition 5 (Stochastic Process) *A stochastic process with index set I and state space E is a family of random variables*

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

defined on a probability space $(\Omega, \mathcal{A}, P)$.

This definition looks very compact and not at all dangerous. We just have a set of random variables which are defined on the same probability space and which have the same state space. For example, we could consider the space of the dice again, i.e. $E = \Omega = \{1, 2, 3, 4, 5, 6\}$, and simply define $X_i(\omega) = \omega$. This means that our process could have six outcomes, namely where all of the random variables show the same value $1, 2, 3, 4, 5$ or 6 of the dice roll. First, the random variables are not independent because knowledge about one random variable implies the same knowledge about every other random variable. Second, it is not a very interesting process. The word *process* suggests that the index set I has usually some interpretation as time ($I = \mathbb{R}^+$) or as time steps ($I = \mathbb{N}$), and that the random variables of the process should depend on each other and develop in time.

Random Walk

Having this motivation in mind, we want to construct another stochastic process on $I = \mathbb{N} = \{0, 1, 2, ...\}$ and $E = \mathbb{Z}$ that is called a random walk on $\mathbb{Z}$. We name the properties that our process, i.e. our random variables $(X_n)_{n\in\mathbb{N}}$, should have.

1. For the initial random variable X_0 it holds $\mathbb{P}[X_0 = 0] = 1$.
2. For every other random variable it holds $X_{j+1} = X_j + Z_j$, where $(Z_j)_{j\in\mathbb{N}}$ are independent random variables with $\mathbb{P}[Z_j = 1] = 1/2$, and $\mathbb{P}[Z_j = -1] = 1/2$.

The random walk does the following. It starts at $X_0 = 0$ and in every step it randomly and independently chooses to increase or decrease the actual value by 1, where each option has the probability $1/2$. This looks much more like something we would call a stochastic process, but following the Definition 5 it is not clear that it is one, yet. The ingredients of a stochastic process are a probability space $(\Omega, \mathcal{A}, P)$, an index set I, a state space E, and the random variables (X_i). We have $I = \mathbb{N}$ and $E = \mathbb{Z}$ already but we have to construct a suitable probability space and random variables that fulfill our criteria. In this case we can use the random variables (Z_j). We can define the space $\Omega_0 = \{-1, 1\}$ and a probability measure with $\mathbb{P}[Z_0 = 1] = 1/2$ and $\mathbb{P}[Z_0 = -1] = 1/2$. Now there is a well-known theorem that there is a probability space $(\Omega, \mathcal{A}, P)$ which generates a sequence of independent copies of Z_0 that we call $(Z_j)_{j\in\mathbb{N}}$. This is exactly the probability space we are looking for because we can recursively define

$$X_{i+1}(\omega) = \sum_{k=0}^{i} Z_k(\omega) = X_i(\omega) + Z_i(\omega), \quad i \in \mathbb{N}$$

with $X_0(\omega) = 1\ \forall\omega \in \Omega$.

Remark 1 *As pointed out before, random variables can be used to construct probability spaces. Here the sequence of random variables (Z_i) have been the tool to define the probability space for the random walk.*

Different Perspectives

In the following we want to consider two classes of stochastic processes, which differ in the choice of the index set I.

1. Time-discrete processes: $I = \mathbb{N} = \{0, 1, 2, ...\}$,
2. Time-continuous processes: $I = \mathbb{R}^+ = [0, \infty)$,

so we include the zero to the index sets. As mentioned above, these cases are very natural and connected to the word *process*, but it also allows to change perspective. By Definition 5 a time-discrete or time-continuous stochastic process is a family of random variables with index set $\mathbb{N}$ or $\mathbb{R}^+$, respectively. On the other hand, one could also think of a process as one random variable X with a different state space, namely

$$\begin{aligned} X &: \Omega \to E^I \\ X(\omega)(i) &= X_i(\omega), \forall i \in I. \end{aligned} \tag{1.2}$$

For $I = \mathbb{N}$ the state space is $E^{\mathbb{N}}$, so the output of X would be a sequence in E and for $I = \mathbb{R}^+$ X would deliver a function $X(\omega) : \mathbb{R}^+ \to E$. These sequences or functions, which are the realizations of X, are also called **paths** or **trajectories**. This point of view allows to ask different questions, e.g. what is $\mathbb{P}[\{\omega | X(\omega) : \mathbb{R}^+ \to E \text{ is continuous}\}]$?
Of course, the space of trajectories is a very rich state space for the analysis of stochastic processes. In many cases one can even identify the space E^I with the probability space Ω itself, but particularly the probability measure on this large state space is difficult to handle. For practical applications like numerical simulation or statistical analysis one will also not be able to generate one complete realization of such a trajectory in the infinite state space E^I. One can only sample sequences of finite length, that is, a random variable Y

$$\begin{aligned} Y &: \Omega \to E^J \\ Y(\omega)(j) &= X_j(\omega), \forall j \in J, \end{aligned} \tag{1.3}$$

where $J \subset I$ is a finite subset of I. For many practical purposes this random variable Y is the filter at hand for the probability space Ω because we can generate and numerically analyze realizations of this random variable. Often the finite set of time steps is chosen to be of the form $J = \{0, \tau, 2\tau, ..., N\tau\}$, that is, a uniform time-discretization with *lag time* $\tau > 0$. Nevertheless, even for this random variable Y the induced probability distribution on E^J is often directly unaccessible for reasonable size of J. Moreover, it not clear

how we can answer questions targeting the space of trajectories by looking at finite sequences only.
In this thesis we will not approach general stochastic processes. We will restrict us to a certain class of processes, namely **Markov processes**. This class is characterized by a very special dependency structure of the random variables (X_i), but before we start to dive into the analysis of Markov processes we have to fix one problem. The word dependency structure indicates that we will need conditional probabilities to describe the essential property of Markov processes, but as pointed out before we do not have a suitable definition of conditional probability that can handle stochastic processes, yet.

Conditional Probability Advanced

We have been able to define a probability of an event A under the condition that event B happened already in Definition 3, but only assuming that $P(B) > 0$. In the analysis of stochastic processes an essential information is the evolution of a trajectory up to some time T, that is, we already know the exact values of the random variables (X_i) for $i \leq T$ and are interested in probabilities of further development. Sadly, this is a too precise background information for which we have

$$\mathbb{P}[X_i = x_i,\ \forall i \leq T] = 0, \quad x_i \in E\ \forall i \leq T.$$

For such events we have not been able to define a conditional probability, yet. We will catch up now.

Definition 6 (Conditional Expectation) *Let X be a random variable on $(\Omega, \mathcal{A}, P)$ and let $\mathcal{F} \subset \mathcal{A}$ be a sub-sigma field of $\mathcal{A}$, i.e. a subset which is a sigma field again. Then, a random variable $Y = \mathbb{E}[X|\mathcal{F}]$ is called **conditional expectation** of X given the sigma field $\mathcal{F}$ if*

- *Y is $\mathcal{F}$-measurable*
- *$\forall A \in \mathcal{F}$ it holds $\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[Y\mathbb{1}_A]$.*

Remark 2 *If $\mathbb{E}[X] < \infty$ exists, the conditional expectation exists and is unique almost surely.*

The title of this section promised "conditional probability advanced". First, we did not define a conditional probability, but rather a conditional expectation. This is no restriction because one can always write for $A \in \mathcal{A}$

$$P(A) = \mathbb{E}[\mathbb{1}_A] \Rightarrow P(A|\mathcal{F}) = \mathbb{E}[\mathbb{1}_A|\mathcal{F}].$$

Second, there are two more things that do not make sense immediately.

1. The conditional probability is taken with respect to a sigma field and not with respect to a certain event.

2. The conditional probability is the random variable $\mathbb{E}[\mathbb{1}_A|\mathcal{F}]$ and not a number between 0 and 1.

At least, we have that

$$P(A|\mathcal{F}) : \Omega \to [0,1]$$

takes values in the interval $[0,1]$, as a probability should. So, instead of the simple conditional probability $P(A|B) \in [0,1]$, we have $P(A|\mathcal{F})(\omega) \in [0,1]$. Actually, this is the secret which solves the problem with the conditional probability depending on events. What we did is, we changed the order of question and answer. $P(A|B)$ has to define a probability a priori, that is, how would we rate the probability for A if we experienced $\omega \in B$? We can hardly answer this question because we cannot imagine that $\omega \in B$ will ever happen if $P(B) = 0$. $P(A|\mathcal{F})(\omega)$ approaches this task the other way around. We first let ω happen and afterwards ask for a conditional probability. Therefore, $P(A|\mathcal{F})(\omega)$ depends on ω. $\mathcal{F}$ defines which events we want to monitor, e.g. $\mathcal{F} = \{\emptyset, \Omega, B, \Omega \setminus B\}$. This defines which information about ω we will have access to, i.e. somebody will tell us for every ω that we randomly generate if $\omega \in F$ for all $F \in \mathcal{F}$. Then, we ask for the probability that also $\omega \in A$ holds.
For example, one can show that

$$\forall F \in \mathcal{F}, P(F|\mathcal{F})(\omega) = \begin{cases} 1, & \omega \in F \\ 0, & \omega \notin F. \end{cases}$$

This makes sense because for an event $F \in \mathcal{F}$ we can tell for sure if it happened or not as soon as ω was generated.
This interpretation of conditional probability will be very useful in the next section because it fits perfectly to the idea of stochastic processes.

1.3 Markov Processes

From now on we want to focus on a particular class of stochastic processes, namely Markov processes. These processes have one characterizing property that is called the **Markov property**. In short, one can say that these processes are memoryless. At every point in time a Markov process is only aware of its present state, but does not remember anything of the past. Mathematically, this is a constraint on the dependency structure of the random variables $(X_i)_{i \in I}$. For discrete time, i.e. $I = \mathbb{N}$, and discrete state space we can write the Markov property as

$$\begin{aligned} \mathbb{P}[X_{n+1} = x_{n+1}|X_n = x_n, ..., X_0 = x_0] = \mathbb{P}[X_{n+1} = x_{n+1}|X_n = x_n] \\ \forall x_i \in E, i = 0, ..., n+1. \end{aligned} \tag{1.4}$$

(1.4) means that the conditional probability distributions of the value of the process at time step $n+1$ given the complete history and given the previous state only are not distinguishable.
Now, we want to give a similar formulation for a time-continuous Markov process with continuous state space. As in (1.4), we have to use an equation involving conditional probabilities which depend on complete observations of a piece of trajectory $\{X_t, t \leq s\}$ until time s or just the value of the final state X_s, respectively. In the previous section we have seen that these kind of conditional probabilities can cause trouble because

$$\mathbb{P}[X_t = x_t, t \leq s] = 0, \tag{1.5}$$

that is, particular pieces of trajectories will have zero probabilities for uncountable state spaces and/or continuous time, in general. We fixed this by adapting the definition of conditional probabilities, which made use of sigma fields. In order to apply this definition we have to construct a suitable sigma field that monitors the trajectory (X_t) up to some time s. That is, it has to consist of all events for which one can tell if they happened or not if one has the concrete information of all values of X_t for $t \leq s$. This sigma field is given by

$$\mathcal{F}_s = \sigma\{X_t^{-1}(A), t \leq s, A \in \mathcal{E}\} \subset \mathcal{A}. \tag{1.6}$$

Obviously, it holds

$$\mathcal{F}_t \subset \mathcal{F}_s \subset \mathcal{A}, \ \forall t < s. \tag{1.7}$$

A family of sigma fields with this property is called **filtration**. Filtrations often appear in the context of stochastic processes because they can be used elegantly to define conditional probabilities with respect to history information of trajectories in the sense of Def. 6. Using the filtration from (1.6) we can define the Markov property in general.

Definition 7 *A stochastic process (X_t) on state space E is called **Markov process** if it holds*

$$\mathbb{P}[X_t \in A|\mathcal{F}_s] = \mathbb{P}[X_t \in A|X_s] \quad \forall t \geq s. \tag{1.8}$$

Here, $\mathbb{P}[X_t \in A|X_s]$ is the abbreviated version of

$$\mathbb{P}[X_t \in A|\sigma(X_s)], \qquad \sigma(X_s) = \sigma\{X_s^{-1}(A), A \in \mathcal{E}\} \subset \mathcal{F}_s. \tag{1.9}$$

*Equation (1.8) is called **Markov property**.*

Def. 7 delivers a proper definition of Markov processes for discrete and continuous time and arbitrary state spaces. Moreover, the formulation (1.8) of the Markov property is as short as possible. This is why filtrations and the advanced conditional probability are so useful in this context. Again,

the interpretation of (1.8) is that we want quantify how likely it is that the process will be in a set A at time t, i.e. $X_t \in A$, and that we have two different spies that provide us with some background information. Spy number one is the sigma field $\mathcal{F}_s$ that can answer every question about the process (X_i) for $i \leq s$, and spy number two, which is the sigma field $\sigma(X_s)$, monitors only the last random variable X_s. (1.8) means that the additional information that the sigma field $\mathcal{F}_s$ delivers does not change the conditional probability. So it is sufficient to monitor the last state only.

Markov Processes on Discrete State Space

We already discussed that one can imagine a stochastic process as a random variable on a very complicated probability space that generates points in a suitable path space E^I. That is, one realization of the random variable will be a sequence or a function in state space E. The limitation from stochastic processes in general to the class of Markov processes has one essential advantage. One can avoid the direct analysis of the unhandy probability measure in path space because it is possible to construct the Markov process, i.e. the probability space and the probability measure, with the use of simpler objects. First, we want to illustrate this in the case of discrete time and state space, where we also call the Markov process a **Markov chain**.

Proposition 1
Let $E \subset \mathbb{N}$ and time $I = \mathbb{N}$ be discrete. Then, for any stochastic matrix P, i.e.

$$P(i,j) \geq 0, \qquad \sum_{j \in E} P(i,j) = 1 \tag{1.10}$$

and any probability distribution ν, there exists a Markov process $(X_n)_{n \in N}$ with

$$\mathbb{P}[X_{n+1} = j | X_n = i] = P(i,j), \qquad P(X_0 = i_0) = \nu(i_0). \tag{1.11}$$

In equation (1.11), the **transition probability** $\mathbb{P}[X_{n+1} = j | X_n = i]$ does not depend on the actual time step n. Such a Markov process is called **homogeneous** and the matrix P its **transition matrix**. We will only consider this class of Markov processes in the following. Note that the state space can be infinite but countable such that the matrix can have infinitely many entries.

Remark 3 *The random walk example of Sec. 1.2 is a homogeneous Markov process with transition matrix $P(i, i+1) = P(i, i-1) = 1/2$ and initial distribution $\nu(0) = 1$.*

Proposition 1 immediately guarantees the existence of a suitable probability space and measure. Moreover, the so called **finite dimensional distributions**, which define the probability measure for the finite random variable

Y from (1.3), can be computed by

$$\begin{aligned}&\mathbb{P}[X_0 = i_0, X_{n_1} = i_1, ..., X_{n_k} = i_k] = \nu(i_0)P^{d_1}(i_0, i_1) \cdot .... \cdot P^{d_k}(i_{k-1}, i_k),\\ &d_j = n_j - n_{j-1}, j = 2, ..., k \quad d_1 = n_1.\end{aligned} \tag{1.12}$$

(1.12) gives the probability distribution of a finite subtrajectory at time steps $n_1, ..., n_k$.

Transfer of probability The transition probabilities $P(i, j)$ generate together with an initial distribution ν the probability space and the Markov chain, but they also define a matrix and hence a linear operator $\mathcal{P} : l^1 \to l^1$

$$(\mathcal{P}v)(j) = \sum_{i \in E} P(i, j)v(i), \tag{1.13}$$

on $l^1 = \{v : E \to \mathbb{R} | \sum_{i \in E} |v(i)| < \infty\}$. When $E = \{1, ..., n\}$ is finite, $l^1 = \mathbb{R}^n$ but for infinite state spaces the restriction to l^1 makes sure that the sum in (1.13) converges such that $\mathcal{P}$ is well-defined. With the definition of the operator $\mathcal{P}$ we enter an algebraic or functional analytic framework, but applying the operator to probability distributions ν,

$$\nu(i) \geq 0, \quad \sum_{i \in E} \nu(i) = 1, \tag{1.14}$$

has still a stochastic interpretation. Assuming that ν is the initial distribution of the Markov chain we can compute

$$\begin{aligned}(\mathcal{P}\nu)(j) &= \sum_{i \in E} P(i, j)v(i)\\ &= \sum_{i \in E} \mathbb{P}[X_1 = j | X_0 = i]\mathbb{P}[X_0 = i] = \mathbb{P}[X_1 = j].\end{aligned} \tag{1.15}$$

Since the Markov chain is homogeneous we could also replace the time indices 0 and 1 by any time steps k and $k+1$. Because of the Markov property we also have

$$(\mathcal{P}^k\nu)(j) = \mathbb{P}[X_k = j]. \tag{1.16}$$

So $\mathcal{P}$ propagates distributions in time i.e. it tells us how the Markov chain will be distributed in later time steps if we start according to some initial distribution.

Invariant measure and reversibility. (1.16) defines a discrete dynamical system on probability distributions and when it comes to the analysis of dynamical systems one of the first objects of interest are fixpoints. In our case this would be a distribution μ which fulfills

$$\mathcal{P}\mu = \mu. \tag{1.17}$$

The interpretation of (1.17) is that if we generate the probability space for our Markov chain with the transition matrix P and the initial distribution μ, the random variables X_k will be identically (but not independently) distributed according to μ for all $k \in \mathbb{N}$. If we do not explicitly provide a different initial distribution, we will consider Markov processes in this **equilibrated probability space** in the following.

From the definition of $\mathcal{P}$ it follows that μ is nothing else than a left eigenvectors of the matrix P with respect to the eigenvalue $\lambda = 1$. Since $P\mathbb{1} = \mathbb{1}$ we know that P has such an eigenvalue and therefore, (1.17) has a non-trivial solution. On the other hand, for an infinite state space it does not necessarily hold $\mu \in l^1$, but we assume that this is the case such that we can also normalize μ to a probability distribution. In [76, 59], for example, one can find exact statements for what kind of processes this assumption might fail. Moreover, we assume in the following for the sake of simplicity that the eigenvalue $\lambda = 1$ is simple (irreducibility) because then the probability distribution μ satisfying (1.17) is unique. It is called **invariant measure** or **stationary distribution**.

We will also concentrate later on a certain class of Markov processes that is called **reversible** because it holds

$$\mu(i)P(i,j) = \mu(j)P(j,i). \tag{1.18}$$

Equation (1.18) is also called detailed balance condition. The interpretation, which is also responsible for the name *reversibility*, is simple. Assume that we know that for a finite realization of the Markov chain $X_0, ..., X_n$ one of the following two events had happened. We know that the Markov chain visited the states $a_0, ..., a_n$, but either we had

$$X_0 = a_0, ..., X_n = a_n \quad \text{or} \quad X_0 = a_n, ..., X_n = a_0. \tag{1.19}$$

Now, for a reversible Markov process started in μ we get by using (1.18) that

$$\mathbb{P}[X_0 = a_0, ..., X_n = a_n | (1.19)] = \mathbb{P}[X_0 = a_n, ..., X_n = a_0 | (1.19)] = \frac{1}{2}.$$

So we cannot tell in which direction the process has visited the states $a_0, ..., a_n$. One also says that one cannot distinguish the process from its time-reversed counterpart.

Transfer operator and spectral decomposition. In (1.13) we defined the linear operator $\mathcal{P}$ that generates the probability distributions of all random variables X_k from the initial distribution ν by propagation. We assumed that the Markov process has a unique positive invariant measure μ and therefore, we can also look at the propagation of a distribution v with

respect to μ, i.e. a vector v with

$$v(i) \geq 0, \quad \sum_{i \in E} v(i)\mu(i) = 1. \tag{1.20}$$

That is, $(v\mu)(i) := v(i)\mu(i)$ is a probability distribution in the usual sense, so we can apply $\mathcal{P}$. On the other hand, $\mathcal{P}(v\mu)$ will also be a usual probability vector and not one with respect to μ as in (1.20), but we can simply fix this by dividing $\mathcal{P}(v\mu)$ entrywise by μ. Combining these steps we can describe the propagation of probability distributions with respect to μ by the operator T,

$$(Tv)(j) = \frac{1}{\mu(j)} \sum_{i \in E} P(i,j)v(i)\mu(i). \tag{1.21}$$

T is well-defined on the space $l^1(\mu) = \{v : E \to \mathbb{R} | \sum_{i \in E} |v(i)\mu(i)| < \infty\}$. Nevertheless, we want to use functional analytical tools for understanding properties of this operator by using a Hilbert space framework. For this purpose, we will restrict the action of T to the Hilbert space $l^2(\mu) = \{v : E \to \mathbb{R} | \sum_{i \in E} v(i)^2\mu(i) < \infty\}$ equipped with the scalar product

$$\langle v, w \rangle = \sum_{i \in E} v(i)w(i)\mu(i). \tag{1.22}$$

We call the operator $T : l^2(\mu) \to l^2(\mu)$ from (1.21) **transfer operator**.

The question is why we make this effort to look at distributions in the μ-weighted Hilbert space. The advantage is demonstrated best for reversible Markov processes because if (1.18) holds one can compute directly that T is a self-adjoint operator, i.e.

$$\langle Tv, w \rangle = \langle v, Tw \rangle. \tag{1.23}$$

One useful consequence is that T has to have real eigenvalues and eigenvectors which form an orthonormal basis of $l^2(\mu)$. In the following we will always order them according to their magnitude, i.e. T has the eigenvalues $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq ...$ and associated eigenvectors $u_0, u_1, u_2, ...$ with

$$\langle u_i, u_j \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{1.24}$$

Note that $u_0 = \mathbb{1}$ represents the invariant measure μ itself in $l^2(\mu)$, and we assumed that the eigenvalue $\lambda_0 = 1$ is simple. Moreover, every $v \in l^2(\mu)$ can be represented as

$$v = \sum_{i=0}^{\infty} \langle v, u_i \rangle u_i, \tag{1.25}$$

which also implies

$$\|v\|^2 = \sum_{i=0}^{\infty} \langle v, u_i \rangle^2. \tag{1.26}$$

Particularly, the representation (1.25) of $T^k v$ yields the spectral decomposition of T

$$T^k v = \sum_{i=0}^{\infty} \lambda_i^k \langle v, u_i \rangle u_i. \tag{1.27}$$

Remember that for an initial distribution v, $T^k v$ is the distribution of the random variable X_k. That is, if T does not have an eigenvalue $\lambda = -1$ (aperiodicity), for every initial distribution v (1.20) we find

$$\|T^k v - \mathbb{1}\| \leq \lambda_1^k \to 0, k \to \infty. \tag{1.28}$$

So the distributions of the random variables X_k will converge towards the stationary distribution and the speed of convergence is bounded by λ_1^k. On the other hand, this convergence speed can be achieved. Let us construct a distribution $\bar{u}_1$

$$\bar{u}_1 = \frac{1}{m_1}(u_1 + m_1 \mathbb{1}), \tag{1.29}$$

where

$$m_1 = -\min_{i \in E}\{u_1(i)\} > 0. \tag{1.30}$$

Then, $\bar{u_1}$ is a probability distribution with respect to μ and

$$T^k \bar{u_1} - \mathbb{1} = \frac{1}{m_1}(\lambda_1^k u_1 + m_1 \mathbb{1}) - \mathbb{1} = \frac{1}{m_1}\lambda_1^k u_1, \tag{1.31}$$

which implies

$$\|T^k \bar{u_1} - \mathbb{1}\| = \frac{1}{m_1}\lambda_1^k. \tag{1.32}$$

Combining with (1.28) we see that $m_1 \geq 1$ and that for $\bar{u}_1$ the convergence speed to the invariant measure is given by λ_1^k. Because there is no initial distribution which yields a slower convergence according to (1.28) one also speaks of the **slowest process in the system**. To imagine what this means think of a very large number of trajectories starting distributed according to $\bar{u}_1$. Note that $\bar{u}_1$ has the same maxima and minima as the eigenvector u_1 itself, so we will have many trajectories starting close to the maxima of the eigenvector u_1 and almost no trajectories around the minima. In the process of equilibration, i.e. convergence to invariant measure, this has to become balanced. So the trajectories have to go from regions in state space, where the eigenvector u_1 is large to regions where u_1 is small. The larger the associated eigenvalue λ_1 is, the slower this balancing process can take place. Of course, we can apply the same argumentation to the other eigenvalues and associated eigenvectors and speak of **slow processes** for the largest eigenvalues, and **fast processes** for eigenvalues close to 0.

Continuous Processes

It is remarkable that things do not change a lot when we move from discrete time and state space to a continuous setting. Note that the transition matrix (1.11) defines for every fixed $k \in \mathbb{N}$ and $i \in E$ a probability distribution $p(k, i, \cdot)$ on E

$$p(k, i, \cdot) := P^k(i, \cdot) \tag{1.33}$$

with

$$p(0, i, j) = \delta_{ij} \qquad p(k + l, i, j) = \sum_{r \in E} p(k, i, r)p(l, r, j). \tag{1.34}$$

This motivates a continuous analog, which also allows to construct a time-continuous Markov process on $E = \mathbb{R}^n$ or another continuous state space that is complete and separable.

Proposition 2
*For every **transition kernel** or **transition function**, i.e. a function $p(t, x, A)$ on $\mathbb{R}^+ \times E \times \mathcal{E}$, with*

$$\begin{aligned} &p(t, x, \cdot) \textit{is a probability measure on } E, \textit{ for all } t \in \mathbb{R}^+, x \in E \\ &p(0, x, \cdot) = \delta_x \\ &p(\cdot, \cdot, A) \textit{ is measurable on } \mathbb{R}^+ \times E \textit{ for all } A \in \mathcal{E} \\ &p(t + s, x, A) = \int_E p(s, y, A)p(t, x, dy). \end{aligned} \tag{1.35}$$

and an initial probability measure ν on E, there exists a Markov process on E and its finite dimensional distributions are given by

$$\begin{aligned} &\mathbb{P}[X_0 \in A_0, X_{t_1} \in A_1, ..., X_{t_n} \in A_n] \\ &= \int_{A_0} \cdots \int_{A_{n-1}} p(t_n - t_{n-1}, y_{n-1}, A_n)p(t_{n-1} - t_{n-2}, y_{n-2}, dy_{n-1}) \cdots p(t_1, y_0, dy_1)\nu(dy_0). \end{aligned} \tag{1.36}$$

A proof can be found in [27], for example. The last equation in (1.35) is also known as the *Chapman-Kolmogorov equation.*

For the sake of simplicity we assume in the following that all introduced measures have a density with respect to the Lebesgue-measure and we abuse notation by calling for a measure ν also its density ν, i.e. we write

$$\nu(A) = \int_A \nu(x)dx. \tag{1.37}$$

From the context it should not be difficult to distinguish since we take ν with respect to a set $A \subset E$ or a certain state $x \in A$, respectively. We can

continue the derivation of the analog for the continuous setting by defining the linear operator

$$(\mathcal{P}_t\nu)(y) = \int_E p(t,x,y)\nu(x)dx. \tag{1.38}$$

If ν is the probability density which provides the initial distribution of X_0, $\mathcal{P}_t\nu$ will describe the distribution of X_t. Again, we assume that the Markov process (X_t) has a unique positive invariant measure μ, i.e. $\mathcal{P}_t\mu = \mu$ and define for every time t the **transfer operator** by

$$(T_t v)(y) = \frac{1}{\mu(y)} \int_E p(t,x,y)v(x)\mu(x)dx \qquad T_t : L^2(\mu) \to L^2(\mu), \tag{1.39}$$

where $L^2(\mu) = \{v : E \to \mathbb{R} | \int_E v(x)^2\mu(x)dx < \infty\}$ is the Hilbert space equipped with the scalar product

$$\langle v, w\rangle = \int_E v(x)w(x)\mu(x)dx. \tag{1.40}$$

(1.35) yields that the family $\{T_t\}_{t\in\mathbb{R}^+}$ forms a semi-group, i.e.

$$T_{t+s} = T_t T_s, \qquad T_0 = Id. \tag{1.41}$$

Generator. In the time-discrete case we needed only one transfer operator T to describe the whole semi-group of operators by $T_k = T^k$. We will now derive a similar feature for time-continuous Markov processes. We assume that the semi-group is strongly continuous, i.e. $\lim_{t\to 0} T_t v = v$. Then, we define the **infinitesimal generator** of the Markov process by

$$Lv = \lim_{t\to 0} \frac{T_t v - v}{t} \tag{1.42}$$

for all $v \in L^2(\mu)$, where the limit exists. One can show [27] that if the limit exists for a density v, it also converges for every density $v_t = T_t v, t > 0$ and we have

$$\frac{d}{dt}v_t = Lv_t = T_t Lv. \tag{1.43}$$

The solution to (1.43) is known to be

$$T_t v = v_t = e^{Lt}v. \tag{1.44}$$

This is why L is called *generator* because by (1.44) one can derive the whole semi-group of transfer operators from the generator.

Brownian Motion and Stochastic Integrals

In the last Section 1.2 we presented an example for a time-discrete Markov chain which was called a *random walk* on $\mathbb{Z}$. We first named the properties the process should have and afterwards constructed a suitable probability space. An important example for a continuous Markov process is **Brownian motion**. It is named after the English botanist Robert Brown, who studied the motion of pollen particles in a liquid. Later, this motion was described by Einstein as a stochastic process $(B_t)_{t\in\mathbb{R}^+}$ which should have the following properties:

- $B_0 = 0$ almost surely (the probability is one).
- the trajectory $t \to B_t$ is almost surely continuous.
- for $0 = t_0 < t_1 < ... < t_k$ the random variables $B_{t_i} - B_{t_{i-1}}$, which are called *increments*, are independent.
- the increments $B_{t_i} - B_{t_{i-1}}$ are normally distributed according to $N(0, (t_i - t_{i-1})Id)$.

It is anything but trivial to verify the existence of a probability space and a process (B_t), which have these properties, but it can be proven, for example, using Kolmogorov's extension theorem and continuity criterion. Details can be found in [65].

Of course, the description of the Brownian motion by the properties above is very abstract, so we are particularly interested in properties which hold for almost every trajectory of the Brownian motion to get a feeling how this process looks like. We know already that realizations of the Brownian motion are almost surely continuous. On the other hand, one can prove that one will never be able to draw such a trajectory because (B_t) is also nowhere differentiable and has infinitely many zeros in every interval $(0, \epsilon), \epsilon > 0$. That is, every path of Brownian motion is changing rapidly and therefore, it is also not of finite variation almost surely, i.e. for every regular sequence of subdivisions

$$\Delta_n = \{0 = t_0 < t_1 < ... < t_n = T\} \qquad |\Delta_n| := \max_{i=1,...,n} |t_i - t_{i-1}| \to 0 \quad (1.45)$$

of the interval $[0, T]$,$\Delta_n \subset \Delta_{n+1}$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{n} |B_{t_i} - B_{t_{i-1}}| = \infty. \quad (1.46)$$

Equation (1.46) causes also problems in the theory of integration and it was the starting point for the famous Ito-calculus. Usually, for a function $A_t : [0, T] \to \mathbb{R}$ with

$$\lim_{n\to\infty} \sum_{i=1}^{n} |A_{t_i} - A_{t_{i-1}}| < \infty, \quad (1.47)$$

i.e. for a function A_t that is of **finite variation** one can easily define an integral, for example, for $f \in C(\mathbb{R})$

$$\int_0^T f(A_s)dA_s = \lim_{n\to\infty} \sum_{i=1}^{n} f(A_{\xi_{i-1}})(A_{t_i} - A_{t_{i-1}}) < \infty \tag{1.48}$$

where the limit does not depend on the choice of $\xi_{i-1} \in [t_{i-1}, t_i]$. This is the usual **Riemann-Stieltjes-Integral**. Moreover, if F is an antiderivative of f, i.e. $F' = f$, and $f \in C^1(\mathbb{R})$, we can calculate the Taylor expansion

$$F(A_{t_i}) = F(A_{t_{i-1}}) + f(A_{t_{i-1}})(A_{t_i} - A_{t_{i-1}}) + \frac{1}{2}f'(A_{\xi_{i-1}})(A_{t_i} - A_{t_{i-1}})^2. \tag{1.49}$$

This yields for all $n \in \mathbb{N}$

$$\begin{aligned} &F(A_T) - F(A_0) = \sum_{i=1}^{n} F(A_{t_i}) - F(A_{t_{i-1}}) \\ &= \sum_{i=1}^{n} f(A_{t_{i-1}})(A_{t_i} - A_{t_{i-1}}) + \sum_{i=1}^{n} \frac{1}{2}f'(A_{\xi_{i-1}})(A_{t_i} - A_{t_{i-1}})^2. \end{aligned} \tag{1.50}$$

Taking the limit $n \to \infty$ we get the **fundamental theorem of calculus**

$$F(A_T) - F(A_0) = \int_0^T f(A_s)dA_s, \tag{1.51}$$

since

$$\begin{aligned} &\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{2}f'(A_{\xi_{i-1}})(A_{t_i} - A_{t_{i-1}})^2 \\ &\leq \frac{1}{2}\|f'\|_\infty \sum_{i=1}^{n} |A_{t_i} - A_{t_{i-1}}| \max_{i=1,\dots,n} |A_{t_i} - A_{t_{i-1}}| \\ &\leq C \max_{i=1,\dots,n} |A_{t_i} - A_{t_{i-1}}| \to 0, n \to \infty. \end{aligned} \tag{1.52}$$

This works for functions A_t of finite variation but trajectories of the Brownian motion, for example, does not have this property. On the other hand, assuming that at least the so called quadratic variation $\langle A \rangle_t$ of a function A_t exists

$$\langle A \rangle_t = \lim_{n\to\infty} \sum_{i=1}^{n} (A_{t_i} - A_{t_{i-1}})^2 < \infty, \tag{1.53}$$

we can calculate that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{2}f'(A_{\xi_{i-1}})(A_{t_i} - A_{t_{i-1}})^2 = \frac{1}{2}\int_0^T f'(A_s)d\langle A \rangle_s. \tag{1.54}$$

Because $\langle A\rangle_t$ is always an increasing function, it has to be of finite variation and therefore, the integral in (1.54) exists in the Riemann-Stieltjes sense as above independently of the points ξ_{i-1}. That is, we have generalized (1.51) to

$$F(A_T) - F(A_0) = \int_0^T f(A_s)dA_s + \frac{1}{2}\int_0^T f'(A_s)d\langle A\rangle_s. \tag{1.55}$$

(1.55) is known as **Ito's formula**, and since $\langle A\rangle_t \equiv 0$ for functions A of finite variation, it really extends (1.51) to the larger class of functions that have finite quadratic variation. Here, it is crucial that $f(A_{t_{i-1}})$ is evaluated at the left point of $[t_{i-1}, t_i]$ in the **Ito-Integral**

$$\int_0^T f(B_s)dB_s = \lim_{n\to\infty}\sum_{i=1}^{n} f(A_{t_{i-1}})(A_{t_i} - A_{t_{i-1}}). \tag{1.56}$$

The limit at the right hand side is not independent of the point of evaluation and other choices lead to different definitions of the integral [65].

Fortunately, the trajectories B_t of the Brownian motion have $\langle B\rangle_t = t$ almost surely such that (1.55) reads

$$F(B_T) - F(B_0) = \int_0^T f(B_s)dB_s + \frac{1}{2}\int_0^T f'(B_s)ds. \tag{1.57}$$

Such stochastic integrals especially appear in the context of stochastic differential equations [43, 4], which we will later consider as examples in the most simple form where $f \equiv \sigma$ is constant.

2

Approximation of Markov Processes

From the previous chapter we have seen that the probability measure on the path space E^I of stochastic processes can be a very difficult object for direct analysis. From a practical, numerical, or statistical point of view we can only get few incomplete realizations on path space, that is, we can only simulate the finite random variable Y. The question now is how to extract information about the Markov process from these realizations. In Section 2.1 we will introduce a well-known approach that is based on so called **Markov State Modeling**. In Section 2.2 we will connect this method to the general framework of **projected transfer operators**. We will also see other examples of methods that fit into this framework, particularly another sophisticated Markov State Modeling approach in Section 2.3. Chapter 3 will finally show how powerful the interpretation in the projected transfer operator sense is.

2.1 Standard Markov State Models

Assume that we want to analyze a time-continuous Markov process $(X_t)_{t\in\mathbb{R}}$ on a continuous or very large discrete state space E. Then, the idea of Markov State Modeling is to construct a Markov chain $(\hat{X}_k)_{k\in\mathbb{N}}$ that lives on a finite state space $\hat{E} = \{1, ..., n\}$ and that reproduces essential characteristics of the original Markov process $(X_t)_{t\in\mathbb{R}}$. In this sense the Markov chain $(\hat{X}_k)$ can be considered to be an approximation of the continuous process.
Obviously, there are two discretizations needed, a discretization of time and space. The time-discretization can be achieved very naturally because for every lag time $\tau > 0$ the time-discrete process $(X_{k\tau})_{k\in\mathbb{N}}$ is again a Markov process on state space E.
In Standard Markov State Modeling the construction of a finite state space $\hat{E}$ is based on a **full partitioning** of state space, i.e. sets $A_1, ..., A_n$ with

$$\bigcup_{j=1}^{n} A_j = E \qquad A_i \cap A_j = \emptyset \ \forall i \neq j, \tag{2.1}$$

with "nice" sets A_j (e.g. with Lipschitz boundary).
Then, we introduce the discrete process $(\tilde{X}_k)_{k\in\mathbb{N}}$ on the finite state space

$\hat{E} = \{1, ..., n\}$ by setting

$$\tilde{X}_k = i \Leftrightarrow X_{k\tau} \in A_i. \tag{2.2}$$

$(\tilde{X}_k)$ describes the snapshot dynamics of the continuous process (X_t) with lag time τ between the sets $A_1, ..., A_n$.
At first glance, the process $(\tilde{X}_k)$ seems to be a good candidate for our approximating chain, but the problem is that this process $(\tilde{X}_k)$ is generally not Markovian, i.e.

$$\mathbb{P}[\tilde{X}_{k+1} = j | \tilde{X}_k = i_k, \tilde{X}_{k-1} = i_{k-1}, ..., \tilde{X}_0 = i_0] \neq \mathbb{P}[\tilde{X}_{k+1} = j | \tilde{X}_k = i_k]. \tag{2.3}$$

As an illustration, why $(\tilde{X}_k)$ is not Markovian in general, we look at the following example.
We take the continuous Markov process that is given by the stochastic differential equation

$$dX_t = -\nabla V(X_t)dt + \sigma dB_t, \tag{2.4}$$

where B_t denotes Standard Brownian Motion and the potential V is shown in Fig. 2.1.

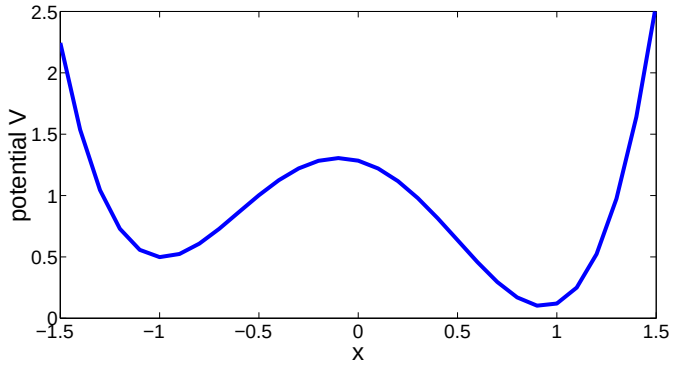


Figure 2.1: potential V

Equation 2.4 is just the abbreviation of

$$X_t = X_0 - \int_0^t \nabla V(X_t)dt + \sigma B_t. \tag{2.5}$$

Note that the deterministic part $X_0 - \int_0^t \nabla V(X_t)dt$ is the solution of the gradient flow $\dot{x} = -\nabla V(x)$. That is, the local minima will attract the process, but noise will disturb this behaviour and eventually lead to jumps between the wells.
The Fokker-Planck equation that governs the propagation of a function f

by the diffusion process is given by $\partial_t u = \mathcal{L}u$, $u(t=0,x) = f(x)$ and in the weighted Hilbert space $L^2(\mu)$ the generator reads

$$\mathcal{L} = -\nabla V(x) \cdot \nabla_x + \sigma^2/2\Delta_x, \tag{2.6}$$

where ∇_x denotes the first derivative wrt. x and Δ_x the associated Laplacian.
We now choose two sets A and B around the local minima that form a full partitioning of state space. If we ask whether our two state switching process between these sets $(\tilde{X}_k)$ has the Markov property, i.e. if (1.8) holds, we have to look at the effect of memory. Let us simply consider a one step memory and compare for a small lag time $\tau = 0.1$ the two probabilities

$$\mathbb{P}[X_{(k+1)\tau} \in A | X_{k\tau} \in B], \qquad \mathbb{P}[X_{(k+1)\tau} \in A | X_{k\tau} \in B, X_{(k-1)\tau} \in A].$$

Because $(X_{k\tau})$ is a Markov process, we can calculate

$$\mathbb{P}[X_{(k+1)\tau} \in A | X_{k\tau} \in B] = \int_{x \in B} (P_\tau v_B)(x) dx,$$
$$\mathbb{P}[X_{(k+1)\tau} \in A | X_{k\tau} \in B, X_{(k-1)\tau} \in A] = \int_{x \in B} (P_\tau v_{BA})(x) dx$$

where v_B is the probability density of $X_{k\tau}$ under the condition $X_{k\tau} \in B$, and v_{BA} is the probability density of $X_{k\tau}$ under the condition $X_{k\tau} \in B$ and $X_{(k-1)\tau} \in A$. These distributions are shown in Fig.2.2.

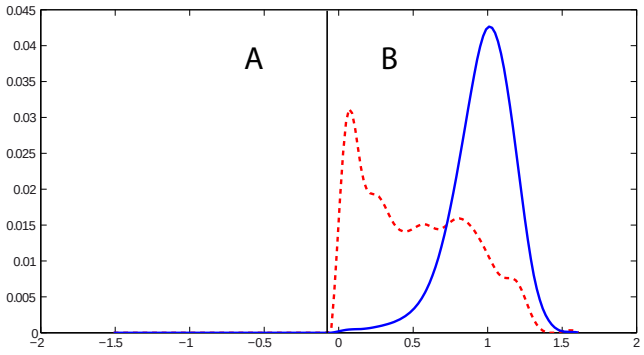


Figure 2.2: solid blue: v_B, dashed red: v_{BA}

If one thinks of an ensemble of trajectories, this picture indicates that the knowledge of $X_{(k-1)\tau} \in A$ implies that at time k most trajectories that arrived in set B are still close to set A because the lag time τ is not too large. That is, they are still inside of the transition region and not close enough to the minimum in set B as it is the case if we only have the knowledge

$X_{k\tau} \in B$. Therefore, we have much more trajectories that will go back to set A if they are distributed according to v_{BA} rather than to v_B. We see this effect in Fig. 2.3.

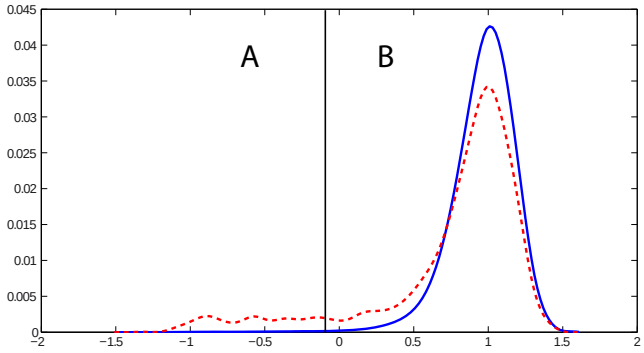


Figure 2.3: solid blue: $P_\tau v_B$, dashed red: $P_\tau v_{BA}$

This results in a difference of probability

$$\mathbb{P}[X_{(k+1)\tau} \in A | X_{k\tau} \in B] = 0.0049$$
$$\mathbb{P}[X_{(k+1)\tau} \in A | X_{k\tau} \in B, X_{(k-1)\tau} \in A] = 0.0994.$$

This issue is called **recrossing problem** because a transition back into set B is much more likely when we introduce memory as above. Therefore, the process $(\tilde{X}_k)$, which describes the switching dynamics between the partitioning sets, is not memoryless and hence no Markov process.

However, Markov State Models attempt to approximate this process via a discrete Markov chain $(\hat{X}_k)_{k\in\mathbb{N}}$ on $\hat{E} = \{1, ..., n\}$ defined by the transition matrix $\hat{P}$ with entries

$$\hat{P}(i,j) = \mathbb{P}[\tilde{X}_1 = j | \tilde{X}_0 = i] = \mathbb{P}[X_\tau \in A_j | X_0 \in A_i]. \tag{2.7}$$

One very essential feature is that one can estimate this matrix from a realization of the process (X_t). Assume that we have a trajectory of N datapoints that we call $x_k, k = 1, ..., N$ where x_k is the realization of the random variable $X_{k\tau}$. Then we can estimate

$$\hat{P}^*(i,j) = \frac{N(i,j)}{n(i)}, \tag{2.8}$$

where $n(i) = \#\{x_k = i\}$ is the number of time steps the process spent in A_i, and $N(i,j) = \#\{x_k = i, x_{k+1} = j\}$ is the number of time steps the process made a transition from set A_i to set A_j. One can even show that $\hat{P}^*$ is a maximum likelihood estimator for the transition matrix $\hat{P}$ and further analyze the statistical error $\|\hat{P} - \hat{P}^*\|$ [55, 66, 75, 48].

Reference. Markov state models have been considered for processes that have metastable dynamics [18, 73, 74, 7], especially in Molecular Dynamics. Recently the interest in MSMs has drastically increased since it could be demonstrated that MSMs can be constructed even for very high dimensional systems [73] and have been especially useful for modeling the interesting slow dynamics of biomolecules [56, 57, 58, 12, 11, 61] and materials [79] (there under the name "kinetic Monte Carlo"). Their approximation quality on large time scale has been rigorously studied, e.g., for Brownian or Glauber dynamics and Ising models in the limit of vanishing smallness parameters (noise intensity, temperature) where the analysis can be based on large deviation estimates and variational principles [29, 84] and/or potential theory and capacities [8, 9]. In these cases the effective dynamics is governed by some MSM with exponentially small transition probabilities and its states label the different attractors of the underlying, unperturbed dynamical systems. Other approaches tried to understand the multi-dimensional setting for complex dynamical systems by generalization of Kramer's approach, e.g., by discussing asymptotic expansions based on the Wentzel-Kramers-Brillouin approximation in semiclassical quantum dynamics, matched asymptotics or similar techniques, see e.g. [46, 62]. Another rigorous approach to the construction of MSM involves the exploitation of spectral properties. The relation between dominant eigenvalues, exit times and rates, and metastable sets has been studied by asymptotic expansions in certain smallness parameters as well as by functional analytic means without any relation to smallness parameters [37, 14, 74, 7, 18]. In real applications with high-dimensional state spaces asymptotic expansions are based on assumptions that typically cannot be checked and often enough are not satisfied, involve quantities that cannot be computed, and/or are rather specific for a certain class of processes. Even if a smallness parameter can be defined we typically cannot check whether we are in the asymptotic regime such that the theoretical results cannot be used for error estimates.

We will follow later the functional analytic approach found in [72, 14, 74] and use the framework of projected transfer operators to answer in Chapter 3 the following questions.

- The switching process between the partitioning sets from (2.2) is not Markovian. Nevertheless, the Markov State Model considers a Markov chain with the transition matrix (2.7). Under which assumptions is this a valid approximation?

- Does this Markov chain reproduce the long-time behaviour of the original process, that is, how well are the dominant eigenvalues of the transfer operator T_τ approximated by the eigenvalues of $\hat{P}$?

2.2 Projected Transfer Operators

Markov processes have the useful property that the probability measure on path space can be generated by a transition matrix in discrete time or a transition kernel for a time-continuous process plus an initial distribution. On the other hand, this transition matrix or transition kernel also defines a semi-group of transfer operators. Many interesting properties of the Markov process can be derived from properties of these linear, bounded operators, particularly in the case of a reversible dynamics (see Sec. 1.3). Moreover, the analysis of linear operators that act on vector spaces seems to be much more feasible than the analysis of measures on abstract probability spaces which are as complicated as the path space E^I. Nevertheless, in practice a transfer operator $T := T_\tau$ for some lag time $\tau > 0$ is often still unavailable. Here, unavailable can have two different meanings. First, we could not have any analytical expression for T because we just do not have any equation for the dynamics of the Markov process. For example, the process (X_t) could be some natural experiment or even a computer experiment, where we cannot formulate an equation for. Second, we could be able to write down some expression for T, like in the case of some stochastic differential equations with $T = e^{L\tau}$, where L is a differential operator, but it could be still impossible to analytically or efficiently numerically compute the interesting quantities of T, e.g. the eigenvalues. As T describes the evolution of probability densities under the dynamics of the Markov process, it is an operator on an infinite dimensional function space. That is, for a numerical treatment it is immediately clear that some sort of finite approximation of T is needed. The most natural way to achieve this is via restriction and bestapproximation.

Definition 8 (Projected Transfer Operator) *Let T be a transfer operator of a Markov process on state space E with unique invariant measure μ. Then, we call any operator of the form*

$$QTQ : D \to D$$

projected transfer operator *if Q is the orthogonal projection in $L^2(\mu)$ onto some subspace D with*

$$D \subset L^2(\mu) \qquad \mathbb{1} \in D.$$

This means that we are only looking at the propagation of densities $v \in D$ by the transfer operator T, so we restrict its domain. Because for $v \in D$ $Tv = TQv$ will not belong to D, in general, we have to approximate the result of propagation by another density in D. We simply choose the bestapproximation of TQv, that is $QTQv$. The constraint $\mathbb{1} \in D$ ensures that the density of the invariant measure is part of the domain of the projected operator and that $QTQ\mathbb{1} = \mathbb{1}$ as well.

In Def. 8 we did not claim that D has to be finite dimensional. For the purpose of a finite approximation that we also call discretization of T, as mentioned above, we will make this additional assumption. On the other hand, we will see later that the general concept of projected transfer operators also connects to well known methods that do not reduce to finite dimensional subspaces.

Note that for the semi-group of transfer operators (T_t) the family of associated projected transfer operators (QT_tQ) does not form a semi-group anymore. That is, they do not describe the ensemble dynamics of a Markov process. We have observed such a Markovianity problem with approximations of Markov processes already in the last Section 2.1. There, the switching process $(\tilde{X}_t)$ between the partitioning subsets $A_1, ..., A_n$ of state space was not Markovian. Actually, this is no coincidence.

Theorem 1
The MSM transition matrix $\hat{P}$ from (2.7) is a matrix representation[1] *of the projected transfer operator QTQ for*

$$D = \{\mathbb{1}_{A_1}, ..., \mathbb{1}_{A_n}\}, \tag{2.9}$$

i.e. the space of stepfunctions, which are piecewise constant on the partitioning sets $A_1, ..., A_n$.

Proof. Let us take the basis $(\psi_i)_{i=1,...,n}$ of probability densities given by

$$\psi_i = \frac{\mathbb{1}_{A_i}}{\mu(A_i)}. \tag{2.10}$$

We can write the orthogonal projection Q as

$$Qv = \sum_{j=1}^{n} \frac{\langle v, \mathbb{1}_{A_j}\rangle}{\mu(A_j)} \mathbb{1}_{A_j}. \tag{2.11}$$

By using the definition of T we get

$$\begin{aligned} QTQ\psi_i = QT\psi_i &= \sum_{j=1}^{n} \frac{\langle T\psi_i, \mathbb{1}_{A_j}\rangle}{\mu(A_j)} \mathbb{1}_{A_j} = \sum_{j=1}^{n} \frac{\langle T\mathbb{1}_{A_i}, \mathbb{1}_{A_j}\rangle}{\mu(A_i)} \psi_j \\ &= \sum_{j=1}^{n} \Big(\frac{1}{\mu(A_i)} \int_{A_i} \mathbb{P}[X_\tau \in A_j | X_0 = x] \, \mu(dx) \Big) \cdot \psi_j. \end{aligned} \tag{2.12}$$

[1] Here and in the following, matrix representation means with respect to multiplying the matrix from the right, so if v has the row vector representation $\hat{v}$, $QTQv$ has the representation $\hat{v}\hat{P}$.

That is,

$$\begin{aligned} QTQ\psi_i &= \sum_{j=1}^{n} \mathbb{P}[X_\tau \in A_j | X_0 \in A_i] \cdot \psi_j \\ &= \sum_{j=1}^{n} \hat{P}(i,j)\psi_j. \end{aligned} \tag{2.13}$$

Therefore, $\hat{P}$ is a matrix representation of QTQ with respect to the basis (2.10). □

Theorem 1 shows that through the linear operator glasses the discrete approximation by Markov State Models is a special case of a projected transfer operator approximation. The advantage of this point of view is that in Chapter 3 several statements about approximation properties of projected transfer operators will be made. These theorems will have direct consequences for the quality of the corresponding approximation method. For example, the question if the Markov chain of a Standard Markov State Model captures the correct dominant timescales can now be investigated by comparing the dominant eigenvalues of T and QTQ with D from (2.9). We will answer this question in Sec. 3.2. Another question was if the approximation of the non-Markovian switching process $(\tilde{X}_{k\tau})$ by a Markov chain $(\hat{X}_k)$ is reasonable. Now, Theorem 1 showed that the propagation of the distributions of these processes is described by the operators QT^kQ and $(QTQ)^k$, respectively. We will analyze the error $\|QT^kQ - (QTQ)^k\|$ in Sec. 3.1, which gives the maximal possible error between the distributions of $(\tilde{X}_{k\tau})$ and $(\hat{X}_k)$ for any initial distribution.

We will see now and in the next section how other methods can be translated into the framework of projected transfer operators.

Averaging Methods

One example for methods that correspond to projections onto infinite dimensional subspaces D are averaging methods. Let us consider for a lag time $\tau > 0$ the time-discretized Markov process $(X_{k\tau})_{k\in\mathbb{N}}$ with transition kernel $p(x, y, \eta, \xi)$ on a state space $E = E_x \times E_y \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Its associated transfer operator can be written as

$$Tf(\eta,\xi) = \frac{1}{\mu(\eta,\xi)} \int_E p(x,y,\eta,\xi) f(x,y) \mu(x,y) d(x,y). \tag{2.14}$$

Now, take the orthogonal projection Q with respect to $L^2(\mu)$ onto the infinite dimensional subspace

$$D = \{f \in L^2(\mu) : \quad f(x,y) \text{ does not depend on } y\}. \tag{2.15}$$

Then, the orthogonal projection Q can be expressed as

$$Qf(x,y) = \frac{1}{\bar{\mu}(x)} \int\limits_{E_y} f(x,\xi)\mu(x,\xi)d\xi, \quad \bar{\mu}(x) = \int\limits_{E_y} \mu(x,\xi)d\xi. \tag{2.16}$$

Indeed, one can calculate that for any $v \in D$, i.e. $v(x,y) = v(x,\xi)$ for any y, ξ, we have

$$\begin{aligned}
\langle Qf, v\rangle &= \int\limits_{E_x}\int\limits_{E_y} (Qf)(x,y)v(x,y)\mu(x,y)dydx \\
&= \int\limits_{E_x}\int\limits_{E_y} \frac{1}{\bar{\mu}(x)} \int\limits_{E_y} f(x,\xi)\mu(x,\xi)d\xi v(x,y)\mu(x,y)dydx \\
&= \int\limits_{E_x}\int\limits_{E_y} \frac{1}{\bar{\mu}(x)} \int\limits_{E_y} f(x,\xi)v(x,\xi)\mu(x,\xi)d\xi\mu(x,y)dydx \\
&= \int\limits_{E_x}\int\limits_{E_y} f(x,\xi)v(x,\xi)\mu(x,\xi)d\xi \frac{1}{\bar{\mu}(x)} \int\limits_{E_y} \mu(x,y)dydx \\
&= \int\limits_{E_x}\int\limits_{E_y} f(x,\xi)v(x,\xi)\mu(x,\xi)d\xi dx = \langle f, v\rangle.
\end{aligned} \tag{2.17}$$

Because $Qf \in D$ does not depend on y,

$$\bar{f}(x) := Qf(x,y) \tag{2.18}$$

is well defined. We can now investigate a Markov process $(\bar{X}_k)$ on state space E_x that comes from averaging over the transition kernel, namely we set

$$\bar{p}(x,\eta) = \frac{1}{\bar{\mu}(x)} \int\limits_{E_y}\int\limits_{E_y} p(x,y,\eta,\xi)\mu(x,y)dyd\xi. \tag{2.19}$$

Then, $\bar{p}$ is a transition function on E_x since it obviously inherits non-negativity and

$$\begin{aligned}
\int\limits_{E_x} \bar{p}(x,\eta)d\eta &= \frac{1}{\bar{\mu}(x)} \int\limits_{E_y}\int\limits_{E} p(x,y,\eta,\xi)d(\eta,\xi)\mu(x,y)dy \\
&= \frac{1}{\bar{\mu}(x)} \int\limits_{E_y} \mu(x,y)dy = 1.
\end{aligned} \tag{2.20}$$

Moreover, its invariant measure is given by $\bar{\mu}$ since

$$\begin{aligned}
\int\limits_{E_x} \bar{\mu}(x)\bar{p}(x,\eta)dx &= \int\limits_{E_x}\int\limits_{E_y}\int\limits_{E_y} p(x,y,\eta,\xi)\mu(x,y)dyd\xi dx \\
&= \int\limits_{E_y} \mu(\eta,\xi)d\xi = \bar{\mu}(\eta).
\end{aligned} \tag{2.21}$$

For example, such averaged transfer operators are considered in the context of so-called Hybrid Monte Carlo methods, see [74, 73]. Using (2.18) the transfer operator for the averaged Markov process $(\bar{X}_k)$ can be interpreted as the projected transfer operator QTQ because

$$\begin{aligned}
&\frac{1}{\bar{\mu}(\eta)} \int\limits_{E_x} \bar{p}(x,\eta)\bar{f}(x)\bar{\mu}(x)dx \\
&= \frac{1}{\bar{\mu}(\eta)} \int\limits_{E_x}\int\limits_{E_y}\int\limits_{E_y} p(x,y,\eta,\xi)\bar{f}(x)\mu(x,y)dyd\xi dx \\
&= \frac{1}{\bar{\mu}(\eta)} \int\limits_{E_y}\int\limits_{E} p(x,y,\eta,\xi)(Qf)(x,y)\mu(x,y)d(x,y)d\xi \\
&= \frac{1}{\bar{\mu}(\eta)} \int\limits_{E_y} (TQf)(\eta,\xi)\mu(\eta,\xi)d\xi = (QTQf)(\eta,\xi')
\end{aligned} \tag{2.22}$$

for arbitrary $\xi' \in E_y$.
This means that we will be able to apply the framework of projected transfer operators to analyze the Markov process on state space E_x which comes from averaging of the transition kernel of the original Markov process (X_t). One immediate consequence is the reversibility of the averaged process $(\bar{X}_k)$ for a reversible process (X_t) because QTQ is obviously self-adjoint if T is.
Before we start to use the projected transfer operator theory for analysis, we derive in the next section another finite dimensional operator that will turn out to be equivalent to a powerful Markov State Modeling technique which is based on a milestoning approach.

2.3 Core Set Approach and Milestoning

In this section we introduce another approach for the discretization of transfer operators. That is, we define another finite dimensional subspace D and its associated projected transfer operator QTQ. In Sec. 2.1 it has been presented how Standard Markov State Models relate to the subspace of stepfunctions which are constant on the MSM partitioning $A_1, ..., A_n$. The projected transfer operator could then be characterized by the transition matrix of the jump process between these sets. Therefore, it is possible to estimate a matrix representation of QTQ from simulations or experiments. We have to keep in mind that this is an essential feature of the subspace D and the projected transfer operator for practical applications. So we will follow the idea of a set oriented discretization of state space and a construction of a suitable subspace D, but now we choose sets $C_1, ..., C_n \subset E$

$$C_i \cap C_j = \emptyset, i \neq j,$$

which do not have to be a full partition of state space anymore and which we call **core sets** in the following. We also define the region that is not assigned to any core set

$$C = E \setminus \bigcup_{k=1}^{n} C_k. \tag{2.23}$$

The next step in the construction of Standard Markov State Models has been the definition of the switching process $(\tilde{X}_t)$ in (2.2). Obviously, this process is not well defined anymore if our sets do not form a full partitioning. Hence, we slightly adjust the definition to the new setting and consider the process

$$\tilde{X}_t = i \Leftrightarrow X_{\sigma(t)} \in C_i, \text{ with } \sigma(t) = \sup_{s \leq t} \left\{ X_s \in \bigcup_{k=1}^{n} C_k \right\}, \tag{2.24}$$

which is called **milestoning process**, cf. [28]. Equation (2.24) means that the process $(\tilde{X}_t)$ stays in state i as long as the last core set that was visited was C_i. Its transition behaviour is illustrated in Fig. 2.4.

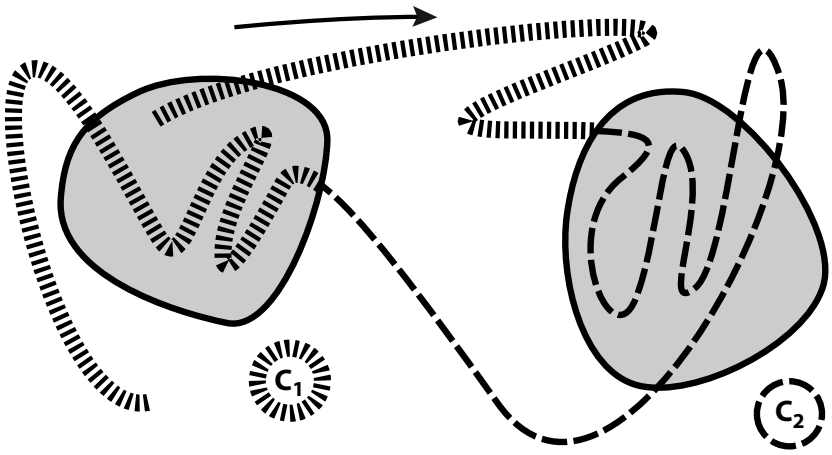


Figure 2.4: Illustration of milestoning process

Now we can start with the construction of a subspace D, of a projected transfer operator, and of a matrix representation that we can estimate from simulations. The key object for this construction is the committor. Assume that the Markov process is in state x at some point in time t, i.e. $X_t = x \in E$. Now, we are interested in the probability that the next core set that the process X_t will enter will be a particular set C_i. We denote this probability with $q_i^+(x)$. Moreover, we could ask for the last core set the process came from. We denote the probability that is was set C_i by $q_i^-(x)$. These two objects, q_i^+ and q_i^-, are called **forward** and **backward committor**, respectively. There are some important facts about the committors

that can be found in [49], for example. First, for a reversible Markov process we have $q_i^+ = q_i^-$. This seems to be clear since reversibility means that the process running backward in time is not distinguishable from the process running forward in time, and since the backward committor is nothing else than the forward committor for this reversed process. We will assume from now on that the Markov process (X_t) is reversible if we do not explicitly state differently and therefore, we will write only $q_i := q_i^+ = q_i^-$ and simply talk about **the committor**.
The next, very useful property of the committor is that it solves a linear system with boundary conditions. Here we have to distinguish between continuous and discrete time. For a time-continuous process $(X_t)_{t\in\mathbb{R}^+}$ the committor solves the equation

$$\begin{aligned} (Lq_i)(x) &= 0, \quad \forall x \in C, \\ q_i(x) &= 1, \quad \forall x \in C_i, \\ q_i(x) &= 0, \quad \forall x \in C_j, j \neq i, \end{aligned} \tag{2.25}$$

where L denotes the generator of the Markov process (X_t). For a discrete $(X_n)_{n\in\mathbb{N}}$ with transfer operator T, q_i fulfills

$$\begin{aligned} ((T - Id)q_i)(x) &= 0, \quad \forall x \in C, \\ q_i(x) &= 1, \quad \forall x \in C_i, \\ q_i(x) &= 0, \quad \forall x \in C_j, j \neq i. \end{aligned} \tag{2.26}$$

Sometimes one also calls $T - Id$ the **discrete generator** of (X_n).
On the other hand, it is also interesting if the equations (2.25) and (2.26) define the committor, i.e. if they are uniquely solvable.

Theorem 2 (Solvability of committor equations)
If the Markov process has a unique invariant measure which is not vanishing on all core sets, then the equations (2.25) and (2.26), respectively, have a unique solution.

Proof. We want to show that

$$\begin{aligned} (Aq_i)(x) &= 0, \quad \forall x \in C, \\ q_i(x) &= 1, \quad \forall x \in C_i, \\ q_i(x) &= 0, \quad \forall x \in C_j, j \neq i \end{aligned} \tag{2.27}$$

has a unique solution if $A = L$ or $A = T - Id$. First, (2.27) is solvable if and only if

$$\Theta A \Theta q_i = -\Theta A \Theta^{\perp} qi = -\Theta A \mathbb{1}_{C_i} \tag{2.28}$$

is solvable, where Θ is the orthogonal projection defined by

$$(\Theta v)(x) = \begin{cases} v(x), & x \in C \\ 0, & \text{else.} \end{cases} \tag{2.29}$$

Now, Fredholm Alternative states that (2.28) and therefore, (2.27) has a unique solution or

$$\Theta A \Theta \tilde{v} = 0 \tag{2.30}$$

has a non-trivial solution, which would imply for $v = \Theta\tilde{v}$

$$\begin{aligned} \Theta A v = 0 \text{ on } C, \\ v = 0, \text{ on } E \setminus C. \end{aligned} \tag{2.31}$$

Now, for $A = T - Id$ this means that $Tv = v$ on C and $v = 0$ on $E \setminus C$. This yields that also $Tv = 0 = v$ on $E \setminus C$ because otherwise we would have $\|Tv\| > \|v\|$. That is, $Tv = v$ on the whole state space E, which tells us that v would be an invariant measure of the process that is vanishing on all core sets. This is a contradiction to the assumption.
For $A = L$ (2.31) can be written as

$$\begin{aligned} \lim_{t\to 0} \frac{\Theta(Id - T_t v)}{t} = 0 \text{ on } C, \\ v = 0, \text{ on } E \setminus C, \end{aligned} \tag{2.32}$$

where $T_t = e^{Lt}$. Now, $Lv \neq 0$ on $E \setminus C$ and (2.32) would imply that there exists $\epsilon > 0$ with

$$\lim_{t\to 0} \left\| \frac{\Theta^{\perp} T_t v}{t} \right\| = \lim_{t\to 0} \left\| \frac{\Theta^{\perp}(Id - T_t v)}{t} \right\| \geq 3\epsilon > 0. \tag{2.33}$$

Because $\left\| \frac{\Theta(Id - T_t v)}{t} \right\| \geq \left\| \frac{\Theta v}{t} \right\| - \left\| \frac{\Theta T_t v}{t} \right\|$ there will be a time $t > 0$, such that

$$\left\| \frac{\Theta v}{t} \right\| - \left\| \frac{\Theta T_t v}{t} \right\| < \epsilon. \tag{2.34}$$

and

$$\left\| \frac{\Theta^{\perp} T_t v}{t} \right\| - \left\| \frac{\Theta^{\perp} v}{t} \right\| = \left\| \frac{\Theta^{\perp} T_t v}{t} \right\| > 2\epsilon. \tag{2.35}$$

The last two equations combine to

$$\|T_t v\| - \|v\| > \epsilon t > 0, \tag{2.36}$$

which cannot be true. That is, $Lv = 0$ on $E \setminus C$ as well, which again implies that v would be an invariant measure of the Markov process vanishing on all core sets. Because this is not allowed by assumption, (2.25) has to have a unique solution. □

It is important to note one thing. We said that the committors solve (2.26) for a time-discrete Markov process and (2.25) for a time-continuous (X_t). On the other hand, for any lag time $\tau > 0$ we could also consider the discretized process $(X_{k\tau})_{k\in\mathbb{N}}$. The associated committors would then be given by equation (2.26) with $T = T_\tau$ instead of (2.25). The difference in interpretation of these committors, which depend on the lag time τ, is the following. A committor q_i always provides the probabilities of hitting a certain core set C_i next rather than the others, but the time-continuous committors and committors for a certain lag time τ have a different resolution of time when it comes to recognizing these hits. The committors from (2.25) can use the whole continuous trajectory $(X_t)_{t\in\mathbb{R}^+}$ to decide whether a core set was hit or not. On the contrary, the committors from (2.26) see only points at discrete time steps of the trajectory, i.e. the process $(X_{k\tau})_{k\in\mathbb{N}}$, and they can not tell if the trajectory connecting these points has hit a core set. They can just decide about reaching core sets by recognizing points inside of core sets at time steps $k\tau$. This behaviour is shown in Figure 2.5.

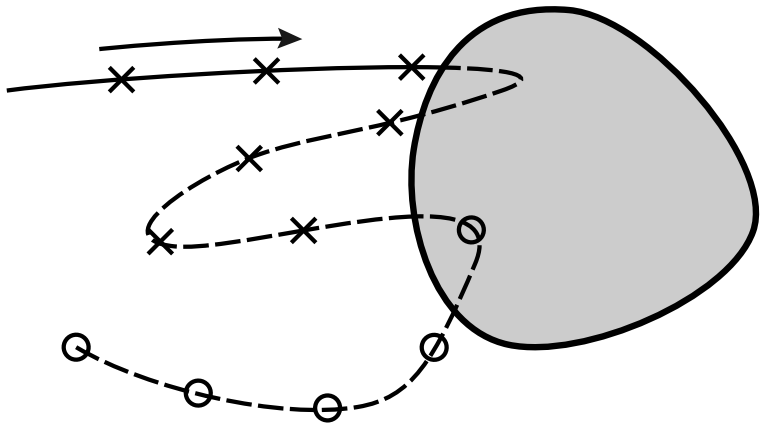


Figure 2.5: Different time-resolution for decision about hitting core sets

This is important because in practical applications we often cannot get a continuous realization of a trajectory of the process (X_t). Therefore, we are only able to talk about hitting core sets at some time-resolution. At this point we will not stress this fact. For a continuous Markov process and n core sets at hand we will assume that the associated committors $q_1, ..., q_n$ solve the continuous equation (2.25) but keep in mind that in praxis there can also be a time-resolution involved.

Now, these committors are obviously linear independent because of the

boundary conditions and we have

$$\sum_{i=1}^{n} q_i = \mathbb{1}. \tag{2.37}$$

Therefore, the n-dimensional space spanned by the committors

$$D = \{q_1, ..., q_n\} \tag{2.38}$$

fulfills $\mathbb{1} \in D$ and is an allowed space for the construction of a projected transfer operator QTQ.
We will now try to understand the nature of QTQ in the case that Q is the orthogonal projection onto our freshly created committor space D from (2.38). In [21, 20] it was possible to show that one can at least compute the eigenvalues of the operator QTQ from a generalized eigenvalue problem, which involves two matrices $\hat{T}$ and M. It was also stated that these matrices have a stochastic interpretation, which allows again statistical estimation. Nevertheless, for a time-continuous process the interpretation in [21, 20] is only valid if we look at the discretized process. This means that the lag time of the transfer operator and the time-resolution for the committors (see discussion above) have to be identical, that is, we can only consider $QT_\tau Q$, where also the space D is spanned by the committors coming from equation (2.26) with $T = T_\tau$. For continuous committors, i.e. an infinitesimally small time-resolution, this is not satisfying because

$$QT_\tau Q \to Id \text{ on } D, \quad \tau \to 0. \tag{2.39}$$

Hence, we will now get rid of this restriction.

Theorem 3 (Matrix representation of committor operator)
Let T be a transfer operator of a Markov process, Q the orthogonal projection onto the space spanned by committors $D = \{q_1, ..., q_n\}$ with respect to some core sets $C_1, ..., C_n$. Then,

$$\hat{T}M^{-1} \qquad \hat{T}_{ij} = \frac{\langle Tq_i, q_j \rangle}{\hat{\mu}(i)} \quad M_{ij} = \frac{\langle q_i, q_j \rangle}{\hat{\mu}(i)} \tag{2.40}$$

with $\hat{\mu}(i) = \sum_{x \in E} q_i(x)\mu(x)$ is a matrix representation of QTQ.

Proof. Since the vectors q_i are linear independent the symmetric matrix

$$S_{ij} = \langle q_i, q_j \rangle \tag{2.41}$$

is invertible and we can write the orthogonal projection Q onto subspace D as

$$Qv = \sum_{i,j=1}^{n} S_{ij}^{-1} \langle v, q_i \rangle q_j. \tag{2.42}$$

For the matrix M from (2.40) we have

$$M_{ij} = \frac{1}{\hat{\mu}(i)} S_{ij} \Rightarrow M_{ij}^{-1} = \hat{\mu}(j) S_{ij}^{-1}. \tag{2.43}$$

Now, take the basis $\{\psi_1, ..., \psi_n\}$ of D, $\psi_i = \frac{1}{\hat{\mu}(i)} q_i$. Then,

$$Qv = \sum_{i,j=1}^{n} M_{ij}^{-1} \langle v, q_i \rangle \psi_j. \tag{2.44}$$

This implies

$$\begin{aligned} QTQ\psi_k = QT\psi_k &= \sum_{i,j=1}^{n} M_{ij}^{-1} \langle T\psi_k, q_i \rangle \psi_j \\ &= \sum_{i,j=1}^{n} M_{ij}^{-1} \frac{\langle Tq_k, q_i \rangle}{\hat{\mu}(k)} \psi_j = \sum_{i,j=1}^{n} M_{ij}^{-1} \hat{T}_{ki} \psi_j \\ &= \sum_{j=1}^{n} (\hat{T} M^{-1})_{kj} \psi_j. \end{aligned} \tag{2.45}$$

That is, $\hat{T}M^{-1}$ is a matrix representation of QTQ with respect to the basis $\{\psi_1, ..., \psi_n\}$. □

Note that we did not use the fact that D is spanned by committors, so Theorem 3 is also valid for any other subspace D which is spanned by a basis $\{q_1, ..., q_n\}$. Now, the interesting question is if this matrix representation has for the committor space a stochastic interpretation that provides insight into the nature of QTQ.

Theorem 4
Let $T := T_\tau$ be a transfer operator of a continuous Markov process for some lag time $\tau > 0$. Let Q be the orthogonal projection onto the space spanned by the continuous committors $D = span\{q_1, ..., q_n\}$ solving (2.25) with respect to some core sets $C_1, ..., C_n$. Let $\hat{T}$ and M be given as in (2.40). Define for every t the stopping time

$$\sigma_t(A) = \inf_{s \geq t} \{X_s \in A\}. \tag{2.46}$$

Then, with $B_j = \bigcup_{k \neq j} C_k$ *we have*

$$M_{ij} = \mathbb{P}[\sigma_t(C_j) < \sigma_t(B_j) | \hat{X}_t = i], \tag{2.47}$$

and

$$\hat{T}_{ij} = \mathbb{P}[\sigma_{t+\tau}(C_j) < \sigma_{t+\tau}(B_j) | \hat{X}_t = i]. \tag{2.48}$$

Proof. Bayes Theorem states that for

$$\hat{q}_i(x) = \frac{q_i(x)\mu(x)}{\hat{\mu}(i)} \tag{2.49}$$

we have

$$\mathbb{P}[X_t \in A | \hat{X}_t = i] = \int_A \hat{q}_i(x)dx, \tag{2.50}$$

so $\hat{q}_i$ is the probability density of X_t under the condition that $\hat{X}_t = i$. Since (X_t) is Markovian and $\{\hat{X}_t = i\} \in \mathcal{F}_t$, the law of total probability yields

$$\begin{aligned} \mathbb{P}[\sigma_t(C_j) < \sigma_t(B_j) | \hat{X}_t = i] &= \int_E \mathbb{P}[\sigma_t(C_j) < \sigma_t(B_j) | X_t = x] \hat{q}_i(x)dx \\ \int_E q_j(x)\hat{q}_i(x)dx &= \int_E q_j(x)\frac{q_i(x)\mu(x)}{\hat{\mu}(i)}dx = M_{ij}. \end{aligned} \tag{2.51}$$

For the interpretation of $\hat{T}$ we note first that

$$\begin{aligned} &\mathbb{P}[\sigma_{t+\tau}(C_j) < \sigma_{t+\tau}(B_j) | X_t = x] \\ &= \int_E \mathbb{P}[\sigma_{t+\tau}(C_j) < \sigma_{t+\tau}(B_j) | X_{t+\tau} = y, X_t = x] \mathbb{P}[X_{t+\tau} \in dy | X_t = x] \\ &= \int_E q_j(y) \mathbb{P}[X_{t+\tau} \in dy | X_t = x]. \end{aligned} \tag{2.52}$$

This implies

$$\begin{aligned} &\mathbb{P}[\sigma_{t+\tau}(C_j) < \sigma_{t+\tau}(B_j) | \hat{X}_t = i] \\ &= \int_E \mathbb{P}[\sigma_{t+\tau}(C_j) < \sigma_{t+\tau}(B_j) | X_t = x] \hat{q}_i(x)dx \\ &= \int_E \int_E q_j(y) \mathbb{P}[X_{t+\tau} \in dy | X_t = x] \hat{q}_i(x)dx \\ &= \int_E \frac{q_j(y)(T_\tau q_i)(y)}{\hat{\mu}(i)} \mu(y)dy = \hat{T}_{ij}. \end{aligned} \tag{2.53}$$

□

Remark 4 *One could also use committors with respect to some time resolution h instead of continuous committors. Then the interpretation of the stopping times (2.46) will change such that it has to be decided from the discrete trajectory with time steps h if a set A has been hit or not. Of course, it is possible to choose h different from τ, i.e. smaller.*

This means that if we have the information that the milestoning process is in state i at some time t, i.e. we know that the last core set that was visited was C_i, M_{ij} gives the conditional probability that C_j will be the next core set that will be hit.
On the other hand, T_{ij} also gives the probability that C_j will be the next core set that will be reached, but with the additional rule that we do not count hits in the time interval $[t, t+\tau]$. This also gives an interpretation of the mapping $QTQ : D \to D$ for a certain class of probability densities in D. The space of probability densities in D is given by

$$D_1 = \{v \in D | v(x) \geq 0 \ \forall x \in E, \int_E v d\mu = 1\}. \tag{2.54}$$

Introducing probability vectors $S = \{r \in \mathbb{R}^n | r_i \geq 0 \ \forall i = 1, ..., n, \sum_{i=1}^{n} r_i = 1\}$ and the basis $\{\psi_i = \frac{q_i}{\hat{\mu}(i)}, i = 1, ..., n\}$ of D we can also write

$$D_1 = \{v \in D | v = \sum_{i=1}^{n} r_i \psi_i, r \in S\}. \tag{2.55}$$

From Theorem 3 it follows immediately

Corollary 1
If $v \in D_1$ has the property

$$v = \sum_{i=1}^{n} a_i \psi_i, a \in S \quad \exists b \in S : \ M^T b = \hat{T}^T a, \tag{2.56}$$

then $QTQv \in D_1$, i.e. it is a probability density again. Moreover, it has the representation

$$QTQv = \sum_{i=1}^{n} b_i \psi_i \tag{2.57}$$

with b from (2.56).

Proof.

$$QTQv = \sum_{i=1}^{n} a_i(QTQ\psi_i) = \sum_{i,j=1}^{n} a_i(\hat{T}M^{-1})_{ij}\psi_j$$
$$= \sum_{j=1}^{n} \psi_j(M^{-T}\hat{T}^T a)_j = \sum_{j=1}^{n} \psi_j b_j.$$

□

Corollary 1 answers the question about the interpretation of the projected transfer operator QTQ. This question is very natural because when we moved from the probability space Ω, the path space E^I, and its complicated probability measure to linear transfer operators, at first sight we entered a functional analytic framework that does not necessarily need to be connected to stochastics anymore. We consider the transfer operator T, which is an operator on $L^2(\mu)$, for example. That is, it can be applied to any function $v \in L^2(\mu)$, and the outcome Tv will be some function in $L^2(\mu)$ again. Neither the function v nor Tv need to have a stochastic meaning, but for a certain class of densities in $L^2(\mu)$, namely probability densities, they do. Now, the same is true for our projected transfer operator QTQ. Corollary 1 means that we take two probability vectors a and b such that the probability to hit any core set C_j next under the initial condition $\mathbb{P}[\hat{X}_0 = i] = b_i$ is the same as the probability to hit core set C_j next under the condition $\mathbb{P}[\hat{X}_0 = i] = a_i$ with the additional rule that hits in the time interval $[0, \tau]$ do not count. Then, a and b are representations for v and QTQ, respectively.

Remark 5 *From the definition of $\hat{T}$ and M it follows immediately that both matrices are stochastic. This implies that the matrix $\hat{T}M^{-1}$, which is the matrix representation of QTQ, is at least pseudostochastic, i.e. its rows sum to one. The non-negativity of its entries depends on the choice of core sets and is not guaranteed, in general, but we will see in examples in the last Section 3.3 that at least for some choices of core sets, $\hat{T}M^{-1}$ will form a fully stochastic matrix.*

Estimation from trajectories. Assuming that the Markov process is ergodic, the stochastic interpretation allows to estimate the matrices $\hat{T}$ and M from trajectories. Assume that we have a realization of the process (X_t) at some time resolution h that provides N datapoints x_k, i.e. x_k is a realization of the random variable X_{kh}. According to Theorem 4 we can approximate

$$M^*_{ij} = \frac{R_{ij}}{r_i}, \tag{2.58}$$

where r_i denotes the number of time steps where the last core set that was hit was C_i, and R_{ij} denotes the number of time steps where the process

came last from C_i and the next core set that was hit was C_j. Moreover,

$$\hat{T}_{ij}^{*} = \frac{R_{ij}^{\tau}}{r_i}, \tag{2.59}$$

where R_{ij}^{τ} denotes the number of time steps where the process came from C_i, and the next core set that was visited was C_j with the additional rule that hits in the time interval $[0, \tau]$ did not count. Algorithmically, one can perform both calculations in the same complexity class, where the effort depends linearly on the length of the trajectory.

3 Analysis of Projected Transfer Operators

In the previous chapter we have seen in 2.1 and 2.2 how Standard Markov State Models can be connected to projected transfer operators by taking a projection onto a finite-dimensional subspace that is constructed from partitioning sets of state space. Moreover, the averaging of the transition kernel along some coordinates could have also been interpreted in terms of projected transfer operators with respect to an infinite-dimensional subspace. In Section 2.3 we finally followed the idea of Standard Markov State Modeling and we introduced another finite-dimensional subspace, namely the space generated by committors. This subspace was also uniquely defined by a finite number of sets in state space that we called core sets.
We will now exploit the abstract framework of projected transfer operators to get insight into the methods by looking at them through linear operator glasses.

3.1 Density Propagation

In Section 2.1 we encountered the problem of non-Markovianity of the switching process $(\tilde{X}_k)$ from (2.2). Nevertheless, in Markov State Modeling one still considers the Markov chain $(\hat{X}_k)$ generated by the transition matrix $\hat{P}$ from (2.7)

$$\hat{P}(i,j) = \mathbb{P}[\tilde{X}_1 = j | \tilde{X}_0 = i] = \mathbb{P}[X_\tau \in A_j | X_0 \in A_i]$$

for some lag time $\tau > 0$ and partitioning sets $A_1, ..., A_n$.
In this section we want to analyze the question when the approximation of the non-Markovian process $(\tilde{X}_k)$ by the Markov chain $(\hat{X}_k)$ is reasonable. Theorem 1 translated this question into the projected transfer operator language. We could deduce that if we distribute $\tilde{X}_0$ and $\hat{X}_0$ equally with an arbitrary initial distribution, the maximal possible error between the distributions of $\tilde{X}_k$ and $\hat{X}_k$ after k time steps is given by

$$E(k) = \|QT^kQ - (QTQ)^k\|. \tag{3.1}$$

So we will investigate under which assumption we can guarantee that this error is small.
For this purpose we will not make a reversibility assumption on the Markov process (X_t). Instead let us assume that T has m real eigenvalues $\lambda_1, ..., \lambda_m \in \mathbb{R}$

$$\lambda_0 = 1 > \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m, \tag{3.2}$$

with an orthonormal system of eigenvectors $(u_j)_{j=1,...,m}$, i.e.

$$Tu_j = \lambda_j u_j, \qquad \langle u_i, u_j \rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{3.3}$$

and $u_0 = \mathbb{1}$. Furthermore we assume that the remainder of the spectrum of T lies within a ball $B_r(0) \subset \mathbb{C}$ with radius $r < \lambda_m$. In order to keep track of the dependence of the eigenvalues on the lag time τ we introduce the associated rates

$$\lambda_j = \exp(-\Lambda_j \tau), \quad r = \exp(-R\tau), \quad r/\lambda_1 = \exp(-\tau(R - \Lambda_1)) = \exp(-\tau\Delta). \tag{3.4}$$

If $T_\tau = e^{L\tau}$, for example, we have

$$Tu = e^{\Lambda\tau} u \Leftrightarrow Lu = \Lambda u. \tag{3.5}$$

The *spectral gap* $\Delta > 0$ will play an essential role later on. We should emphasize that the notion "spectral gap" is usually used differently. It usually designates a situation in which an entire interval of the real axis does not contain any eigenvalues, whereas the intervals above and below show a significantly denser population of eigenvalues. Despite the obvious difference of our case, we will adopt the name *spectral gap* for Δ since it plays a similar role in finding upper bounds as usual spectral gaps.

Based on the above assumptions we can write

$$Tv = T\Pi v + T\Pi^\perp v = \sum_{j=0}^{m} \lambda_j \langle v, u_j \rangle u_j + T\Pi^\perp v, \tag{3.6}$$

where Π is the orthogonal projection onto $U = \text{span}\{u_0, ..., u_m\}$

$$\Pi v = \sum_{j=0}^{m} \langle v, u_j \rangle u_j \tag{3.7}$$

and $\Pi^\perp = \text{Id} - \Pi$ is the projection error with

$$\|T\Pi^\perp\| \leq r < \lambda_m, \qquad \text{spec}(T) \setminus \{1, \lambda_1, ..., \lambda_m\} \subset B_r(0) \subset \mathbb{C}. \tag{3.8}$$

Furthermore, we assume that the subspace U and the remaining subspace don't mix under the action of T:

$$\Pi T \Pi^\perp = \Pi^\perp T \Pi = 0 \tag{3.9}$$

and therefore the dynamics can be studied by considering the dynamics of both subspaces separately

$$T^k = (T\Pi)^k + (T\Pi^\perp)^k \ \forall k \geq 0, \tag{3.10}$$

where the operator $T\Pi$ is self-adjoint because of (3.3). Note that $\Pi^\perp T\Pi = 0$ in (3.9) is always true, but $\Pi T\Pi^\perp = 0$ is an assumption. For sure reversible processes have this property (see Remark 6 below). Nevertheless, it is not completely clear which other classes of processes might match the condition (3.9).
In addition we also define the orthogonal projection Π_0 as

$$\Pi_0 v := \langle v, u_0\rangle u_0 = \langle v, \mathbb{1}\rangle \mathbb{1}. \tag{3.11}$$

According to the above we have the asymptotic convergence rate

$$\|T^k - \Pi_0\| = \lambda_1^k \text{ for all } k \in \mathbb{N}. \tag{3.12}$$

Remark 6 *The assumptions (3.2), (3.3), (3.8), and (3.9) are definitely satisfied if T is sufficiently ergodic and is self-adjoint (T is self-adjoint if the underlying original Markov process (X_t) is reversible). But it may also be sufficient if, e.g., (X_t) is sufficiently ergodic and has a dominant self-adjoint part as it is the case for second-order Langevin dynamics with not too large friction [33], or for thermostatted Hamiltonian molecular dynamics or stochastically perturbed Hamiltonian systems [74, 17]. Reversible or not, the property of being "sufficiently ergodic" seems to be central in any case. We will now give sufficient conditions for a reversible process. These results and their generalizations to non-reversible cases can be found in [36, 74].*

- *A reversible, and μ-irreducible process (X_t) is sufficiently ergodic if one of the following scenarios holds:*

  (i) (X_t) is V-ergodic or geometrically ergodic, see [74].

  (ii) The stochastic transition function $p(t, x, \cdot) = p_a(t, x, \cdot) + p_s(t, x, \cdot)$ associated with (X_t), where p_a denotes the absolutely continuous part and p_s the singular part, satisfies the following two conditions: (a) $p_a \in L^r(\mu \times \mu)$, for some $2 < r < \infty$, and (b) $Sv(y) = \int v(x)p_a(t, x, y)\mu(dy)$ satisfies $\|S\|_{2,\mu} > 0$.

  The above conditions mainly guarantee that the essential spectrum of T is contained in some circle with radius strictly smaller than 1.

- *There are many processes for which these conditions can be shown to be valid; an example is a diffusion process in a smooth energy landscape V with $V \to \infty$ for $\|x\| \to \infty$ fast enough; in this case the spectrum is known to be discrete and real-valued. Comparable results (discrete and real-valued dominant spectrum) can be found in [33] for second-order Langevin dynamics with not too large friction.*

The following Lemma shows the inheritance of first ergodicity properties.

Lemma 1
For every $k \in \mathbb{N}$ we have

$$\|(QTQ)^k - \Pi_0\| \leq \|(TQ)^k - \Pi_0\| \leq \lambda_1^k. \tag{3.13}$$

Proof. Because of $\Pi_0 Q = Q\Pi_0 = \Pi_0$ and $\|T - \Pi_0\| = \lambda_1$ we have for $k = 1$:

$$\|TQ - \Pi_0\| = \|(T - \Pi_0)Q\| \leq \lambda_1. \tag{3.14}$$

Since furthermore $T\Pi_0 = \Pi_0$, and $T\Pi$ is self-adjoint we find for arbitrary $v \in L^2(\mu)$:

$$\begin{aligned} \Pi_0 T v &= \langle Tv, \mathbb{1}\rangle \mathbb{1} = \langle T\Pi v, \mathbb{1}\rangle \ \mathbb{1} + \langle T\Pi^{\perp} v, \Pi\mathbb{1}\rangle \ \mathbb{1} \\ &= \langle v, T\Pi \, \mathbb{1}\rangle \mathbb{1} + \langle \Pi T \Pi^{\perp} v, \mathbb{1}\rangle \mathbb{1} = \langle v, \mathbb{1}\rangle \mathbb{1} = \Pi_0 v, \end{aligned} \tag{3.15}$$

where the identity before the last follows from (3.9). Therefore

$$\Pi_0 T = T\Pi_0 = \Pi_0. \tag{3.16}$$

From this and $Q\Pi_0 = \Pi_0 Q = \Pi_0$ it follows that $(TQ - \Pi_0)^k = (TQ)^k - \Pi_0$ and thus with (3.14)

$$\begin{aligned} \|(QTQ)^k - \Pi_0\| &= \|Q(TQ)^k - Q\Pi_0\| \leq \|(TQ)^k - \Pi_0\| \\ &= \|(TQ - \Pi_0)^k\| \leq \|TQ - \Pi_0\|^k \leq \lambda_1^k, \end{aligned} \tag{3.17}$$

which was the assertion. □

Lemma 1 immediately implies that the error (3.1) decays exponentially,

$$\begin{aligned} E(k) &= \|QT^kQ - (QTQ)^k\| \leq \|QT^kQ - \Pi_0\| + \|(QTQ)^k - \Pi_0\| \\ &\leq \|Q(T^k - \Pi_0)Q\| + \|(QTQ)^k - \Pi_0\| \leq 2\lambda_1^k, \end{aligned} \tag{3.18}$$

independent of the choice of the subspace D, as long as $\mathbb{1} \in D$. Since we want to understand for a Markov State Model how the choice of the sets and other parameters like the lag time τ influence the approximation quality we have to analyze the pre-factor in much more detail.

Theorem 5
Let $T = T_\tau$ be a transfer operator for lag time $\tau > 0$ with properties as described above, in particular (3.2), (3.3), (3.8), and (3.9).
Let $D \subset L^2(\mu)$ be a subspace with $\mathbb{1} \in D$ and define

$$\|Q^{\perp} u_j\| =: \delta_j \leq 1 \ \forall j, \qquad \delta := \max_{j=1,\ldots,m} \delta_j \tag{3.19}$$

where $Q^\perp = \mathrm{Id} - Q$ denotes the projection onto the orthogonal complement of D in $L^2(\mu)$.
Furthermore set

$$\eta(\tau) := \frac{r}{\lambda_1} = \exp(-\tau\Delta) < 1, \quad \text{with } \Delta > 0. \tag{3.20}$$

Then the error (3.1) is bounded from above by

$$E(k) \leq \min\Big[2\,;\, C(\delta, \eta(\tau), k)\Big] \cdot \lambda_1^k, \tag{3.21}$$

with a leading constant of following form

$$\begin{aligned} C(\delta, \eta, k) &= (m\delta + \eta)\Big[C_{\text{space}}(\delta, k) + C_{\text{spec}}(\eta, k)\Big] && (3.22)\\ C_{\text{space}}(\delta, k) &= m^{1/2}(k-1)\,\delta && (3.23)\\ C_{\text{spec}}(\eta, k) &= \frac{\eta}{1-\eta}(1-\eta^{k-1}). && (3.24) \end{aligned}$$

In order to proof Theorem 5, we first observe that the error in (3.1) at time k consists of the $k-1$ projection errors that are propagated until time k is reached, as direct calculation shows.

$$QT^kQ - (QTQ)^k = \sum_{i=1}^{k-1} QT^iQ^\perp(TQ)^{k-i}. \tag{3.25}$$

By this expression we can estimate the approximation error $E(k)$ by observing that it consists of two different parts. Because of $Q^\perp Q^\perp = Q^\perp$ we have

$$\|QT^kQ - (QTQ)^k\| \leq \sum_{i=1}^{k-1} \|QT^iQ^\perp\| \|Q^\perp(TQ)^{k-i}\|. \tag{3.26}$$

The first term $\|QT^iQ^\perp\|$ describes the propagation of the projection error in i steps and the second term $\|Q^\perp(TQ)^{k-i}\|$ measures how large a projection error can be in the $(k-i)$-*th* iteration of applying operator QTQ. So the *i-th* summand explains the effect of propagation of error that is made in the $(k-i)$-*th* iteration.

We will estimate the overall error by looking at both parts of error separately. Let us prepare this with the following lemma.

Lemma 2
For the first part of the error we have the upper bound

$$\|QT^kQ^\perp\| \leq \sqrt{m}\lambda_1^k\delta + r^k. \tag{3.27}$$

Proof. Let v be arbitrary with $\|v\| = 1$. Because $u_0 = \mathbb{1}$ and $Q^\perp u_0 = 0$,

$$(T\Pi)^k Q^\perp v = T^k \Pi Q^\perp v = \sum_{j=1}^{m} \lambda_j^k \langle Q^\perp u_j, v\rangle u_j, \tag{3.28}$$

which leads to

$$\|(T\Pi)^k Q^\perp v\|^2 = \sum_{j=1}^{m} \lambda_j^{2k} \langle Q^\perp u_j, v\rangle^2 \overset{(3.19)}{\leq} m\lambda_1^{2k}\delta^2. \tag{3.29}$$

and therefore

$$\|Q(T\Pi)^k Q^\perp\| \leq \sqrt{m}\lambda_1^k \delta. \tag{3.30}$$

Now we can estimate

$$\begin{aligned}\|QT^k Q^\perp\| &\overset{(3.10)}{\leq} \|Q(T\Pi)^k Q^\perp\| + \|Q(T\Pi^\perp)^k Q^\perp\| \overset{(3.30)}{\leq} \sqrt{m}\lambda_1^k\delta + \|T\Pi^\perp\|^k \\ &\overset{(3.8)}{\leq} \sqrt{m}\lambda_1^k\delta + r^k.\end{aligned} \tag{3.31}$$

□

Now we can proof Theorem 5.

Proof. First recall that the first argument 2 in the minimum taken in (3.21) comes from (3.18). Moreover, recall (3.26), that is,

$$\|QT^kQ - (QTQ)^k\| \leq \sum_{i=1}^{k-1} \|QT^iQ^\perp\| \|Q^\perp (TQ)^{k-i}\|. \tag{3.32}$$

Because $Q^\perp \Pi_0 = 0$ we can write

$$\|Q^\perp (TQ)^{k-i}\| = \|Q^\perp TQ(TQ)^{k-i-1}\| = \|Q^\perp TQ((TQ)^{k-i-1} - \Pi_0)\|. \tag{3.33}$$

Moreover

$$\|Q^\perp TQ\| \leq \|Q^\perp T\Pi Q\| + \|Q^\perp T\Pi^\perp Q\| \leq \|Q^\perp T\Pi Q\| + r \tag{3.34}$$

and for v with $\|v\| = 1$

$$\|Q^\perp T\Pi Q v\|^2 = \sum_{i,j=1}^{m} \langle Qv, u_i\rangle\langle Qv, u_j\rangle \lambda_i\lambda_j \langle Q^\perp u_i, Q^\perp u_j\rangle \leq m^2\lambda_1^2\delta^2\|v\|^2. \tag{3.35}$$

We use Lemma 1 to get

$$\|Q^\perp TQ((TQ)^{k-i-1} - \Pi_0)\| \leq (m\lambda_1\delta + r)\lambda_1^{k-i-1}. \tag{3.36}$$

Inserting (3.36) and Lemma 2 into (3.26) yields

$$E(k) = \|QT^kQ - Q(TQ)^k\| \leq (m\lambda_1\delta + r)\sum_{i=1}^{k-1}(\sqrt{m}\lambda_1^i\delta + r^i)\lambda_1^{k-i-1}. \quad (3.37)$$

Now we have

$$\sum_{i=1}^{k-1}(\sqrt{m}\lambda_1^i\delta + r^i)\lambda_1^{k-i-1} = \sqrt{m}\delta(k-1)\lambda_1^{k-1} + \lambda_1^{k-1}\sum_{i=1}^{k-1}\eta^i \quad (3.38)$$

and

$$\sum_{i=1}^{k-1}\eta^i = \frac{1-\eta^k}{1-\eta} - 1 = \frac{\eta-\eta^k}{1-\eta} = \frac{\eta}{1-\eta}(1-\eta^{k-1}). \quad (3.39)$$

□

The theorem shows that the overall error can be made arbitrarily small by making the factor $[C_{\text{space}}(\delta, k)+C_{\text{spec}}(\eta, k)]$ small. In order to understand the role of these two terms, consider for now $k \geq 2$ to be fixed. It can then be observed that:

1. The pre-factor C_{space} depends on the choice of the subspace D only. For a Standard Markov State Model this means that it depends on the choice of sets $A_1, \ldots, A_n$ where the projection error $\|Q^\perp u_i\|$ measures how well the eigenvector u_i can be approximated by a stepfunction on the partitioning sets. Therefore, it can be made smaller than any tolerance by choosing the sets appropriately and the number of sets, n large enough.

2. The pre-factor C_{spec} is independent of the set definition and depends on the spectral gap Δ and the lag time τ only. While the spectral gap is given by the problem, the lag time may be chosen and thus C_{spec} can also be made smaller than any tolerance by choosing τ large enough. However, the factor C_{spec} will grow unboundedly for $\tau \to 0$ and $k \to \infty$, suggesting that using a large enough lag time is essential to obtain an MSM with good approximation quality, even if the sets are well chosen.

Note that there is a trade-off between the projection error δ and the spectral part of the error that can be modulated by varying the number of resolved eigenfunctions, m. Theorem 5 is valid for every number m, as long as the assumptions on T are fulfilled, i.e. the first m eigenvalues have to be real and the corresponding eigenvectors have to be orthogonal. That is, for a reversible Markov process, for example, Theorem 5 provides a bound for the error $E(k)$ for every $m \in \mathbb{N}$. When increasing m, more eigenvectors are taken into account and the minimal projection error that can be obtained

with a fixed number of stepfunctions, n, will increase. On the other hand, the spectral part of the error will decrease, as growing m increase the spectral gap Δ. This means that increasing m and thus Δ will allow to decrease the lag time τ without changing the spectral part of the error. Moreover, for a reversible Markov process for any fixed $\tau > 0$ we can make the bound in Theorem 5 as small as possible by only refining our discretization. This is clear since we have just discussed that for a reversible process we can always choose a number m^* that is large enough to make the spectral part of the error smaller than any given threshold. Then, we have to choose our partitioning sets $A_1, ..., A_n$ such that the projection error of the first m^* eigenvectors is small enough. This also justifies algorithmic strategies that finely partition state space by using clustering algorithms that have been employed by several researchers in the field [42, 64, 57].

Metastability. [74] gives the following theorem in which smallness of the projection error δ is related to the metastability of a subdivision $A_1, \ldots, A_n$ of state space:

Theorem 6
Let T be a self-adjoint transfer operator with lag time τ and properties as described above, in particular (3.2), (3.3), (3.8), and (3.9). The metastability of an arbitrary decomposition $A_1, \ldots, A_n$ of the state space is bounded from below and above by

$$\begin{aligned} 1 + (1-\delta_1^2)\lambda_1 + \ldots + (1-\delta_{n-1}^2)\lambda_{n-1} + c \leq \sum_{j=1}^{n} \mathbb{P}[X_\tau \in A_i | X_0 \in A_i] \\ \leq 1 + \lambda_1 + \ldots + \lambda_{n-1} \end{aligned} \tag{3.40}$$

where, as above, $\delta_j = \|Q^\perp u_j\|$, and $c = -r\ (\delta_1^2 + \ldots + \delta_{n-1}^2)$.

This result tells us that the minimization of δ with n sets (for n eigenvalues) corresponds to identifying the subdivision with maximum joint metastability.

Example: Markov State Model for double-well potential

The results and concepts from above will first be illustrated on a one-dimensional diffusion in a double-well potential. We consider the process

$$dX_t = -\nabla V(X_t)dt + \sigma dB_t \tag{3.41}$$

with some $\sigma > 0$. The potential V and its unique invariant measure are shown in Fig.3.1.

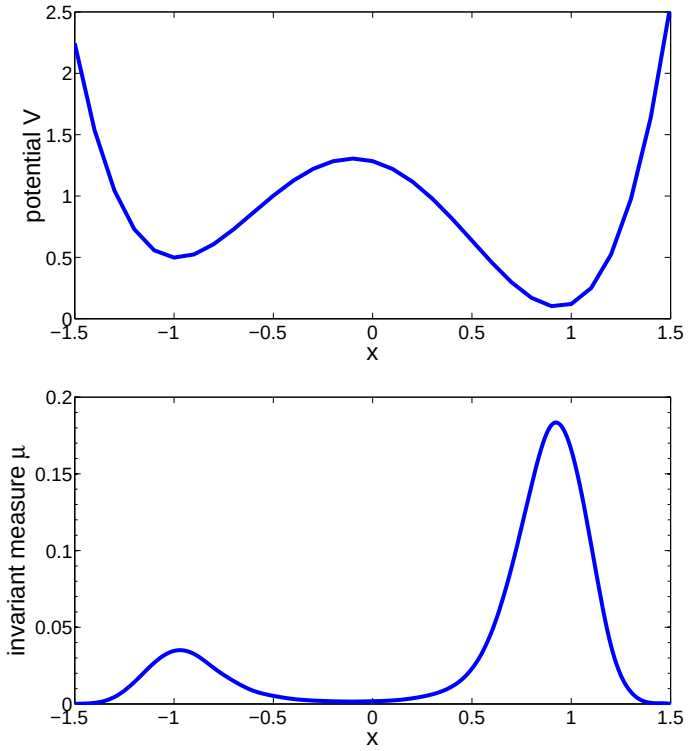


Figure 3.1: (a) The potential V and (b) the associated invariant measure.

This process satisfies all necessary assumptions and by resolving only the slowest process ($m = 1$), the following spectral values are obtained:

$$\Lambda_1 = 0.201, \quad R = 16.363, \quad \Delta = 16.162.$$

The eigenvector u_1 is given in the middle panel of Figure 3.2. It is seen that it is almost constant on the two wells of the potentials and changes sign close to where the local maximum is located. Now, we will build several different Standard Markov State Models and discuss the approximation quality of the switching process $(\tilde{X}_k)$ by the Markov chain $(\hat{X}_k)$ with respect to the results from above.

Projection error δ. Let us first choose the lag time $\tau = 0.1$. Then $\lambda_1 = 0.9801$ and $r = 0.1947$. Fig. 3.2 shows the values of the projection error δ for $n = 2$ and sets of the form $A_1 = (-\infty;\ x]$ and $A_2 = (x;\ \infty)$ depending on the position of the dividing surface, x.

One can see that it is optimal for the boundary between the two sets to lie close to the local maximum of the potential, where the second eigenvector is strongly varying. Next we want to decrease the projection error δ even further and hence optimize the approximation quality of the Markov State

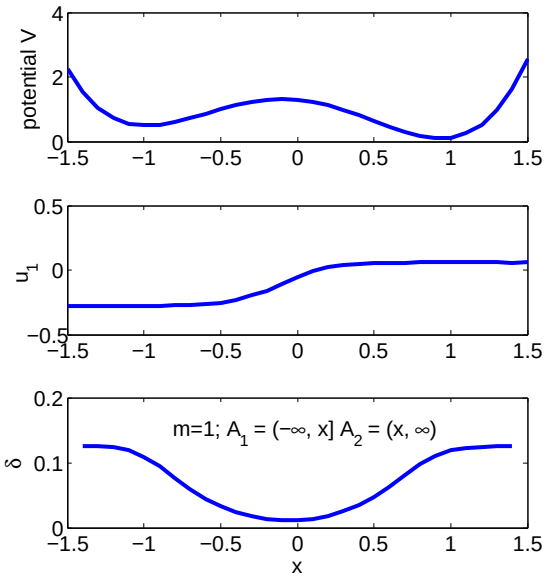


Figure 3.2: Upper panel: Potential V. Middle panel: Eigenvector u_1. Lower Panel: Projection error δ for different sets $A_1 = (-\infty;\ x]$ and $A_2 = (x;\ \infty)$ plotted against x.

Model. We will compare two approaches. On the one hand, we choose $A_1, ..., A_n$ simply as a uniform discretization of the interval $[-1.5, 1.5]$ and include the rest of state space, i.e. the intervals $(-\infty, 1.5)$ and $(1.5, \infty)$ to the outer sets. On the other hand, we will consider a simple adaptive refinement strategy. Here, for the case $n = 2$, the dividing surface is placed so as to minimize the δ error (see Fig. 3.2). For $n = 3$, another dividing surface is introduced at a point that minimized the resulting δ-error, and so on. Fig. 3.3 shows the projection error δ for increasing number of sets.

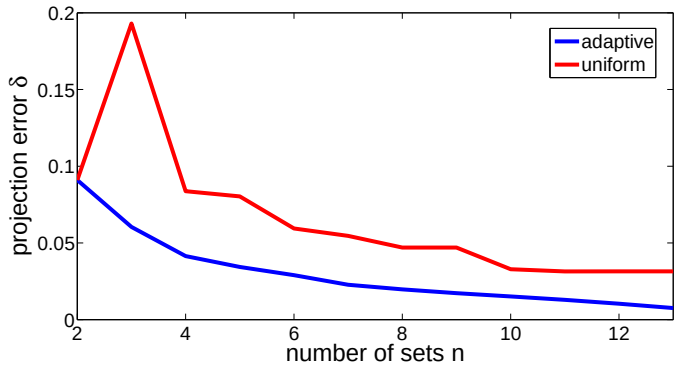


Figure 3.3: Projection error δ against number of sets n for uniform and adaptive discretization.

For the uniform discretization the projection error δ does not monotonically decrease with increasing n. This means that the approximation of the switching process $(\tilde{X}_k)$ by a Markov chain can even get worse while uniformly refining the grid. This is why using a uniform discretization should be avoided. The adaptive refinement strategy, although it does not yield an optimal discretization for $n > 2$, guarantees that the error decreases monotonically with increasing n. Fig.3.4 shows the bestapproximating step-functions for both methods and the first non-trivial eigenvector u_1 for $n = 5$.

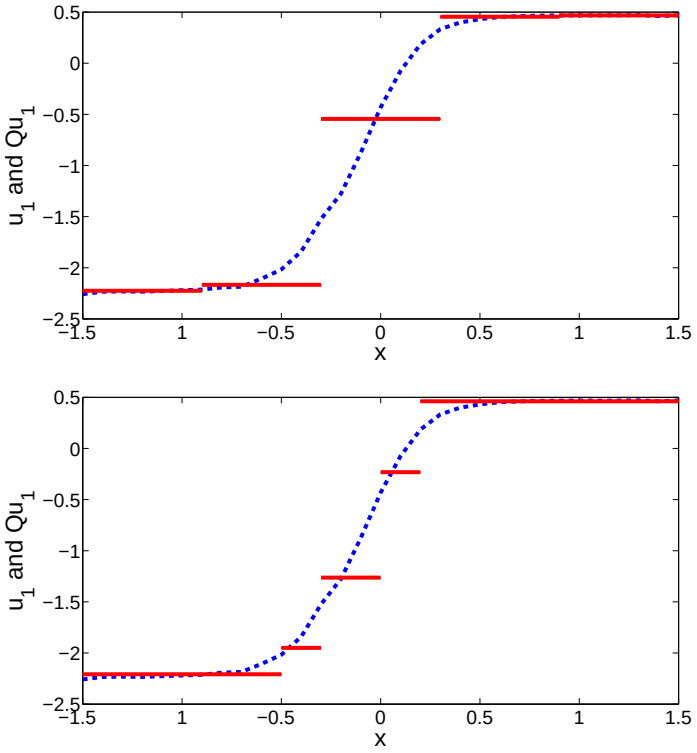


Figure 3.4: Galerkin approximation Qu_1 of second eigenvector. Upper panel: uniform grid of $n = 5$ sets. Lower panel: adaptive grid with $n = 5$ sets.

The adaptive refinement is concentrated on the transition region between the minima of the potential, since most of the projection error is made in this region resulting from the strong variation of the eigenvector.

Effect of the lag time. Next let us study the effect of different lag times τ. Fig. 3.5 shows the bound on the MSM approximation error $E(t)$ from Theorem 5 compared to the exact approximation error $E(t)$ computed via extensive direct numerical simulation for $n = 3$ adaptive sets. Here $E(t)$ is defined as $E(k)$, where $k\tau = t$. Upon increasing the lag time from $\tau = 0.1$ to $\tau = 0.5$ the bound from Theorem 5 becomes much sharper, see Fig. 3.6.

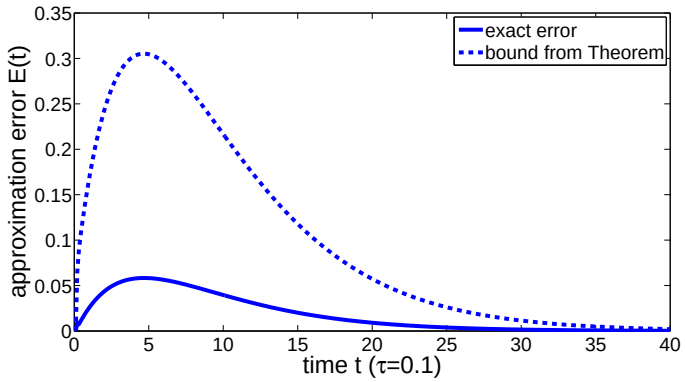


Figure 3.5: Bound and exact error $E(t)$ for $\tau = 0.1$ on adaptive grid with $n = 3$ adaptive sets.

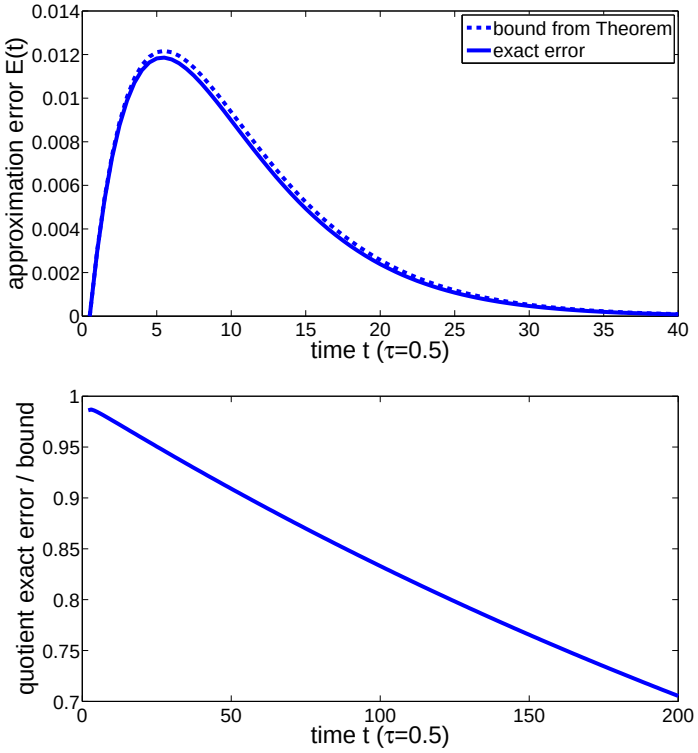


Figure 3.6: Upper panel: bound $B(t)$ from Theorem and exact error $E(t)$ for $\tau = 0.5$ on adaptive grid with $n = 3$. Lower panel: the quotient $\frac{E(t)}{B(t)}$.

The lower panel of Fig. 3.6 additionally shows that the exponential decay of both, the real error $E(t)$ and the upper bound $B(t)$, does not hide some strong discrepancy between $E(t)$ and $B(t)$ for growing t.

Furthermore, Fig. 3.7 exhibits that the approximation quality of the MSM becomes significantly better when the lag time is increased.

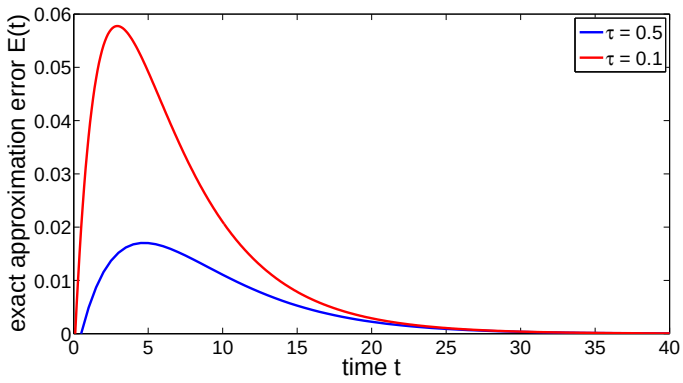


Figure 3.7: Exact error E for different lag times ($\tau = 0.1$ and 0.5) on adaptive grid with $n = 3$.

Finally, Fig. 3.8 compares exact errors and bounds for $n = 3$ sets with uniform and adaptive grid with lag time $\tau = 0.5$ exhibiting a dramatic advantage of the adaptive over the uniform discretization for longer lag times.

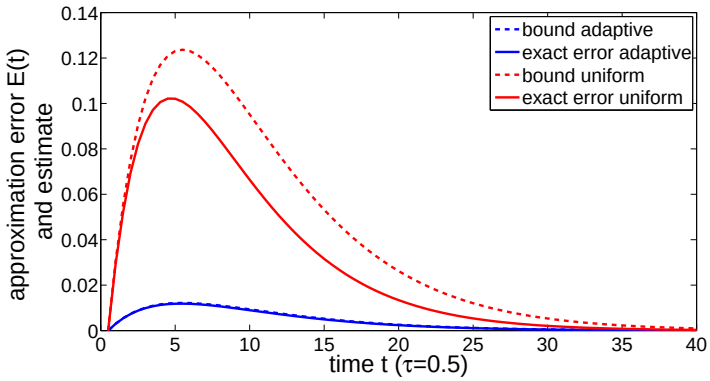


Figure 3.8: Exact error and bound for uniform and adaptive grid, $n = 3, \tau = 0.5$.

Double well potential with diffusive transition region

In the last example we have learned that in order to decrease the projection error δ we had to adaptively find a finer discretization of the transition region between the two wells of the potential. Now we will consider another one-dimensional diffusion (again equation 3.41) in a different potential with two

wells that are connected by an extended transition region with substructure. The new potential V and its unique invariant measure are shown in Fig.3.9.

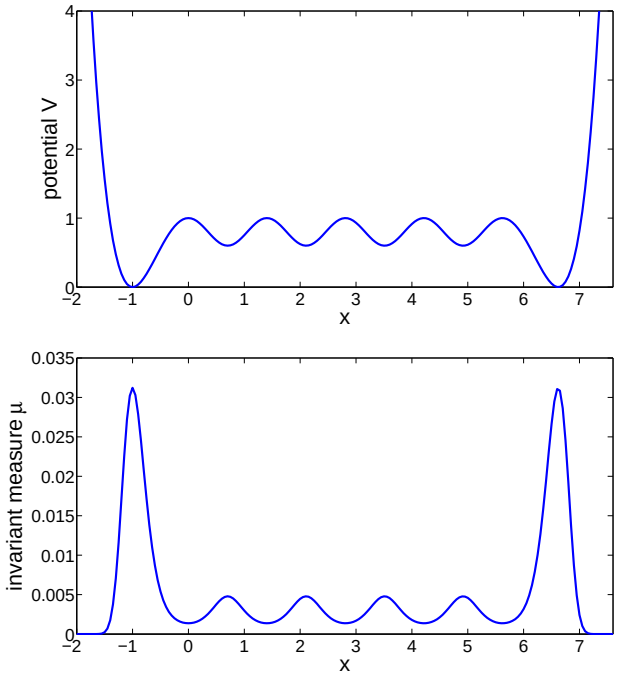


Figure 3.9: The potential V with extended transition region and the associated invariant measure for $\sigma = 0.8$.

We observe that the transition region between the two main wells now contains four smaller wells that will have their own, less pronounced metastability each. When considering the semigroup of transfer operators associated with this dynamics we find the dominant eigenvectors as shown in Fig. 3.10.

The eigenvectors all are almost constant on the two main wells but are non-constant in the transition region. The dominant eigenvalues take the following values (in the form of lag time-independent rates as introduced above):

Λ_0	Λ_1	Λ_2	Λ_3	Λ_4	Λ_5	Λ_6	Λ_7
0	-0.0115	-0.0784	-0.2347	-0.4640	-0.7017	-2.9652	-3.2861

The main metastability has a corresponding timescale $|1/\Lambda_1| \approx 87$ related to the transitions from one of the main wells to the other. Four other, minor metastable timescales related to the interwell switches between the main and the four additional small wells exist in addition.

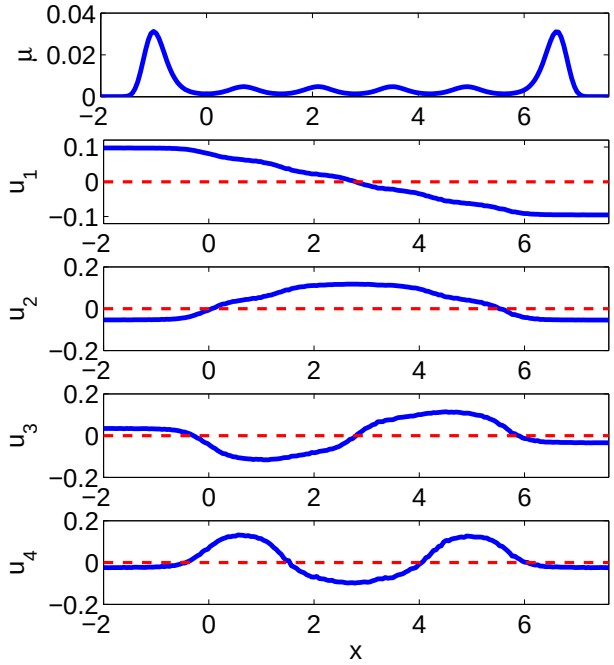


Figure 3.10: Invariant measure and eigenvectors u_j, $j = 1, \ldots, 4$ for Brownian motion in the potential V with extended transition region from Figure 3.9 for $\sigma = 0.8$.

Adaptive subdivisions and projection error. Let us first fix $m = 2$ because for this example it provides the smallest bound (for detailed discussion see [68]), and lag time $\tau = 0.5$ and study how the decay of the projection error depends on the number n of sets in the respective optimal adaptive subdivision. To this end we first observe that adaptive subdivisions will have to decompose the transition regions finer and finer, see Figure 3.11 for an example for $n = 20$.

The decay of the projection error δ with n is shown in Figure 3.12. Figure 3.12 also includes the comparison of the decay of δ with n and the decay of the total propagation error of the underlying MSMs. We observe that the two curves decay in a similar fashion as suggested by our error bound $E(k)$ on the propagation error.

So, as in the previous example the discretization has to be refined because the dominant eigenvectors are not constant in the transition region. In this example we had to increase the number of sets dramatically to achieve a small projection error δ. Especially, if we are dealing with higher dimensional state spaces this property will be critical because the number of sets we have to introduce will scale exponentially with dimension. Therefore, we will work

on this issue in the next sections.

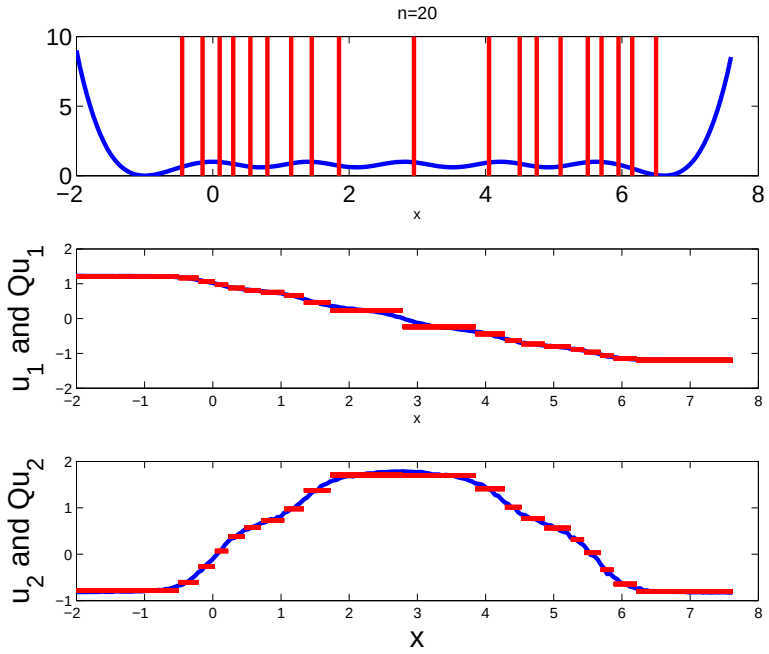


Figure 3.11: Potential and eigenvectors u_j, $j = 1, 2$ and their stepfunction approximation Qu_j for $n = 20$ adaptive sets. The resulting projection error is $\delta = 0.052$.

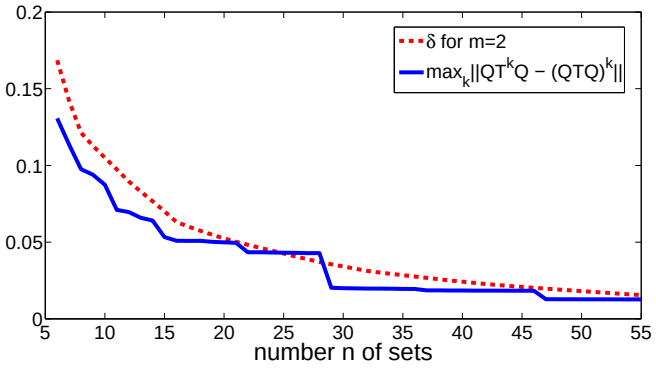


Figure 3.12: Decay of δ and of the maximal propagation error $\max_k \|QT^kQ - (QTQ)^k\|$ with the number n of sets in the optimal adaptive subdivision for $m = 2$.

3.2 Timescales

The next question is how well the eigenvalues of the projected transfer operator QTQ approximate the original eigenvalues of T. In this section we will only consider reversible Markov processes again. Then, because of self-adjointness of the transfer operator we can use the results from [40, 41] to show

Theorem 7
Let $1 = \lambda_0 > \lambda_1 > ... > \lambda_{m-1}$ be the m dominant eigenvalues of T, i.e. for every other eigenvalue λ it holds $\lambda < \lambda_{m-1}$. Let $u_0, u_1, ..., u_{m-1}$ be the corresponding normalized eigenvectors, $D \subset L^2(\mu)$ a subspace with

$$\mathbb{1} \in D \qquad \dim(D) =: n \geq m \tag{3.42}$$

and Q the orthogonal projection onto D.
Moreover, let $1 = \hat{\lambda}_0 > \hat{\lambda}_1 > ... > \hat{\lambda}_{m-1}$ be the dominating eigenvalues of the projected operator QTQ. Then

$$E(\delta) = \max_{i=1,...,m-1} |\lambda_i - \hat{\lambda}_i| \leq \lambda_1 (m-1)\delta^2, \tag{3.43}$$

where

$$\delta = \max_{i=1,...,m-1} \|Q^{\perp} u_i\|$$

is the maximal projection error of the eigenvectors to the space D.

Proof. The eigenvector of T w.r.t. the trivial eigenvalue $\lambda_0 = 1$ is known: $u_0 = \mathbb{1}$. Therefore

$$u_0 \in D \Rightarrow Qu_0 = u_0. \tag{3.44}$$

This implies that u_0 is also eigenvector of QTQ w.r.t. its largest eigenvalue $\hat{\lambda}_0 = 1$.
Now define

$$\Pi_0 v = \langle v, u_0 \rangle u_0, \tag{3.45}$$

set again $\Pi_0^{\perp} = Id - \Pi_0$, and consider the operator $T\Pi_0^{\perp} = T - \Pi_0$. Since T is self-adjoint, its eigenvectors $u_0, u_1, ...$ are orthogonal, which implies that

$$T\Pi_0^{\perp} u_j = Tu_j - \Pi_0 u_j = Tu_j = \lambda_j u_j \qquad \forall j > 0$$

and $T\Pi_0^{\perp} u_0 = 0$, that is, the operator $T\Pi_0^{\perp}$ has the same eigenvalues with the same corresponding eigenvectors as T, just the eigenvalue $\lambda_0 = 1$ changed to a zero eigenvalue.
Moreover,

$$\Pi_0 T \Pi_0^{\perp} = 0, \quad \text{and therefore} \quad T\Pi_0^{\perp} = \Pi_0^{\perp} T \Pi_0^{\perp},$$

which implies self-adjointness of the operator $T\Pi_0^\perp$.
Now set $U = \text{span}\{u_0, ..., u_{m-1}\}$, and let Π be the orthogonal projection onto U. Then, the operator $\Pi T \Pi_0^\perp \Pi$ has exactly the eigenvalues $\lambda_1, ..., \lambda_{m-1}$ and an additional eigenvalue zero, that corresponds to the eigenvector u_0.
From (3.44) it follows that $Q\Pi_0 Q = \Pi_0$ and hence

$$QT\Pi_0^\perp Q = QTQ - \Pi_0.$$

The same argument as above shows that the operator $QT\Pi_0^\perp Q$ has the same spectrum as QTQ, just the corresponding eigenvalue of u_0 changed from $\hat{\lambda}_0 = 1$ to zero.
Using the results from [40], we find for the error (3.43)

$$E(\delta) = \max_{i=1,...,m-1} |\lambda_i - \hat{\lambda}_i| \leq (\lambda_1 - \lambda_{\min(U+D)}) \max_i \sin^2(\theta_i(U, D)), \quad (3.46)$$

with $\Theta = \Theta(U, D) = \{\theta_0, ..., \theta_{m-1}\}$, a vector of principal angles between the subspaces U and D. $\lambda_{\min(U+D)}$ is the smallest eigenvalue of the operator ZTZ, where Z is an orthogonal projection on the space $U + D$. In our case this means $\lambda_{\min(U+D)} = 0$. Let $\sigma_i(A)$ and $\Lambda_i(B)$ denote the i-th singular value of operator A and i-th eigenvalue of operator B, respectively. The principal angles are defined as $\cos(\theta_i) = \sigma_i(Q\Pi)$. Moreover, the definition of singular values yields

$$\sigma_i^2(Q\Pi) = \Lambda_i((Q\Pi)^* Q\Pi) = \Lambda_i(\Pi Q \Pi), \quad (3.47)$$

where $(Q\Pi)^*$ denotes the Hermitian transpose of $(Q\Pi)$. We get

$$\sin^2(\theta_i) = 1 - \cos^2(\theta_i) = 1 - \Lambda_i(\Pi Q \Pi) = \Lambda_i(\Pi - \Pi Q \Pi) = \Lambda_i(\Pi Q^\perp \Pi). \quad (3.48)$$

As in (3.47),

$$\Lambda_i(\Pi Q^\perp \Pi) = \sigma_i^2(Q^\perp \Pi) \leq \|Q^\perp \Pi\|^2. \quad (3.49)$$

Now let $v, \|v\| = 1$ be arbitrary. If we define $\hat{v} \in \mathbb{R}^{m-1}$ as

$$\hat{v}_j = \langle v, u_j \rangle, j = 1, ..., m-1,$$

it is well known for the usual p-norms on $\mathbb{R}^{m-1}$

$$\sum_{j=1}^{m-1} |\langle v, u_j \rangle| = \|\hat{v}\|_1 \leq \sqrt{m-1} \|\hat{v}\|_2 = \sqrt{m-1} \left(\sum_{j=1}^{m-1} \langle v, u_j \rangle^2 \right)^{1/2} \leq \sqrt{m-1}. \quad (3.50)$$

Since $Q^\perp u_0 = 0$,

$$\begin{aligned} \|Q^\perp \Pi v\| = \left\| \sum_{j=1}^{m-1} \langle v, u_j \rangle Q^\perp u_j \right\| &\leq \sum_{j=1}^{m-1} |\langle v, u_j \rangle| \|Q^\perp u_j\| \\ &\leq \sum_{j=1}^{m-1} |\langle v, u_j \rangle| \delta \leq \sqrt{m-1} \cdot \delta. \end{aligned} \quad (3.51)$$

Combining (3.48), (3.49) and (3.51)

$$\sin^2(\theta_i) \leq \|Q^{\perp}\Pi\|^2 \leq (m-1)\delta^2. \tag{3.52}$$

Putting everything together gives (3.43). □

Remark 7 *Inserting (3.5) into (3.43), we get the lag time depended eigenvalue estimate*

$$E(\tau,\delta) = \max_{i=1,\ldots,m-1} |\lambda_i - \hat{\lambda}_i| \leq e^{\Lambda_1 \tau}(m-1)\delta^2, \tag{3.53}$$

where (λ_i) *are the dominant eigenvalues of the transfer operator* T_τ *and* $(\hat{\lambda}_i)$ *the dominant eigenvalues of the projection* $QT_\tau Q$.
Since $\Lambda_1 < 0$,

$$E(\tau,\delta) \to 0, \text{ for } \tau \to \infty. \tag{3.54}$$

In the last Section 3.1 the same projection error δ of the dominant eigenvectors played already a key role. We have seen that the jumps of a Markov process between partitioning sets $A_1, ..., A_n$ of state space can be well approximated by a Markov chain if this factor δ is small enough and the lag time τ is chosen appropriately. For a reversible Markov process, as we consider it in the next sections only, we could even get the approximation error below every threshold for arbitrarily small lag time τ by taking care of the projection error $\|Q^{\perp}u_i\|$ for enough eigenvectors u_i. Now, Theorem 7 shows that this factor δ can also guarantee a good approximation of the original longest timescales, i.e. the dominant eigenvalues of T, by the projected transfer operator QTQ. For a Standard Markov State Model the eigenvalues of QTQ would simply be the eigenvalues of the matrix $\hat{P}$ from (2.7), which describes the transition probabilities between the partitioning sets. On the other hand, the diffusion example with the slightly more complicated potential with extended transition region (Fig. 3.9) revealed problems in finding a discretization of state space in order to optimize the important δ. We had to introduce many small sets inside of the transition region to guarantee a small error. Therefore, we will now investigate the projected transfer operator QTQ which was introduced in Section 2.3. This operator was constructed with the use of the milestoning process (2.24) with respect to core sets $C_1, ..., C_n$. The subspace for projection D was then given by the associated committors.

Our hope is that this subspace might not require a refinement of the transition region because core sets do not need to form a full partition of state space anymore. The plan that we have in mind is to introduce core sets only at the deep main wells of the potential where the eigenvectors are almost constant and leave the problematic transition region undefined. Note that the projection of an eigenvector Qu with respect to the committors will always be constant on the core sets, but in the undefined region C (2.23) it will be the solution of a linear equation with boundary conditions.

Extended transition region and committor approximation

We consider again the diffusion process

$$dX_t = -\nabla V(X_t)dt + \sigma dB_t \tag{3.55}$$

with B_t denoting Brownian motion in the potential V (Fig. 3.9) with two wells that are connected by an extended transition region with noise intensity $\sigma = 0.8$.

Two core sets

In the following paragraphs we will compare the eigenvalues and ITS of the original process to the ones resulting from different Markov State Models. More precisely, we first choose a lag time τ and consider the transfer operator T_τ. Because of (3.5) we can compute the implied timescale

$$|1/\Lambda_1| = -\frac{\tau}{\ln(\lambda_{1,\tau})}, \tag{3.56}$$

where $\lambda_{1,\tau} < 1$ is the largest non-trivial eigenvalue of T_τ.
The minima in the two main wells are located at $x_1 = -1$ and $x_2 = 6.62$, the respective local maxima that separate the main wells from the rest of the landscape at $x_1^{\pm} = x_1 \pm 1$, and $x_2^{\pm} = x_2 \pm 1$, respectively.
We said that we want to choose two core sets of the form $C_1^s = (-\infty, x_1 + s]$ and $C_2^s = [x_2 - s, \infty)$ for some parameter s around the deep main wells of the potential. Then we compare the ITS from (3.56) to the one, which corresponds to the largest non-trivial eigenvalue $\hat{\lambda}_{i,\tau}$ of the projected operator $QT_\tau Q$

$$|1/\hat{\Lambda}_1| = -\frac{\tau}{\ln(\hat{\lambda}_{1,\tau})}. \tag{3.57}$$

Since the process under investigation is just one-dimensional, we can compute the committor functions from finite element discretization of $\mathcal{L}$ because it is a differential operator (2.6), and very accurate approximations of $\hat{T}_\tau$ and M from Theorem 3, which provide the matrix representation of $QT_\tau Q$. Figure 3.13 shows the dependence of the non-trivial eigenvalue and the projection error $\delta = \|Q^\perp u_1\|$ on the core set size s for $\tau = 1$.

We observe that for small enough core sets the approximation of the exact first non-trivial eigenvalue of T_τ, $\exp(\tau\Lambda_1)$, is good, while for too large core sets the approximation quality decreases. Moreover, Theorem 7 connected this error to the projection error $\|Q^\perp u_1\|$ and Fig. 3.13 also shows that this error behaves exactly like the approximation quality of the eigenvalue.

For $m = 2$, that is, if we just consider the first non-trivial eigenvalue, we can also study the relative error

$$E_{rel}(\tau, \delta) = \frac{|\lambda_{1,\tau} - \hat{\lambda}_{1,\tau}|}{\lambda_{1,\tau}} \tag{3.58}$$

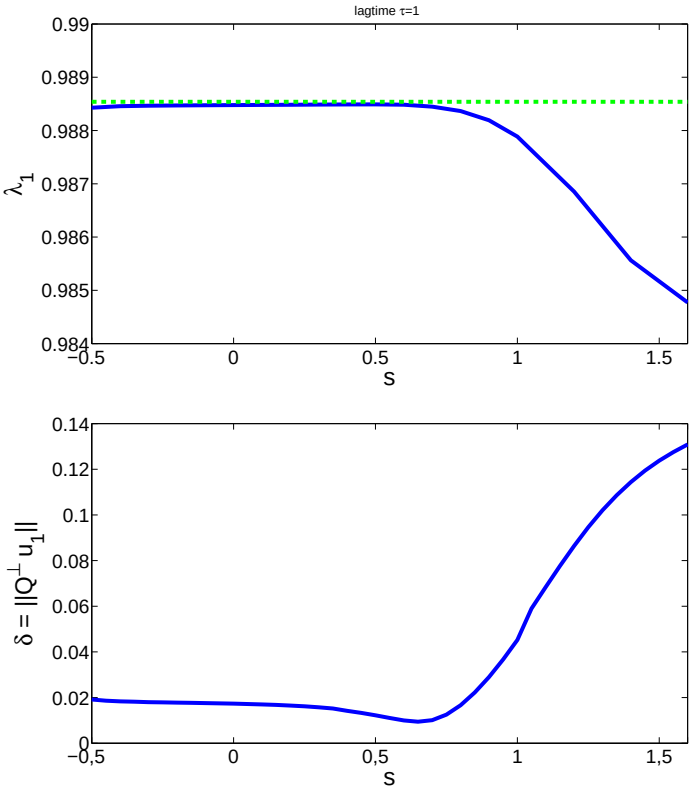


Figure 3.13: Top: Non-trivial eigenvalues $\lambda^s_{1,\tau} < 1$ of $QT_\tau Q$ versus cores set size parameter s for lag time $\tau = 1$ in comparison to the exact first non-trivial eigenvalue $\exp(\tau\Lambda_1)$. Bottom: Projection error $\|Q^\perp u_1\|$ dependent on the size of core sets, i.e. the parameter s.

for different core set sizes s. Theorem 7 provides an upper bound by the τ-independent square of the projection error $\delta = \|Q^\perp u_1\|$. In Fig. 3.14 we observe that for small lag times the real relative error is significantly smaller than δ^2 but for larger lag times the upper bound and the real error are very close. As to be expected from Fig. 3.13 (bottom) the error for good core sets ($s = 0.5$) is two orders of magnitude smaller than the "not so good" core sets for $s = 2$.

Estimation from data

The computation of the committor functions will only be possible via finite element discretization of the generator, which is infeasible in higher dimensions. Fortunately, Theorem 4 from Section 2.3 provided a possibility to estimate the matrices $\hat{T}$ and M, which form the matrix representation of $QT_\tau Q$, from realizations of the process.
Therefore, we study the milestoning process $(\hat{X}_{n\tau})$ on state space $\{1, 2\}$

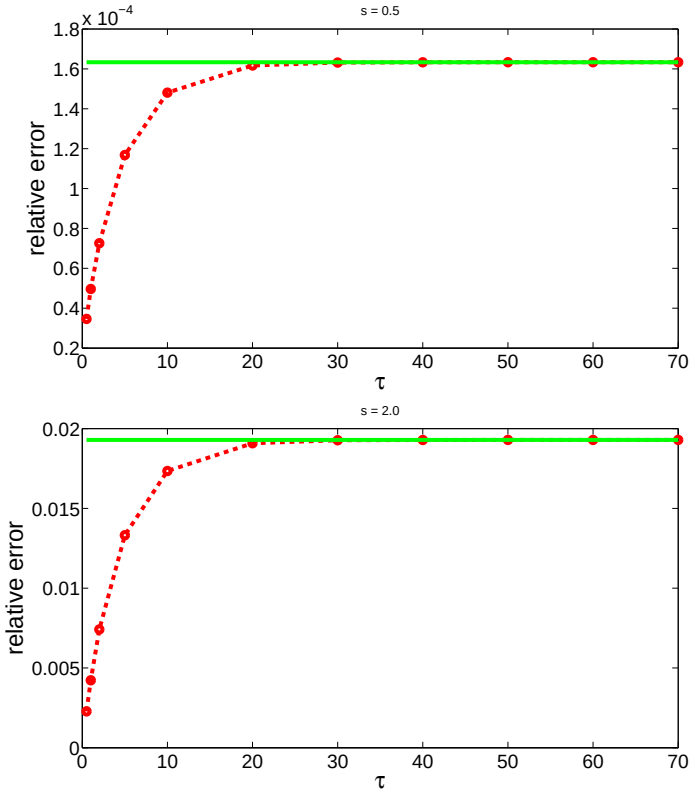


Figure 3.14: Relative error $E_{rel}(\tau, s)$ versus lag time τ (dashed red line) compared to the upper bound δ^2 given by Theorem 7 (green solid line), for $s = 0.5$ (top) and $s = 2$ (bottom).

induced by the time-discrete process given by T_τ and the cores sets C_i^s, $i = 1, 2$.

Then, we compute a very long trajectory $x(t)$, $t \in [0, t_{max}]$ of the diffusion process (for example based on Euler-Maruyama discretization of the SDE (3.55)). From this, we get discrete trajectories of the process $X_{n\tau}$ and of the milestoning process $\hat{X}_{n\tau}$, $n = 0, \ldots, N_\tau$ with $N_\tau = \lfloor t_{max}/\tau \rfloor$. Here, this was done based on a trajectory $x(t)$ in the time interval $[0, t_{max}]$ with $t_{max} = 50000$. Then we can estimate $\hat{T}$ and M by $\hat{T}^*_{N_\tau}$ and $M^*_{N_\tau}$ respectively as described in Sec. 2.3. In this example we also choose the time resolution for the committors to be the lag time τ. The resulting non-trivial eigenvalues $\hat{\lambda}_1^*$ of the generalized eigenvalue problem $\hat{T}^*_{N_\tau} r = \hat{\lambda}^* M^*_{N_\tau} r$, which gives the eigenvalues of the matrix $\hat{T}^*_{N_\tau} M^{*-1}_{N_\tau}$, are compared to the ones of $\hat{T}r = \hat{\lambda} M r$, where $\hat{T}$ and M come from the finite element discretization, and to the exact first non-trivial eigenvalue $\lambda_1 = \exp(\tau \Lambda_1)$ in Fig. 3.15.

We observe that the trajectory-based eigenvalues are overestimating the "exact" eigenvalues of the generalized eigenvalue problem, and that the ap-

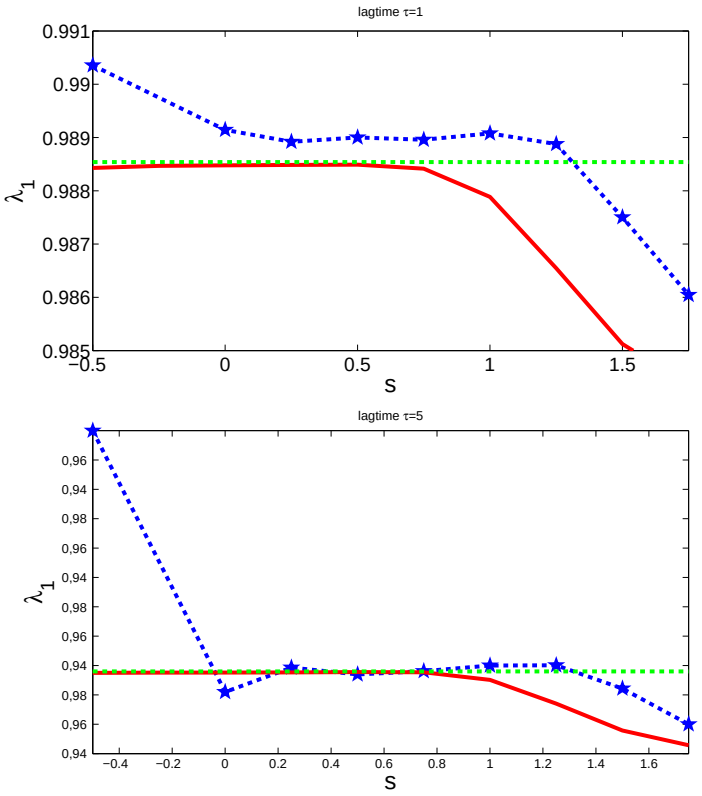


Figure 3.15: Comparison of the non-trivial eigenvalues λ_1^* of the trajectory-based generalized eigenvalues problem $\hat{T}_{N_\tau}^* r = \hat{\lambda} M_{N_\tau}^* r$ (blue, dashed, stars), the ones of $\hat{T}r = \hat{\lambda}Mr$ (red, solid line) and the exact first non-trivial eigenvalue $\lambda_1 = \exp(\tau\Lambda_1)$ (green, straight dashed line) in dependence on the core size parameter s for different lag time $\tau = 1$ (top) and $\tau = 5$ (bottom).

proximation is getting worse for small values of s, especially for larger lag times. This is not surprising since for $s < 0$ and sparse undersampling of the trajectory for large lag times, we will miss events in which the process stays close to the minima x_i without entering the cores for some time which is not long compared to the lag time.

Despite the good approximation quality of the trajectory-based generalized eigenvalues we should not forget that they are subject to an unknown statistical sampling error resulting from the finiteness of the trajectory.

Comparison to full partition of state space Let us fix $m = 2$ and observe how the relative eigenvalue error E_{rel} as defined in (3.58) above behaves in the case of a full partition of state space, especially how it changes for different full subdivisions of the state space and different lag times. From Theorem 7 we know that, as above, the bound on the relative eigenvalue

error is given by the square of the projection error δ. First we choose $n = 2$ and the subdivision $A_1 = (-\infty, x]$ and $A_2 = (x, \infty)$. Figures 3.16 and 3.17 show the bound δ^2 compared to the relative error $E_{rel}(\tau, \delta)$, for two different subdivisions, i.e., different values of x. We can see that the error converges to δ^2 for increasing τ. Also, a better choice of the subdivision results not only in a smaller relative error, but in its faster convergence to the bound.

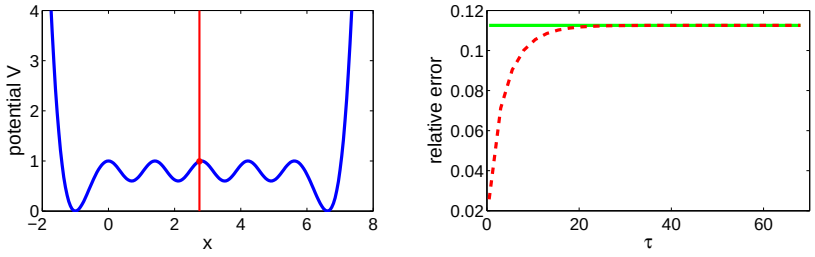


Figure 3.16: Relative error for eigenvalues and bound for $\tau = 0.5$, $n = 2$ and $x = 2.75$

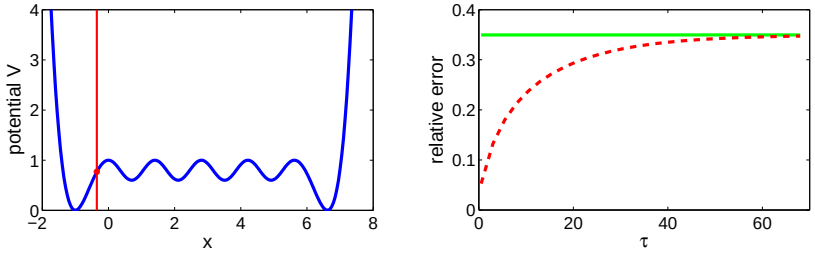


Figure 3.17: Relative error for eigenvalues and bound for $\tau = 0.5$, $n = 2$ and $x = -0.35$

Now we consider the full partition of a state space into $n = 6$ sets. The sets are chosen in such a way that every well belongs to one set. This choice of sets results in a smaller bound and faster convergence of the relative error to this bound, which can be seen in Figure 3.18.

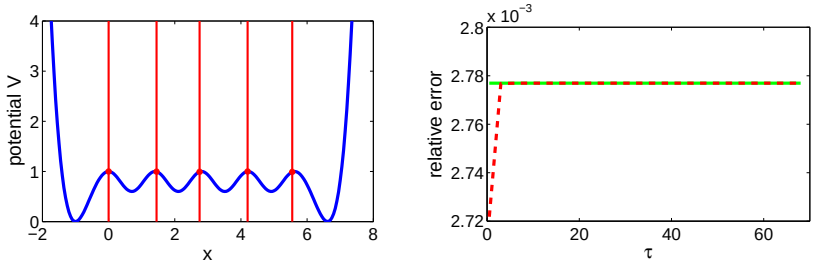


Figure 3.18: Relative error for eigenvalues and bound for $\tau = 0.5$ and $n = 6$

If we finally compare the results for full subdivisions to the approximation via two core sets, we observe the following: Even the optimal full subdivision into $n = 2$ sets cannot compete with the approximation quality of the approximation based on two "reasonable/good" core sets. Good core sets result in an approximation error that is even better than the one for the optimal full subdivision into $n = 6$ sets which already resolves the well structure of the energy landscape. Thus, Markov State Models based on fuzzy ansatz spaces resulting from appropriate core sets and associated committor ansatz functions seem to lead to superior approximation quality than comparable Standard full subdivision Markov State Models.

Projections on Infinite-Dimensional Subspaces

In Theorem 7, one important assumption on the subspace D is finite dimensionality. We have seen already that the projected transfer operator approach might also be interesting for particular infinite-dimensional subspaces D. In Sec. 2.2, for example, we connected an averaging method to the subspace of functions which are independent of some variable. Theorem 7 would not be directly applicable to this subspace and therefore, we could not make a rigorous statement about the timescale approximation of the underlying averaging method. We will now derive an analog of Theorem 7, where the proof shows a quite general approach for the extension to infinite-dimensional subspaces. In Section 3.3, this will be useful again.

Theorem 8
Let $1 = \lambda_0 > \lambda_1 > ... > \lambda_{m-1}$ be the m dominant eigenvalues of T, i.e. for every other eigenvalue λ it holds $\lambda < \lambda_{m-1}$. Let $u_0, u_1, ..., u_{m-1}$ be the corresponding normalized eigenvectors. Assume that the state space E has the form $E = E_x \times E_y$ and let $D \subset L^2(\mu)$ be the infinite dimensional subspace

$$D = \{v(x, y) \in L^2(\mu) | v(x, y) = \tilde{v}(x)\}, \tag{3.59}$$

that is, $v \in D$, if v does not depend on y.
Let Q be the orthogonal projection onto D and let $1 = \hat{\lambda}_0 > \hat{\lambda}_1 > ... > \hat{\lambda}_{m-1}$ be the dominating eigenvalues of the projected operator QTQ. Then

$$E(\delta) = \max_{i=1,...,m-1} |\lambda_i - \hat{\lambda}_i| \leq \lambda_1 (m-1)\delta^2, \tag{3.60}$$

where

$$\delta = \max_{i=1,...,m-1} \|Q^{\perp} u_i\|.$$

is the maximal projection error of the eigenvectors to the space D.

Proof. Let $\epsilon > 0$ be arbitrary and $\hat{u}_0, ..., \hat{u}_{m-1}$ be the normalized eigenvectors of QTQ w.r.t. the eigenvalues $1 = \hat{\lambda}_0 > \hat{\lambda}_1 > ... > \hat{\lambda}_{m-1}$. $\hat{u}_i \in L^2(\mu)$ and since $u_i \in L^2(\mu)$, also $Qu_i \in L^2(\mu)$, so there must be a compact set $K_x \subset E_x$ with

$$\int\limits_{E\setminus(K_x\times E_y)} Qu_i^2 d\mu \leq \epsilon \qquad \int\limits_{E\setminus(K_x\times E_y)} \hat{u}_i^2 d\mu \leq \epsilon \qquad \forall i = 0, ..., m-1. \quad (3.61)$$

Qu_i and $\hat{u}_i$ can also be arbitrarily well approximated by stepfunctions and because $Qu_i(x, y), \hat{u}_i(x, y) \in D$, they do not depend on y, that is, there is a partitioning $A_1^x, ..., A_N^x$ of K_x, i.e.

$$A_i^x \cap A_j^x = \emptyset \qquad \bigcup_{j=1}^{N} A_j^x = K_x,$$

such that

$$\int\limits_{K_x\times E_y} (Qu_i - P_N Qu_i)^2 d\mu \leq \epsilon \qquad \int\limits_{K_x\times E_y} (\hat{u}_i - P_N\hat{u}_i)^2 d\mu \leq \epsilon \qquad \forall i = 0, ..., m-1, \quad (3.62)$$

where P_N is the orthogonal projection onto the space

$$V_N = span\{\mathbb{1}_{A_1}, ..., \mathbb{1}_{A_N}, \mathbb{1}_{E\setminus K}\}$$

with

$$A_i = A_i^x \times E_y \qquad K = K_x \times E_y.$$

We obviously have

$$QP_N = P_N. \quad (3.63)$$

This implies also

$$P_N = P_N^* \overset{(3.63)}{=} (QP_N)^* = P_N^* Q^* = P_N Q \quad (3.64)$$

and

$$P_N^\perp u_i = u_i - P_N u_i = u_i - Qu_i + Qu_i - P_N u_i = Q^\perp u_i + Qu_i - P_N Qu_i.$$

Because $Q^\perp Q = Q^\perp P_N = 0$ we get

$$\|P_N^\perp u_i\|^2 = \|Q^\perp u_i\|^2 + \|Qu_i - P_N Qu_i\|^2. \quad (3.65)$$

We now denote with $\lambda_i(A)$ the ith largest eigenvalue of an operator A. Then we have

$$\begin{aligned} |\lambda_i - \hat{\lambda}_i| &= |\lambda_i(T) - \lambda_i(QTQ)| = |\lambda_i(T) - \lambda_i(P_N TP_N) + \lambda_i(P_N TP_N) - \lambda_i(QTQ)| \\ &\leq |\lambda_i(T) - \lambda_i(P_N TP_N)| + |\lambda_i(P_N TP_N) - \lambda_i(QTQ)| \\ &\overset{(3.63),(3.64)}{=} |\lambda_i(T) - \lambda_i(P_N TP_N)| + |\lambda_i(P_N QTQP_N) - \lambda_i(QTQ)|. \end{aligned} \quad (3.66)$$

As P_N is a projection onto a finite dimensional subspace V_N and $\mathbb{1} \in V_N$, we can apply Theorem 7 to get estimates for each of the summands. First, for every $i = 1, ..., m-1$

$$\begin{aligned} |\lambda_i(T) - \lambda_i(P_N T P_N)| &\leq \lambda_1(m-1) \max_{k=1,...,m-1} \|P_N^\perp u_k\|^2 \\ &\overset{(3.65)}{=} \lambda_1(m-1) \max_{k=1,...,m-1} \left[\|Q^\perp u_k\|^2 + \|Qu_k - P_N Q u_k\|^2\right] \\ &\overset{(3.61),(3.62)}{\leq} \lambda_1(m-1)(\max_{k=1,...,m-1} \|Q^\perp u_k\|^2 + 2\epsilon) \end{aligned} \tag{3.67}$$

Using the same calculation for the second summand we get

$$\begin{aligned} &|\lambda_i(P_N Q T Q P_N) - \lambda_i(QTQ)| \leq \hat{\lambda}_1(m-1) \max_{k=1,...,m-1} \|P_N^\perp \hat{u}_k\|^2 \\ &\leq \hat{\lambda}_1(m-1) 2\epsilon. \end{aligned} \tag{3.68}$$

Inserting (3.67) and (3.68) into (3.66) yields for every $i = 1, ..., m-1$

$$|\lambda_i - \hat{\lambda}_i| \leq \lambda_1(m-1)(\max_{k=1,...,m-1} \|Q^\perp u_k\|^2 + 2\epsilon) + \hat{\lambda}_1(m-1)2\epsilon.$$

Since this inequality holds for all $\epsilon > 0$, it also holds for $\epsilon = 0$, which gives the proposition. □

Multiscale Core Set Approach

A natural problem that arises, when one tries to approximate the dynamics of a continuous Markov process by a low-dimensional Markov State Model, i.e. which has only few states, is the approximation of very different timescales at the same time. As a motivation we consider again the diffusion in the extended double-well potential. Assume we are now interested in an approximation of the corresponding transfer operator T, which has for a certain set of diffusion parameters and a lag time $\tau > 0$ the dominating eigenvalues

λ_0	λ_1	λ_2	λ_3	λ_4
1.0000	0.9885	0.9247	0.7911	0.6289

As an approximation we want to find a projected transfer operator QTQ, where Q denotes the orthogonal projection onto the subspace D of committor functions that belong to core sets $C_1, ..., C_n$.

First eigenvalue. Let us first try to choose two core sets such that the first non-trivial eigenvalue is well approximated by the only non-trivial eigenvalue of QTQ.

Theorem 7 tells us that we just have to find core sets such that $\delta = \|Q^\perp u_1\|$ is small. We have found such core sets already in the previous example. Fig 3.19 shows core sets that yield $\delta = 0.0164$ and therefore an eigenvalue error $|\lambda_1 - \hat{\lambda}_1| < 10^{-4}$.

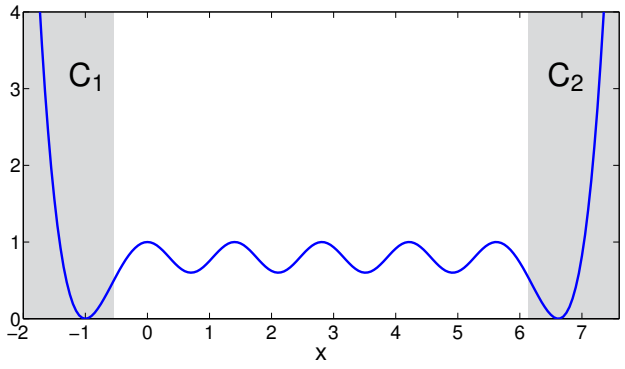


Figure 3.19: Two good core sets to approximate first timescale.

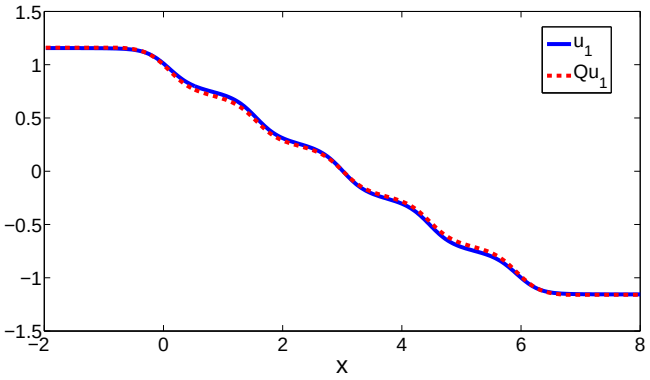


Figure 3.20: Eigenvector u_1 and its projection Qu_1.

Another eigenvalue and a problem. Now assume that we also want to approximate the second non-trivial eigenvalue λ_2. From Theorem 7 we know that we have to choose 3 core sets such that $\delta = \max\{\|Q^\perp u_1\|, \|Q^\perp u_2\|\}$ is small. But even in this simple one-dimensional example we will see immediately that it is problematic to simultaneously make $\|Q^\perp u_1\|$ and $\|Q^\perp u_2\|$ small. If we look at the shape of the first non-trivial eigenvector u_1, we see that we cannot introduce a large set inside of the transition region because the eigenvector is varying in this region and its projection will be constant on core sets. So, a larger core set in the transition region will definitely yield a larger error for $\|Q^\perp u_1\|$. This effect is demonstrated by comparing two different sizes of a third core set as shown in the following figures.

For the small third core set we get

$$\|Q^\perp u_1\| = 0.0339 \qquad \|Q^\perp u_2\| = 0.1024$$

and exactly the other way around for the large third core set

$$\|Q^\perp u_1\| = 0.1302 \qquad \|Q^\perp u_2\| = 0.0444.$$

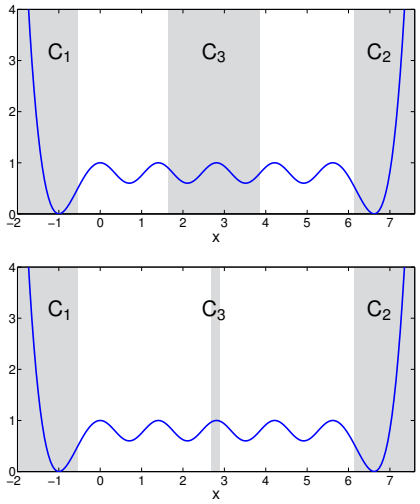


Figure 3.21: Take a larger or smaller third core set?

The controversial effect can also be seen in Fig. 3.22.

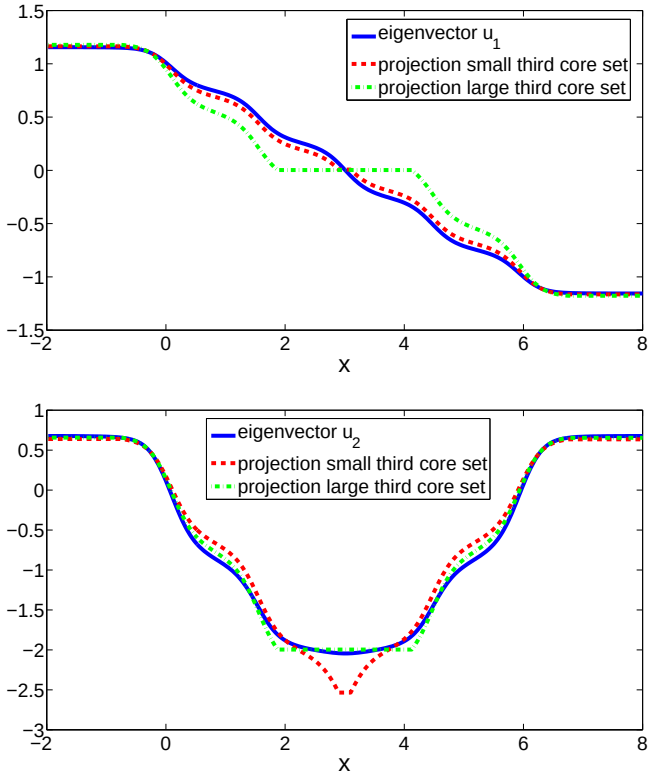


Figure 3.22: Controversial projection error when projecting eigenvectors.

This is an issue because in both cases Theorem 7 cannot guarantee a good approximation of the second timescale because $\max\{\|Q^\perp u_1\|, \|Q^\perp u_2\|\}$ is too large. Nevertheless, we look at the eigenvalues of the projected transfer operators QTQ for both types of third core sets.

$$\text{Small third core set: } \hat{\lambda}_1 = 0.9883, \hat{\lambda}_2 = 0.9203$$
$$\text{Large third core set: } \hat{\lambda}_1 = 0.9847, \hat{\lambda}_2 = 0.9235$$

The larger third core set gives a more appropriate result for the second non-trivial eigenvalue λ_2, but we pay by losing accuracy in the approximation of the slowest timescale. Theorem 7 can not explain this effect that the larger third core set gives a better estimate for λ_2 since $\max\{\|Q^\perp u_1\|, \|Q^\perp u_2\|\}$ is larger for this setting. Therefore, we need

Theorem 9
Let T be a self-adjoint transfer operator and Q the orthogonal projection to a subspace D with $\mathbb{1} \in D$. Let λ be an eigenvalue of T and u the corresponding normalized eigenvector and set $\delta = \|Q^\perp u\|$. Then there exists an eigenvalue $\hat{\lambda}$ of the projected transfer operator QTQ with

$$|\lambda - \hat{\lambda}| \leq \lambda_1 \delta (1 - \delta^2)^{-\frac{1}{2}}.$$

Proof. For $\lambda = 1$ it is trivial, so $\lambda < 1, u \neq \mathbb{1}$. Since T is self-adjoint, also QTQ is self-adjoint on a finite dimensional space. Therefore, we have an orthonormal basis of eigenvectors $\hat{u}_1, ..., \hat{u}_n$ and real eigenvalues $\hat{\lambda}_1, ..., \hat{\lambda}_n$ and

$$QTQu = \sum_{i=1}^{n} \hat{\lambda}_i \langle u, \hat{u}_i \rangle \hat{u}_i.$$

On the other hand we have

$$\begin{aligned} QTQu &= QTu - QTQ^\perp u = \lambda Qu - QTQ^\perp u \\ &= \lambda \sum_{i=1}^{n} \langle u, \hat{u}_i \rangle \hat{u}_i - QTQ^\perp u. \end{aligned}$$

Putting both equations together we get

$$QTQ^\perp u = \sum_{i=1}^{n} (\lambda - \hat{\lambda}_i) \langle u, \hat{u}_i \rangle \hat{u}_i.$$

Therefore,

$$\begin{aligned}\|QTQ^{\perp}u\|^2 &= \sum_{i=1}^{n}(\lambda-\hat{\lambda}_i)^2\langle u,\hat{u}_i\rangle^2 \\ &\geq \min_{i=1,\ldots,n}\{(\lambda-\hat{\lambda}_i)^2\}\sum_{i=1}^{n}\langle u,\hat{u}_i\rangle^2 = \min_{i=1,\ldots,n}\{(\lambda-\hat{\lambda}_i)^2\}\|Qu\|^2 \\ &= \min_{i=1,\ldots,n}\{(\lambda-\hat{\lambda}_i)^2\}(1-\delta^2)\end{aligned}$$

So, there exists an eigenvalue $\hat{\lambda}$ with

$$(\lambda-\hat{\lambda})^2 \leq \|QTQ^{\perp}u\|^2(1-\delta^2)^{-1}.$$

Moreover,

$$\|QTQ^{\perp}u\|^2 \leq \|QTQ^{\perp}\|^2\|Q^{\perp}u\|^2 \leq \lambda_1^2\delta^2,$$

since $Q^{\perp}u_0 = Q^{\perp}\mathbb{1} = 0$. Taking the square root completes the proof. □

Theorem 9 gives us the opportunity to approximate each timescale completely separately from each other. Sadly, we do not get a second order dependence on δ like in Theorem 7, in general, if we do not assume anything on the projection error of other eigenvectors. On the other hand, in the example above we could also use the following two core sets (Figure 3.23) in order to get the same projection error $\|Q^{\perp}u_2\|$, when we just focus on the timescale that belongs to λ_2.

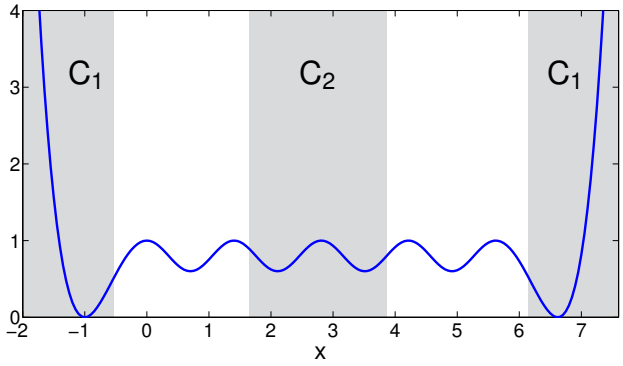


Figure 3.23: Just two core sets to approximate the second timescale.

The projected transfer operator QTQ would have the two eigenvalues

$$\hat{\lambda}_0 = 1 \qquad \hat{\lambda}_1 = 0.9235.$$

An advantage of taking two-dimensional approximations is that in this special case we come back to the second order dependence on δ.

Theorem 10
Let T be a self-adjoint transfer operator and Q the orthogonal projection to a two-dimensional subspace D with $\mathbb{1} \in D$. Let λ be an eigenvalue of T and u the corresponding normalized eigenvector and set $\delta = \|Q^\perp u\|$. Let the smallest negative eigenvalue of T be given by λ^-.
Then for the non-trivial eigenvalue $\hat{\lambda}$ of the projected transfer operator QTQ it holds

$$|\lambda - \hat{\lambda}| \leq \max\{\lambda_1 - \lambda, \lambda - \lambda^-\}\delta^2(1-\delta^2)^{-1}.$$

Proof. From the proof of Theorem 9 we get

$$QTQ^\perp u = (\lambda - \hat{\lambda})\langle u, \hat{u}\rangle \hat{u} = (\lambda - \hat{\lambda})Qu,$$

where $\hat{u}$ is the eigenvector to the only non-trivial eigenvalue $\hat{\lambda}$. On the other hand,

$$\langle Qu, \mathbb{1}\rangle = \langle u, Q\mathbb{1}\rangle = \langle u, \mathbb{1}\rangle = 0,$$

which means, that $\{1, \frac{Qu}{\|Qu\|}\}$ is an orthonormal basis of D. Therefore,

$$QTQ^\perp u = \langle TQ^\perp u, \mathbb{1}\rangle \mathbb{1} + \frac{1}{\|Qu\|^2}\langle TQ^\perp u, Qu\rangle Qu = \frac{1}{\|Qu\|^2}\langle TQ^\perp u, Qu\rangle Qu.$$

Combination with the first equation yields

$$\begin{aligned}
\lambda - \hat{\lambda} &= \frac{1}{\|Qu\|^2}\langle TQ^\perp u, Qu\rangle = \frac{1}{\|Qu\|^2}\langle Q^\perp u, Q^\perp TQu\rangle \\
&= \frac{1}{\|Qu\|^2}(\langle Q^\perp u, Q^\perp Tu\rangle - \langle Q^\perp u, Q^\perp TQ^\perp u\rangle) \\
&= \frac{1}{\|Qu\|^2}(\lambda\|Q^\perp u\|^2 - \langle Q^\perp u, TQ^\perp u\rangle) \\
&\leq \frac{1}{\|Qu\|^2}(\lambda - \lambda^-)\|Q^\perp u\|^2 = (\lambda - \lambda^-)\delta^2(1-\delta^2)^{-1}.
\end{aligned}$$

Moreover,

$$\hat{\lambda} - \lambda = \frac{1}{\|Qu\|^2}(\langle Q^\perp u, TQ^\perp u\rangle - \lambda\|Q^\perp u\|^2) \leq (\lambda_1 - \lambda)\delta^2(1-\delta^2)^{-1}.$$

□

Statistically, when it comes to the estimation of a matrix representations of the operator QTQ, there is also an advantage in joining the two outer core

sets and considering a two-dimensional approximation. We namely have to count transitions in the sense of milestoning between the core sets and on one hand, we will simply observe more transitions between the two core sets when we treat the outer sets as one core set, and on the other hand, we only have to estimate two instead of six entries of the stochatic matrices $\hat{T}$ and M.
In general, if we were only interested in one particular timescale, i.e an eigenvalue λ and an eigenvector u, we would know already a perfect subspace D with $\mathbb{1} \in D$, which is only two-dimensional, namely $D = \{\mathbb{1}, u\}$. Here is $\delta = \|Q^{\perp}u\| = 0$, which implies that the projected transfer operator QTQ would have the eigenvalues $\hat{\lambda}_0 = 1, \hat{\lambda}_1 = \lambda$. In the example above it was possible to approximate the eigenvalue u well by two core sets and their corresponding committors but this was only possible because of the symmetric nature of the potential. Therefore, u_2 took the same values in the outer left and outer right minimum. Of course, we cannot expect this in general.

3.3 Consequences for Markov State Modeling

In the last two sections we have discussed the approximation properties of projected transfer operators QTQ. In Section 3.1 we focused on the question if the discrete semi-group $\{(QTQ)^k\}_{k\in\mathbb{N}}$ is a valid simplification of the family $\{QT_{k\tau}Q\}_{k\in\mathbb{N}}$ which is no semi-group. In Section 3.2 we analyzed the inheritance of the longest timescales of the system by the projected operator, i.e. we compared the dominant part of the spectrum of T with the largest eigenvalues of QTQ. Both questions led to the projection error

$$\delta = \max_{i=1,\dots,m} \|Q^{\perp}u_i\| \tag{3.69}$$

of the dominant eigenvectors to the subspace D. We have seen that independently of the lag time $\tau > 0$ it is essential that the subspace D is chosen such that these first m eigenvectors can be well approximated by functions in D. This has direct consequences for the projected transfer operators that come from Markov State Modeling.

For a Standard Markov State Model we have to make sure that our sets $A_1, ..., A_n$, which have to partition, i.e. cover the whole state space, are chosen such that the considered dominant eigenvectors are almost constant on these sets. Now, this criterion is not very constructive because we do not know how constant an eigenvector is on a certain set A_i because we cannot compute them. Nevertheless, we have seen that for processes in potentials which are not completely trivial we will most certainly not be able to find only few large sets with this property. Particularly, a threshold on δ will soon lead to exponential growth of the number of sets n with the dimension of state space. Therefore, for processes in higher dimensional state spaces a

construction of a Standard Markov State Model and the associated subspace D can quickly become very costly.

In [82, 44, 83] a variant has been developed where

$$D = span\{f_1, ..., f_n\} = span\{u_0, ..., u_{n-1}\} \Rightarrow Q = \Pi, \delta = 0 \tag{3.70}$$

with non-negative ansatz functions $f_1, ..., f_n$. In [19] it is also shown how to optimally compute this basis. However, for this approach the eigenvectors have to be known exactly.

Hence, we considered the subspace D spanned by committors with respect to core sets. In our examples this method seemed to be superior than Standard Markov State Modeling. For the double-well potential with extended transition region we needed only two sets around the main minima to get $\delta \approx 10^{-2}$. Of course, now we also have to answer the questions: why does this happen, and how do we find core sets that lead to a good approximation, in general? As for Standard Markov State Models our subspace D depends on the choice of these sets that not necessarily partition the state space anymore. The answer that they have to be chosen such that δ is minimized has become at first glance even more ridiculous. In contrast to the full partition approach not only the eigenvectors are unknown but we also cannot compute the committors, which form the subspace for projection. That is, the new core set approach and the δ criterion seem to be a maximally unpractical combination because no object in $\|Q^{\perp}u_i\|$ can be computed.

Therefore, it is surprising that we can analyze $\|Q^{\perp}u_i\|$ further to understand under which conditions on the core sets this error will be small for the dominant eigenvectors. We start with

Theorem 11
Let $1 = \lambda_0 > \lambda_1 > ... > \lambda_{m-1}$ be the m dominant eigenvalues of T, i.e. for every other eigenvalue λ it holds $\lambda < \lambda_{m-1}$. Let $u_0, u_1, ..., u_{m-1}$ be the corresponding normalized eigenvectors and $D \subset L^2(\mu)$ an infinite dimensional subspace of the form

$$D = \{v \in L^2(\mu) | v(x) = c_j \forall x \in C_j, c_j \in \mathbb{R}, j = 1, ..., n\} \tag{3.71}$$

for a finite number of arbitrary, but fixed disjoint sets $C_j \subset E$. That is, $v \in D$, if v is constant on each set C_j.
Let Q be the orthogonal projection onto D and let $1 = \hat{\lambda}_0 > \hat{\lambda}_1 > ... > \hat{\lambda}_{m-1}$ be the dominating eigenvalues of the projected operator QTQ. Then

$$\max_{i=1,...,m-1} |\lambda_i - \hat{\lambda}_i| \leq \lambda_1 (m-1)\delta^2, \tag{3.72}$$

where

$$\delta^2 = \max_{i=1,...,m-1} \|Q^{\perp} u_i\|^2 = \sum_{j=1}^{n} \int_{C_j} \left(u_i - \frac{1}{\mu(C_j)} \int_{C_j} u_i d\mu \right)^2 d\mu.$$

is the maximal projection error of the eigenvectors to the space D.

Proof. Let $\epsilon > 0$ be arbitrary and $\hat{u}_0, ..., \hat{u}_{m-1}$ be the normalized eigenvectors of QTQ w.r.t. the eigenvalues $1 = \hat{\lambda}_0 > \hat{\lambda}_1 > ... > \hat{\lambda}_{m-1}$. As $u_i \in L^2(\mu)$ and $\hat{u}_i \in L^2(\mu)$ there must be a compact set K with $C_1, ..., C_n \subset K$ and

$$\int_{E \setminus K} u_i^2 d\mu \leq \epsilon \qquad \int_{E \setminus K} \hat{u}_i^2 d\mu \leq \epsilon \qquad \forall i = 0, ..., m-1.$$

Let us define $K_C := K \setminus \left(\bigcup_{j=1}^{n} C_j \right)$. Then, because $u_i \in L^2(\mu)$ and $\hat{u}_i \in L^2(\mu)$, these eigenvectors can be arbitrarily well approximated by stepfunctions, that is, there is a partitioning $A_1, ..., A_N$ of K_C, i.e.

$$A_i \cap A_j = \emptyset \qquad \bigcup_{j=1}^{N} A_j = K_C,$$

such that

$$\int_{K_C} (u_i - P_N u_i)^2 d\mu \leq \epsilon \qquad \int_{K_C} (\hat{u}_i - P_N \hat{u}_i)^2 d\mu \leq \epsilon \qquad \forall i = 0, ..., m-1, \tag{3.73}$$

where P_N is the orthogonal projection onto the finite dimensional space

$$V_N = span\{\mathbb{1}_{A_1}, ..., \mathbb{1}_{A_N}, 1_{C_1}, ..., 1_{C_n}, \mathbb{1}_{E \setminus K}\}.$$

Moreover, for this projection we obviously have

$$QP_N = P_N, \tag{3.74}$$

because a function $v \in V_N$ is always constant on the sets $C_1, ..., C_n$ and therefore in D.
This implies also

$$P_N = P_N^* \overset{(3.74)}{=} (QP_N)^* = P_N^* Q^* = P_N Q, \tag{3.75}$$

We now denote with $\lambda_i(A)$ the ith largest eigenvalue of an operator A. Then we have

$$\begin{aligned}|\lambda_i - \hat{\lambda}_i| &= |\lambda_i(T) - \lambda_i(QTQ)| = |\lambda_i(T) - \lambda_i(P_NTP_N) + \lambda_i(P_NTP_N) - \lambda_i(QTQ)| \\ &\leq |\lambda_i(T) - \lambda_i(P_NTP_N)| + |\lambda_i(P_NTP_N) - \lambda_i(QTQ)| \\ &\overset{(3.74),(3.75)}{=} |\lambda_i(T) - \lambda_i(P_NTP_N)| + |\lambda_i(P_NQTQP_N) - \lambda_i(QTQ)|.\end{aligned} \tag{3.76}$$

As P_N is a projection onto a finite dimensional subspace V_N and $\mathbb{1} \in V_N$, we can apply Theorem 7 to get estimates for each of the summands. First, for every $i = 1, ..., m-1$

$$\begin{aligned}&|\lambda_i(T) - \lambda_i(P_NTP_N)| \leq \lambda_1(m-1) \max_{k=1,...,m-1} \|P_N^{\perp} u_k\|^2 \\ &= \lambda_1(m-1) \max_{k=1,...,m-1} \Bigg[\sum_{j=1}^{n} \int_{C_j} (u_k - P_N u_k)^2 d\mu + \int_{K_C} (u_k - P_N u_k)^2 d\mu \\ &+ \int_{S \setminus K} (u_k - P_N u_k)^2 d\mu \Bigg] \\ &\leq \lambda_1(m-1) \max_{k=1,...,m-1} \left[\sum_{j=1}^{n} \int_{C_j} (u_k - Qu_k)^2 d\mu + 2\epsilon \right] \\ &= \lambda_1(m-1) (\max_{k=1,...,m-1} \|Q^{\perp} u_k\|^2 + 2\epsilon)\end{aligned} \tag{3.77}$$

Using the same calculation for the second summand we get

$$\begin{aligned}&|\lambda_i(P_NQTQP_N) - \lambda_i(QTQ)| \leq \hat{\lambda}_1(m-1) \max_{k=1,...,m-1} \|P_N^{\perp} \hat{u}_k\|^2 \\ &= \hat{\lambda}_1(m-1) \max_{k=1,...,m-1} \Bigg[\sum_{j=1}^{n} \int_{C_j} (\hat{u}_k - P_N \hat{u}_k)^2 d\mu + \int_{K_C} (\hat{u}_k - P_N \hat{u}_k)^2 d\mu \\ &+ \int_{E \setminus K} (\hat{u}_k - P_N \hat{u}_k)^2 d\mu \Bigg] \leq \hat{\lambda}_1(m-1) 2\epsilon,\end{aligned} \tag{3.78}$$

because $P_N \hat{u}_k = \hat{u}_k$ on the sets $C_1, ..., C_n$, since $\hat{u}_k \in D$.
Inserting (3.77) and (3.78) into (3.76) yields for every $i = 1, ..., m-1$

$$|\lambda_i - \hat{\lambda}_i| \leq \lambda_1(m-1)(\max_{k=1,...,m-1} \|Q^{\perp} u_k\|^2 + 2\epsilon) + \hat{\lambda}_1(m-1) 2\epsilon.$$

As ϵ was arbitrary and

$$(Qu_k)(x) = \begin{cases} u_k(x), & \text{if } x \notin \bigcup_{j=1}^{n} C_j, \\ \frac{1}{\mu(C_j)} \int_{C_j} u_k d\mu, & \text{if } x \in C_j \end{cases}$$

the proof is complete. □

We will further need the following

Lemma 3
Let $v \in Ł^2(\mu)$ be the solution of

$$\begin{aligned} Av &= 0, \ \text{on } C \\ v &= g, \ \text{on } E \setminus C, \end{aligned} \tag{3.79}$$

with $A = L$ or $A = Id - T_t$ and $g \in L^\infty(\mu), g \neq 0$ on $E \setminus C$. Then,

$$\|v\|_\infty := \max_{y \in E} |v(y)| \leq \max_{y \in E \setminus C} |g(y)|.$$

Proof. The linear system above is equivalent to

$$\Theta A \Theta v = -\Theta A \Theta^\perp v = -\Theta A \Theta^\perp g \tag{3.80}$$

with

$$(\Theta v)(x) = \begin{cases} v(x), & x \in C \\ 0, & \text{else.} \end{cases} \tag{3.81}$$

As in the proof of Theorem 2 it has to be uniquely solvable because otherwise we could construct an invariant measure that vanishes on $E \setminus C$.

Take now $A = L$. Dynkin formula [51] applied to the solution v yields for $x \in C$ and the stopping time $\tau = \inf_{t \geq 0}\{X_t \in E \setminus C\}$

$$\mathbb{E}[v(X_\tau)|X_0 = x] = v(x) + \mathbb{E}[\int_0^\tau (Lv)(X_s)ds|X_0 = x] = v(x)$$

because $X_s \in C$ for all $s \in (0, \tau)$ and therefore $(Lv)(X_s) = 0$. On the other hand, we have

$$X_\tau \in E \setminus C \Rightarrow v(x) = \mathbb{E}[v(X_\tau)|X_0 = x] = \mathbb{E}[g(X_\tau)|X_0 = x].$$

Obviously, it holds

$$|\mathbb{E}[g(X_\tau)|X_0 = x]| \leq \max_{y \in E \setminus C} |g(y)| \ \forall x \in C,$$

which proves the assertion.

For $A = (Id - T)$ we find with the discrete version of Dynkin formula [50] and $\tau \in \mathbb{N}$

$$\mathbb{E}[v(X_\tau)|X_0 = x] = v(x) + \mathbb{E}[\sum_{k=0}^{\tau-1} (-Av)(X_k)|X_0 = x] = v(x)$$

and $\|v\|_\infty \leq \max_{y \in E \setminus C} |g(y)|$ with the same reasoning as above. □

Now we assume that we have some core sets $C_1, ...C_n$ and the corresponding committors, that solve

$$\begin{aligned} Tq_i &= q_i \text{ on } C, \\ q_i(x) &= 1, \text{ on } C_i \\ q_i(x) &= 0, \text{ on } C_k, k \neq i. \end{aligned}$$

The space spanned by the committors can be written as

$$D = span\{q_1, ..., q_m\} = \{v \mid v \text{ is constant on each set } C_j, Tv = v \text{ on } C\}.$$

Then we can estimate the projection error $\|Q^{\perp}u\|$ for any eigenvector u.

Theorem 12
Take the setting from above, i.e. let D be the space spanned by committors with respect to n core sets $C_1, ..., C_n$. Denote with Q the orthogonal projection onto D. Let λ be an eigenvalue of T and u the corresponding, normalized eigenvector. Then

$$\|Q^{\perp}u\| \leq p(u) + 2\mu(C)p_{max}(u) + r(C)(1-\lambda)\left(\int_C u^2 d\mu\right)^{\frac{1}{2}}$$

with

$$\begin{aligned} r(C) &= \sup_{\substack{\|v\|=1, \\ v=0 \text{ on } E\setminus C}} \left(\frac{1}{\int_C (v - Tv)^2 d\mu}\right)^{1/2} \\ p(u) &= \|P^{\perp}u\| \\ p_{max}(u) &= \|P^{\perp}u\|_{\infty} \\ (Pu)(x) &= \begin{cases} u(x), & \text{if } x \in C, \\ \frac{1}{\mu(C_j)} \int_{C_j} u d\mu, & \text{if } x \in C_j. \end{cases} \end{aligned} \tag{3.82}$$

Proof. Take the projection P onto the space $V = \{v \in L^2(\mu)) | v(x) = c_j \forall x \in C_j, c_j \in \mathbb{R}, j = 1, ..., n\}$ of functions, which are constant on the core sets.
First, $\|Q^{\perp}u\| = \|u - Qu\|\| \leq \|u - q\|$ for every $q \in D$, as Qu is the bestapproximation. Take the interpolating $q \in D$, that is a solution of

$$\begin{aligned} Tq &= q \text{ on } C, \\ q &= Pu, \text{ on } E \setminus C. \end{aligned} \tag{3.83}$$

As $q \in V$ we have $Pq = q$. Therefore (3.83) is equivalent to

$$\begin{aligned} PTPq &= q \text{ on } C, \\ q &= Pu, \text{ on } E \setminus C. \end{aligned} \tag{3.84}$$

Moreover, for the projection Pu

$$PTPu = PTu - PTP^{\perp}u = \lambda Pu - PTP^{\perp}u.$$

Therefore the error $e := Pu - q$ solves

$$\begin{aligned} (Id - PTP)e &= (1-\lambda)Pu + PTP^{\perp}u \text{ on } C, \\ e &= 0, \text{ on } E \setminus C. \end{aligned}$$

This means, $e \in E_\Theta = \{v | v(x) = 0, x \in E \setminus C\} \subset E$ fulfills

$$\Theta(Id - PTP)\Theta e = (1-\lambda)\Theta Pu + \Theta PTP^{\perp}u \tag{3.85}$$

with

$$\Theta v(x) = \begin{cases} v(x), & x \in C \\ 0, & x \in E \setminus C \end{cases}.$$

Obviously it holds $P\Theta = \Theta P = \Theta$. Thus, (3.85) is equivalent to

$$Re := \Theta(Id - T)\Theta e = (1-\lambda)\Theta u + \Theta TP^{\perp}u \tag{3.86}$$

Now R has to be invertible on E_Θ because if it was not, there would be some $v \in E_\Theta$ satisfying

$$Rv = 0,$$

which would imply

$$\begin{aligned} Tv &= v \text{ on } C, \\ v &= 0, \text{ on } E \setminus C. \end{aligned}$$

But then it must hold $Tv = 0$ on $E \setminus C$, because otherwise we would have $\|Tv\| > \|v\|$. This would imply $Tv = v$ on E, which is a contradiction to the unique, positive invariant measure.
So we can write

$$e = R^{-1}(1-\lambda)\Theta u + R^{-1}\Theta TP^{\perp}u.$$

R is self-adjoint, because T is, and therefore $\|R^{-1}\| = \frac{1}{\kappa}$, where κ is the smallest eigenvalue of R, i.e. there is a vector $v \in E_\Theta, \|v\| = 1$ with

$$\Theta(Id - T)\Theta v = \kappa v.$$

Now we have

$$\begin{aligned} \kappa^2 &= \int_E (\kappa v)^2 d\mu = \int_E (\Theta(Id - T)\Theta v)^2 d\mu \\ &= \int_C ((Id - T)v)^2 d\mu. \end{aligned}$$

This implies

$$\|R^{-1}\| \leq \frac{1}{\min\limits_{v \in E_\Theta, \|v\|=1} \left(\int\limits_C (v - Tv)^2 d\mu \right)^{\frac{1}{2}}}.$$

Moreover, $\Theta T P^\perp = \Theta(Id - T)P^\perp$, which gives

$$\|R^{-1}\Theta T P^\perp u\| = \|R^{-1}\Theta(Id - T)P^\perp u\| = \|\Theta f\|,$$

where Θf solves

$$R\Theta f = \Theta(Id - T)\Theta f = \Theta(Id - T)P^\perp u \Leftrightarrow \Theta(Id - T)(\Theta f - P^\perp u) = 0.$$

That is, $w := \Theta f - P^\perp u$ is the solution of

$$\begin{aligned} (Id - T)w &= 0 \text{ on } C \\ w &= -P^\perp u \text{ on } E \setminus C \end{aligned}$$

Lemma 3 now implies that $\|w\|_\infty = \|P^\perp u\|_\infty$ and therefore

$$\|R^{-1}\Theta T P^\perp u\| = \|\Theta f\| \leq \mu(C)\|P^\perp u + w\|_\infty \leq 2\mu(C)\|P^\perp u\|_\infty.$$

So,

$$\begin{aligned} \|e\| &= \|R^{-1}(1-\lambda)\Theta u + R^{-1}\Theta T P^\perp u\| \leq \|R^{-1}\|\|(1-\lambda)\Theta u\| + \|R^{-1}\Theta T P^\perp u\|) \\ &= r(C)(1-\lambda)\left(\int\limits_C u^2 d\mu \right)^{\frac{1}{2}} + 2\mu(C)\|P^\perp u\|_\infty. \end{aligned}$$

Further,

$$\|Q^\perp u\| = \|u - Qu\| \leq \|u - q\| \leq \|u - Pu\| + \|Pu - q\| = p(u) + \|e\|.$$

Putting all together completes the proof. □

In Theorem 12 the committors are solutions of

$$\begin{aligned} T_t q_i &= q_i \text{ on } C, \\ q_i(x) &= 1, \text{ on } C_i \\ q_i(x) &= 0, \text{ on } C_k, k \neq i \end{aligned}$$

where t denotes the time resolution at which we can see whether a core set was hit or not. We can also formulate the theorem for time-continuous committors.

Theorem 13
Let L be the generator of the process and let D be the space spanned by time-continuous committors with respect to n core sets $C_1, ..., C_n$, i.e.

$$D = span\{q_1, ..., q_n\}$$

$$\begin{aligned} Lq_i &= 0 \text{ on } C, \\ q_i(x) &= 1, \text{ on } C_i \\ q_i(x) &= 0, \text{ on } C_k, k \neq i \end{aligned}$$

Denote with Q the orthogonal projection onto D. Let Λ be an eigenvalue of the generator L and u the corresponding, normalized eigenvector. Then,

$$\|Q^{\perp}u\| \leq p(u) + 2\mu(C)p_{max}(u) - r(C)\Lambda \left(\int_C u^2 d\mu \right)^{\frac{1}{2}}$$

with

$$\begin{aligned} r(C) &= \sup_{\substack{\|v\|=1, \\ v=0 \text{ on } E\setminus C}} \left(\frac{1}{\int_C (Lv)^2 d\mu} \right)^{1/2} \\ p(u) &= \|P^{\perp}u\| \\ p_{max}(u) &= \|P^{\perp}u\|_{\infty} \\ (Pu)(x) &= \begin{cases} u(x), & \text{if } x \in C, \\ \frac{1}{\mu(C_j)} \int_{C_j} u d\mu, & \text{if } x \in C_j. \end{cases} \end{aligned} \tag{3.87}$$

The proof is analog to the proof of Theorem 12.
Note that

$$(P^{\perp}u)(x) = \begin{cases} 0, & \text{if } x \in C, \\ u(x) - \frac{1}{\mu(C_j)} \int_{C_j} u d\mu, & \text{if } x \in C_j. \end{cases}$$

That is, the terms $p(u)$ and $p_{max}(u)$ measure how constant the eigenvector on the core sets is. If the eigenvector is not bounded, i.e. $u \notin L^{\infty}$, we assume that we do not consider core sets, where the eigenvector is growing unboundedly such that $p_{max}(u)$ does not exist.

Interpretation of the inequality

First, except for $p(u)$ and $p_{max}(u)$ the inequality consists of objects that depend on the choice of $\bigcup_{i=1}^{n} C_i$ only, that is, it is important if a point in state space belongs to any core set or not but the clustering, i.e. the specific assignment to one core set, only enters into the error $p(u)$ and $p_{max}(u)$.

In order to achieve a small bound we have to make sure that the two summands

$$-r(C)\Lambda \left(\int_C u^2 d\mu \right)^{\frac{1}{2}} \qquad p(u) + 2\mu(C) p_{max}(u) \tag{3.88}$$

are small. Remember that for a Markov process which is distributed at time t according to the measure with density v_t we have

$$\frac{d}{dt} v_t = L v_t. \tag{3.89}$$

If we start the process only distributed in the region C, there has to be an infinitesimal change of probability because of the flow from C into the core sets. Even if the process was perfectly equilibrated within C, this change in probability distribution could not be avoided. This is exactly what the factor $r(C)$ measures. It will be small if the process leaves the region C "quickly enough". Now, the first term in (3.88) tells us what quickly enough means. It compares the attractiveness of the core sets $r(C)$ with the eigenvalue of the corresponding timescale that we want to approximate. So if we start outside of the core sets, the timescale at which probability has to flow back into the core sets should be shorter as the timescale of interest. This also implies that the more timescales we want to approximate, the larger the region of core sets has to be in order to increase the overall attractiveness. Having found an appropriate set C we have to cluster the region $E \setminus C$ into core sets $C_1, ..., C_n$ such that on each core set the dominant eigenvectors, which we want to approximate, are almost constant in order to guarantee small $p(u)$ and $p_{max}(u)$.

Theoretical considerations using diffusion examples

We want to think about the identification of core sets which have the properties that have been motivated by Theorem 13. As we have seen, we can perform this identification in two steps. First, we have to split the state space into sets C and $E \setminus C$, and second, we have to cluster the set $E \setminus C$ into core sets $C_1, ..., C_n$. Note that once we sorted out the region C the second clustering step should not be very difficult because the core sets will always be dynamically well separated. So we will focus on the question

which states do we have to include into core sets, and which states can be considered as transition region.

Assume that we have a diffusion in an energy landscape with noise intensity σ that we call "process 1". Now, we look at the same process with increased noise intensity, e.g. where σ is multiplied with the factor $\sqrt{2}$, that will be named "process 2". Sets which are metastable and attractive with respect to process 2 are even more metastable and attractive with respect to process 1. Moreover, if process 1 had the invariant measure with density μ, the invariant measure of process 2 would be given by the density

$$\mu^*(x) = \frac{1}{Z}\sqrt{\mu(x)}, \tag{3.90}$$

where Z is a normalization constant, in this example. The density μ^* has the same structure, e.g. the same local minima and maxima but it is less peaked. So, if we take this density as initial distribution for process 1 and let it propagate the ensemble, it will have to converge to invariant measure μ again. The way how μ^* is being propagated towards equilibrium will provide the information we are looking for.

First, take the diffusion in the simple double-well potential from Fig. 3.1. We choose a time step $\alpha = 0.1$ and look at the density μ^* and its propagation under $T = T_\alpha$ in the following figure.

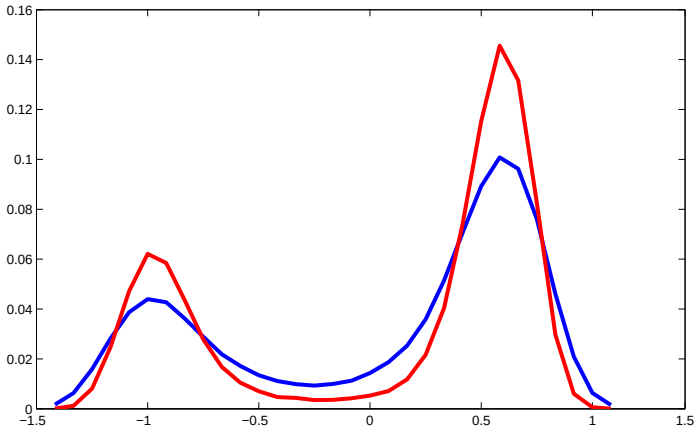


Figure 3.24: Blue: μ^*, Red: $T\mu^*$

The set

$$C^\alpha = \{x \in E, T_\alpha \mu^*(x) > \mu^*(x)\} \tag{3.91}$$

obviously identifies the two regions around the wells of the potential. This is not surprising because the ensemble distributed according to μ^* is slightly too uniform on state space and has to relax from the transition region to the

minima. The transfer from the minima back to the transition region and between the wells is very small on that timescale. So the set from (3.91) would be a good candidate for $E \setminus C$, i.e. core sets, because by construction this set is attractive within the time step α.

Let us now consider the more comprehensive example of the double-well potential with extended transition region as shown in Fig. 3.9. Again we compare the density μ^* to its propagation for a certain time step. The results for three different time steps $\alpha = 1, 10, 20$ are illustrated below.

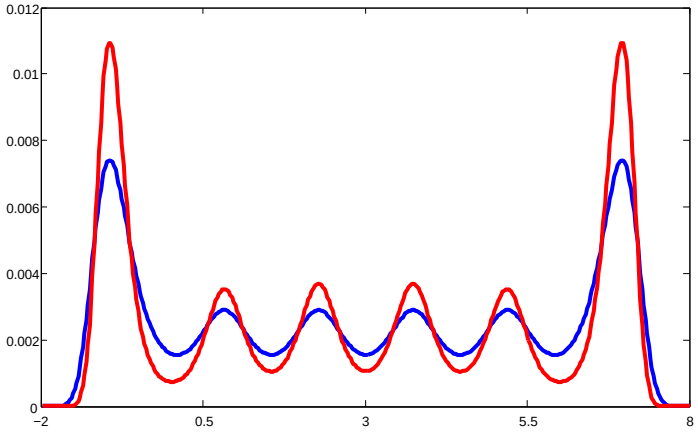


Figure 3.25: Blue: μ^*, Red: $T_\alpha \mu^*$, $\alpha = 1$

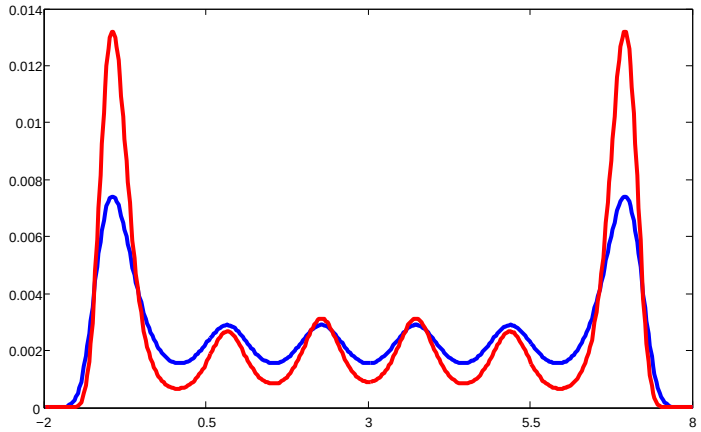


Figure 3.26: Blue: μ^*, Red: $T_\alpha \mu^*$, $\alpha = 10$

For the shortest time step $\alpha = 1$ all peaks of the invariant measure are identified (Fig. 3.25). It is clear that this will always happen for short

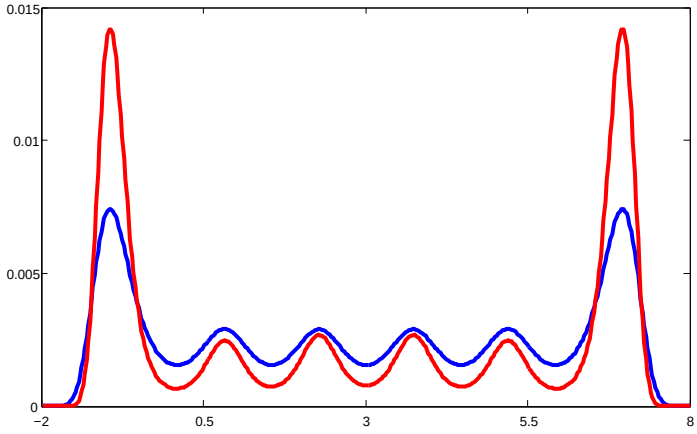


Figure 3.27: Blue: μ^*, Red: $T_\alpha \mu^*$, $\alpha = 20$

enough time steps because locally the density μ^* is too flat at peaks and will increase on a very short timescale. In Fig. 3.26 and Fig. 3.27 we increase the time step and as expected we see the effect of the larger timescales. For $\alpha = 10$ there is still some region in the middle of the potential that would belong to the candidate for core sets as in (3.91) and for $\alpha = 20$ only the two main wells would satisfy equation (3.91) anymore.
Here the approach connects to our inequality from Theorem 13. We have seen that the core sets should be attractive sets but this attractiveness was relative to the timescale of interest $(-r(C)\Lambda)$. So if we only want to approximate the slowest dynamics, the core sets must be attractive just on a slower timescale than the one which is implied by Λ_1. $\alpha = 20$ would be enough for this goal and the set $\{x \in S, T_\alpha \mu^*(x) > \mu^*(x)\}$ with $\alpha = 20$, which gives the main wells, would be a good candidate for core sets. With the obvious clustering of set C^α (3.91) into 2 core sets we can now compute the eigenvalues of T and QTQ for time-continuous committors. The implied dominant timescales are given by

$$-1/\Lambda_1 = 86.7546 \qquad -1/\hat{\Lambda}_1 = 86.3772.$$

The next implied timescale is $-1/\Lambda_2 = 12.49$. In order to approximate this timescale, too, we have to decrease the time step α in order to measure attractiveness. We have seen already that for $\alpha = 10$ a set in the middle of transition region is introduced. This is reasonable because it is where the maximum of the second non-trivial eigenvector u_2 is located. Therefore, this set becomes attractive on the shorter timescale which is recognized in the sense of (3.91). For the shortest time step $\alpha = 1$ we finally get core sets which lead to a good approximation of the first five eigenvalues and the implied timescales.

	$-1/\Lambda_1$	$-1/\Lambda_2$	$-1/\Lambda_3$	$-1/\Lambda_4$	$-1/\Lambda_5$
T	86.75	12.76	4.26	2.15	1.42
QTQ	84.30	12.49	4.16	2.10	1.41

Note that we lose approximation quality again for the largest eigenvalue compared to the two core sets (Fig. 3.27). Theorem 13 claims to assign more states of state space to core sets for the approximation of shorter timescales. On the other hand, eigenvectors have to be constant on every core set and for the very dominant eigenvectors this will be a constraint that introduces a larger error if we have too many or too large core sets. We discussed this problem already in Section 3.2 and proposed to consider different levels of discretization. That is, we could calculate the spectrum of the operator QTQ for $\alpha = 1$ with 6 core sets first, and then increase the time step to $\alpha \geq 20$ in order to get an even better approximation of the slowest process.

3.4 Simulation based Algorithm: Building Markov State Models using Core Sets

Finally, we want to perform a complete analysis of the dominant timescales of a Markov process based on simulation. That is, we want to identify the region of core sets $E \setminus C$, cluster this region into n sets $C_1, ..., C_n$, and estimate a matrix representation of the projected transfer operator QTQ with respect to the associated committors. We have seen already how we can realize the last task for given core sets if we have a realization of the process $(X_{kh}), k = 1, ..., N$ at some time resolution h with N data points that we call (x_k). Then, we can estimate the matrices $\hat{T}$ and M (2.40) as described in Section 2.3. In order to develop an algorithmic approach for the identification of the core sets, the first observation is that we only have to tell for the data points (x_k) if they belong to core sets or not. That is, we do not have to split the whole state space E into C and $E \setminus C$ if we want to use the trajectory (x_k) afterwards for the estimation of $\hat{T}$ and M. For this purpose, it would be equivalent to simply find out which pieces of the trajectory (x_k) lie in core sets and which do not. In the last Section 3.3 we formulated the properties core sets should have in order to imply a good approximation result. In short, the cores of the most metastable regions should be included and they have to be relatively attractive compared to the timescales of interest. In principle, there are two possibilities how to analyze which points of the trajectory (x_k) fulfill these conditions. If we have additional knowledge about the Markov process, we could try to analytically gain insight into the dynamics around the points (x_k). This will usually end up in the construction of a network between the points (x_k) which should make it possible to measure metastability and attractivity in the sense of Section 3.3. Here, we will follow a different approach. We do not even

assume that we have an expression for the dynamics of the process that we can analyze further. We rather assume that the Markov process $(X_t) = (X_t^\xi)$ depends on a parameter $\xi > 0$, which controls the overall metastability of the process, i.e. for smaller ξ the metastability should increase. Then, we use the following algorithm for the identification of the core sets along the trajectory (x_k).

Algorithm 1: Principle Core Set Identification

Data: $(x_k), k = 1, ..., N; \alpha > 0; \beta \in \mathbb{N}; \rho > 0$
Result: $\hat{T}, M$
$y_k := x_{k\beta}, k = 1, ..., N/\beta$;
for $k = 1$ **to** N/β **do**
  simulate $z_k := X_\alpha^{\xi/2}$ with $X_0^{\xi/2} = y_k$;
end
for $k = 1$ **to** N **do**
  if $\#\{z_k \in B_\rho(x_k)\} > \#\{y_k \in B_\rho(x_k)\}$ **then**
    $x_k \to E \setminus C$;
  else
    $x_k \to C$;
  end
end
cluster$(E \setminus C)$;
estimate $\hat{T}, M$ as in Section 2.3;

So, we perform the identification in the following way. First, we take some data points $y_k = x_{k\beta}, k = 1, ..., N/\beta$ for a fixed $\beta \in \mathbb{N}$ out of the trajectory. Since we assume that the Markov process is ergodic, in the limit $N \to \infty$ we have

$$\mu(A) = \lim_{N\to\infty} \frac{\#\{y_k \in A\}}{N/\beta}, \tag{3.92}$$

for every measurable set $A \subset E$. Then, we consider points z_k, $k = 1, ..., N/\beta$, where z_k is a realization of the process $X_\alpha^{\xi/2}$ at time $\alpha > 0$ and started in y_k. Again, in the limit $N \to \infty$ it holds

$$(\tilde{T}_\alpha \mu)(A) = \lim_{N\to\infty} \frac{\#\{z_k \in A\}}{N/\beta}, \tag{3.93}$$

where $(\tilde{T}_t)$ is the semi-group of transfer operators for the process $(X_t^{\xi/2})$. This process has an increased metastability compared to (X_t). Approximately we can tell for every data point x_k and a small ball $B_\rho(x_k)$ of radius ρ around it if $(\tilde{T}_\alpha \mu)(B_\rho(x_k)) > \mu(B_\rho(x_k))$ by comparing the number of points y_k and z_k in $B_\rho(x_k)$. Therefore, we assume that a point x_k for $k = 1, ..., N$ has to belong to the region of core sets under the condition

$$x_k \in E \setminus C \Leftrightarrow \#\{z_k \in B_\rho(x_k)\} > \#\{y_k \in B_\rho(x_k)\}. \tag{3.94}$$

In this case, namely, these points became more attractive while increasing metastability. Again, attractivity is measured relatively to a timescale of interest, which enters the algorithm through α. We will make a remark (Rem. 8) on the choice of α in the examples below. Often, one will choose $\alpha < \beta h$, which implies that the effort of core set identification is smaller than the effort for the simulation of the trajectory (x_k) itself. Having identified the region $(E \setminus C) \cap \{x_k\}$ we have to split it into the final core sets $C_1, ..., C_n$. This task is usually less difficult than the identification because a constantness of the dominant eigenvectors on core sets implies a very uniform dynamical behaviour inside of core sets, e.g. a small diffusion distance. Without assuming anything, one can cluster the points $\{y_k \in E \setminus C\}$ first by analyzing the milestoning process (2.24) with respect to their voronoi tessellation [78], e.g. with spectral clustering methods like PCCA [81, 19]. Afterwards, one can include the $x_k \in E \setminus C$ with respect to the clustering of the points y_k in their neighbourhood.
In the end we estimate the matrices $\hat{T}$ and M from the same trajectory (x_k) without additional sampling such that the overall effort is dominated by the effort for simulating the trajectory (x_k).

Numerical results for two examples

One dimension. In this subsection we will test the algorithm for two examples. Both examples will be diffusions in potentials. Here, the parameter $\xi := \sigma$ is given by the noise intensity. Then, the algorithm directly connects to the theoretical considerations in Section 3.3. We start with our main example, the one dimensional diffusion in the extended potential with noise intensity $\sigma = 0.8$. Using the Euler-Maruyama scheme we simulate a trajectory of length $N = 5 \cdot 10^6$ at a time resolution $h = 0.01$.

Figure 3.28 shows the first 10.000 data points of the trajectory and the data points which have been identified as core sets for different timescale parameters α and $\beta = 1000$.

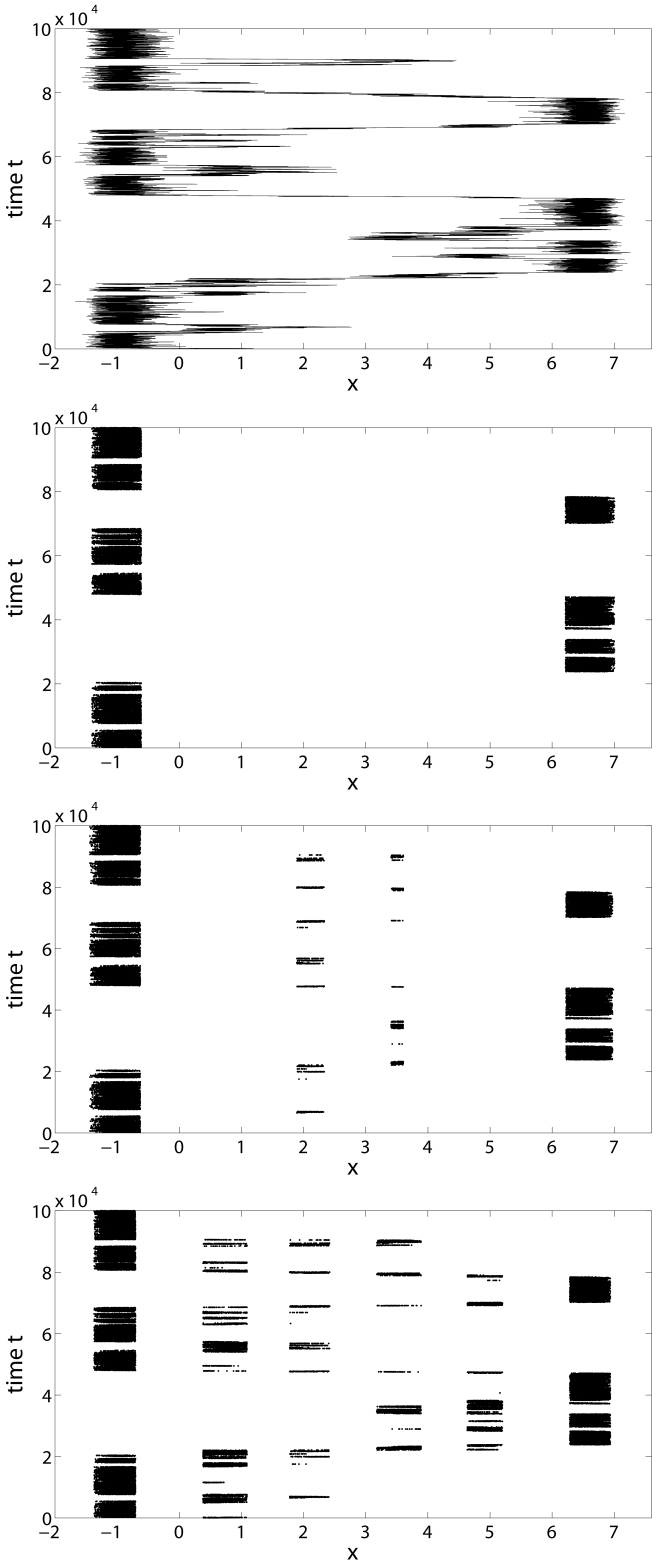


Figure 3.28: Piece of trajectory (top) and data points that are identified as core sets for $\alpha = 50$(second figure), $\alpha = 30$(third figure), $\alpha = 1$(bottom). 99

We observe that exactly the regions are identified that we had in mind since the theoretical discussion above.

$\alpha = 50$. With the largest parameter α we only get the two main wells as core sets and the following estimates for the matrix representation of $QT_\tau Q$, where we choose $\tau = 1$.

$$\hat{T} = \begin{pmatrix} 0.8719 & 0.1281 \\ 0.1388 & 0.8612 \end{pmatrix} \qquad M = \begin{pmatrix} 0.8764 & 0.1236 \\ 0.1346 & 0.8654 \end{pmatrix}$$

$$\hat{T}M^{-1} = \begin{pmatrix} 0.9939 & 0.0061 \\ 0.0057 & 0.9943 \end{pmatrix}.$$

We have seen already that the matrix representation $\hat{T}M^{-1}$ will always be a pseudostochastic matrix and that it has a stochastic interpretation for a class of probability vectors. Here, $\hat{T}M^{-1}$ is even stochastic such that the projected transfer operator will conserve the probability constraints for every probability distribution $v \in D$.

Moreover, the spectrum of $\hat{T}M^{-1}$ is given by

$$\hat{\lambda}_0 = 1, \hat{\lambda}_1 = 0.9883.$$

We remember that the dominant eigenvalue of T was computed by a finite element method as $\lambda_1 = 0.9885$.
So we found a good 2×2 Markov chain that preserves nicely the slowest dynamics of the original continuous Markov process. Although the example is fairly simple this result would have been completely impossible to achieve for a Standard Markov State Model as we have witnessed in Sec. 3.1. On top of that, we can investigate how our results change with the lag time τ. Figure 3.4 shows the first non-trivial eigenvalue of $\hat{T}M^{-1}$ plotted against the lag time τ that enters $\hat{T}$ ranging from $\tau = 0.1$, which is almost as short as possible with respect to the resolution of the numerical Euler scheme, to $\tau = 20.1$. It shows a perfect exponential decay over all lag times. This yields that the implied estimates for the eigenvalues of the generator and the implied timescales are very robust against the choice of the lag time. Fig. 3.4 backs this up. Again, this is a very desirable property as the original eigenvalues of the transfer operator T decay perfectly exponentially but something that is hardly to achieve with a Standard MSM, except for very fine discretizations.

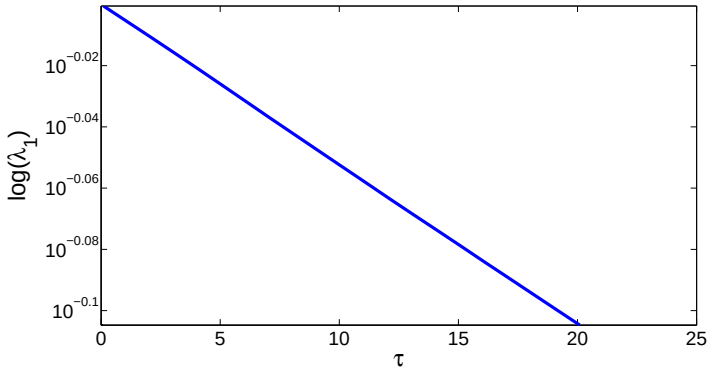


Figure 3.29: Logarithmic plot of the eigenvalue λ_1 of $\hat{T}M^{-1}$ estimated from the trajectory over lag time τ.

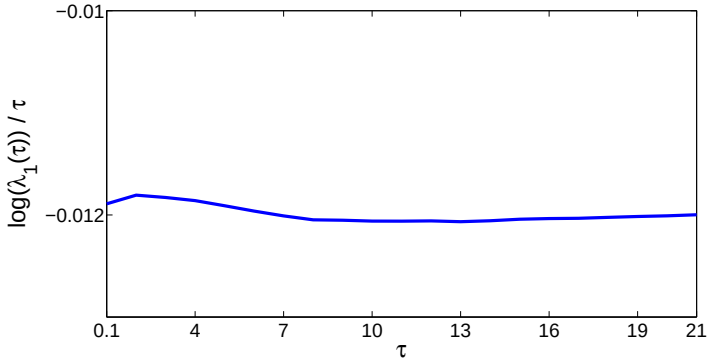


Figure 3.30: Generator eigenvalue estimated from $\hat{T}M^{-1}$ over lag time τ.

$\alpha = 30, \alpha = 1$. With the additional core set in the middle of the transition region ($\alpha = 30$) we find

$$\hat{T} = \begin{pmatrix} 0.9198 & 0.0802 & 0 \\ 0.1112 & 0.7711 & 0.1177 \\ 0 & 0.0767 & 0.9233 \end{pmatrix} \qquad M = \begin{pmatrix} 0.9342 & 0.0658 & 0 \\ 0.0902 & 0.8137 & 0.0961 \\ 0 & 0.0629 & 0.9371 \end{pmatrix},$$

$$\hat{T}M^{-1} = \begin{pmatrix} 0.9827 & 0.0193 & -0.0020 \\ 0.0279 & 0.9432 & 0.0288 \\ -0.0018 & 0.0184 & 0.9834 \end{pmatrix}$$

and the eigenvalues $\hat{\lambda}_i$ compared to the finite element results λ_i

i	0	1	2
$\hat{\lambda}_i$	1.0000	0.9849	0.9244
λ_i	1.0000	0.9885	0.9247

We notice that $\hat{T}M^{-1}$ is still almost stochastic but slightly negative entries have been introduced. As expected, the approximation quality of the second non-trivial eigenvalue is very good at the expense of the largest timescale. Nevertheless, taking the two levels of discretization for $\alpha = 50$ for λ_1 and $\alpha = 30$ for λ_2 together would provide a precise insight into the two dominant timescales of the underlying process.

Remark 8 *The proposed identification algorithm is very compatible with the multilevel idea. In order to find the core sets for a certain α we have to realize the process $X_t^{\xi/2}$ up to time α. This means, that we have also realizations of random variables $X_t^{\xi/2}$ for $0 < t < \alpha$. That is, we have associated point clouds z_k already on different time levels α and therefore automatically multiple core set proposals for multiple timescale approximations. One can also start with an over- or underestimated α and learn from the estimated timescales about the size of α, a posteriori.*

Finally, for $\alpha = 1$ all wells are resolved and we estimate

$$\hat{T} = \begin{pmatrix} 0.9465 & 0.0525 & 0.0010 & 0 & 0 & 0 \\ 0.1933 & 0.6395 & 0.1597 & 0.0074 & 0 & 0 \\ 0.0075 & 0.1635 & 0.6576 & 0.1647 & 0.0066 & 0 \\ 0 & 0.0069 & 0.1615 & 0.6569 & 0.1671 & 0.0076 \\ 0 & 0 & 0.0077 & 0.1652 & 0.6413 & 0.1858 \\ 0 & 0 & 0 & 0.0019 & 0.0458 & 0.9523 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.9798 & 0.0202 & 0 & 0 & 0 & 0 \\ 0.0743 & 0.8785 & 0.0472 & 0 & 0 & 0 \\ 0 & 0.0491 & 0.8972 & 0.0537 & 0 & 0 \\ 0 & 0 & 0.0549 & 0.8916 & 0.0535 & 0 \\ 0 & 0 & 0 & 0.0542 & 0.8742 & 0.0716 \\ 0 & 0 & 0 & 0 & 0.0179 & 0.9821 \end{pmatrix}$$

$$\hat{T}M^{-1} = \begin{pmatrix} 0.9631 & 0.0376 & -0.0009 & 0.0001 & 0 & 0 \\ 0.1429 & 0.7169 & 0.1403 & -0.0001 & 0 & 0 \\ -0.0034 & 0.1461 & 0.7166 & 0.1416 & -0.0011 & 0.0001 \\ 0 & 0.0003 & 0.1359 & 0.7196 & 0.1471 & -0.0030 \\ 0 & 0 & -0.0001 & 0.1414 & 0.7222 & 0.1365 \\ 0 & 0 & 0 & 0.0002 & 0.0326 & 0.9672 \end{pmatrix}$$

i	0	1	2	3	4	5
$\hat{\lambda}_i$	1.0000	0.9875	0.9230	0.7903	0.6187	0.4862
λ_i	1.0000	0.9885	0.9247	0.7911	0.6289	0.4957

Two dimensions. The last example demonstrated that one can efficiently identify few sets from simulation data which can be used in order to construct small Markov chains that approximate the longest timescales of a continuous Markov process very well. In one dimension. Interesting is how the effort to reach similar results changes if we increase the dimensionality of state space. If we want to achieve the same discretization level with a Standard Markov State Model the number of sets usually grows exponentially with the dimension of state space. Hence, one will need a large number of simulations for reliable estimates of transition probabilities between these sets.

We will consider a diffusion in a two dimensional potential with noise intensity $\sigma = 1.1$. A contour plot of the potential is shown in Fig. 3.4.

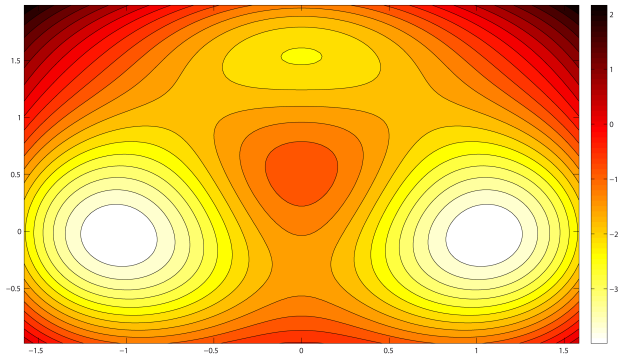


Figure 3.31: Three-well potential in 2D.

Now, we do not change the effort of our algorithm. We take again a trajectory of length $N = 5 \cdot 10^6$, computed with the Euler-Maruyama scheme with resolution $h = 0.01$. We also set $\beta = 1000$ and $\tau = 1$. Figure 3.32 shows a piece of the trajectory and two results of the identification algorithm for different parameters α.

For $\alpha = 10$ we identify two core sets, which are located in the region of the deepest local minima of the potential. For $\alpha = 2$ another small core set is introduced in the third less deep well.
Running the algorithm to the end we find the following estimates.

$\alpha = 10$:

$$\hat{T} = \begin{pmatrix} 0.9257 & 0.0743 \\ 0.0820 & 0.9180 \end{pmatrix} \qquad M = \begin{pmatrix} 0.9629 & 0.0371 \\ 0.0427 & 0.9573 \end{pmatrix},$$

$$\hat{T}M^{-1} = \begin{pmatrix} 0.9596 & 0.0404 \\ 0.0427 & 0.9573 \end{pmatrix}$$

and

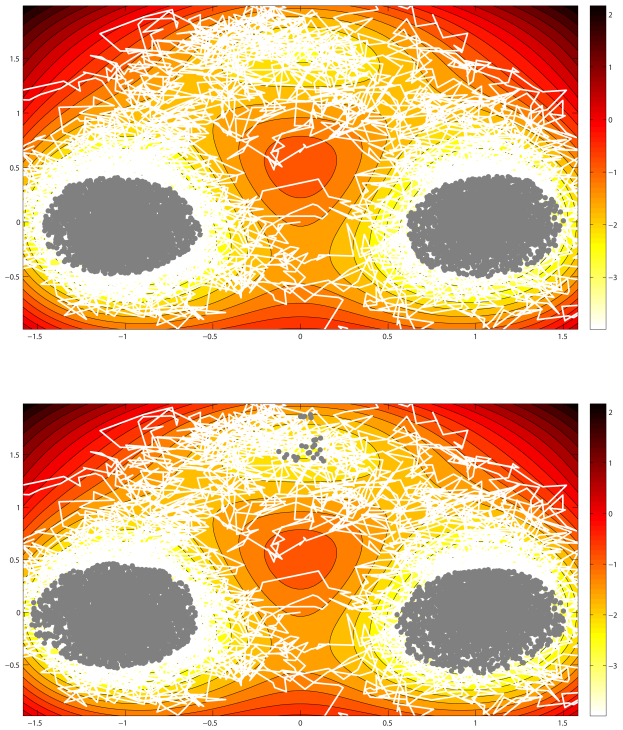


Figure 3.32: First 10.000 sample points. top: $\alpha = 10$ tracks the region around the two deepest wells, bottom: $\alpha = 2$ identifies a small region inside of the less pronounced well.

i	0	1
$\hat{\lambda}_i$	1.0000	0.9169
λ_i	1.0000	0.9215

$\alpha = 2$:

$$\hat{T} = \begin{pmatrix} 0.9282 & 0.0266 & 0.0452 \\ 0.3611 & 0.3185 & 0.3204 \\ 0.0476 & 0.0272 & 0.9252 \end{pmatrix} \qquad M = \begin{pmatrix} 0.9733 & 0.0141 & 0.0126 \\ 0.1890 & 0.6443 & 0.1667 \\ 0.0137 & 0.0142 & 0.9721 \end{pmatrix},$$

$$\hat{T}M^{-1} = \begin{pmatrix} 0.9494 & 0.0199 & 0.0308 \\ 0.2738 & 0.4830 & 0.2432 \\ 0.0315 & 0.0207 & 0.9478 \end{pmatrix}$$

and

i	0	1	2
$\hat{\lambda}_i$	1.0000	0.9174	0.4627
λ_i	1.0000	0.9215	0.4569

Again, we observe that for the very dominant timescales $\hat{T}M^{-1}$ is a stochastic matrix, which implies that QTQ will always transfer probability distributions. So, also in the two dimensional case it was possible to estimate Markov chains on the state spaces $\hat{E} = \{1, 2\}$ and $\hat{E} = \{1, 2, 3\}$ respectively, which approximate the slowest transition behaviour of the continuous Markov process well. Moreover, the effort that was needed to find these results in two dimensions was not increased against the effort in one dimension.

Note that the important idea is to identify the core sets just with respect to the trajectory that we want to use in order to estimate $\hat{T}$ and M later on. This identification is realized by comparing the two measures μ and $\tilde{T}_\alpha\mu$ in a small region along the trajectory. In continuous or large discrete state spaces this will result in a sampling problem. If we just wanted to identify core sets without having in mind that we want to estimate $\hat{T}$ and M from a trajectory, this could quickly cause problems in higher dimensions. Therefore, we first start the sampling for $\hat{T}$ and M, which cannot be avoided, and restrict our analysis to the resulting trajectory. At the same time this trajectory serves as a sample for μ in its own neighbourhood. This also implies that if the Markov process lives on a lower dimensional manifold, for example, we will automatically work on this manifold without complete identification of this object. So, in order to find the core sets one has to overcome this sampling problem, but with the proposed construction we can expect the effort to be much smaller than the sampling effort to estimate the MSM itself.

In the next section, we will demonstrate the core set identification for Markov jump processes on finite state space. In this case we do not have to sample probabilities because we can compute all measures directly. This will also underline that in continuous state spaces the core set identification reduces mainly to the sampling problem. Moreover, we will recognize that this Markov State Modeling technique will immediately lead to an interesting new approach to fuzzy clustering problems of Markov chains and Markov jump processes.

3.5 An Approach to Fuzzy Clustering

We start this section with the task to build a Markov State Model based on milestoning for a Markov jump process $(X_t)_{t\in\mathbb{R}}$ on finite state space $E = \{1, ..., N\}$. We assume that (X_t) is reversible, has a generator L and a unique positive invariant measure μ. In this case, L can be expressed by a $n \times n$ rate matrix, i.e.

$$L(x, y) \leq 0 \ \forall x \neq y \qquad L(x, x) = -\sum_{\substack{y\in E \\ y\neq x}} L(x, y) < 0. \tag{3.95}$$

We have briefly reviewed a result of [36] in Theorem 6. It states that the construction of a Standard Markov State Model based on a full partition with n sets corresponds to a metastable clustering problem. This relationship between full partition Markov State Models, the projection error $\delta = \max_{i=1,...,n-1} \|Q^{\perp} u_i\|$, and metastability has been exploited to develop methods that cluster the state space with respect to metastability [18]. Nevertheless, we have seen that from the MSM perspective it can be very useful to refine the partition in transition regions, where eigenvectors are strongly varying. This makes also sense from the clustering point of view because for such states it is not totally clear to which cluster they should be assigned. Most of the time there is no satisfying answer to this question. On the other hand, in the context of clustering one usually tries to avoid to introduce many of such small cluster. A possible way out of this dilemma is a so called **fuzzy clustering**. That is, we have to find for every cluster $i \in \{1, ..., n\} = \hat{E}$ an affiliation function f_i

$$f_i : E \to [0, 1], \qquad \sum_{i=1}^{n} f_i(x) = 1 \ \forall x \in E. \tag{3.96}$$

The properties in (3.96) imply that $(f_i(x))_{i \in \hat{E}}$ is a probability distribution on the cluster space $\hat{E}$. This allows for an interpretation of a fuzzy clustering as a randomly generated normal clustering. That is, every state x will be assigned to cluster i with probability $f_i(x)$ independently of the other states. One can think of a fuzzy clustering as an ensemble of deterministic ones. On the other hand, the word clustering claims that there should be some structure in these randomly generated sets because the goal of clustering is separation. That is, at least some states should always be separated, i.e. be assigned to different cluster in all realizations. Therefore, we assume that for every cluster i there should exist a set of states

$$C_i = \{x \in E | f_i(x) = 1\}, \tag{3.97}$$

which we call core of the cluster.

At this point we should notice an analogy. When we considered Markov State Models based on milestoning instead of full partition MSM, we chose a different subspace D for projection of the operator T. That is, instead of the space $span\{\mathbb{1}_{A_1}, ..., \mathbb{1}_{A_n}\}$ we introduced a subspace D spanned by the committors $D = span\{q_1, ..., q_n\}$. Now, these committors are affiliation functions (3.96), so they define a fuzzy clustering, and the cores of this clustering would contain the core sets of the MSM. On the other hand, enlarging the core sets to the sets (3.97) with respect to the committors would not change the projection and provide an equivalent Markov State Model. This means that we have a connection between core set MSM and fuzzy clusterings as we had before between Standard MSM and deterministic

clusterings. Moreover, we have understood that the region of all core sets $E \setminus C$ should be metastable and attractive on the dominant timescales. This would justify the use of methods like PCCA+ [19], optimal fuzzy aggregation [69], or other metastable fuzzy cluster approaches to identify the sets (3.97) and therefore to construct a core set MSM.

The drawback is that fuzzy cluster problems are computationally much harder to solve than deterministic problems, in general. However, we approached the task of the construction of a core set MSM already in the last section. We still want to exploit the connection between Markov State Models and cluster problems but the other way around. The idea is not to use a fuzzy cluster method to define a Markov State Model, but rather let the construction of the MSM define a fuzzy cluster method. In Section 3.3 we have seen that one can split the task of constructing a MSM into two steps. First, we have to find the region $E \setminus C = \bigcup_{i=1}^{n} C_i$ and second, we have to split this region into the core sets $C_1, ..., C_n$. The last Section 3.4 introduced an algorithmical idea to solve the first issue for a parameter dependent Markov process. In the case of a finite state space we even do not have to face the sampling problem. We can take the invariant measure μ^* of the process with reduced metastability, calculate for $\alpha > 0$ the set

$$C^\alpha = \{x \in E | (e^{L\alpha}\mu^*)(x) > \mu^*(x)\}, \tag{3.98}$$

and simply set $E \setminus C = C^\alpha$. In the second step, we have to split $E \setminus C$ into n sets $C_1, ..., C_n$. That is, the calculation of (3.98) transformed the fuzzy cluster problem on the set E into a deterministic cluster problem on the smaller set $E \setminus C$. Usually, this can be solved efficiently since the problematic region C is not involved anymore. $E \setminus C$ should consist only of states, which are dynamically well separated. For Markov jump processes, we propose to study the dynamics of the embedded Markov chain w.r.t. the original jump process in terms of milestoning, where we treat every state in $E \setminus C$ as single core set (cf. example below). Then, we can apply a spectral deterministic clustering method, e.g. PCCA [18]. The advantage is that this process between the states in $E \setminus C$ does not depend on a lag time $\tau > 0$. Having clustered the set $E \setminus C$, the core set MSM is defined. It immediately implies a fuzzy clustering of the state space E by the associated committors $q_1, ..., q_n$. Note that this type of affiliation is also very natural. Once we know where the cores of the cluster (3.97) are, we have to define a probability for all states in C to become assigned to one of these cluster. Letting the committors define these affiliation functions means that we let the dynamics of the Markov process itself decide. That is, for a state x and a cluster i we simply choose the affiliation $f_i(x)$ to be the probability that starting in x the Markov process will reach the cluster core C_i next.

Application to Networks

We will now apply this method to an example. We consider a network, i.e. a set of nodes $V = \{1, ..., N\}$ and a set of edges with adjacency matrix $a(x, y)$,

$$a(x,y) = \begin{cases} 1, & \text{edge between nodes } x, y \text{ exists} \\ 0, & \text{else.} \end{cases} \tag{3.99}$$

The example network is illustrated in Fig. 3.33.

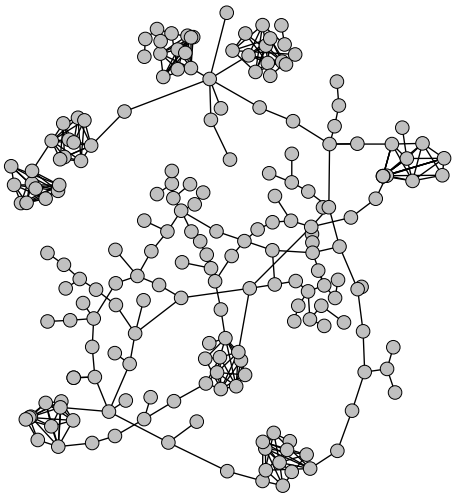


Figure 3.33: Example Network

This network consists of $N = 200$ nodes and 360 edges. It is well known that there is a strong connection between networks and Markov processes. We will follow the approach of [67] and define a class of random walks on the network, namely continuous Markov jump processes with generators $L_\xi, \xi \in \mathbb{N}_0$

$$L_\xi(x,y) = \begin{cases} -\frac{1}{d(x)^\xi}, & x = y \\ \frac{a(x,y)}{d(x)^{\xi+1}}, & x \neq y, \end{cases} \tag{3.100}$$

where

$$d(x) = \sum_{y \in V} a(x,y) \tag{3.101}$$

is the degree of a node x. One can directly compute that the embedded Markov chain of these jump processes is given by the transition matrix

$$P(x,y) = \frac{a(x,y)}{d(x)}, \tag{3.102}$$

independently of the parameter ξ. Moreover, the processes are reversible and have for connected networks the unique invariant measure

$$\mu_\xi(x) = \frac{1}{Z_\xi} d(x)^{\xi+1}. \tag{3.103}$$

If we look at the example network in Fig. 3.33, we notice that there are some nodes which are strongly interconnected, and parts which have only few edges. The family of random walks in (3.100) is constructed such that the interconnected sets become metastable in the sense of the random walk. Moreover, the expected waiting times of the random walk in each node is given by $-1/l(x,x) = d(x)^\xi$. That is, for increasing ξ the Markov process will jump from node to node faster, relatively to the other jump rates, if the nodes are only loosely connected. This means that we can apply our algorithm to this class of parameter dependent Markov jump processes if we are interested in computing a fuzzy clustering, where rather interconnected nodes will become the cores of the cluster (3.97).

For our example, we take the generator $L := L_1$. The spectrum of the generator is shown in Fig. 3.34.

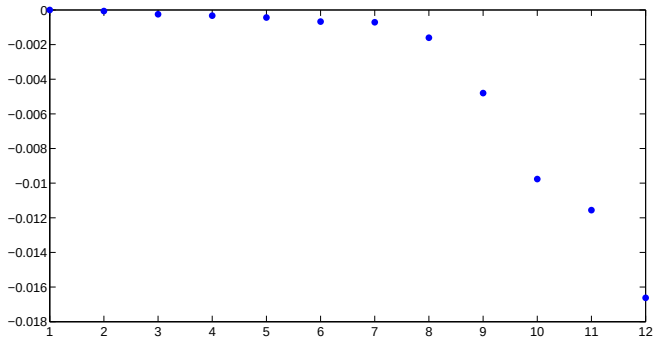


Figure 3.34: First 12 eigenvalues of the generator L.

From this picture, we would guess for a spectral clustering method that we should consider 7 or 8 cluster. On the other hand, we do not have to decide on the number of cluster at this point. We will see that this is a big advantage of the whole approach. At the moment, we only have to specify an α in order to erase the transition region C.

Influence of α and cluster number. Let us start with $\alpha = 1000$. Compared to the implied timescales of the jump process, this is a rather large choice because $-1/\Lambda_7 \approx 1000$. We know that larger α will only show core sets, which are attractive on the long timescales. So with a large choice of α we implicitly reduce the resolution of the clustering, i.e. we will get rather

few cores in the clustering. The invariant measure of the less metastable and less attractive jump process with generator L_0 is given by

$$\mu^*(x) = \frac{1}{Z} d(x) = \frac{1}{Z'} \sqrt{\mu(x)}, \tag{3.104}$$

where μ is the invariant measure of our main Markov process. So, we compute C^α as in (3.98), and as proposed above the lag time independent transition matrix P_α on C^α

$$P_\alpha(x, y) = \sum_{z \in V} P(x, z) q_y(z), \quad x, y \in C^\alpha, \tag{3.105}$$

where $\{q_y\}_{y \in C^\alpha}$ are the committors treating every node in C^α as single core set $C_y = \{y\}$. That is, $P_\alpha(x, y)$ describes the probability that starting in node x the next node that is reached will be y, ignoring the waiting times. In our example, the spectrum of this matrix is visualized in Fig. 3.35.

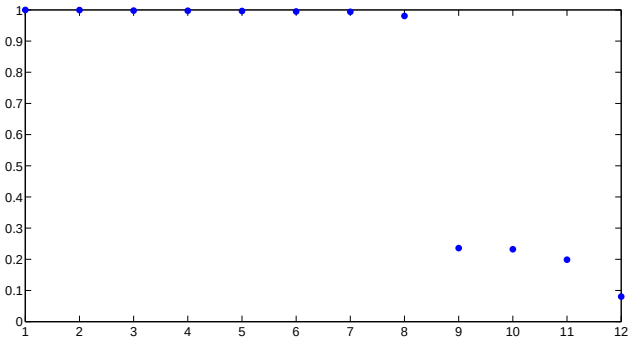


Figure 3.35: First 12 eigenvalues of P_α for $\alpha = 1000$.

It is not surprising that the gap in the spectrum after 8 eigenvalues became much clearer because we expected the core sets on the dominant timescales to be dynamically well separated. On the other hand, we fixed the cores of the cluster but still did not decide on a cluster number. Now, this is again a useful feature of this approach breaking down a fuzzy clustering into these multiple steps. We can look at hierarchical structures inside of a clustering for fixed cluster cores by simply choosing different cluster numbers for the second deterministic cluster part. On page 114, we see the results of the cluster analysis for the example network. First, we see the identified region C^α for $\alpha = 1000$ and a hierarchical splitting into 6,7, and 8 sets. Then, we decrease α to a value of 150 and find the following spectrum of P_α (Fig. 3.36).

Now, we could consider up to 10 cluster, although the last two eigenvalues indicate a lower metastability. On page 115, we immediately see where

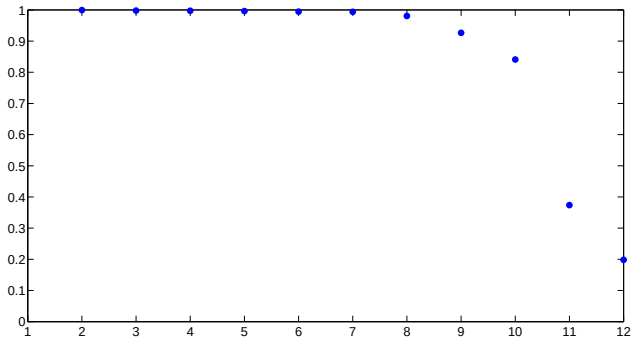


Figure 3.36: First 12 eigenvalues of P_α for $\alpha = 150$.

this comes from. Because of the lowered α value another small, less interconnected region was marked to be treated as a core set. Then, the spectrum of the associated transition matrix P_α suggested to cluster them separately rather than to merge them with existing cores.

The choice of a cluster number is usually one of the most challenging tasks in cluster analysis. In many situations there is even no unique best solution. Our approach handles this issue differently. One the one hand, it tries to make the choice of the most interesting cluster number easier by concentrating on the most separated sets. Therefore, the spectrum of the matrices of the random walks becomes easier to interpret. For example, usually one considers a random walk on the network that is described by the Markov chain with transition matrix P (3.102). Its spectrum is shown in Fig. 3.37, which does not at all display the structure we identified.

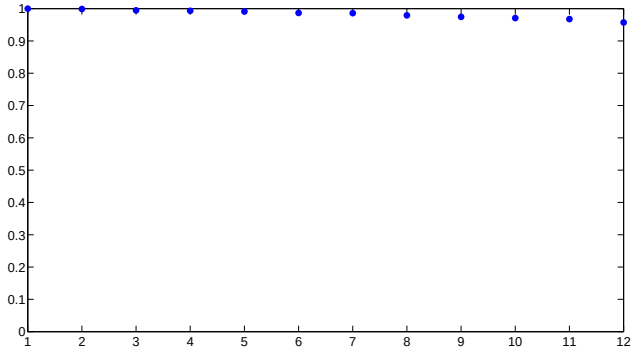


Figure 3.37: First 12 eigenvalues of P (3.102).

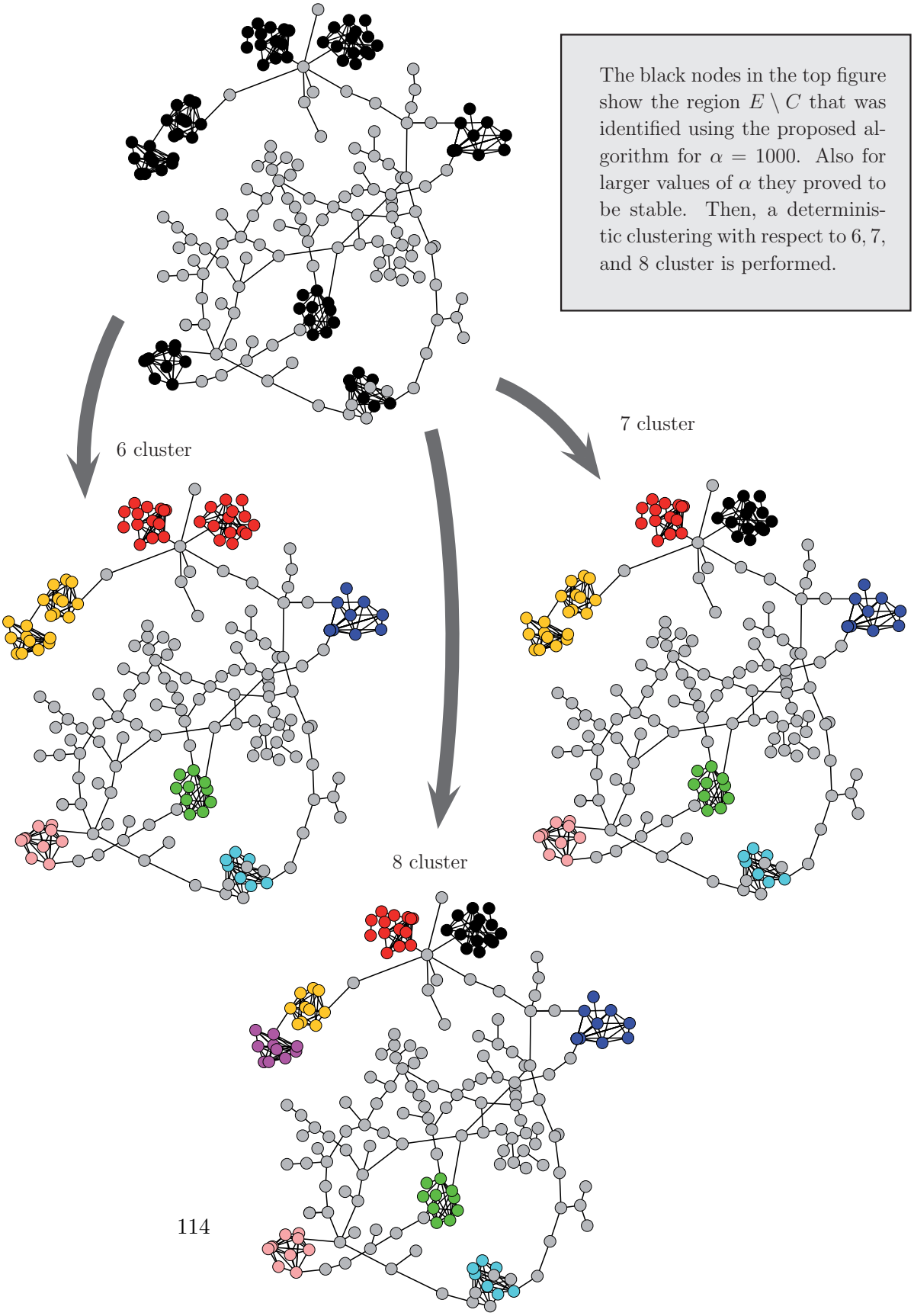
On the other hand, our method provides a high level of control and interpretation to different layers of the clustering. That is, we can choose between two types of hierarchical refinement of the cluster. For fixed α, we can analyze different resolutions within the cluster cores, or we can decrease the value of α in order to introduce new core sets, which will have to correspond to less pronounced cluster. Finally, we should note that the effort to compute such a fuzzy clustering boils down to the effort of the deterministic cluster part w.r.t. the Markov chain described by P_α (3.105). Before that, we have to compute C_α, which needs to evaluate $e^{L\alpha}\mu^*$. Nevertheless, we never use the matrix exponential again, so we can also use algorithms like [3] to directly compute the action of $e^{L\alpha}$ to μ^*. In the last step, we only have to solve the linear equation (2.25) to achieve a full fuzzy clustering on E.

Reference to cluster methods

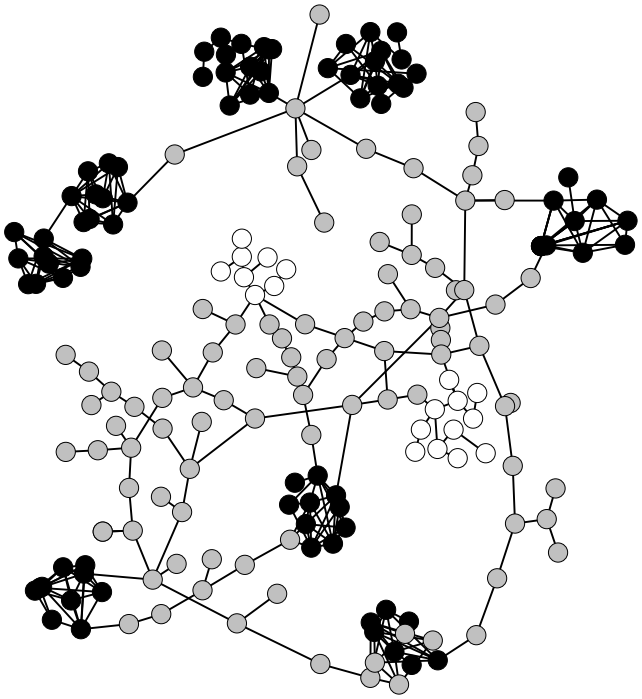
The literature contains many techniques targeting network partitioning. First, one has to distinguish between deterministic and fuzzy cluster approaches. Deterministic clusterings are very popular and there are a lot of different approaches, methods and variants. Some ideas are based on a random walk approach [18, 47, 6, 22, 45, 23] and often use spectral decompositions. Others use measures directly on the graph like betweenness [30], modularity [53], congestion [16], etc. On the other hand, in real applications one usually encounters some nodes, which do not belong to a certain community but rather have an affiliation to several cluster. This is why one often prefers fuzzy cluster approaches. There are several fuzzy versions of deterministic cluster methods which analyze random walks like [19, 69]. We find techniques based on Gaussian mixture models [31, 32], Bayesian network models [38], and many other fuzzy clustering algorithms [34]. Most of these algorithms solve optimization problems with respect to a functional that measures the quality of the clustering in some sense. For example, the methods in [54, 13, 2, 26, 1] use a specific statistical model , i.e. they assume that the associated graph of the network was constructed randomly with respect to some parameterized probabilistic model. Then, they optimize the likelihood with respect to the parameters.

We have seen that our approach is also connected to random walks, but the analysis of the associated Markov process is performed very differently. We do not assume that the graph is generated according to a particular underlying model and we do not solve an optimization problem. We rather use the Markov State Modeling technique to find the cores of the cluster directly. Then, we analyze another random walk, who just lives on these core sets, aiming at a deterministic clustering. Having erased the nodes which have affiliation to different communities, the application of one of the many available and fast deterministic cluster methods is reasonable again.

To avoid confusion we want to point out that our approach has nothing to do with so called random networks, i.e. graphs that are generated according to a probabilistic rule, which is also described by a transition matrix. As mentioned above, some cluster methods are based on the assumption that a network of interest is the realization of such a random network. Our approach uses a probabilistic framework only in terms of a random walk on a given network but does not analyze the stochastic generation of networks.

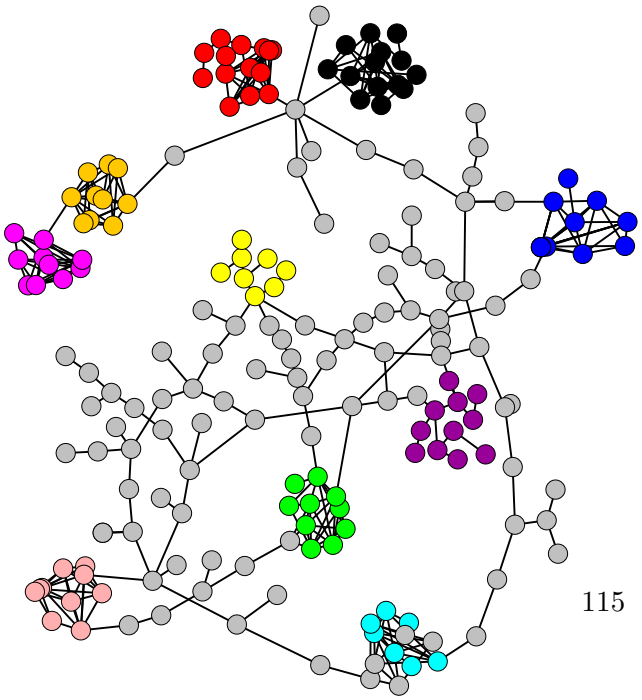


The black nodes in the top figure show the region $E \setminus C$ that was identified using the proposed algorithm for $\alpha = 1000$. Also for larger values of α they proved to be stable. Then, a deterministic clustering with respect to 6, 7, and 8 cluster is performed.



We decrease α to a value of 150. This implies that the exit times out of region C, which is not assigned to cluster cores, have to become shorter. This is why a region (white) is added to the set $E \setminus C$.

10 cluster



Summary

For 15 years, so called Markov State Models (MSM) have been used successfully to simplify large complex systems, which are described by Markov processes. The goal of Markov State Modeling is the approximation of a Markov process by a Markov chain on a small finite state space. Until 2005, Markov State Models have been always constructed via full partitions of state space, i.e. sets that cover the whole state space. That is, one calculated the transition probabilities for the approximating Markov chain from the transition probabilities of the original Markov process between partitioning sets. We also called this class of methods classical or Standard MSM. Then, [19] introduced the idea of rather using fuzzy affiliation functions instead of a deterministic clustering into sets. Two years ago, in [11] an approach was proposed that avoided full partitions of state space and constructed a fuzzy MSM variant by defining small disjoint sets in the most dominant metastable regions. Recently, we introduced another method, which is based on these so called core sets, in [70].

In this thesis, we developed a mathematical framework that is applicable to all of these former MSM techniques. We used a general, functional analytic approach and understood Markov State Models as best approximations of the original transfer operator in terms of discretization. For core set based MSM this led to a new construction via conditional stopping times instead of transition probabilities. From this point of view we could also prove several statements about the approximation quality of the models, which are also valid for classical MSM and other methods that refer to projected operators. For example, we connected the reproduction of the dominant timescales of the system by the Markov State Model to projection errors of the associated eigenvectors. From error estimates for these projection errors, we further understood how to choose the discretization, i.e. the core sets, in order to ensure a good approximation quality. Moreover, we used these results to construct an algorithm for the estimation of the MSM from a realization of the original Markov process. One should emphasize that the estimation of an appropriate discretization itself is also part of this method. Finally, for finite state spaces we could connect the core set based MSM variant to so called fuzzy cluster problems. That is, we used the construction of the Markov State Model to develop a novel fuzzy clustering approach and we demonstrated its properties by application to a sample network. We think that this ending is very appealing because it shows the broad impact of the developed mathematical framework. We started with MSM, i.e. discretizations for Markov processes on continuous state spaces, and using the results from this analysis we ended with a proposal for the fuzzy clustering of networks.

Zusammenfassung

Seit 15 Jahren werden so genannte Markov State Modelle (MSM) erfolgreich eingesetzt um komplexe Systeme zu vereinfachen, die von Markovprozessen beschrieben werden. Das Ziel eines MSM ist die Approximation eines solchen Markovprozesses durch eine Markovkette auf möglichst kleinem Zustandsraum. Dabei werden die Übergangswahrscheinlichkeiten der Markovkette aus Übergangswahrscheinlichkeiten des ursprünglichen Prozesses zwischen Teilmengen des Zustandsraums berechnet. Bis 2005 wurden hierbei immer vollständige Zerlegungen betrachtet, d.h. die Teilmengen überdeckten den kompletten Zustandsraum. Solche Modelle nannten wir daher auch klassische oder Standard MSM. Mit [19] kam die Idee auf, die Punkte des Zustandsraumes nicht deterministisch in Mengen aufzuteilen, sondern so genannte weiche/fuzzy Zuordnungen zu verwenden. Vor ungefähr zwei Jahren wurde dann in [11] vorgeschlagen, ebenfalls vollständige Zerlegungen zu vermeiden und eine fuzzy MSM Variante durch vereinzelte, disjunkte Mengen in den Regionen der höchsten Metastabilität zu konstruieren. Auf diesen so genannten Core Sets beruht auch der bisher letzte, von uns in [70] vorgestellte Ansatz.

In dieser Arbeit ist nun ein mathematischer Rahmen entstanden, der alle bisherigen MSM Methoden einschließt. Dazu wählten wir einen funktionalanalytischen Zugang und verstanden Markov State Modelle als Bestapproximationen des ursprünglichen Transferoperators durch Diskretisierung. In Bezug auf Core Set basierte MSM entstand daraus eine neuartige Konstruktion durch bedingte Stoppzeiten anstelle von einfachen Übergangswahrscheinlichkeiten. Wir erarbeiteten aus diesem Blickwinkel mehrere Resultate über die Approximationsgüte der Methode, welche sogar für die klassischen MSM und andere Verfahren gültig sind, die sich auf Projektionen von Operatoren zurückführen lassen. Wir konnten unter anderem zeigen, dass die dominanten Zeitskalen des Markovprozesses durch das MSM korrekt wiedergegeben werden, falls die Projektionsfehler der dazugehörigen Eigenvektoren klein genug sind. Fehlerabschätzungen für diese Projektionsfehler ließen sogar die Einsicht zu, wie ein solches Core Set basiertes Markov State Modell zu konstruieren ist. Wir konnten dies nutzen, um einen Algorithmus zur Schätzung des Modells aus einer Simulation des ursprünglichen Prozesses anzugeben. Besonders hervorzuheben ist, dass die Methode auch die Schätzung der Core Sets, also der Diskretisierung selbst, beinhaltet. Für endliche Zustandsräume konnten wir die Ideen zur Konstruktion der MSM nutzen um einen neuartigen Ansatz zur Clusteranalyse herzuleiten. Die Methode wurde dann an einem Netzwerkbeispiel verdeutlicht. Wir finden, dass dies ein harmonisches Ende ist, da es die weitreichenden Einsatzmöglichkeiten des entwickelten mathematischen Hintergrundes aufzeigt. Wir begannen mit einer Analyse von MSM, d.h. Diskretisierungen von Markovprozessen auf kontinuierlichen Zustandsräumen, und nutzen die Resultate um die Arbeit mit einem Vorschlag zu fuzzy Clusterverfahren abzuschließen.

Bibliography

[1] E. Airoldi. *Bayesian Mixed-Membership Models of Complex and Evolving Networks*. PhD thesis, Carnegie Mellon University, 2006.

[2] E. Airoldi, D. Blei, S. Fienberg, and E. P. Xing. Mixed membership stochastic block model. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[3] A. H. Al-Mohy and N. J. Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *To appear in SIAM Journal on Scientific Computing. MIMS EPrint, 2010.30, The University of Manchester*, 2010.

[4] L. Arnold. *Stochastic Differential Equations: Theory and Applications*. Wiley, New York, 1974.

[5] A. L. Beberg, D. L. Ensign, G. Jayachandran, S. Khaliq, and V. S. Pande. Folding@home: Lessons from eight years of volunteer distributed computing. *ipdps*, IEEE International Symposium on Parallel&Distributed Processing:1–8, 2009.

[6] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 6:1373–1396, 2003.

[7] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability and low lying spectra in reversible markov chains. *Comm. Math. Phys.*, 228:219–255, 2002.

[8] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6:399–424, 2004.

[9] A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. II. precise asymptotics for small eigenvalues. *J. Eur. Math. Soc. (JEMS)*, 7:69–99, 2005.

[10] P. Bremaud. *Markov chains, Gibbs fields, Monte Carlo simulation, and queues*. Springer Verlag New York, 2001.

[11] N. V. Buchete and G. Hummer. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, 112:6057–6069, 2008.

[12] J. Chodera, N. Singhal, V. S. Pande, K. Dill, and W. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *Journal of Chemical Physics*, 126, 2007.

[13] J. J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs: A variational approach. *Technical Report 4, Statistics for systems biology group, INRA, Jouy-en-Josas, France*, 2007.

[14] E. B. Davies. Spectral properties of metastable markov semigroups. *J. Funct. Anal.*, 52:315–329, 1983.

[15] B. de Groot, X. Daura, A. Mark, and H. Grubmüller. Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Bio.*, 301:299–313, 2001.

[16] M. Dellnitz and R. Preis. Congestion and almost invariant sets in dynamical systems. In *Symbolic and Numerical Scientific Computation*, volume 2630 of *Lecture Notes in Computer Science*, pages 255–284. Springer Berlin / Heidelberg, 2003.

[17] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. Computation of essential molecular dynamics by subdivision techniques. In *Lecture Notes in Computational Science and Engineering*, volume 4, pages 98–115. Springer, 1999.

[18] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315:39–59, 2000.

[19] P. Deuflhard and M. Weber. Robust perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398 Special issue on matrices and mathematical biology:161–184, 2005.

[20] N. Djurdjevac, M. Sarich, and C. Schütte. Estimating the eigenvalue error of markov state models. *Multiscale Modeling & Simulation (Submitted)*, 2010.

[21] N. Djurdjevac, M. Sarich, and C. Schütte. On Markov state models for metastable processes. *Proceeding of the ICM 2010 as invited lecture*, 2010. Preprint download via http://publications.mi.fu-berlin.de/991/.

[22] D. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data,. *PNAS*, 100:5591–5596, 2003.

[23] W. E, T. Li, and E. V. Eijnden. Optimal partition and effective dynamics of complex networks. *Proc. Nat. Acad. Sci.*, 105:7907–7912, 2008.

[24] W. E and E. Vanden-Eijnden. Towards a theory of transition paths. *Journal of statistical physics*, 123:503–523, 2006.

[25] W. E and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annual Review of Physical Chemistry*, 61:391–420, 2010.

[26] E. A. Erosheva and S. E. Fienberg. Bayesian mixed membership models for soft clustering and classification. In C. Weihs and W. Gaul, editors, *Classification - The Ubiquitous Challenge*. Springer, Heidelberg, 11-26, 2005.

[27] S. N. Ethier and T. G. Kurtz. *Markov Processes - Charaterization and Convergence*. John Wiley and Sons, 2005.

[28] A. K. Faradjian and R. Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120:10880–10889, 2004.

[29] M. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*. Springer, New York, 1998.

[30] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99:7821–7826, 2002.

[31] G.McLachlan and K. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[32] G.McLachlan and D. Peel. *Finite mixture models*. Wiley, New York, 2000.

[33] F. Herau, M. Hitrik, and J. Sjoestrand. Tunnel effect for Kramers-Fokker-Planck type operators: Return to equilibrium and applications. *International Mathematics Research Notices*, article ID rnn057, 2008.

[34] F. Höppner, F. Klawonn, R. Kruse, and T. Runkle. *Fuzzy cluster analysis*. John Wiley and Sons, New York, 1999.

[35] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich. Understanding ensemble protein folding at atomic detail. *Proc. Natl. Acad. Sci. USA*, 103(47):17747–17752, November 2006.

[36] W. Huisinga. *Metastability of Markovian Systems A transfer operator based approach in application to molecular dynamics*. Phd thesis, Fachbereich Mathematik und Informatik, FU Berlin, 2001.

[37] W. Huisinga, S. Meyn, and C. Schütte. Phase transitions and metastability for Markovian and molecular systems. *Ann. Appl. Probab.*, 14:419–458, 2004.

[38] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

[39] M. E. Karpen, D. T. Tobias, and Brooks. Statistical clustering techniques for analysis of long molecular dynamics trajectories. *Biochemistry*, 32:412–420, 1993.

[40] A. Knyazev and M. E. Argentati. Rayleigh-ritz majorization error bounds with applications to fem. *SIAM Journal on Matrix Analysis and Applications*, 31:1521, 2010.

[41] A. V. Knyazev and M. E. Argentati. Principal angles between subspaces in an a-based scalar product: Algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

[42] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Nat. Acad. Sci. USA*, 101:14766–14770, 2004.

[43] B. Øksendal. *Stochastic Differential Equations*. Springer, 2007.

[44] S. Kube and M. Weber. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.*, 126:024103–024113, 2007.

[45] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1393–1403, 2006.

[46] R. Maier and D. Stein. Limiting exit location distribution in the stochastic exit problem. *SIAM J. Appl. Math.*, 57, 1997.

[47] M. Meila and J. Shi. A random walks view of spectral segmentation. *AI and Statistics (AISTATS)*, 2001.

[48] P. Metzner, F. Noé, and C. Schütte. Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data. *Phys. Rev. E*, 2008.

[49] P. Metzner, C. Schütte, and E. Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219, 2009.

[50] S. P. Meyn and R. L. Tweedie. Stability of markovian processes i: Criteria for discrete-time chains. *Advances in Applied Probability*, 24(3):pp. 542–574, 1992.

[51] S. P. Meyn and R. L. Tweedie. Stability of markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):pp. 518–548, 1993.

[52] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a -sheet miniprotein. *Proteins*, 70:1185–1195, 2007.

[53] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69 (026113), 2004.

[54] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *PNAS*, 104:9564–9569, 2007.

[55] F. Noé. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.*, 128:244103, 2008.

[56] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, 18:154–162, 2008.

[57] F. Noé, I. Horenko, C. Schütte, and J. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126:155102, 2007.

[58] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium trajectories. *PNAS*, 106(45):19011–19016, 2009.

[59] J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

[60] A. C. Pan and B. Roux. Building markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.*, 129(6):064107+, August 2008.

[61] A. C. Pan and B. Roux. Building Markov state models along pathways to determine free energies and rates of transitions. *Journal of Chemical Physics*, 129, 2008.

[62] I. Pavlyukevich. *Stochastic Resonance*. PhD thesis, HU Berlin, 2002.

[63] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. . Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134, 174105, 2011.

[64] F. Rao and A. Caflisch. The protein folding network. *J. Mol. Bio.*, 342:299–306, 2004.

[65] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Springer Verlag, 1999.

[66] S. Roeblitz. *Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics*. PhD thesis, FU Berlin, 2008.

[67] M. Sarich, N. Djurdjevac, S. Bruckner, T. O. Conrad, and C. Schütte. In preparation. 2011.

[68] M. Sarich, F. Noé, and C. Schütte. On the approximation quality of markov state models. *Multiscale Modeling and Simulation*, 8(4):1154–1177, 2010.

[69] M. Sarich, C. Schütte, and E. Vanden-Eijnden. Optimal fuzzy aggregation of networks. *Multiscale Modeling and Simulation*, 8 (4):1535–1561, 2010.

[70] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden. Markov state models based on milestoning. *J. Chem. Phys*, 134 (20), 2011.

[71] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan. Extracting markov models of peptide conformational dynamics from simulation data. *J. Chem. Theory Comp.*, 1:515–526, 2005.

[72] C. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Applications to Biomolecules*. Habilitation thesis, Fachbereich Mathematik und Informatik, FU Berlin, 1998.

[73] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comp. Physics Special Issue on Computational Biophysics*, 151:146–168, 1999.

[74] C. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In *Handbook of Numerical Analysis*, pages 699–744. Elsevier, 2003.

[75] N. Singhal and V. S. Pande. Error analysis in Markovian state models for protein folding. *Journal of Chemical Physics*, 123, 2005.

[76] S.Meyn and R. Tweedie. *Markov Chains and Stochastic Stability.* Springer, 1993.

[77] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9. *J. Am. Chem. Soc.*, 132(5):1526–1528, February 2010.

[78] M. G. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J. Reine Angew. Math.*, 134:198–287, 1908.

[79] A. Voter. Introduction to the kinetic Monte Carlo method. In *Radiation Effects in Solids*. Springer, NATO Publishing Unit, Dordrecht, The Netherlands, 2005.

[80] D. J. Wales. *Energy Landscapes*. Cambridge University Press, Cambridge, 2003.

[81] M. Weber. Improved perron cluster analysis. *ZIB Report*, 03-04, 2003.

[82] M. Weber. *Meshless Methods in conformation dynamics*. PhD thesis, FU Berlin, 2006.

[83] M. Weber and S. Kube. Preserving the markov property of reduced reversible markov chains. volume 1048 of *Numerical Analysis and Applied Mathematics*, pages 593–596. Int. Conf. on Num. Analy. and Appl. Math. 2008, AIP Conference Proceedings, Kos, 2008.

[84] Weinan E and E. Vanden Eijnden. Metastability, conformation dynamics, and transition pathways in complex systems. In *Multiscale Modelling and Simulation*, pages 38–65. Springer, 2004.