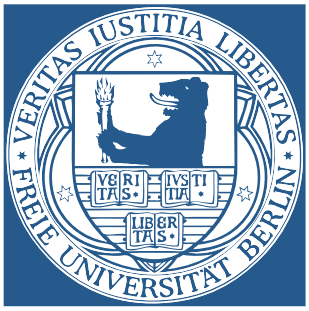
Freie Universität Berlin



Fachbereich Mathematik und Informatik

Importance sampling for metastable dynamical systems in molecular dynamics

Dissertation eingereicht am Fachbereich Mathematik und Informatik der Freien Universität Berlin zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

von

Jannes Quer

Berlin, Juli 2018

Betreuer	PD Dr. Marucs Weber Leiter der Arbeitsgruppe Computational Molecular Design Numerische Mathematik Zuse Institut Berlin
Gutachter	PD Dr. Marcus Weber Zuse Institut Berlin Prof. Dr. Carsten Hartmann BTU Cottbus - Senftenberg
Tag der Disputation	23. November 2018

Abstract

The behaviour of molecules is determined by rare events and these rare events matter. For example, a large conformational change in a molecule can lead to complete different behaviour of this molecule. But these rare events also affect the numerical simulation of molecules. They can cause a high variance of certain estimators. This is why it is important to develop effective and reliable numerical tools for the sampling of these rare events.

The problems caused by rare events in the effective sampling of the different quantities are caused by the stochastic behaviour of the dynamical system and a phenomenon called metastability. Metastability means that a dynamical system remains in a certain area for a comparatively long time before hopping rapidly into another metastable area. Therefore, metastability is one of the most challenging problems for effective sampling.

This thesis is about importance sampling strategies for metastable dynamical systems. The main idea of this thesis is to decrease the metastability to get estimators with a lower variance and reduce the sampling effort.

After an introduction and a presentation of the relevant theory we explore in Chapter 3 an idea of global optimization to decrease the metastability in the dynamical system. We show how the approach can be used for sampling thermodynamic and dynamic quantities and support the results with numerical experiments.

In Chapter 4 we use a local approach to decrease the metastability and thus build an importance sampling scheme for dynamic quantities. We use the experience of well-known MD algorithms to build good local perturbations. For the importance sampling scheme the algorithms have to be assimilated and combined with a result from stochastic analysis. The resulting algorithm is tested in different numerical settings.

In Chapter 5 we consider two different methods (Gradient descent and Cross-Entropy method) which have been proposed for finding the optimal perturbation in terms of variance reduction. For the gradient descent we develop different gradient estimators and for the Cross-Entropy method we develop a non-parametric representation of the optimal perturbation. The results are supported by numerical examples.

The thesis finishes with a summary of our findings and an outlook on future research.

Acknowledgements

There are many people who have supported me the last years and without them I would not be here. First of all I would like to thank PD Dr. Marcus Weber and Prof. Dr. Carsten Hartmann for giving me the possibility to write this thesis and introducing me in this interesting and interdisciplinary field of research. I would like to thank Tony Lelièvre for my research stay in Paris at École des Ponts. It was a stimulating experience to visit the institute and discuss with so many researchers. Furthermore, I would like to thank Prof. Dr. Manfred Opper for sharing his enthusiasm and his advices on Gaussian Processes.

Furthermore, I would like to thank Han Lie and Wei Zhang for the many discussions we had and their answers to many of my questions. Lara Neureither, Gottfried Hastermann and Adam Nielson for walking this way together and sharing thoughts. Franziska Erlekam and Luzie Helfmann for reading parts of my thesis.

I would like to thank my parents and my sister for their constant support and advice.

Finally, I would like to thank Miriam for being there whenever I needed it. Without you writing this thesis would not have been possible.

This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 "Scaling Cascades in Complex Systems", A05 (Probing scales in equilibrated systems by optimal nonequilibrium forcing).

Contents

1 Introduction

Molecular Dynamics simulation (MD) is a widely used tool in many different research areas, e.g material sciences or computational drug design. It is used to perform simulations to validate different claims or to predict certain phenomena. These simulations can give interesting insights into specific behaviour of the molecule beyond the experimental understanding. This can be used to understand certain behaviour within the molecule very deeply and leads to the opportunity to design molecules which have very special features; see e.g. [80]. In the last years MD has become more and more important and this is why MD is sometimes called the third way of science besides theory and experiments.
Even though Molecular Dynamics has brought some great achievements in the last years there are still some unsolved problems. The simulations of a molecule involve an integration over a very long time scale and a very large number of particles. To capture the small bond vibrations one has to discretize the equations of motion at the atomistic level in the order of one femtosecond (10^{-15}). The time scale of interesting phenomena is in the range of microseconds to hours ($\sim 10^{-6} - 10^{3}$) depending on the application; see e.g.[78]. This shows that molecules have an inherent multi-scale behaviour which arises from this variety of space and time scales and the interaction between them. Furthermore, many phenomena emerge from a collective behaviour of the particles and strongly depend on the number of particles and the time horizon. This is why, for example stable, integration schemes for large systems are of interest for the MD community. The latest results on this topic can be found in e.g. [55].
Apart from stable integration schemes and other problems the evaluation of different quantities of interest is a problem. Due to the large time scale and the large number of particles involved the quantities of interest cannot be calculated analytically. This is why these quantities of interest are approximated by empirical averages. The problem which arises in this approximation context is often called the sampling problem and we will focus on this in the thesis.

Sampling Problem

The sampling problem arises from the fact that the interesting quantities, expressed as averages, cannot be calculated analytically and thus have to be sampled. Due to the analytically intractability of the large non-linear dynamical system these quantities are approximated by empirical averages over a number of realizations

of the underlying dynamical system. The quantities of interest in MD are either thermodynamic quantities or dynamic quantities. Before explaining the different quantities in more detail and also the different strategies which can be applied to sample these, let us have a closer look what the difficulty of the sampling problem is.

One of the main problems in the sampling of these dynamical systems is the discrepancy between microscopic and macroscopic scales. The most important reason of this discrepancy is related to a phenomenon which is called metastability and the resulting rare events of the dynamical system. Metastability means the system remains in some region of the configuration space for a long time (called metastable set or metastable area) before it changes rapidly into another metastable set. This metastable behaviour implies that the transitions between metastable sets are rare events and thus not often observed. But these rare events are very important because they characterize the dynamical system. Furthermore, metastability causes the slow convergence of the empirical approximations of the interesting quantities.
When sampling metastable dynamical systems in order to observe the rare events the metastable sets are explored in detail while the transition area is only explored tenuously. So much of the sampling effort is used to investigate the metastable sets. In order to observe the relevant transitions which are rare events a long time has to be simulated to observe them. Therefore, the calculation of the different quantities is very much affected by the rare events. Metastability is also the reason that the approximation of the quantities of interest by empirical averages (also called Monte Carlo methods) have a large variance. Furthermore, the large variance of the empirical averages causes a large statistical error. Usually, the statistical error gets worse the smaller the probability of the rare event is. Let us consider an example of a metastable dynamical system to give an intuitive understanding before describing the different quantities of interest in more detail.

Consider a particle moving in a bistable potential V (see 1.1). The time evolution of the position of the particle, denoted by x_t, is described by

$$dx_t = -\nabla V(x_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad x_0 = x. \tag{1.1}$$

A typical trajectory is shown in 1.1. The position of the particle shows a metastable behaviour. It remains in the area around the left minimum for quite a long time before it hops into the area around the right minimum. The time of the transition is, compared to the time spent in the minimum, very short. The metastability arises from a so-called energetic barrier. In order to migrate from one metastable region into the other one a local maximum has to be crossed.

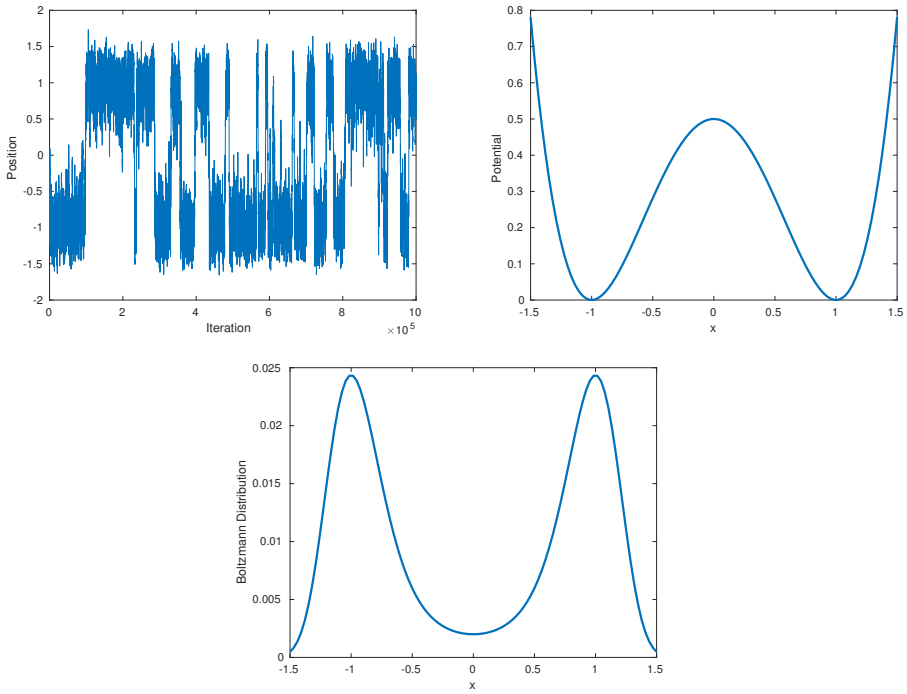


Fig. 1.1: Typical trajectory of a metastable dynamical system (upper left), bistable potential (upper right) and the resulting Boltzmann distribution (down middle).

On a more abstract level metastability is related to the multimodality of the underlying probability given by the dynamics and the separation of the modes by small probability regions. A multimodal distribution is a distribution with two or more modes (see figure 1.1). Metastability now means that there are high probable regions which are disconnected by low probable regions. In MD, for example, the stable confirmation of a molecule corresponds to a state of low energy and thus of high probability while a transition region corresponds to a state of high energy and thus of low probability. For a conformational change the state of low energy has to be left and a barrier has to be crossed. So observing a transition from one mode into the other mode in a metastable dynamical system a region of low probability has to be crossed.

In general there are two types of metastability: energetic and entropic. The energetic metastability occurs when the dynamical system has to cross level sets of the potential function V. The entropic metastability occurs because of steric constraints. The main difference between the energetic and entropic metastability can be seen when the temperature in the system is changed. For the energetic metastability the typical time scale to leave the metastable region grows exponentially when the temperature decreases. For the entropic barrier the change of temperature is asymptotically equivalent to a linear rescaling in time cf. [58].

Due to the strong influences of the metastability on the convergence of the estimators it is at the heart of the numerical challenges. Let us now have a closer look at the different quantities of interest in MD.

Quantities of interest

After describing the main cause of the sampling problem we are going now to have a closer look at the different quantities of interest and how the sampling is affected by the metastability. In molecular dynamics there are two different kinds of quantities which are of interest: thermodynamic quantities and dynamic quantities.

Thermodynamic quantities

Thermodynamic quantities are averages of observables of the dynamical system at equilibrium with respect to the probability measure of the configuration space. Consider a dynamical system which is only described by its position x. Furthermore, let us consider an observable $\varphi(x)$. Then the thermodynamic quantity is given by

$$\mathbb{E}_\nu[\varphi] = \int_{\mathcal{D}} \varphi(x)\nu(x)dx \tag{1.2}$$

where the probability measure is given by the Boltzmann-Gibbs measure

$$\nu(x)dx = Z_\nu^{-1} \exp(-\beta V(x))dx, \quad Z_\nu^{-1} = \int_{\mathcal{D}} \exp(-\beta V(x))dx. \tag{1.3}$$

Thermodynamic quantities are difficult to approximate because of the large state space $\mathcal{D}$ which is included in this calculation. One possible strategy of sampling these thermodynamic quantities is to replace the average over the state space by an average over time. In order to do this it is necessary that the dynamical system is ergodic with respect to the equilibrium probability measure (we will explain ergodicity in the next chapter). The state space integral can then be expressed as a time integral

$$\int_{\mathcal{D}} \varphi(x)\nu(x)dx = \lim_{t\to 0} \int_0^t \varphi(x_s)ds. \tag{1.4}$$

In general it is very hard for a large dynamical system to check that the system is ergodic.

The main problem which is caused by metastability in order to sample thermodynamic quantities is that the metastability hinders the state space exploration. As already seen in the example the dynamical system will spend the most time of the sampling in the metastable set. So to explore the state space sufficiently many long time simulations have to be calculated. Even though expressing the thermodynamic quantity as a time integral is very elegant, this approach does not overcome the slow state space exploration due to metastability.

Dynamic quantities

The other quantities of interest in MD are dynamic quantities. These quantities are given by an expectation over some path functional φ

$$\mathbb{E}_{\mathbb{P}}[\varphi(x_{0:\tau})] \tag{1.5}$$

where $x_{0:\tau}$ is a random trajectory (or a path) of length τ and $\mathbb{P}$ is a path measure. Dynamical quantities, for example, are transition probabilities or exit times from metastable sets. These quantities explicitly depend on the dynamical system. Let us consider as an example the exit time from a metastable region $\mathcal{S}$ of the dynamical system given in (1.1). The exit time is again a random variable defined as

$$\tau = \inf\{t > 0, x_t \notin \mathcal{S}\}.$$

It is implicitly assumed that $x_0 \in \mathcal{S}$. In order to sample τ it is necessary to sample paths $(x_t)_{0\leq t\leq \tau}$. So calculating dynamic quantities is again a sampling problem.
The main problem in sampling dynamic quantities is that due to the metastability the trajectory has to go through a low probability region. The dynamical system in (1.1) is following a negative gradient descent. So for a small diffusion the system will go into the next minimum and stay there until the random perturbation will kick it out. If the diffusion is much smaller than the drift term it is unlikely to sample one of these trajectories leaving the metastable set.

Difference

The main difference of the two quantities of interest is the probability measure and the random variable which are considered. In the case of thermodynamic quantities the $x \in \mathbb{R}^n$ are mircostates of the dynamical system. These are usually real valued random variables e.g. positions of the atoms. The considered probability measure is the Boltzmann-Gibbs measure defined on the configuration space as shown in equation (1.2).
For dynamic quantities the random variables are paths of the dynamical system $x_{0:\tau} \in \mathcal{C}([0,\tau] : \mathbb{R}^n)$. The considered probability measure is the path measure defined on the space of all possible trajectories namely the Wiener measure.
One general strategy to overcome the sampling problem is importance sampling which is introduced in the next section.

Importance sampling

In the literature one can find different approaches how the problems of rare event sampling can be overcome, e.g control variates, splitting methods and importance sampling; see [68] for details on all mentioned methods. We will mainly focus on importance sampling because it is the tool used in this thesis.

The main idea of importance sampling is to sample the quantity of interest from a different probability distribution and correct the error which has been introduced because of this substitution. The resulting importance sampling estimator can again expressed as an expectation and so Monte Carlo methods can be used again to approximate these estimators (details are given in Chapter 2). Let us consider as an example the importance sampling estimator for thermodynamic quantities. It is given by

$$\mathbb{E}_\nu[\varphi] = \int_{\mathcal{D}} \varphi(x)\frac{\nu(x)}{\mu(x)}\mu(x)dx \tag{1.6}$$

where μ is now the new probability measure. We clearly see that this approach only works if μ is never zero for any random variable where ν is positive. If this assumption holds the two considered measures are said to be absolute continuous with respect to each other. Since we are interested to reduce the variance the first question which arises is how can we choose μ such that we have a variance reduction. For this we have to calculate the variance of the empirical estimator denoted by $\bar{\varphi}$. We assume that the different simulations (say we have N simulations) of the quantity of interest are independent from each other (we assume them to be i.i.d.). The variance of the estimator is then $1/N$ times the variance of one realization of the quantity of interest denoted by $\bar{\varphi}_i$. The variance is given by

$$\begin{aligned} N\text{Var}[\bar{\varphi}_i] &= \int_{\mathcal{D}} \left(\varphi(x)\frac{\nu(x)}{\mu(x)} - \mathbb{E}_\nu[\varphi]\right)^2 \mu(x)dx \\ &= \int_{\mathcal{D}} \left(\frac{\varphi(x)^2\nu(x)^2}{\mu(x)} - 2\mathbb{E}_\nu[\varphi]\varphi(x)\nu(x) + \mathbb{E}_\nu[\varphi]^2\mu(x)\right)dx \\ &= \int_{\mathcal{D}} \left(\frac{\varphi(x)^2\nu(x)^2}{\mu(x)}\right)dx - \mathbb{E}_\nu[\varphi]^2 \end{aligned}$$

To find now the optimal bias in terms of variance reduction we see from the above equation that we have to minimize the integral in the last expression. Doing this we find

$$\begin{aligned} \int_{\mathcal{D}} \left(\frac{\varphi(x)^2\nu(x)^2}{\mu(x)}\right)dx &= \int_{\mathcal{D}} \left(\frac{\varphi(y)^2\nu(y)^2}{\mu(y)^2}\right)\mu(y)dy \\ &\geq \left(\int_{\mathcal{D}} \frac{|\varphi(y)|\nu(y)}{\mu(y)}\mu(y)dy\right)^2 = \left(\int_{\mathcal{D}} |\varphi(y)|\nu(y)dy\right)^2 \end{aligned}$$

where we have used Jensen's inequality. For the minimization we can drop the square. We know that we will only have equality if and only if the random variable is almost surely constant. So the integral term will only be optimal for $|\varphi(y)|\nu(y)$ being constant which can only be true if $\frac{|\varphi(x)|\nu(x)}{\mu(x)}$ is almost surely constant. From this we see that the optimal probability measure is given by

$$\nu_{opt}(x) = \frac{\varphi(x)\nu(x)}{\int_{\mathcal{D}} |\varphi(y)|\nu(y)dy}. \tag{1.7}$$

This optimal probability measure would lead to a zero variance estimator. Unfortunately, the optimal probability measure depends on the quantity we would like to sample. So from this short calculation we see the main dilemma of importance sampling. In order to sample something without variance we have to know what we want to sample.

Even though we have only shown the problems for importance sampling of thermodynamic quantities similar problems arise also for dynamic quantities and the different probability measures. Furthermore, we will see in the next chapter that the optimal bias for dynamic quantities, which would give a zero variance estimator can be expressed as the solution of a non-linear partial differential equation.

Related Works

The problems of importance sampling are very well-known in the literature cf. [10] or [68] and especially in the computational physics community many algorithms and methods have been proposed to deal with the sampling problem in the MD context. In this paragraph we are going to mention the related works which have been done so far very briefly. We are going to elaborate the methods relying on importance sampling ideas further in the next chapter and refer to other articles giving more details about the other methods.

Thermodynamic quantities

Problems in sampling thermodynamic quantities caused by metastability have been known for a long time in the computational physics community. Methods to overcome these problems using biasing techniques are for example Metadynamics [53], Varaitional approach [86], Adaptive Biasing Force [17], Hyperdynamics [89] or Wang Landau [91]. All methods have in common that they introduce a bias which reduces the metastability in the dynamical system. Since the bias changes the configuration space also the equilibrium measure is changed. This is why importance sampling techniques can be used to correct the sampling results.

Another well-known technique for overcoming problems with metastability is simulated annealing [84]. The idea is to sample the dynamical system for a higher temperature such that the system is less affected by the potential barriers. Due to the fact that the Boltzmann distribution depends on the temperature the high temperature sampling can be again related to a low temperature using importance sampling approaches. Since both methods presented so far can use importance sampling we will elaborate on these in the next chapter.
Also Replica exchange was invented to sample thermodynamic quantities and state space exploration cf. [83] or [90]. The idea of this method is to interchange information coming from a high temperature sampling with a low temperature sampling. For this the system is sampled at different temperatures in parallel and every now and then the positions of the atoms are changed. This way the low temperature sampling can explore much more from the state space without artificial forcing and there is no need to correct the sampling results. More details on replica exchange are presented in chapter 3.

Dynamical quantities

The methods which have been designed for the sampling of dynamic quantities can be divided into two classes, on the one hand splitting methods (like Forward Flux Sampling [1], Adaptive Multilevel Splitting [13] or Milestoning [29]) and on the other hand importance sampling e.g. [95]. The importance sampling approach can be further divided into general approaches and approaches in the large deviation context e.g. [20]. Let us quickly summarize the main ideas behind the different approaches.
The main idea of the splitting method is to introduce intermediate levels in the sampling. In this way only parts of the whole trajectory are sampled and the computational effort is focused on the low probable part of the trajectory. For more details on the different methods see the above mentioned citations and the references therein.
Importance sampling for path dependent quantities for stochastic differential equations has already been proposed by Milstein in 1954 [63]. Milstein suggested to perturbed the drift term of the stochastic differential equation and correct the expectation with the so-called Girsanov transformation. The Girsanov transformation or Girsanov's theorem gives an explicit way how the change of the drift term of the SDE has an effect on the underlying path measure (we will present the theorem in chapter 2). Furthermore, Milstein showed that one can find a zero variance estimator, if a solution of a associated Bellman equation is used as an additional drift. In the continuous case this is related to the non-linear Hamilton-Jacobi-Bellman equation (also presented later). But this connection also shows how difficult it is to design a bias giving a zero variance estimator.

In [95] the method proposed by Milstein was combined with a dimensionality reduction technique. But in order to apply the method the equilibrium measure has to be known. In [38] an optimization framework was proposed to find the optimal bias. The optimal feedback control is projected into some space of ansatz functions so only feedback controls of the form $u_t = \sum_{i=1}^{m} a_i b_i(x)$ are considered. The problem of finding the optimal bias can then be reformulated as an optimization problem of finding the optimal weights a_i. A detailed derivation of the optimization framework can be found in Chapter 5.
In the large deviations context different methods have been proposed by [20, 22, 21, 23]. The here presented importance sampling schemes all rely on asymptotic arguments that the noise of the dynamical system is going to zero. Since we are going to develop importance sampling schemes which do not rely in these asymptotic arguments we are going to briefly summarize this approach in the next chapter and point to the relevant literature for details.

Research questions

From the above presentation we see that designing a good bias or a good method for the effective sampling of different quantities of interest in MD is a difficult task. We have also seen that different approaches have been presented in the literature but there are still many open questions. In this thesis we address the following research questions:

- **Observation:** The metastability is cased by the energetic or entropic barriers.
  - **Question:** How can barriers which cause metastability be decreased without apriori knowledge of the system?
  - **Question:** Can this approach be used for the sampling of thermodynamic or dynamic quantities?
- **Observation:** In the MD community much effort has been made to overcome the problem of metastability for the sampling of thermodynamic quantities.
  - **Question:** Can these algorithms be used for the sampling of path dependent quantities?
- **Observation:** In theory there is a zero variance estimator. In the literature an optimization framework was developed and two algorithms have been proposed: the gradient descent method and Cross-Entropy method.
  - **Question:** Is there an efficient way how the gradient can be calculated or estimated for the gradient descent?
- **Observation:** There are other ways in the literature how function approximations can be realized e.g. Gaussian Process.

– **Question:** Can we rephrase the Cross-Entropy method such that we can connect the method with these approaches?

Structure of the thesis

This thesis is structured as followed. In Chapter 2 the relevant theory of the thesis is given. We are going to summarize how molecular dynamics and stochastic differential equations can be connected. After a brief presentation of some relevant stochastic analysis, Monte Carlo methods and importance sampling are elaborated further. We will have a closer look how the connection between optimal control and importance sampling is derived and show how the previously presented well-known MD methods for thermodynamic quantities can be connected to importance sampling.
Chapter 3 addresses the first research question. In order to lower the barriers of the dynamical system without apriori knowledge we are going to explore if a global perturbation can be used to decrease the barriers. The idea is based on the convolution approach of global optimization. After numerical tests that show the decreased metastability we will show how the convolution approach can be used for the sampling of thermodynamic quantities and integrate this approach into the replica exchange algorithm. Furthermore, we will show that the convolution can also be understood in the linear response theory. In the end of the chapter we are going to use the approach for the sampling of dynamic quantities and support our results with numerical tests.
Chapter 4 addresses the second research question. As we have already seen many different algorithms have been invented for the sampling of thermodynamic quantities. This is why we use this algorithms in a assimilated way to change the barrier locally. Combing these assimilated methods with Girsanov's theorem we can build effective importance sampling methods for the sampling of dynamic quantities. The resulting algorithm is an importance sampling algorithm in path space and independent from the used bias the estimator is unbiased. We show that Girsanov's theorem can be applied under some assumptions and that the method does not change the ergodicity of the system. In the end of the chapter the algorithm is tested for different numerical examples and we conclude with a discussion of our findings.
The last chapter is devoted to the optimization of the bias. As already mentioned above there is a bias which would give a zero variance estimator. In the literature there are two methods based on an optimization framework to approximate the optimal bias. In the first part of the chapter we are going to address the third research question and try to find good gradient estimators for the gradient descent method. We first use the so-called Malliavin gradient descent approach to develop a gradient descend in one-dimension and then develop two gradient estimators for high-dimensional problems. We support our results with numerical tests. In the

second part we address the research question four. For this we derive a kernelized version of the Cross-Entropy method and show how this method can been interpreted as a Gaussian Process approach. In the end another numerical example is given to show the application of the method.
In the end we will have a summary and a discussion of the findings followed by an outlook about future research.

2 Theory

In this chapter we are going to review the relevant theory for this thesis. In the first part the relevant models for Molecular Dynamics Simulations and the connection to stochastic differential equations are described. In the second part the theory of stochastic processes is briefly summarized. Then the main idea of importance sampling and the connection to optimal control is presented. In the end we present some of the well-known methods for the effective sampling of thermodynamic and dynamic quantities and discuss their connection to importance sampling.

2.1 From MD to SDEs

In MD molecules are considered to be mechanical systems. The most precise model for mechanical systems at the smallest scales is given by quantum mechanics. In quantum mechanics one considers the Schrödinger equation to describe behaviour of matter. The equation cannot be solved analytically and even numerical solutions to this equation are difficult because of the many dimensions involved in the calculation. The most serious problem is the computational complexity which grows extremely fast with the number of atoms which are involved in the simulation. Only Butane has 34 electrons and 14 nuclei. This is why the quantum mechanics approach is only used for very small molecules or in combination with a coarser model in which only a specific region is investigated quantum-mechanically [55].

A classical model can be derived from the Schrödinger equation by using the Born-Oppernheimer approximation [7]. For this one can assume that heavy nuclei dominates the movement such that the relatively light electrons do not have to be simulated. With this approximation the motion of molecules is only described on an atomic level. But even this model can give enough inside information to study interesting phenomena. The equations of motion are now equations of a classical mechanical system given by Newton's second law (force = mass $\times$ acceleration)

$$M\frac{\partial^2}{\partial t^2}p = -\nabla V(p) \tag{2.1}$$

where M is a diagonal mass matrix, p is the position of the atoms and V is the potential which models the atom-atom interaction. The interaction potential is sometimes obtained by solution from the Schrödinger equation for fixed nuclei

position cf. [55]. But since these calculations are very expensive, empirical potentials are often used; see [55] or [58]. Furthermore, these potentials encode the physics of the model because they model the atomic interaction of the particles in the system.

In the literature there are different descriptions of mechanical systems. In molecular dynamics Hamiltonian mechanics is often used. In Hamiltonian mechanics the dynamical system is described in position p and momenta q, also called phase space. The total energy of a system is given by the sum of the kinetic energy and potential energy

$$H(p,q) = \frac{1}{2} q^T M^{-1} q + V(p). \tag{2.2}$$

The evaluation of the system is given by

$$\dot{p} = M^{-1} q \quad \dot{q} = -\nabla V(p). \tag{2.3}$$

Since the potential only describes the internal interactions of the atoms the equations of motion model an energetically closed system. Furthermore, the total energy is preserved under the dynamics. But in molecular dynamics an isolated molecular system is only of partial interest. In general, it would be desirable to simulate a system which is in contact with a heat bath to model energy transfer.

An appropriated model is given in the statistical physics framework. The molecular system in equilibrium in contact with a heat bath generates a probability density function

$$\mu(dp, dq) = Z^{-1} \exp(-\beta H(p,q)) dq dp, \quad Z = \int \exp(-\beta H)$$

where H is the Hamiltonian given in (2.2), $\beta = \frac{1}{k_B T} > 0$ called the inverse temperature and $Z < \infty$ is the normalization constant [58].

Due to the energy preservation of the Hamiltonian mechanics the system is not ergodic cf. [79]. So in order to sample the canonical distribution the environment or solvent has to be taken into account. The environment is necessary to model the energy transfer which enables the system to explore the entire state space. But it also inflates up the system size which has to be simulated. However, to simulate a system that is too large is in general not possible. There are several ideas to overcome this problem like thermostats and other techniques cf. [55].

Another approach is to replace the deterministic molecular model in a solvent by a stochastic model where the energy transfer from the solvent is modelled as a stochastic force. The usage of stochastic models is natural in MD because finite Hamiltonian models suppress microscopic interaction between the molecular system and the environment due to simplification cf. [55]. Furthermore, the results are

more reliable, the computational difficulty is reduced and ergodicity can be achieved; see e.g. [55]. A partial justification for modelling mechanical systems by stochastic differential equations can be given by deriving the Langevin equation for a small simple mechanical model which is in contact with a heat bath. There are different approaches in the literature e.g [69, 76] or [96]. We are going to summarize the approach presented in [69] to motivate the connection between MD and SDEs.

Let us consider a one-dimensional particle in contact with a heat bath which is assumed to be a system with infinite heat capacity at temperature β^{-1}. We assume the system to be at thermodynamic equilibrium at time $t = 0$. We will model the heat bath as a system of infinitely many harmonic oscillators. We know that a collection of harmonic oscillators can be expressed as a Hamiltonian system with a corresponding Boltzmann-Gibbs distribution. An extension of the finite dimensional system of harmonic oscillators can be done by considering the wave equation. This wave equation can be seen as a infinite-dimensional Hamiltonian system. But the extension of the Boltzmann-Gibbs distribution into a infinite dimensional space has to be done carefully because the Lebesgue measure does not exist in this infinite dimensional space. However, the Hamiltonian of the wave equation is a quadratic function such that the corresponding Boltzmann-Gibbs distribution is Gaussian and by considering a Hilbert space the theory holds.
We assume that the dynamical system (the particle coupled to the heat bath) is described by the Hamiltonian

$$H(p, q, \varphi, \theta) = H(p, q) + H_{HB}(\varphi, \theta) + H_I(p, \theta) \tag{2.4}$$

where H is the Hamiltonian for the particle in position p and momenta q, H_{HB} is the Hamiltonian of the heat bath (wave equation) given by

$$H_{HB}(\varphi, \theta) = \frac{1}{2} \int_{\mathbb{R}} \left(|\varphi(x)|^2 + |\frac{\partial}{\partial x}\theta(x)|^2 \right) dx \tag{2.5}$$

where (φ, θ) is a field and H_I denotes the interaction of the particle and the heat bath. Furthermore, we assume that the coupling of the heat bath and the particle is only through the position and θ.
Assuming that the particle is moving in a confining potential $V(p)$ (a definition is given later) the Hamiltonian is given by

$$H(p, q) = \frac{q^2}{2} + V(p). \tag{2.6}$$

For the coupling we assume that is is linear given by

$$H_I(p, \theta) = \int_{\mathbb{R}} \theta(x)\rho(x - p)dx \tag{2.7}$$

where ρ is the so-called charge density; see [69] and the references therein for details. Since the position of the particles does not change too much because it is moving in a confining potential we can use a Taylor expansion. This together with a integration by party gives

$$H_I(p,\theta) \approx p \int_{\mathbb{R}} \frac{\partial}{\partial x}\theta(x)\rho(x)dx. \tag{2.8}$$

Putting all this together we get

$$H(p,q,\varphi,\theta) \approx \frac{q^2}{2} + V(p) + \frac{1}{2}\int_{\mathbb{R}} \left(|\varphi(x)|^2 + |\frac{\partial}{\partial x}\theta(x)|^2\right)dx + p\int_{\mathbb{R}} \frac{\partial}{\partial x}\theta(x)\rho(x)dx \tag{2.9}$$

as the Hamiltonian of the particle couple to a infinite heat bath. From this we can derive the equations of motion

$$\dot{p} = q \quad \dot{q} = -V'(p) - \int_{\mathbb{R}} \frac{\partial}{\partial x}\theta(x)\rho(x)dx \tag{2.10}$$

$$\frac{\partial}{\partial t}\theta = \varphi \quad \frac{\partial}{\partial t}\varphi = \frac{\partial^2}{\partial x^2}\theta + p\frac{\partial}{\partial x}\rho. \tag{2.11}$$

From this we can solve the equations for the field (2.11) and substitute the solution into the equations for the particle (2.10) in order to derive a closed-form equation for the particle. After a lengthy calculation which we are going to skip one can derive a so-called generalized Langevin equation (GLE) which is a stochastic integro-differential equation given by

$$\ddot{p} = -V'(p) - \int_0^t \gamma(t-s)\dot{p}(s)ds + F(t) \tag{2.12}$$

where $F(t)$ is a mean-zero stationary Gaussian process with autocorrelation function. The GLE for the particle is Newton's equation of motion augmented by a linear dissipation term which depends on the history of the particle and a stochastic forcing. The noise term is Gaussian and stationary because we have assumed that the heat bath is at equilibrium at time $t = 0$. The GLE is equivalent to the full Hamiltonian dynamics with random initial conditions distributed according to the Boltzmann-Gibbs measure on the Hilbert space [69]. But due to the history-dependent term in the integral the GLE is not an easy model to work with. However, the GLE approach is a very appealing approach for high-dimensional dynamical systems and this is why it is still a very lively field of research; cf. [40] and the references therein.
In the second step we can approximate the non-Markovian GLE by the Langevin equation. For this one considers a vanishing correlation time of $\gamma(t) \to \delta(t)$. This corresponds to a localization of the coupling in the full Hamiltonian system $\rho(x) \to \delta(x)$.

The Langevin dynamics is given by

$$dp_t = M^{-1}q_t dt \tag{2.13}$$

$$dq_t = -V'(p_t)dt - \gamma M^{-1}q_t dt + \sqrt{2\gamma\beta^{-1}}M^{1/2}dB_t \tag{2.14}$$

where q is the momentum of the system and p is the position, $\gamma > 0$ is the so-called friction term, V is an interaction potential describing the atom-atom interaction and M^{-1} is the mass matrix.
From the Langevin equation one can derive an even more simplified model by considering a large γ limit or scale γ with the inverse of the mass (replace γ by γM^{-1}) which is also known as the zero mass limit. For this derivation we are going to follow the presentation in [55]. We assume that the inertial dynamics is dominated by collisional effects. Furthermore we assume that the acceleration can be neglected and we denote $u = M^{-1}q$. The Langevin equation reduces to

$$dp_t = u_t dt \quad 0 = -V'(p_t)dt - \gamma M u_t dt + \sqrt{2\gamma\beta^{-1}}M^{1/2}dB_t.$$

By solving now the second equation for u we find

$$dp_t = -\gamma M^{-1}V'(p_t)dt + \sqrt{2\gamma^{-1}\beta^{-1}}M^{1/2}dB_t \tag{2.15}$$

The transition from Langevin to overdamped Langevin is often referred to as Kramers- to Smoluchowski- limit.
The case $\gamma = 1$ and $M = I$ where I is the identity matrix is often considered for simplification. The above model is called the overdamped Langevin equation which is also known as 'Brownian Dynamics'. The multidimensional formula is given by

$$dx_t = -\nabla V(x_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad x_0 = x. \tag{2.16}$$

Above, x_t denotes the state of the system at time $t \geq 0$, $\beta > 0$ is a scaling factor for the noise associated with the temperature and the Boltzmann constant, often called the inverse temperature, and B_t is a standard n-dimensional Brownian motion with respect to the probability measure $\mathbb{P}$ on some probability space $(\Omega, \mathbb{P}, \mathcal{F})$, and $V : \mathbb{R}^n \to \mathbb{R}$ is a sufficiently smooth (e.g. $\mathcal{C}^\infty$) potential for technical reasons. This simplified stochastic model is going to be the root model of this thesis.
This change of model also introduces a change of the point of view on molecular dynamics. One does not consider a time evolution of a partial differential equation (PDE) any more now, one considers paths in some finite-dimensional Euclidean space.

Let us next give a brief introduction into molecular potentials for completeness. Since the topic of molecular potentials is a research field on its own we just want to present the main idea without going into too much detail.

Molecular potentials
The potential describes the energy of a molecule. It consists of a sum of different interatomic potentials which describe the interaction within 2, 3 or 4 atoms

$$V(p_i, p_j, p_k, p_l) = \sum_{i,j} V_{short}(p_i, p_j) + \sum_{i,j,k} V_{inter}(p_i, p_j, p_k) + \sum_{i,j,k,l} V_{long}(p_i, p_j, p_k, p_l)$$

where p_i is the position vector of atom i. A typical expression given in a detailed way may satisfy

$$V(p_i, p_j, p_k, p_l) = \sum_{bonds} \frac{1}{2} k_b (p_{ij} - p_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2$$
$$\sum_{torsion} \frac{U_n}{2}(1 + \cos(n\phi - \phi_0)) + \sum_{LJ} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^2}{p_{ij}^2} - \frac{\sigma_{ij}}{p_{ij}} \right)^6 + \sum_{elec} \frac{r_i r_j}{p_{ij}}$$

where $p_0, \theta_0, U_n, \phi_0, \epsilon, \sigma, r$ are bound specific constants, $p_{ij} = \|p_i - p_j\|$ is the distance between two atoms, ϕ is an angle spanned between three atoms and θ is an angle spanned within four atoms. The different interatomic potentials are designed to mimic a certain behaviour which was observed in experiments. The parameters have to be adjusted to match the properties of the system known from experiments or quantum mechanical modelling. So the behaviour of the molecule is a complex interplay of all different atomic interactions; see e.g. [78], [85] or [33] for further details.

2.2 Stochastic differential equations

As we have seen in the first section molecular dynamics can be described by stochastic models. We have seen how the root model for this thesis, namely the overdamped Langevin equation, was derived. The overdamped Langevin equation is a stochastic differential equation (SDE). In the next paragraph we are going to give a brief summary about the theory of SDEs, Monte Carlo methods and the connection of importance sampling in paths space and optimal control.

2.2.1 Properties of SDEs

Consider a stochastic differential equations in a more general form satisfying

$$dx_t = b(t, x_t)dt + \sigma(t, x_t)dB_t, \quad x_0 = x \tag{2.17}$$

where $b(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are some Borel-measurable functions and B_t is a standard Brownian motion with respect to the probability measure $\mathbb{P}$ on some probability space $(\Omega, \mathbb{P}, \mathcal{F})$. The function b is called the drift term of the SDE and σ is called the

diffusion term. Uniqueness and existence of the SDE are given under some Lipschitz condition on the drift and the diffusion as we will see in the next theorem.

Theorem 1 ([67]). *Let $T > 0$ and $b(\cdot,\cdot) : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma(\cdot,\cdot) : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ be measurable functions satisfying*

$$|b(t,x)| + |\sigma(t,x)| \leq C(1+|x|); \qquad x \in \mathbb{R}^n, t \in [0,T]$$

for some constant C, (where $|\sigma|^2 = \sum |\sigma_{i,j}|^2$) and such that

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \leq D(1+|x-y|); \qquad x,y \in \mathbb{R}^n, t \in [0,T]$$

for some constant D. Let Z be a random variable which is independent of the σ-algebra generated by $B_s, s > 0$ and such that

$$\mathbb{E}[|Z|^2] < \infty.$$

The 2.17 with $x_0 = Z$ has a t-continuous solution.

For the model (2.16) this means that we have a strong solution, if the gradient of the potential is locally Lipschitz continuous. The considered diffusion term is always Lipschitz continuous because for the considered model it is state and time independent.

The SDE (2.17) is associated with a differential operator given by

$$\mathcal{L} = b \cdot \nabla + \frac{1}{2}\sigma\sigma^T : \nabla^2 \tag{2.18}$$

which is called the infinitesimal generator of the process x_t. We will consider generators to be defined on appropriate dense subsets of the domain of the generator such as $\mathcal{C}^\infty$ functions with compact support or whose derivatives grow at most polynomially.
In the following lemma the link between (2.17) and (2.18) is made explicitly

Lemma 1 ([58]). *Let φ be a compactly supported $\mathcal{C}^\infty$ function. Then*

$$\frac{\partial}{\partial t}[\mathbb{E}^x[\varphi(x_t)]]\Big|_{t=0} = \mathcal{L}\varphi(x) \tag{2.19}$$

where the expectation is taken over the Brownian paths.

Let us assume that the law of x_t at time t admits a density ρ with respect to the Lebesgue measure (see [58] for details). Then x_t is distributed according to $\rho(t,x)dx$.

The initial distribution will be denoted as $\rho(0,x)dx = \rho_0(x)$. It is wel- known in the literature that ρ satisfies the Fokker-Planck equation

$$\frac{\partial}{\partial t}\rho(t,x) = \mathcal{L}^\dagger \rho(t,x), \quad \rho(0,x) = \rho_0(x) \tag{2.20}$$

where $\mathcal{L}^\dagger$ denotes the L^2 adjoint of the operator $\mathcal{L}$ which is given by

$$\mathcal{L}^\dagger = -\text{div}(b\cdot) + \frac{1}{2}\nabla^2 : (\sigma\sigma^T). \tag{2.21}$$

Under some assumptions on the drift (see below), the long-term evolution of the law of the process (2.17) will converge to a equilibrium state. This equilibrium distribution is also called the stationary or invariant distribution.

Definition 1 ([69]). *A potential V is called confining if $\lim_{|x|\to\infty} V(x) = +\infty$ and*

$$\exp(-\beta V(x)) \in L^1(\mathbb{R}^n) \tag{2.22}$$

for all $\beta \in \mathbb{R}^+$.

Proposition 1 ([69]). *Let V be a smooth confining potential. The unique invariant distribution of the process* (2.16) *is the Gibbs distribution*

$$\rho_\beta(x) = \frac{1}{Z}\exp(-\beta V(x)) \tag{2.23}$$

where the normalization factor Z is the partition function

$$Z = \int_{\mathbb{R}^n} \exp(-\beta V(x))dx. \tag{2.24}$$

Furthermore, the process (2.16) *with the generator $\mathcal{L} = -\nabla V\nabla + \beta^{-1}\nabla^2$ is ergodic.*

Ergodicity means that for an ergodic dynamical system the average over long-time trajectories is the same as the average over phase space averages with respect to the underlying probability measure. The ergodicity of the dynamics implies that

$$\lim_{t\to\infty}\int_0^t \varphi(x_s)ds = \int_\Omega \varphi(x)\rho_\beta(x)dx \tag{2.25}$$

holds in a well-defined limit for some observable φ.
Ergodicity is a very important aspect for many methods developed in MD. In general observables which are averages with respect to the invariant distribution are very hard to calculate because of the high dimension of the state space. Often the invariant distribution is approximated by an empirical mean which is only sufficiently accurate for long-time samplings as we have seen in the introduction. In this way it is possible to calculate ensemble averages as averages over trajectories.

Another connection of SDEs and PDEs is given by the Feynam-Kac formula. The formula expresses solutions of certain PDEs as averages over paths of stochastic processes. Thus it is possible to solve PDEs by simulating SDEs.

Theorem 2 ([58]). *Let $\mathcal{S}$ be a C^∞ bounded domain of $\mathbb{R}^n$ and let $f : \mathcal{S} \to \mathbb{R}^n$, $v_0 : \mathcal{S} \to \mathbb{R}^n$ and $\varphi.\partial\mathcal{S} \to \mathbb{R}$ be three C^∞ functions. Let $v(t, y)$ be a smooth solution (e.g. C^1 in t and C^2 in y) to the boundary value problem*

$$\begin{aligned} \partial_t v &= \mathcal{L}v + fv \quad \text{for } t \geq 0,\ y \in \mathcal{S} \\ v &= \varphi \qquad \text{on } \partial\mathcal{S} \\ v(0, y) &= v_0(y). \end{aligned}$$

The smooth solution can be expressed as a conditional expectation $\mathbb{E}[\cdot|x_0 = x] = \mathbb{E}^x[\cdot]$

$$\begin{aligned} v(t, y) = \mathbb{E}^x\Big[\varphi(x_{\tau_\mathcal{S}}) \exp\Big(\int_0^{\tau_\mathcal{S}} f(x_s)ds\Big)\mathbf{1}_{\tau_\mathcal{S}<t}\Big] \\ +\mathbb{E}^x\Big[v_0(x_t) \exp\Big(\int_0^{t} f(x_s)ds\Big)\mathbf{1}_{\tau_\mathcal{S}\geq t}\Big], \end{aligned}$$

where $(x_t)_{t\geq 0}$ is the process satisfying (2.16) *and $\tau_\mathcal{S}$ is the first exit time of $(x_t)_{t\leq 0}$ from $\mathcal{S}$*

$$\tau_\mathcal{S} = \inf\{t \geq 0, x_t \notin \mathcal{S}\}. \tag{2.26}$$

Proof. Fix a time $t > 0$ and consider $u(s, y) = v(t - s, y)$ for $s \in [0, t]$. The function u satisfies

$$\begin{aligned} \partial_s u + \mathcal{L}u + fu = 0 \qquad & \text{s} \in [0, t], y \in \mathcal{S}, \\ u = \varphi \qquad & \text{on } \partial\mathcal{S} \\ u(t, y) = v_0(y). & \end{aligned}$$

We now use Itô calculus on $u(s, x_s) \exp(\int_0^s f(x_r)dr)$ for all $s \in [0, \min(\tau_\mathcal{S}, t)]$. In order to calculate the stochastic differential we are going to consider the function as $z(s)u(s, x_s)$ with $z(s) = \exp(\int_0^s f(x_r)dr)$ and so the differential is given by

$$d(z(s)u(s, x_s)) = zdu + udz + dudz. \tag{2.27}$$

By normal differentiation

$$dz(s) = z(s)f(x_s)ds. \tag{2.28}$$

By Itô's Lemma

$$du(s, x_s) = (\partial_s u + \mathcal{L}u)ds + \sigma\nabla u dB_t. \tag{2.29}$$

The term $dzdu$ can be neglected because it goes to zero very quickly (it is often written $dudz = 0$) cf. [67]. Thus the stochastic differential is given by

$$u(s, x_s) \exp(\int_0^s f(x_r)dr) =$$
$$u(0, x_0) + \int_0^s (\partial_s u(s, x_s) + \mathcal{L}u(s, x_s))z(s)ds + \sigma \nabla u(s, x_s)z(s)dB_s + z(s)f(x_s)u(s, x_s)ds$$

After rearranging terms we see

$$= u(0, x_0) + \int_0^s (\partial_s u(s, x_s) + \mathcal{L}u(s, x_s) + f(x_s)u(s, x_s))z(s)ds + \sigma \nabla u(s, x_s)z(s)dB_s$$
$$= u(0, x_0) + M_s$$

where

$$M_s = \int_0^s \sigma \nabla u(s, x_s)z(s)dB_s$$

is a local martingale. And since u and f are C^∞ and x_r lives in the bounded domain $\mathcal{S}$ up to time $\min(\tau_S, t)$, we conclude that

$$\begin{aligned} v(t, y) = u(0, y) =& \mathbb{E}^x \left[u(\min(\tau_\mathcal{S}, t), x_{\min(\tau_\mathcal{S}, t)}) \exp\left(\int_0^{\min(\tau_\mathcal{S}, t)} f(x_r)dr \right) \right] \\ =& \mathbb{E}^x \left[\mathbf{1}_{\tau_\mathcal{S} < t} u(\tau_\mathcal{S}, x_{\tau_\mathcal{S}}) \exp\left(\int_0^{\tau_\mathcal{S}} f(x_r)dr \right) \right] \\ &+ \mathbb{E}^x \left[\mathbf{1}_{\tau_\mathcal{S} \geq t} u(t, x_t) \exp\left(\int_0^t f(x_r)dr \right) \right] \\ =& \mathbb{E}^x \left[\mathbf{1}_{\tau_\mathcal{S} < t} \varphi(\tau_\mathcal{S}, x_{\tau_\mathcal{S}}) \exp\left(\int_0^{\tau_\mathcal{S}} f(x_r)dr \right) \right] \\ &+ \mathbb{E}^x \left[\mathbf{1}_{\tau_\mathcal{S} \geq t} v_0(x_t) \exp\left(\int_0^t f(x_r)dr \right) \right]. \end{aligned}$$

This concludes the proof. □

Example

Let us consider the mean exit time of a process satisfying (2.17) from a domain $\mathcal{S}$. Assume that $x_0 \in \mathcal{S}$ the first exit time

$$\tau_\mathcal{S} = \inf\{t \geq 0 : x_t \notin \mathcal{S}\}. \tag{2.30}$$

Now defining the mean first exit time as

$$\tau(x) = \mathbb{E}^x[\tau_\mathcal{S}] = \mathbb{E}[\inf\{t \geq 0 : x_t \notin \mathcal{S}\} | x_0 = x \in \mathcal{S}] \tag{2.31}$$

It is possible to derive from the general formulation of the Feynman-Kac formula that the mean first exit time can be calculated by solving a boundary value problem

$$\begin{aligned} \mathcal{L}\tau &= -1, \quad x \in \mathcal{D} \\ \tau &= 0 \quad x \in \partial\mathcal{D} \end{aligned}$$

where $\mathcal{L}$ is the generator of the diffusion process (2.16) cf. [69].

The last theorem of this section is Girsanov's theorem. The theorem is significant for the importance sampling strategies in path space presented later in this thesis. The theorem states that changing the drift of an Itô diffusion with a non-degenerate diffusion coefficient does not change the law of the process dramatically. The new law will be absolutely continuous with respect to the old law and the Radon-Nikodym derivative can be calculated explicitly. We will formulate a proposition based on Girsanov's theorem such that it is easier to see why it can be used for importance sampling and give a brief summary of the proof. A discrete derivation of the theorem can be found in the appendix; see Appendix 7.

Theorem 3 ([67]). *Let $x_t \in \mathbb{R}^n$ and $y_t \in \mathbb{R}^n$ be an Itô diffusion and an Itô process, reprecisely, of the from*

$$\begin{aligned} dx_t &= b(t, x_t)dt + \sigma(t, x_t)dB_t, \quad t \leq T,\ x_0 = x \\ dy_t &= (c(t, \omega) + b(t, x_t))dt + \sigma(t, y_t)dB_t, \quad t \leq T,\ y_0 = x \end{aligned}$$

where $b : \mathbb{R}^n \to \mathbb{R}$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ satisfy the necessary condition such that existence and uniqueness can be guaranteed. Suppose there exists a process $u(t, \omega)$ such that

$$\sigma(y_t)u(t, \omega) = c(t, \omega) \tag{2.32}$$

and we assume that u satisfies Novikov's condition

$$\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{1}{2}\int u^2(s, \omega)ds\right)\right] < \infty. \tag{2.33}$$

Then we define

$$M_t = \exp\left(-\int_0^t u(s, \omega)dB_s - \frac{1}{2}\int u^2(s, \omega)ds\right), \quad t \leq T, \tag{2.34}$$

$$\hat{B}_t = \int_0^T u(s, \omega)ds + B_t, \quad t \leq T \tag{2.35}$$

and a new probability measure $\mathbb{Q}$

$$d\mathbb{Q}(\omega) = M_T(\omega)d\mathbb{P}(\omega) \text{ on } \mathcal{F}_T \tag{2.36}$$

where F_T is the filtration given by B_t. Then we can define a new stochastic process

$$dy_t = b(y_t)dt + \sigma(y_t)d\hat{B}_t. \tag{2.37}$$

Therefore, the $\mathbb{Q}$-law of y_t^x is the same as the $\mathbb{P}$-law of x_t^x; $t \leq T$.

Let us formulate a proposition of this theorem such that it is easier to see why Girsanov's theorem can be used for importance sampling.

Proposition 2. *Let $x_t \in \mathbb{R}^n$ and $y_t \in \mathbb{R}^n$ be an Itô diffusion and an Itô process of the form*

$$dx_t = b(x_t)dt + \sigma(x_t)dB_t, \quad t \leq T, \quad x_0 = x \tag{2.38}$$

$$dy_t = (u(y_t) + b(y_t))dt + \sigma(y_t)dB_t, \quad t \leq T, \quad y_0 = x \tag{2.39}$$

where $b : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ satisfy Lipschitz conditions such that we can guarantee uniqueness and existence of the solution and the time $T < \infty$. Furthermore, we define for an adapted measurable process $a : \mathbb{R}^n \to \mathbb{R}$ the stochastic process

$$M_t = \exp(-\int_0^t a(y_s)dB_s - \frac{1}{2}\int_0^t a(y_s)^2 ds), \tag{2.40}$$

for all $t \in [0, T]$ and $\sigma(y_s)a(y_s) = u(y_s)$. Then, given that Novikov's condition

$$\mathbb{E}[\exp(\frac{1}{2}\int_0^T |a(y_t)|^2 dt)] < \infty \tag{2.41}$$

holds, for any function $f \in \mathcal{C}_0(\mathbb{R}^n)$ and any stopping time τ adapted to the filtration $\mathcal{F}_T$ (the filtration associated to the Brownian motion B in (2.38) *and* (4.7)*) we have*

$$\mathbb{E}^x_{\mathbb{P}}[f(x_{0:\tau})] = \mathbb{E}^x_{\mathbb{P}}[M_\tau f(y_{0:\tau})]. \tag{2.42}$$

Proof. We are going to give a short sketch of the proof here for completeness.

We define a new probability measure

$$d\mathbb{Q} := M_T d\mathbb{P} \quad \text{on } \mathcal{F}_T.$$

Then,

$$\hat{B}_t := \int_0^t a(y_s)ds + B_t \quad t \leq T$$

is a Brownian motion with respect to $\mathbb{Q}$, and in terms of $\hat{B}_t$ the process y_t can be represented by

$$dy_t = b(y_t)dt + \sigma(y_t)\hat{B}_t, \quad y_0 = x,\ t \leq T.$$

Therefore, the $\mathbb{Q}$-law of y_t with $y_0 = x$ is the same as the $\mathbb{P}$-law of x_t with $x_0 = x$ for $t \leq T$. This follows directly from the weak uniqueness of solutions of stochastic differential equations (see [67] p.71 Lemma 5.3.1). Due to the absolute continuity of the two probability measures $\mathbb{Q}$ and $\mathbb{P}$, we can use a change of measure to rewrite the expectation. Thus, for any function $f \in \mathcal{C}_0(\mathbb{R}^n)$ and any stopping time $\tau \leq T$ which is adapted to the filtration $\mathcal{F}_T$ we can write

$$\mathbb{E}_{\mathbb{P}}[f(x_{0:\tau})] = \mathbb{E}_{\mathbb{Q}}[f(y_{0:\tau})] = \mathbb{E}_{\mathbb{P}}[M_\tau f(y_{0:\tau})]$$

which gives us the desired result. □

By setting $b(\cdot) = -\nabla V(\cdot)$ and $\sigma = \sqrt{2\beta^{-1}}$ we have the metastable SDE model (2.16) and can use proposition 2 to reduce the metastability in the dynamical system.

Furthermore, it is possible to derive another reweighting formula (2.40), if $u(\cdot)$ is of gradient form ($u(\cdot) = \nabla v(\cdot)$). Then, one can use Itô's formula and calculate another expression for the stochastic integral term in (2.40) as it is done in [58] p.838. Applying Itô's formula to v we get

$$\begin{aligned} v(y_T) - v(y_0) = \int_0^T \frac{1}{\beta}\nabla^2 v(y_s) - \nabla V(y_s)\cdot\nabla v(y_s) + |\nabla v(y_s)|^2 ds \\ + \sqrt{2\beta^{-1}}\int_0^T \nabla v(y_s) dB_s \end{aligned} \tag{2.43}$$

where ∇^2 is the Laplacian. Now rearranging terms we get a new expression for the stochastic integral which can be used in (2.40) to derive

$$\begin{aligned} M_T = \exp\Bigg(&\frac{1}{2\beta^{-1}}(v(y_T) - v(y_0)) + \\ &\frac{1}{2\beta^{-1}}\int_0^T \left(\nabla V(y_s)\cdot\nabla v(y_s) + \frac{1}{2}|\nabla v(y_s)|^2 - \beta^{-1}\nabla^2 v(y_s) ds\right)\Bigg). \end{aligned} \tag{2.44}$$

This expression is still stochastic because y_s is a stochastic process. From the first point of view it seems that this term could be treated more easily numerically compared to the stochastic integral. We will investigate this in the numerical examples shown in Chapter 4 and 5.

2.2.2 Sampling methods and Monte-Carlo

Since only a few stochastic differential equations can be solved analytically numerical methods have to be used. An overview about different numerical methods is given in [48]. We give a short overview about the most popular methods.

The best-known numerical method for stochastic differential equations is the Euler-Maruyama method. The method is an explicit $\frac{1}{2}$ order method.

Euler-Maruyama
Suppose we have an initial value problem for the SDE given in (2.16). In order to calculate the solution we discretize the time interval $[0, T]$ into $n \in \mathbb{N}$ timesteps $0 = t_1 < t_2 \ldots < t_n = T$ with $t_k = k\Delta t,\ k = 1, \ldots n$ and $\Delta t = \frac{T}{n}$. Furthermore, we discretize the stochastic differential

$$\Delta B_k = B_{t_{k+1}} - B_{t_k}, \quad k = 0, \ldots, n-1. \tag{2.45}$$

It follows that the ΔB_k are independent normally distributed with mean 0 and variance Δt. Then the SDE (2.16) can be approximated by

$$d\hat{x}_{k+1} = d\hat{x}_k + b(\Delta t, \hat{x}_k)\Delta t + \sigma(\Delta t, \hat{x}_k)\Delta B_k, \quad \hat{x} = x_0 \quad k = 0, \ldots, n-1. \tag{2.46}$$

Convergence results can be found in e.g. [39] or [48] and the references therein.

Examples for higher order methods are Milstein [63] or Runge Kutta methods [48]. The Milstein method can only be used in the case in which the diffusion term is state dependent because it requires a derivative of the diffusion term. For SDE without state dependent diffusion the Milstein method boils down to an Euler-Maruyama method. But higher order schemes for these SDEs have been developed in [77], for example.
For the case that the SDE is only used as a sampling device for the stationary distribution it has been shown that a slight variation of the Euler-Maruyama has better properties cf. [55].

2.3 Monte Carlo and Importance sampling

As we have already seen in the introduction many quantities of interest in MD are expressed as expectations. For examples, thermodynamic quantities are given as

$$\mathbb{E}_\nu[\varphi] = \int_{\mathcal{D}} \varphi(x)\nu(x)dx$$

where φ is some function expressing the quantity of interest and $\nu(x)dx$ is the Boltzmann-Gibbs measure on the considered space $\mathcal{D}$. Since these expectations cannot be calculated analytically we approximate them by a Monte Carlo estimator.

For these different realizations of the quantity of interest are generated independently and at random from the distribution and their average (empirical mean) is taken

$$\mathbb{E}_\nu[\varphi] \approx \frac{1}{N}\sum_{i=1}^{N}\varphi(x^i) := \kappa_N(\varphi) \tag{2.47}$$

where x^i are samples from the distribution $\nu(x)$. By the strong law of large numbers we have that the absolute error will go to zero

$$\mathbb{P}\Big(\lim_{N\to\infty}|\kappa_N(\varphi) - \mathbb{E}_\nu[\varphi]| < 0\Big) = 1 \tag{2.48}$$

cf. [68]. If we further assume that the quantity of interest has finite variance $\mathrm{Var}(\varphi) < \infty$ one can show for the i.i.d. sampling that the Monte Carlo estimator is unbiased

$$\lim_{N\to\infty}\mathbb{E}[\mu_N] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\varphi(x^i)] = \mathbb{E}_\nu[\varphi] \tag{2.49}$$

cf. [68]. But often the variance of these Monte Carlo estimators behaves badly, especially if the quantity of interest is a rare event. In order to improve the Monte Carlo estimator one can take more samples but if the sampling is computationally expensive the Monte Carlo approximation also gets very costly. This is why other strategies have been suggested in the literature, e.g. stratified sampling, control variates or importance sampling. In this thesis we are going to focus on importance sampling; see [68] and the references therein for details on the other methods.
The importance sampling strategy is to sample the quantity of interest from a different distribution and then correct the wrong sampling. The fundamental theory for doing this is given by the following theorem.

Theorem 4 ([25]). *Let μ and ν be probability distributions on the probability space Ω. If μ is absolutely continuous with respect to ν, then there exists an almost everywhere strictly positive function ρ on Ω, such that for any function f for which $\mathbb{E}_\mu[f]$ exists and is finite,*

$$\mathbb{E}_\nu[f] = \mathbb{E}_\mu[f\rho], \tag{2.50}$$

where

$$\mathbb{E}_\mu[f\rho] = \int_\Omega f(x)\rho(x)\mu(dx).$$

The function ρ is called the 'Radon-Nikodym derivative' or likelihood ratio of ν with respect to μ and is denoted by $\frac{d\nu}{d\mu}$.

The resulting importance sampling estimator can be again written as an expectation

$$\mathbb{E}_\nu[\varphi] = \int_{\mathcal{D}}\varphi(x)\frac{\nu(x)}{\mu(x)}\mu(x)dx = \mathbb{E}_\mu\Big[\varphi\frac{\partial\nu}{\partial\mu}\Big]. \tag{2.51}$$

The importance sampling estimator can now again be approximated by a Monte Carlo estimator

$$\mathbb{E}_\mu\Big[\varphi\frac{\partial\nu}{\partial\mu}\Big] \approx \frac{1}{N}\sum_{i=1}^{N}\varphi(x^i)\frac{\nu(x^i)}{\mu(x^i)} \tag{2.52}$$

where now $x^i \sim \mu(x)dx$. Because of the absolute continuity we find

$$\begin{aligned}\mathbb{E}_\mu\Big[\frac{\varphi(x)\mu(x)}{\nu(x)}\Big] &= \int_{\mathcal{D}}\frac{\varphi(x)\nu(x)}{\mu(x)}\mu(x)dx \\ &= \int_{\mathcal{D}}\varphi(x)\nu(x)dx = \mathbb{E}_\nu\Big[\varphi(x)\Big].\end{aligned}$$

In principle the importance sampling estimator should satisfy two properties. First it should be easy to sample from the importance sampling distribution $\nu(x)dx$ and second the variance of the resulting estimator should be lower than the variance of the original Monte Carlo estimator. In general, there is no guarantee that it is easy to sample from the importance sampling distribution and that sampling from any other distribution decreases the variance of the sampling. As we have already seen in the introduction the optimal distribution in terms of variance reduction depends on the quantity itself. This is why the design of a nefficient importance sampling distribution is sometimes called the 'art of importance sampling'.
From the literature it is well-known that Monte Carlo is not affected by the dimension. The convergence of Monte Carlo methods scales with $1/\sqrt{N}$. But for importance sampling the likelihood ratio reveals a dimension effect. The variance of the likelihood ratio grows exponentially with the dimension [68].
We have only presented the idea of importance sampling for thermodynamic quantities for simplicity. All of the here presented theory is also valid for the sampling of dynamic quantities. The main difference are the measures and the spaces which are considered but the general framework of importance sampling is the same. This is why we do not discuss the framework again for dynamical quantities but show the connection to importance sampling of dynamic quantities and optimal control in the SDE context in the next section.

2.4 Importance sampling and optimal control

The general idea for importance sampling in path space is equivalent to the already presented idea of importance sampling. Instead of sampling with respect to the difficult probability measure one tries to sample with respect to another probability measure which is easier to sample. As we have already seen in the case of two stochastic processes the Radon-Nikodym derivative is explicitly given by Girsanov's theorem. One can show that there exists an optimal change of drift (also called change of measure). This optimal change of measure results in a zero variance estimator. Furthermore, the optimal change of drift is the solution of an optimal

control problem. Thus, in order to find the optimal change of drift a Hamilton-Jacobi-Bellman (HJB) equation has to be solved. In this section we derive the HJB equation for the optimal bias by using the explicit formula of the Radon-Nikodym derivative and the Feynman-Kac relation following [58].
The connection of optimal control and variance reduction for Monte Carlo methods was first proposed by [63]. Milstein developed the discrete version of the optimal control approach. He proposes to use Girsanov's theorem and add an additional drift to the SDE. He then derives that there exists an optimal drift which satisfies the Bellman equation, such that the variance of the estimators is zero.

We consider the SDE given by (2.17) with deterministic starting conditions $x_0 \in \mathcal{S}$ in some sufficiently smooth and bounded set $\mathcal{S}$. We are interested in observables of the

$$I = \mathbb{E}_{\mathbb{P}}^{x}\left[\exp\left(\int_0^\tau f(x_s)ds + g(x_\tau)\right)\right] \tag{2.53}$$

for some given functions $f : \mathcal{S} \to \mathbb{R}$ and $g : \mathcal{S} \to \mathbb{R}$ and the stopping time

$$\tau = \inf\{t \geq 0, x_t \in \mathcal{T}\} \tag{2.54}$$

where $\mathcal{T}$ is some given target set (e.g. $\mathcal{T} = \mathcal{S}^c$ the complement of a bounded set). On the one hand these observables can be used to describe free energy cf. [12] and interesting quantities like exit times or transition probabilities and on the other hand these observables can be connected to stochastic control theory [58]. For example, choosing $g = 0$ and $f = \lambda$ then I is the moment generating function of the exit time. Typically I is approximated by a Monte Carlo estimator

$$I \approx \frac{1}{N}\sum_{i=1}^{N} I_n \tag{2.55}$$

where I_n are independent realizations of (2.53). If the drift term is metastable, it is very hard to sample trajectories which reach the target set. Furthermore, the statistical error

$$\sqrt{\frac{\mathrm{Var}(I_n)^2}{I}} \tag{2.56}$$

grows, if the probability we are trying to estimate is very small. In order to reduce the variance of the estimator an importance sampling technique is used. For this we sample the biased dynamics

$$dy_t = -\nabla(V + \bar{V})(y_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad y_0 = x \tag{2.57}$$

where $\bar{V} : \mathcal{S} \to \mathbb{R}$ is a C^∞ biasing potential. In order to correct the expectations calculated from the biased dynamics we use the Radon Nikodym derivative from

Girsanov's theorem. As we have seen the importance sampling estimator constructed with Girsanov's theorem is unbiased

$$I = \mathbb{E}[I_{\bar{V}}] \tag{2.58}$$

where

$$\begin{aligned} I_{\bar{V}} &= \exp\left(\int_0^{\bar{\tau}} f(y_s)ds + g(y_\tau)\right) \\ &\times \exp\left(\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\bar{\tau}} \nabla\bar{V}(y_s)dB_s - \frac{1}{4\beta^{-1}}\int_0^{\bar{\tau}} |\nabla\bar{V}|^2(y_s)ds\right) \end{aligned} \tag{2.59}$$

with $\bar{\tau}$ being the stopping time

$$\bar{\tau} = \inf\{t \geq 0, y_t \in \mathcal{T}\}. \tag{2.60}$$

So Girsanov's theorem provides a way of importance sampling in path space for any bias potential which satisfies the conditions of Girsanov's theorem. Using the alternative expression of the Girsanov weight given in equation (2.44) and rearranging terms we find an expression for the stochastic integral in (2.59). Using the above formula we find a second formula for the estimator

$$\begin{aligned} I_{\bar{V}} &= \exp\left(\int_0^{\tau} f(y_s)ds + g(y_\tau) + \frac{1}{2\beta^{-1}}(\bar{V}(y_\tau) - \bar{V}(x_0))\right) \\ &\left(\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\bar{\tau}} \nabla V(y_s)\cdot\nabla\bar{V}(y_s) + \frac{1}{2}|\nabla\bar{V}|^2(y_s) - \beta^{-1}\nabla^2\bar{V}(y_s)ds\right) \end{aligned}$$

We will now derive the bias potential which gives a zero variance estimator. Let us consider the bias

$$\bar{V} = -2\beta^{-1}\log u \tag{2.61}$$

where $x \in \mathcal{S}$

$$u(x) = \mathbb{E}^x_{\mathbb{P}}\left[\exp\left(\int_0^{\tau} f(x_s)ds + g(x_\tau)\right)\right]. \tag{2.62}$$

Recall that $(x_s)_{s\geq 0}$ satisfies (2.17) with $x_0 = x$. From the Feynman-Kac formula the function $u : \mathcal{S} \to \mathbb{R}$ satisfies a partial differential equation

$$\begin{aligned} \mathcal{L}u + fu &= 0 \quad \text{in } \mathcal{S} \\ u &= \exp(g) \quad \text{on } \partial\mathcal{S} \end{aligned}$$

where $\mathcal{L}$ is the infinitesimal generator of (2.17). Therefore we know that $\bar{V} = -2\beta^{-1}\log u$ satisfies

$$\nabla V\nabla\bar{V} - \beta^{-1}\nabla^2\bar{V} + \frac{1}{2}|\nabla|^2 + 2\beta^{-1}f = 0 \quad \text{in } \mathcal{S} \tag{2.63}$$

$$\bar{V} = -2\beta^{-1}g \quad \text{on } \partial\mathcal{S} \tag{2.64}$$

Using the biasing potential $\bar{V} = -2\beta^{-1} \log u$ we obtain

$$\begin{aligned} I_{\bar{V}} &= \exp\left(g(y_\tau) + \frac{1}{2\beta^{-1}} (\bar{V}(y_\tau) - \bar{V}(x_0) \right) \\ &= \exp\left(- \frac{1}{2\beta^{-1}\bar{V}(x_0)} \right) = u(x_0) = I. \end{aligned}$$

From the above equation we see that the estimator $I_{\bar{V}}$ is almost surely a zero variance estimator. Thus, the estimator is optimal in terms of bias. As always in importance sampling the optimal result requires the quantity which we want to calculate from the very beginning.
The above result shows on the one hand that it is quite difficult to construct the optimal bias because it is a solution of a non-linear PDE. On the other hand by approximating the optimal bias we can use it to bias the dynamics to get better estimators (better in terms of variance). The closer the approximation is to the real solution the better is the associated Monte Carlo procedure. More details on the connection of importance sampling and optimal control in the MD context can be found in [58] and on stochastic optimal control in [30, 37].

2.5 Related works

In the last part of the theory chapter we are going to mention some related works for describing the metastability in a dynamical system and algorithms for the efficient sampling of quantities. At first we are going to give a brief summary about mathematical ways how metastability can be described. Then we are going to present the importance sampling approach based on large deviation context and at the end we are going to summarize the different algorithmic approaches for the efficient sampling of thermodynamic quantities.

Mathematical approaches to metastability

Large deviations
The modern mathematical approach to metastability was given by M. Freidlin and A. Wentzell in the late 1960s. Freidlin and Wentzell introduced the theory of large deviations on path space to analyse the long-term behaviour of a deterministic dynamical system perturbed by a weak stochastic noise. Their approach is often called the pathwise approach to metastability. The main advantage of this approach is that it gives very detailed information on the metastable behaviour. By minimizing an action functional, which is called the rate function, crossing times and other

information can be determined. Identifying and controlling the rate function are non trivial and thus limits the application; see e.g. [32] for details.

Spectral approach

An axiomatic approach to metastability was introduced by Davis in the 1980s based on the spectral properties of the generator of the dynamical system (he considered a reversible Markov process). He showed that the related process exhibits a metastable behaviour if the spectrum has a cluster of very small real eigenvalues which are separated by a comparatively wide gap from the rest of the spectrum. Under some further assumptions on the corresponding eigenfunctions he showed that the state space can be decomposed into metastable sets and that the motion of the Markov process between these sets is slow. The limitation of this approach is that it is difficult to verify the necessary assumption on the spectrum; see e.g. [19] for details.

Potential theory

The potential-theoretic approach was introduced in 2001 by Bovier et al. Instead of identifying the paths between the metastable sets, the metastability is interpreted as a sequence of visits of the path to different metastable sets. In this way, it focuses on the analysis of the hitting probabilities and hitting times based on potential theory. From a different point of view this approach tries to understand the metastable behaviour of the Markov process to the study of equilibrium potentials and capacities in networks; see e.g. [8] for details.

Quasi-stationary distribution

Another tool of partial differential equations for studying metastable behaviour are quasi-stationary distributions (QSD). The quasi-stationary distribution can be used to analyse the exit event from a metastable set S. The QSD describes the long term behaviour of the process conditioned to not leaving S. It is attached to the metastable set by the first eigenvector of the Fokker-Planck operator with homogeneous Dirichlet boundary conditions on ∂S. This technique is also known as the Fleming-Viot process and can be used to analyse algorithms like the Replica exchange method ; see e.g. for details [57].

Log Sobolev inequalities

The metastability can also be described by the rate of convergence of the law converging to the stationary distribution. The rate of convergence is described by the relative entropy of the time evolution of the probability measure and the equilibrium probability measure. The relative entropy of the law at time t and the stationary distribution is bounded by the relative entropy of the starting distribution and the stationary distribution times an exponential factor depending on some constant R. In general, one can say that the smaller R the more metastable the dynamical system.

Determining the sharp estimates of the constant R is actually very challenging; see e.g. [57] for details.

Importance sampling in the large deviations context

Large deviation arguments have become a very important tool for importance sampling strategies and the efficient design of Monte Carlo algorithms; see e.g. [3]. For example, exponential change of measure have been proposed for an efficient rare event sampling. Considering the connection between optimal control and importance sampling different strategies were developed in the large deviation context; see e.g. [20, 22, 21]. By a scaling of diffusion a first order Hamilton-Jacobi-Bellman can be derived. Based on this first order HJB equation Dupuis et al. derive importance sampling schemes for different situations. For this the temperature is sent to zero ($\beta^{-1} \to \infty$). The resulting first order HJB equation is given by

$$\begin{aligned} \nabla V \cdot \nabla \bar{V} + \frac{1}{2}|\nabla \bar{V}|^2 + \bar{f} = 0 \quad &\text{in } \mathcal{S} \\ \bar{V} = -\bar{g} \text{ on } &\partial\mathcal{S} \end{aligned}$$

where $\nabla \bar{V}$ is used to bias the drift of the considered system. These ideas have also been developed for applications in Molecular Dynamics by [88]. In their work van den Eijden and Weare proposed a technique based on the solution of the deterministic control problem associated to the sampling problem for dynamical quantities.
Furthermore, advanced importance sampling strategies for MD relevant situations have been studied in [23]. Here an importance sampling scheme for resting points was developed. The numerical examples of the article show that the importance sampling scheme constructed for this situation is better than the scheme which does not take the resting point into account. In order to build such importance sampling scheme a lot of knowledge on the dynamical system is necessary, but this results in a better variance reduction.
In [82] K. Spiliopoulos developed a performance measure for importance sampling scheme related to small noise diffusion processes which give the possibility to compare the different importance sampling schemes analytically.

Algorithms to overcome metastability

In the MD community more algorithmic approaches to overcome problems with metastability for thermodynamic quantities have been proposed. We summarize the some well-known methods and show, if and why they can be seen as importance sampling schemes. In the end of this section we present a method which was proposed for the sampling of dynamic quantities.

Due to the many interacting particles in the molecule the function V is unknown. So in order to calculate thermodynamic quantities the stationary distribution has to be approximated. In order to do this the function V has to be explored effectively. But again the exploration of the function V is difficult because of the many minima and the resulting metastable behaviour. In order to overcome these problems many algorithms have been developed in the past years. These algorithms which have been designed for this problem are often called enhanced sampling techniques. Many of these methods are based on ideas from importance sampling (see e.g. [60]) but also other ideas and techniques are used in enhanced sampling approaches; see e.g. [6, 59, 15] and the references therein for details.

Adaptive methods

Adaptive methods are quite popular methods which have been designed to enhance the sampling of stationary distributions. There are three very popular examples: the histogram approach by [91], the adaptive biasing force method [18] and Metadynamics [53]. All these methods share the same principle which is to modify the potential (or the force and thus implicitly the potential) in order to remove or decrease the metastability. These algorithms are often applied on a so-called reaction coordinate which is a low-dimensional representation of the high-dimensional dynamical system. Adaptive methods can be seen as adaptive importance sampling methods since the stationary distribution is determined by the potential as we have seen in (1.2). So changing the potential with a biasing potential U will also result in a different distribution function

$$\nu(x) = \frac{1}{Z_U} \exp\left(\frac{-(V(x)+U(x))}{k_B T}\right)$$

where Z_U is the new partition function for the modified potential. So the Radon-Nikodym derivative between the original and the perturbed distribution is given by

$$\frac{\nu(x)}{\mu(x)} = \frac{\frac{1}{Z}\exp\left(\frac{-V(x)}{k_B T}\right)}{\frac{1}{Z_U}\exp\left(\frac{-(V(x)+U(x))}{k_B T}\right)} = \frac{Z_U}{Z}\exp\left(\frac{U(x)}{k_B T}\right).$$

The biasing potential can be calculated based on the history of the trajectory. In this way the biasing potential can be made problem dependent. This makes adaptive methods interesting for problems for which no apriori information about the dynamical system is available. There has been lots of research for adaptive methods in the last years. For example, convergence results have been found for variants of Metadynamics [53] or adaptive biasing force [18]. This area is still a very active field and new ideas have been proposed recently; see e.g. [86] or [87] and the references therein.

Simulated annealing

In simulated annealing the temperature of the simulated system is changed; see [45]. As we have seen the temperature has an impact on the stationary distribution such that the change of temperature changes the stationary distribution. This can be interpreted as change of measure and this is why the method can be seen as an importance sampling method for thermodynamic quantities. The stationary distribution for a system at temperature T_1 is given by

$$\nu(x) = \frac{1}{Z} \exp\left(\frac{-V(x)}{k_B T_1}\right).$$

Suppose we do a second sampling with temperature T_2, then one can relate the two different measures by a Radon Nikodym derivative

$$\frac{\nu(x)}{\mu(x)} \frac{\exp\left(\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\frac{V(x)}{k_B}\right)}{\int \exp\left(\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\frac{V(x)}{k_B}\right) dx}. \tag{2.65}$$

The quantity of interest for the temperature T_1 can be calculated by sampling the dynamical system for temperature T_2 and reweight with the Radon Nikodym derivative (2.65).

Replica exchange

In the case of Replica exchange methods, e.g. [83], a different strategy than importance sampling is used. The method is designed for the effective sampling of thermodynamic quantities and stationary distributions. The main idea of this method is to start different trajectories with different temperatures and to interchange the current position of the trajectories according to some rule. Often a Metropolis Hastings rule is used as the rule to interchange the current positions. In such a way the low-temperature trajectory visits states which it would not have visited without these switchings. If the dynamical system is ergodic the thermodynamic quantities of the system can be sampled in this way.

From a theoretical point of view one could built an importance sampling scheme with different temperatures involved. If, for example, a temperature replica exchange is applied, then this can be interpreted as a multistage simulated annealing, and in principle one could compute the reweighting factors for each trajectory segment. However, one has to track the interchange of the trajectories very precisely in order to correctly reweight the trajectory segments. A similar strategy has been proposed by [61].

Multilevel Splitting

Multilevel splitting is a technique to sample path dependent quantities in contrast to the techniques presented before. The main idea of multilevel splitting techniques

like adaptive multilevel splitting is to decompose the whole path which has to be sampled into much smaller parts. The intermediate sampling goals are much easier to reach and thus the sampling is much faster. Furthermore, the sampling of the different parts can be parallelized. Famous examples are e.g [13] or [24] but many other approaches exist.

3 Convolution approach

As we have seen in the introduction metastability which is caused by the energetic or entropic barriers has an impact on both thermodynamic quantities and on dynamical quantities. Therefore, one of the main research questions is how these barriers and thus the metastability can be decreased. If information about the dynamical system and the location of the metastable sets are available, local techniques can be used as we will see in the next chapter. If this is not the case, a lot of sampling effort has to be used to explore the state space of the system to get this relevant information. Especially for the sampling of thermodynamic quantities this is necessary because the stationary distribution which depends on the whole state space has to be known to calculate them. In order to approximate the equilibrium distribution the state space has to be explored such that the empirical approximation is sufficient. So decreasing the metastability of the dynamical system without apriori knowledge will help to explore the state space more rapidly and thus reduce the sampling effort. To do so we propose a global perturbation of the potential by a convolution. The main idea of this approach is to sample a slightly perturbed dynamical system in which the metastability is reduced. The intuition of this method is that the convolution first smoothes out the small minima which hinder the trajectory from moving in the potential freely. The convolution also decreases the large barriers such that the metastability is decreased and thus the sampling is accelerated. Since the metastability of the system is affected globally, the convolution approach can be used for effective sampling of the state space. Furthermore, the perturbation of the potential also has an impact on the equilibrium distribution and we can use importance sampling techniques for the sampling of thermodynamic quantities.
The convolution approach can be especially helpful for dynamical systems with an entropic barrier. Normally simulated annealing techniques are used to overcome barriers. But for entropic barriers these techniques fail since these barriers are not affected by the temperature or, even worse, a higher temperature increases the barrier. Since the convolution approach changes the potential directly and is independent from any system parameter, it will have an effect on the entropic barrier.

The convolution approach is motivated by an idea of global optimization. In global optimization one is interested to find the global minimum of a function with many local minima. The general strategy following the steepest descent until one has found a possible candidate fails. Gradient descent methods will often get stuck in

a local minimum not finding the global minimum. The convolution is used here to smooth out the local minima and thus find the global minimum. This approach is often combined with a restart technique. The convolution approach in global optimization was introduced in a series of papers [49] and [64] etc. The authors used a convolution with a Gaussian kernel. This convolution is connected to the heat equation because the fundamental solution can be expressed in this way.
Recently, the convolution approach was applied in the field of deep learning. In deep learning weights of a neuronal network have to be optimized. This can be phrased as a optimization problem and the objective function is often high-dimensional and thus finding the global minimum is difficult. It is often too expensive to evaluate the full gradient and thus the gradient is only evaluated on some smaller subspace (or mini batch) or stochastic gradient methods are used. If a stochastic gradient descent method is used, the gradient descent can be modelled as an overdamped Langevin equation. This is why problems with metastability also occur in this research area. The convolution approach was used in order to overcome these metastability problems in the optimization of deep neuronal networks and was compared to other approaches cf. [14].

The chapter is structured as follows. First, we present the motivating idea from Scheraga's original work. We are going to present then different approximations schemes for the convolution of high-dimensional potentials. The presentation of the different approximation schemes is accompanied by one-dimensional examples visualizing the intuition of the convolution approach and how the metastability is affected. After this we are going to show results for the application of the approach to Butane. Butane is especially interesting because the high-dimensional dynamical system can be fully expressed in a one-dimensional reaction coordinate. We used this example to understand if the convolution in the high-dimensional space influences the metastability in the low-dimensional reaction coordinate. Furthermore, we would like to understand in this example how the metastability is influenced by the convolution. After this very general investigation we use the convolution approach for rapid state space exploration and the sampling of thermodynamic quantities. For this we are going to integrate the convolution approach in the Replica exchange algorithm. Normally, different temperatures are used in the Replica exchange method and we propose that different convolution parameters can also be used in order to generate more information about the dynamical system under investigation. Then, we are going to explore how the convolution approach can be combined with importance sampling. As the convolution changes the potential the equilibrium distribution also changes. So by quantifying the difference of the different Boltzmann-Gibbs distributions we can build an importance sampling estimator for thermodynamic quantities. In the following section we integrate the convolution approach into the Linear Response theory. The Linear Response theory is a way to describe the reaction of a dynamic system on a small external force. Our

result can be used to quantify the impact of the convolution on thermodynamic quantities. In the last part of the chapter we are going to use the global perturbation to sample dynamic quantities. First, we develop an extrapolation scheme for mean first exit times and then generalize the approach to other dynamic quantities. A short summary and a discussion concludes the chapter.

3.1 Decreasing metastability by convolution

In this section we are going to explore how the convolution approach decreases metastability. This global approach can be extremely useful, if there is no information about the dynamic system under investigation available. In this section we are going to investigate how the metastability changes under the convolution of the potential. Furthermore, we show how the convolution can be realized, if it cannot be calculated analytically. The different approximation schemes are accomplished with one-dimensional numerical examples to show the decreasing impact of the convolution approach. In the second part of this section we apply the approach to the high-dimensional example of Butane. We used Butane because the dynamics of Butane can be expressed in a one-dimensional reaction coordinate by describing the whole system only with the dihedral angle. With our numerical examples we would like to understand how the convolution of the high-dimensional interaction potential influences the dynamic in the low-dimensional reaction coordinate. We will stick to the numerical testing of the convolution approach because of two reasons. First, there are different definitions in the literature how metastability can be determined for a dynamical system. Each of the definitions has other assumptions and so it is unclear if all of these definitions are equivalent. The second reason is the analytical difficulty. Let us consider for example the Log Sobolev constants as a measure of metastability as suggested in [57]. It is very difficult to precisely calculate the constant even for a small system and mostly impossible to calculate it for a large system like Butane. Let us first have a look at the motivating example of Scheraga's work before showing the different approximation schemes and the one-dimensional examples.

Scheraga was the first who applied the convolution approach in MD [71]. His aim was to find the global minimum of a potential. In his article a transformation is proposed to destabilize any potential well. He motivates his approach with a one-dimensional example by considering the function $V \in \mathcal{C}^2 : \mathbb{R} \to \mathbb{R}$. He proposes the following transformation operator given by

$$V^{[1]}(x) := V(x) + \xi V''(x) \quad \text{for } \xi > 0. \tag{3.1}$$

He argues that this transformation operator does not change the inflection point since $V'' = 0$ but the convex part of the function goes up and the concave part goes down. So the existing extrema are destabilized and barriers are lowered (barrier = max V- min V in some closed connected subset). A repetition of this transformation leads to the idea to use a convolution with Gaussian kernel which can easily be generalized to a multidimensional setting; see [71] or [49] for details. This kind of convolution is also known as smoothing and satisfies

$$V_\lambda(x) = (4\pi\lambda)^{\frac{-n}{2}} \int_{\mathbb{R}^n} V(y) \exp\Big(- \frac{(x-y)^2}{2\lambda^2} \Big) dy. \tag{3.2}$$

From now on we will use $V_\lambda(x)$ to denote the convolution of the function V.

It is known from the literature that the equation (3.2) is the fundamental solution of the heat equation; see e.g. [26]. So the transformation can be also expressed as PDE given by

$$\nabla^2 V_\lambda(x) = \frac{\partial}{\partial\lambda} V_\lambda(x), \quad V_0(x) = V(x) \tag{3.3}$$

where ∇^2 is the Laplacian and λ is normally referred to as 'time'. The main difference between the heat equation and the here introduced convolution approach is that we will not consider λ as time since we are not interested in the λ evolution of the potential. The key idea of the convolution approach is to solve the convolution for a fixed parameter λ and use the resulting potential for sampling and this is why we will call λ the smoothing parameter. We assume that a higher smoothing parameter has a bigger impact on the metastability; see Figure 3.1. Furthermore, we are interested in quantities of the unconvoluted potential and so we are interested in $\lambda \to 0$. Because of this one could now get the impression that we are interested in solving the heat equation backwards in 'time'. But this is not true here. The boundary condition is set to $V_0(x) = V(x)$ and is thus well-defined. Solving the heat equation backwards in time is actually a hard inverse problem; cf. [26].

We have seen in equation (3.2) that in order to convolute the potential either a heigh-dimensional integral or a PDE has to be solved. Both ways of calculating the convolution are difficult and thus limit the application of this approach. So before investigating the influence of the convolution on the metastability let us summarize different approaches how the convolution can be efficiently realized for high-dimensional problems. Furthermore, we are going to test the different approximation schemes for one-dimensional examples to show the intuition of the convolution approach.

Approximation schemes

In this section we are summarizing different approaches from the literature how the convolution of the high-dimensional potential can be approximated efficiently. In total we are going to present three different methods. The first can be applied, if the potential is a polynomial function. The convolution can then be written in an explicit form. The second approach is a Monte Carlo approach. Monte Carlo is the standard technique to approximate high-dimensional integrals. The third approach is based on the introduction of an additional fast variable and the usage of a homogenization method. All the methods will be tested at an example which indicates that the approximation of the convolution has the decreasing impact on the dynamical system.

Polynomial case

Let us first explore the case that the potential is of polynomial form. The convolution can then be calculated in an explicit way and in fact this method is not an approximation. But its application is limited to polynomial potential. The result was first proposed by [49] in the context of global optimization.

Theorem 5. *Consider a polynomial function $v : \mathbb{R}^n \to \mathbb{R}$ of the form*

$$v(x) = \sum_{i_1, i_2 \dots i_d} a_{i_1, i_2 \dots i_d} x_1^{i_1} x_2^{i_2} \dots x_n^{i_n}. \tag{3.4}$$

Then the convolution is given by the following formula

$$v_\lambda(x) = \sum_{i_1, i_2 \dots i_d} a_{i_1, i_2 \dots i_d} W_{i_1}(x_1, \lambda) W_{i_2}(x_2, \lambda) \dots W_{i_n}(x_n, \lambda) \tag{3.5}$$

where

$$W_n(x, \lambda) = \sum_{k=0}^{\lfloor n/2 \rfloor} a_{i_1, i_2 \dots i_d} \frac{n!}{k!(n-2k)!} \lambda^k x^{n-2k}. \tag{3.6}$$

Proof. [49] Since the convolution is the fundamental solution of the heat equation

$$\nabla^2 V_\lambda(x) = \frac{\partial}{\partial \lambda} V_\lambda(x), \quad V_0(x) = v(x) \tag{3.7}$$

then the solution can be expressed as

$$\begin{aligned} V_\lambda(x) &= \exp(\lambda \nabla^2) v(x) = \Pi_{j=1}^n \exp(\lambda(\partial^2 / \partial x_j^2)) v(x) \\ &= \sum_{i_1, i_2 \dots i_d} a_{i_1, i_2 \dots i_d} \Pi_{j=1}^n \exp(\lambda(\partial^2 / \partial x_j^2)) x_j^{i_j} \\ &= \sum_{i_1, i_2 \dots i_d} a_{i_1, i_2 \dots i_d} W_{i_1}(x_1, \lambda) W_{i_2}(x_2, \lambda) \dots W_{i_n}(x_n, \lambda) \end{aligned}$$

where the $W_n(x,\lambda)$ is a finite sum

$$W_n(x,\lambda) = \exp(\lambda(\partial^2/\partial x^2))x^n = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}(\partial^{2k}/\partial x^{2k})x^n$$
$$= \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{n!}{k!(n-2k)!}\lambda^k x^{n-2k}$$

□

Example

The following example shows clearly how the barrier is decreased and how the low probability region which is connecting the two metastable states is raised.

Consider the one-dimensional asymmetric bistable potential given by

$$V(x) = 8x^4 - 44/3x^3 + 2x^2 + 11/3x + 1. \tag{3.8}$$

The convolution can be calculated by (3.5) and is given by

$$V_\lambda(x) = V(x) + 96(\lambda^2/2)^2 + (4 - 88x + 96x^2)(\lambda^2/2). \tag{3.9}$$

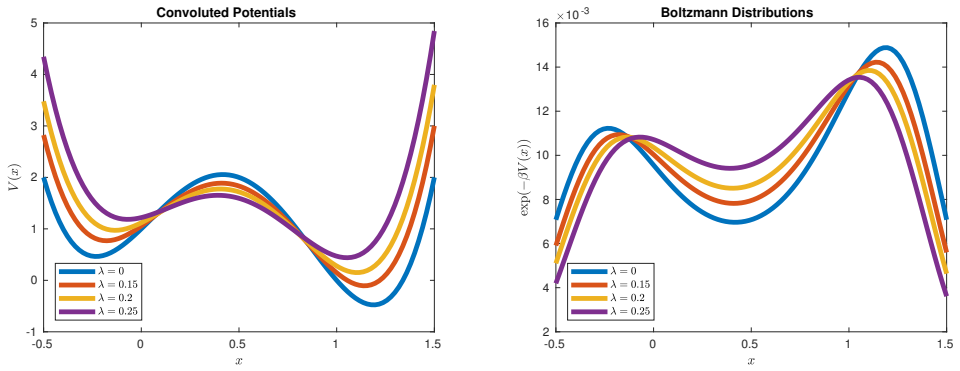


Fig. 3.1: left: Potentials for different smoothing parameters, in blue the original potential is shown. right: Resulting Boltzmann distributions for different smoothing parameters. The inverse temperature $\beta = 3$

One clearly sees that the area around the minimum is raised while the area around the local maximum is decreased. So the barrier heights decrease for bigger smoothing parameters. The stationary distribution of a stochastic process which satisfies (2.16) is given by the Boltzmann distribution which depends on the potential. So the convolution also has an impact on the stationary distribution. The resulting Boltzmann distributions of the convoluted potentials show that the probability of the metastable states is lowered. Furthermore, the probability of the transition region increases. So it is more likely that a transition occurs in the convoluted potential as in the original potential. All of the results for the one-dimensional example show the

decreasing effect of the convolution on the metastability. Let us consider a typical trajectory following equation (2.16) and a time evolution of a trajectory moving in the convoluted potential to indicate that the sampling in a convoluted potential is easier.

We sample one trajectory of length 1000000 time steps in the original potential and in the smoothed potential in order to approximate the stationary distribution. The temperature was set to $\beta = 5$ and the temporal discretization was $dt = 0.001$. A standard Euler-Mayurama scheme was used for the numerical approximation of the SDE. The starting point of the SDE was chosen $x_0 = -0.25$ for all simulations.

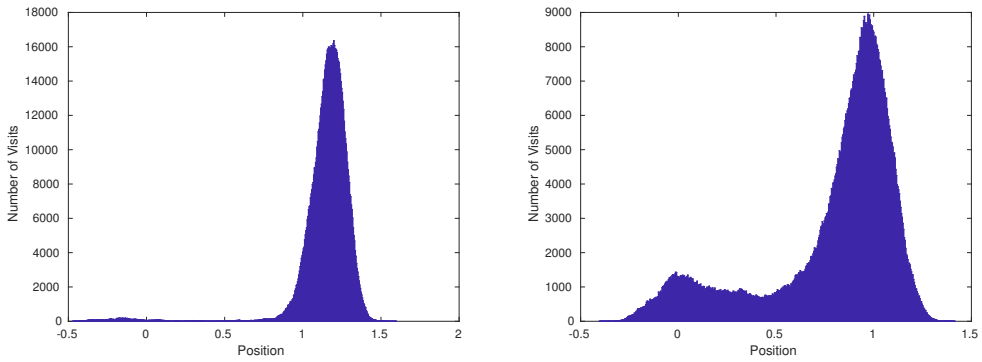


Fig. 3.2: Histogram of the sampled region in the original potential (left) and the sampled region for the convoluted potential (right).

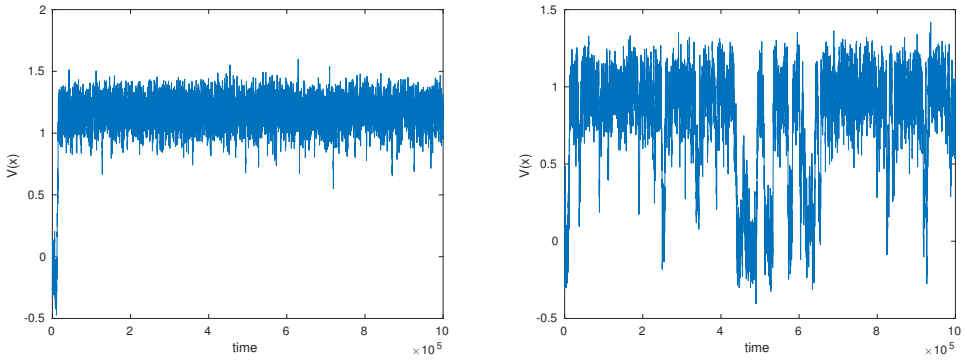


Fig. 3.3: Evolution of the trajectory in the original potential (left) and evolution of the trajectory in the convoluted potential (right).

This example shows that the barrier is decreased by the convolution and so the transition region is better explored compared to the sampling in the original potential. The histogram of the convoluted potential shows much more visits in the transition region and less visits in the metastable regions. Comparing the visualization of the typically trajectories we see that in the sampling of the convoluted potential much more transitions occur and that the exploration of the metastable sets is wider compared to the sampling in the original potential.

Monte Carlo approximation

The convolution integral can also be interpreted as an expectation and so one can use a Monte Carlo approximation method to approximate the integral. The integral

$$V_\lambda(x) = (4\pi\lambda)^{\frac{-n}{2}} \int_{\mathbb{R}^n} V(y) \exp\left(-\frac{\|x-y\|}{2\lambda^2}\right) dy \tag{3.10}$$

is an expectation with respect to a normal distribution $\mathcal{N}(x, \lambda)$. So using a standard Monte Carlo method the integral can be approximated by

$$V_\lambda(x) \approx \frac{1}{M} \sum_{i=1}^{M} V(x + \epsilon_i) \tag{3.11}$$

where ϵ are Gaussian random variables with mean x and variance λ. So in order to evaluate this sum we have to explore the neighbourhood of the point x. But this can be done very sufficiently by drawing Gaussian random variables from the Gaussian density with mean x and variance λ. The evaluation of the potential at all of these sampling points can become very costly, if many sampling points are used (large M). For a small number of sampling points the approximation can be computed very efficiently.

Let us again consider a typical time evolution of a trajectory in the original potential and in the convoluted potential. This time we are going to use a symmetric bistable potential given by

$$V(x) = \frac{1}{2}(x^2 - 1)^2. \tag{3.12}$$

We again sample one trajectory of length 1000000 in the original potential and in the smoothed potential in order to approximate the stationary distribution. The temperature was set to $\beta = 5$ and the temporal discretization was $dt = 0.001$. A standard Euler-Mayurama scheme was used for the numerical approximation of the SDE. The starting point of the SDE was chosen $x_0 = -1$ for all simulations. We used $N = 20$ to approximate the convoluted potential.

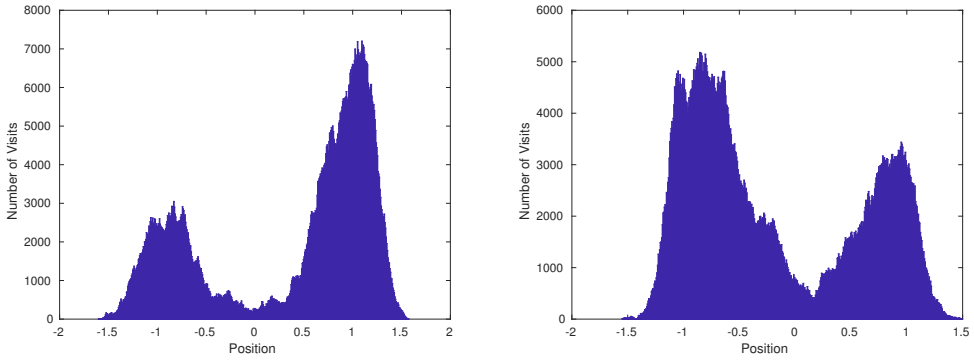


Fig. 3.4: Histogram of the sampled region in the original potential (left) and the sampled region for the convoluted potential (right).

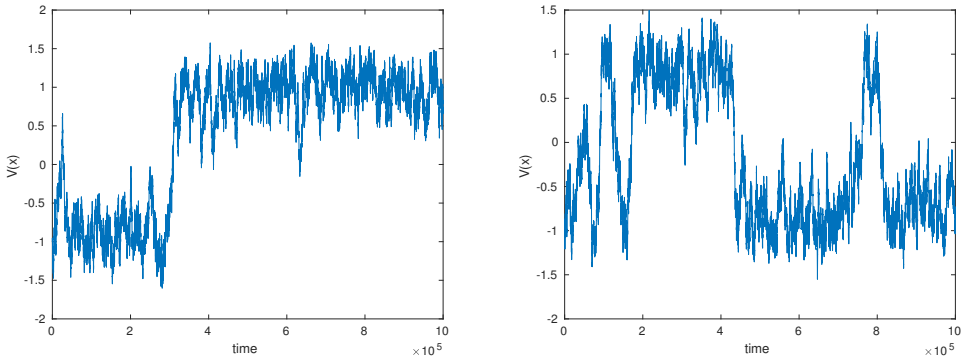


Fig. 3.5: Evolution of the trajectory in the original potential (left) and evolution of the trajectory in the convoluted potential (right).

Similar to the previous example of the asymmetric bistable potential we see that the barrier is also decreased in the Monte Carlo approximation of the convolution. Again the transition region is better explored compared to the sampling in the original potential. The histogram of convoluted potential shows much more visits in the transition region and less visits in the metastable regions. Comparing the visualization of the typical trajectories we see that in the sampling of the convoluted potential a lot more transitions occur and that the exploration of the metastable sets is wider compared to the sampling in the original potential.

Slow fast system

The convolution of the potential can also be realized by introducing an artificial fast variable and using homogenization techniques [70]. We will give a brief summary of the theory and then present a one-dimensional example as a proof of concept.

We introduce an artificial fast dynamical system

$$dx_s = -\nabla V(x_s - y_s)ds + \sqrt{2\beta^{-1}}dB_s \tag{3.13}$$

$$dy_s = -\frac{1}{\epsilon\gamma}y_s ds + \frac{1}{\sqrt{\epsilon\alpha}}dB_s \tag{3.14}$$

In this case it follows that in the limit $\alpha \to 0$ the dynamical system x_s in (3.13) converges to

$$dX_s = \bar{h}(X_s)ds + \sqrt{2\beta^{-1}}dB_s \tag{3.15}$$

where $\bar{h}$ is the homogenized vector field for X is defined as the average against the invariant measure of y_s. So if we choose the equation of motion of y_s such that the invariant distribution for the fast variable is $\rho_\infty(y, X) = G_{\alpha^{-1}\gamma}(y)$ then the homogenized dynamics is given by

$$dX_s = -\nabla V(X_s) * G_{\alpha^{-1}\gamma}ds + \sqrt{2\beta^{-1}}dB_s \tag{3.16}$$

The resulting dynamics system corresponds to a Gaussian averaging of the gradients [14, 70].

Example

We test the approximation scheme based on the homogenization in a one-dimensional setting. We simulate a stochastic process satisfying (2.16) in a symmetric bistable potential V as given in the previous example.

We sample one trajectory of length 1000000 time steps in the original potential and in the smoothed potential in order to approximate the stationary distribution. The temperature was set to $\beta = 5$ and the temporal discretization was $dt = 0.001$. A standard Euler-Mayurama scheme was used for the numerical approximation of the SDE. The starting point of the SDE was chosen $x_0 = -1$ for all simulations. The smoothing parameter was set to $\alpha = 50$ in the simulation of the convoluted potential.

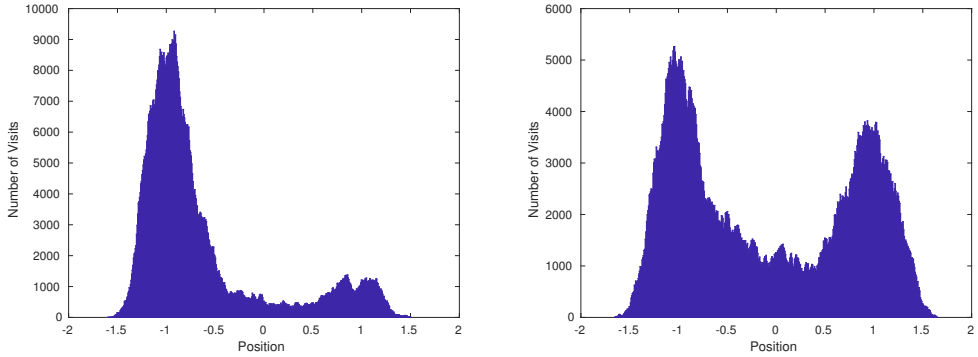


Fig. 3.6: Histogram of the sampled region in the original potential (left) and the sampled region for the convoluted potential (right).

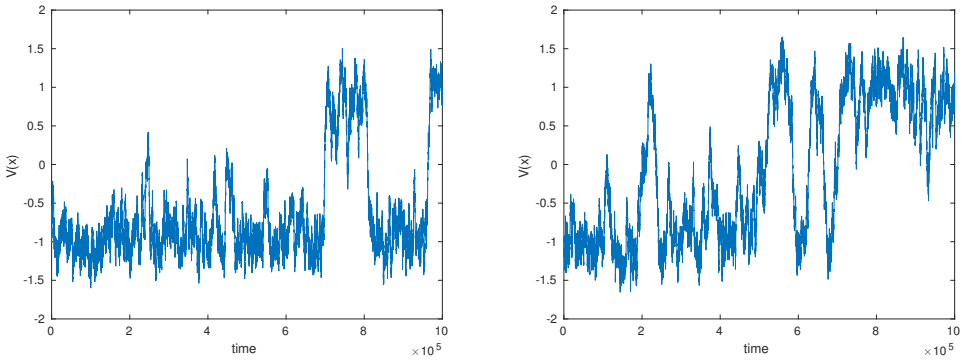


Fig. 3.7: Evolution of the trajectory in the original potential (left) and evolution of the trajectory in the convoluted potential (right).

Similar to the previous examples we see that the approximation of the convolution by this approach decreases the barrier. Again the transition region is better explored compared to the sampling in the original potential. The histogram of the convoluted

potential shows a lot more visits in the transition region and less visits in the metastable regions. Comparing the visualization of the typical trajectories we see that in the sampling of the convoluted potential a lot more transitions occur and that the exploration of the metastable sets is wider compared to the sampling in the original potential.

So far we have seen different approximations for the convolution which are independent from the dimension of the potential. Furthermore, we have shown different one-dimensional examples, all showing that the convolution has a decreasing impact on the barriers and so the metastability of the dynamical system is decreased. Let us next have a look on an application of the approach on a high-dimensional potential. Before presenting the results we are going to introduce the transfer operator because the eigenvalues of this operator are a numerical tool to describe the metastability of a system.

Spectral Gap and Eigenvalues

In the numerical community the eigenvalues of the transfer operator and the spectral gap are often used to indicate the metastability, e.g. [42]. Heuristically the number of eigenvalues close to one show how many metastable sets the dynamical system has. So in order to investigate if the convolution influences the metastability we are going to look at the eigenvalues of the transfer operators for different convolutions. But first we are briefly going to introduce the transfer operator; see e.g. [79] for further details.

The transfer operator $\mathcal{T} : L^1(\mu) \to L^1(\mu)$ is an operator that propagates probability densities μ. Since for large molecular systems the operator is analytically intractable, different approximation schemes are used. One approximation was introduced in [79] and is given by

$$\mathcal{T}^\tau := \exp(\tau\mathcal{L}) \tag{3.17}$$

where τ is a so-called lag time and $\mathcal{L}$ is the generator of the considered dynamical system. The transfer operator describes how the dynamical system evolves in time τ. For the above one-dimensional example the generator can be approximated by a finite difference scheme. So the transfer operator can be approximated very accurately. For high-dimensional systems the transfer operator is usually approximated by discretizing the state space into the metastable regions; see for example [79] and the reference therein for details.

Let us again consider the potential of the last one dimensional example and the resulting transfer operator. To see how the convolution influences the metastability we are going to investigate the behaviour of the eigenvalues under the convolution. The eigenvalues for the convoluted potentials are shown in figure 3.8.

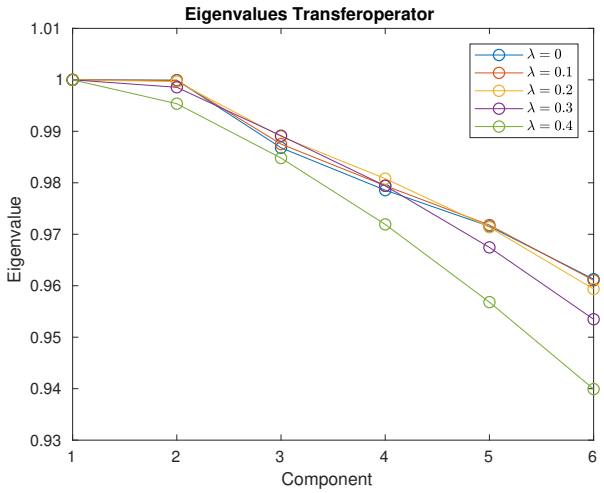


Fig. 3.8: Eigenvalues of the transfer operator for different convolution parameters λ

The eigenvalues are decreasing for a larger smoothing parameter for the bistable system. The larger the smoothing parameter, the lower the second eigenvalue is. Due to this behaviour of the second eigenvalue we can conclude that the metastability is decreased. By looking at the behaviour of the potential this is exactly what we would expect; see 3.1. The bistable system would turn into a system with only one metastable state for very high smoothing parameters. So if the convolution is done in this way that the metastability is decreased, but the number of metastable states is conserved, the sampling will get enhanced and not too much information about the original system is lost.

Example: Butane

In order to test the convolution approach for a high-dimensional system we test the approach for Butane. This work was done by Lisa Brust in her master thesis under my supervision [9]. We will only present parts of the master thesis here to show that the convolution approach can be applied for high-dimensional sampling problems. The details of the shown simulation can be found in [9].

Butane is a small molecule consisting of 4 Carbon atoms and 10 Hydrogen atoms. So the full potential function is in a 42 dimensional space. Butane has three metastable states which can be expressed in a one-dimensional reaction coordinate (torsion angle). For the simulations the Matlab package *trajlab* was used. For simplification we only considered the backbone for the simulations taking the Hydrogen atoms into account by changing the mass of the Carbon atoms. For all simulations the same random number generator was used for better comparison of the results.

To investigate the effect of the convolution on the metastability we simulated long term trajectories of Butane to approximate the stationary distribution of the torsion angle for different smoothing parameters. In the first experiment the whole potential was convoluted. Furthermore, the transfer operator was approximated such that we could calculate its eigenvalues and investigate their behaviour under the convolution.

Comparing the resulting projected stationary distribution of the torsion angle one can see that the convolution of the whole potential has a decreasing effect on the metastability.

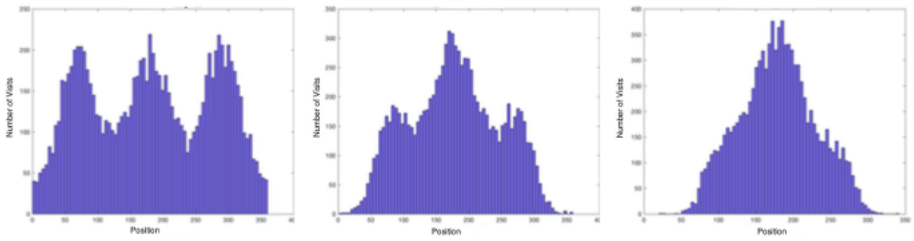


Fig. 3.9: Stationary distribution for different smoothing parameters approximated by a long term simulation. The left figure shows the stationary distribution for $\lambda = 0.15$, the middle figure shows the stationary distribution for $\lambda = 0.3$ and the right figure the stationary distribution for $\lambda = 0.4$ is shown.

The projected stationary distributions of the convoluted potential of Butane show a similar behaviour as the distributions of the one-dimensional example. So the convolution of the whole potential has an effect on the metastability. Furthermore, one can also see that a larger smoothing parameter has a larger impact on the metastability. The left figure shows the resulting distribution for the smoothing parameter $\lambda = 0.15$. The distribution shows three separated modes which are all more or less equally likely. The middle figure shows the distribution for $\lambda = 0.3$. Here one also sees three different modes but the two outer modes are not so often visited. Since the same random number generator was used, it gives the impression that the metastability of these states has really changed. Comparing the transition regions with each other one also sees that they are more often visited if the smoothing parameter is larger. The right figure shows the resulting distribution for $\lambda = 0.4$. The distribution only has one mode. So the two outer states, which were metastable before, are now not metastable any longer. The two outer metastable states are not as often visited as in the less convoluted potentials.

As we have seen Butane has three metastable states. Comparing the eigenvalues of the transfer operator one sees the decreasing impact of the convolution because the second and the third eigenvalue are lowered. One also sees that a larger smoothing parameter has a larger impact on the metastability since the eigenvalues are lower. All in all, these results show that the convolution approach can influence

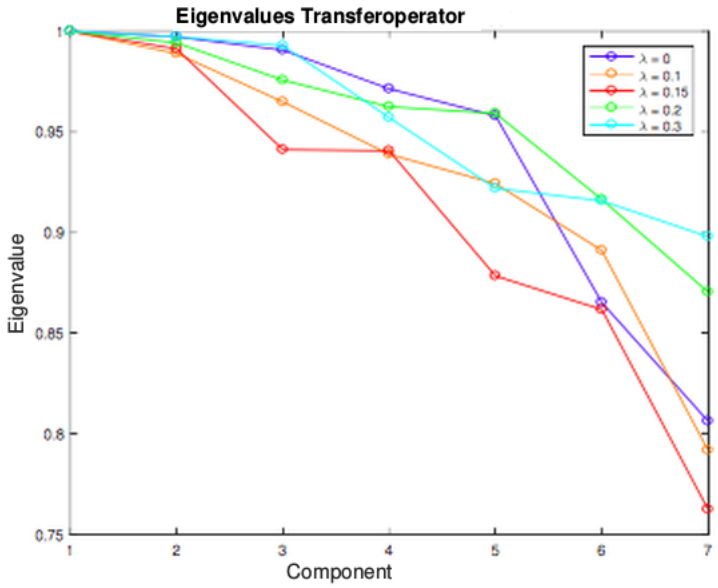


Fig. 3.10: Eigenvalues of the resulting transfer operator for differently smoothed potentials. The transfer operator has been approximated by the algorithm Metastable.

the metastability of a high-dimensional system. Furthermore, the results indicate that the metastability is decreased and thus the exploration of the state space is faster than in the original potential.

In a second test we wanted to see if convoluting only certain interatomic interactions of the potential can also have an effect on the metastability. In order to investigate this we only convoluted the potential describing the torsion angle and calculated long-term trajectories to approximate the stationary distribution. We also compared the different trajectories with each other to see if more transitions were observed. To have a better comparison of the individual trajectories the same random numbers for each individual simulation were used.

Comparing the individual trajectories shows that the convolution increases the number of transitions in other metastable regions. It seems that the convolution lowers the barriers and the trajectory can move more freely in the state space. But the metastability of the deepest minimum is also increased. This can be seen by the time which the trajectory spends in the middle metastable state. While the time which was spent in the outer metastable states was reduced the time spent in the middle metastable state increases. This again indicates the effect of the convolution on the metastability. It seems that the convolution enables the trajectory to explore the state space more freely. In the convoluted potential more transitions can be observed.

Comparing the stationary distributions for the resulting potentials we observe similar effects as in the example before. For different smoothing parameters the modes of the

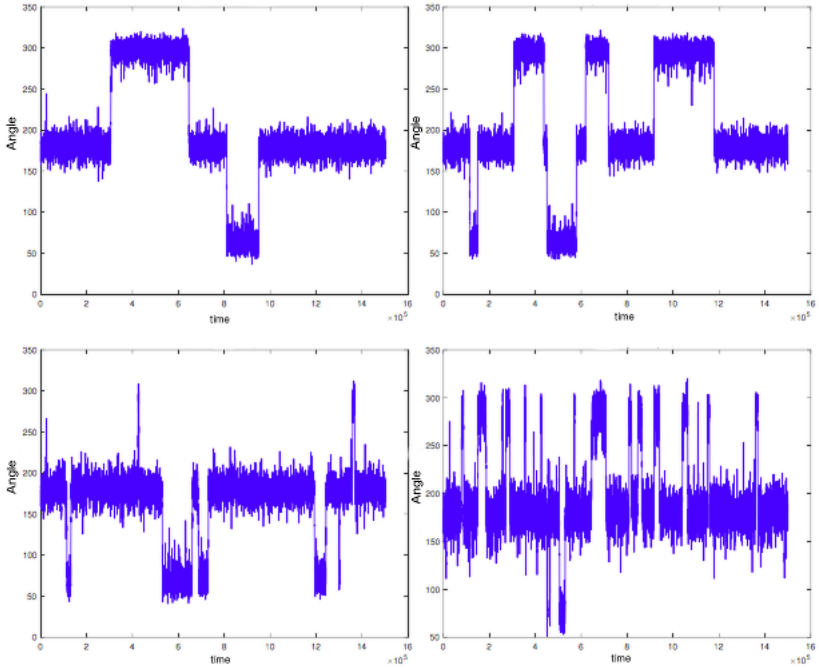


Fig. 3.11: Time evaluation of different trajectories simulated for different smoothing parameters; $\lambda = 0$ (upper left), $\lambda = 0.1$ (upper right), $\lambda = 0.3$ (lower left), $\lambda = 0.5$ (lower right)

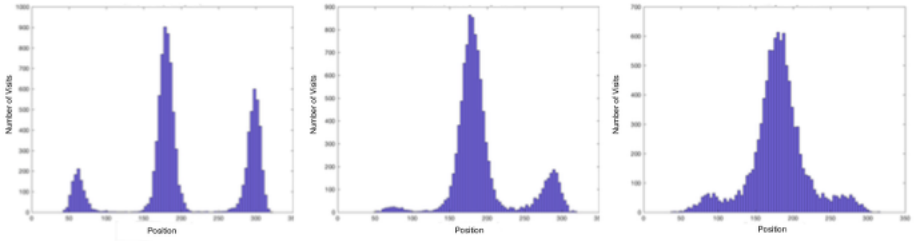


Fig. 3.12: Stationary distributions for different smoothing parameters. $\lambda = 0.15$ (left) $\lambda = 0.3$ (middle), $\lambda = 0.4$ (right)

stationary distribution are getting closer together. So the transition region is visited more often. This indicates that more transitions will occur during the sampling. One can also observe the tendency of the stationary distribution to become unimodal for a high smoothing parameter.

In this section we have seen that the global change of the potential by a convolution has a decreasing effect on the metastability. We have also seen that the convolution when performed in a moderate manner (with a small smoothing parameter) does not change the dynamic systems too much. So the simulation of the convoluted potential can approximate certain insights of the original dynamical system, e.g. the number of minima in the system. Since the metastability is decreased, the exploration of the system is faster compared to a plain simulation of the dynamical system and a lot more transitions can be observed. However, we are sampling not the original system

and so it is not possible to really sample thermodynamic quantities. This is why we are going to consider in the next section how the convolution approach can be used for the sampling of thermodynamic quantities of the original system.

3.2 Convolution for thermodynamic quantities

In this section we are going to show how the convolution approach can be used for the sampling of thermodynamic quantities. In order to do this we are going to integrate the convolution approach into well-known sampling algorithms from MD. First, we are going to show how the convolution approach can be integrated into the replica exchange algorithm. Replica exchange is a method for efficient state space exploration and the sampling of stationary distributions. As we saw in the preliminary examples the convolution influences the barrier heights and thus simplifies the state space exploration similar to annealing techniques. But in contrast to the annealing techniques the convolution approach directly influences the potential. This is why the convolution approach could be superior to the annealing techniques when entropic barriers are present. Entropic barriers are not really affected by an increasing temperature and thus simulated annealing techniques do not really help in this situation.
Second, we are going to combine the convolution approach with the importance sampling idea for thermodynamic quantities. We have already seen that the convolution has an impact on the stationary distribution and so we can use similar importance sampling strategies as presented in Chapter 2.

3.2.1 Replica exchange

In this paragraph we are going to connect the convolution approach and the Replica exchange algorithm. The Replica exchange algorithm is a very well-known algorithm in the MD community. The algorithm which is sometimes called the 'parallel tempering method' is a Monte Carlo method which was invented for the efficient sampling of potentials with many local minima. As we saw earlier the local minima tend to slow down the exploration of the state space. To overcome these problems the algorithms resemble simulated annealing techniques and exchange information from the high temperature sampling into the low temperature sampling. We are going to introduce Replica exchange before showing how the convolution approach can be integrated.

We are interested in sampling the state space of a system at low temperature. Due to the metastability the sampling is very inefficient; see, for example, figure 3.17). The idea of Replica exchange is now to use some information of the system at a different thermodynamic state to explore the state space more efficiently. The main idea is

to swap the atom positions of the different systems every now and then. To control the swapping of the positions a Metropolis Hastings ratio is used. This ensures detailed balance in the swapping and thus that the reverse swapping is equally likely. In the application often more than two different systems are considered to enable as many swaps as possible. The systems often differ in temperature, but not necessarily. Other system parameters can also be used to generate systems in a different thermodynamic state; see e.g. [6] for a review on different Replica exchange algorithms. The Replica exchange methods can also be used to calculate thermodynamic quantities; see e.g. [62].
In order to describe the algorithm let us consider two systems at different temperatures for simplification. The system at high temperature will efficiently sample the state space because it is not so much affected by the barriers but will not sample the area in a metastable state very accurately. On the contrary the system at low temperature explores this region but does not explore the rest of the state space efficiently. The main idea of the Replica exchange algorithm is now to include trail moves which try to interchange the information of the high temperature system and the low temperature system. Therefore, the algorithm tries to swap the positions of the two different systems according to the Metropolis Hastings ratio. Due to the swapping the low temperature sampling explores much more of the state space. Since the swapping probability is very low, if the temperature difference of two systems is very high, many intermediate systems are introduced. Therefore, a good swapping rate can be guaranteed and thus a good exploration of the state space is achieved. The swap moves are not very expensive because they do not involve additional calculations. Furthermore, the Boltzmann distribution corresponding to a particular system is not changed by the swapping. The temperature of the individual trajectory is not changed, only the position. So the stationary distribution can be easily obtained as in an ordinary sampling. A visualization of a Replica exchange algorithm is given in 3.13.

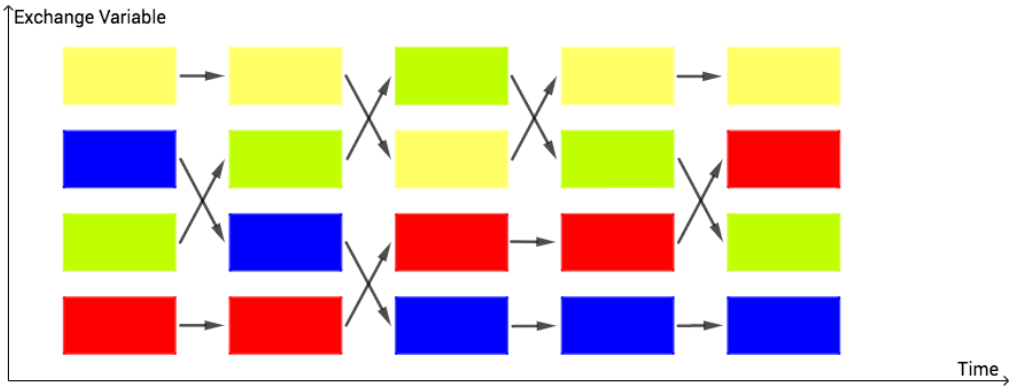


Fig. 3.13: Example scheme of a Replica exchange algorithm.

The convolution approach can also be integrated into the Replica exchange algorithm. Like in the original Replica exchange algorithm we run different trajectories

but we use the convolution in order to generate different systems. The smoothing decreases the barrier and thus the trajectory can explore the state space much faster similar to the high temperature sampling. For smaller smoothing parameters the metastable states are better explored. This variant of Replica exchange might be very useful in a situation where an entropic barrier is present for the reasons explained above. **Example**
As a proof of concept we tested the convolution Replica exchange for a one-dimensional and a two-dimensional example. For the sampling an Euler-Maruyama discretization was used. We compared our results with a temperature Replica exchange.

For the one-dimensional example we calculated a trajectory of 10,000 steps. We used the asymmetric bistable potential given in (3.8). In this case the convolution is given by (3.9). For simplification we only used two different potentials for the convolution Replica exchange and the temperature Replica exchange. For the convolution Replica exchange $\lambda_1 = 0$ and $\lambda_2 = 0.04$ was used. The inverse temperature $\beta = 4$ was chosen. For the temperature Replica exchange $\beta = 4$ and $\beta = 1$ were chosen. The time step was set to $dt = 0.001$ for all simulations. The algorithm tries to swap the position of the particle in every step.

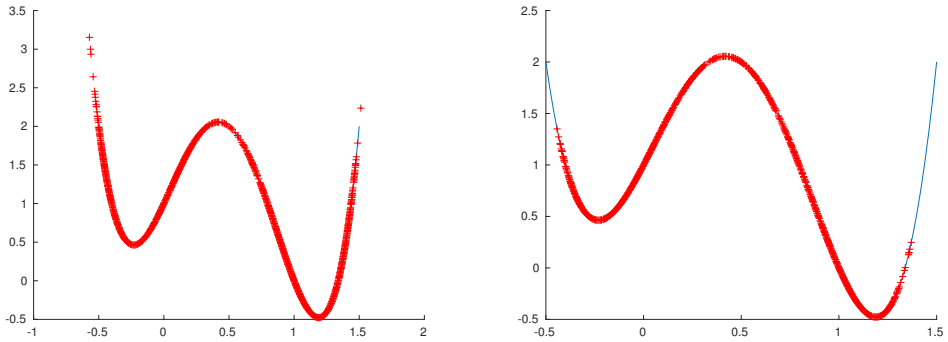


Fig. 3.14: Sampled region for the temperature Replica exchange (left) and the convolution Replica exchange (right).

In the one-dimensional example one clearly sees a difference between the two approaches. For the temperature Replica exchange the whole potential in the region [-0.5,1.5] has been sampled. Due to the high temperature the particle also visited very unlikely states. In the convolution Replica exchange the metastable states were explored very well. The particle did not visit states very far outside compared to the temperature Replica exchange. The sampling is much more focused around the local maximum. In the temperature Replica exchange the swapping took place 4,437 times and in the convolution Replica exchange the position was swapped 4,863 times.

The algorithm was also tested for a two-dimensional example. We consider the following potential

$$V(x,y) = 3\exp(-x^2-(y-1/3)^2) - 3\exp(-x^2-(y-5/3)^2) \\ - 5\exp(-(x-1)^2-y^2) - 5\exp(-(x+1)^2-y^2) + 1/5x^4 + 1/5(y-1/3)^4. \tag{3.18}$$

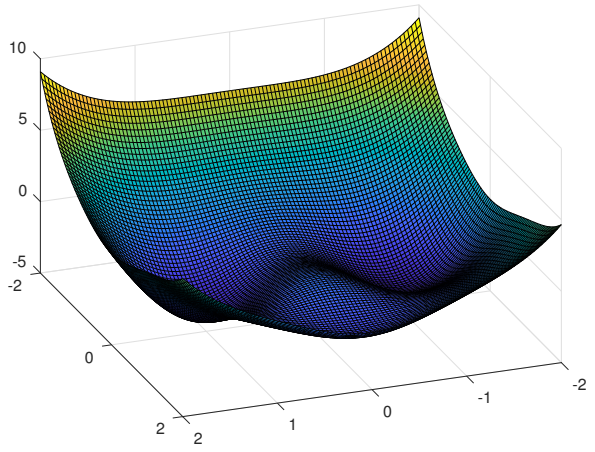


Fig. 3.15: Visualization of the two-dimensional three well potential.

The potential has two equally deep minima at approximately $(0,1)$ and $(0,-1)$ and one less deep minima at $(1.5,0)$. A local maximum is approximately at $(0.75,0)$. There are two different transition paths which connect the two deep minima. There is the direct connection and a second path is given by first going into the less deep minimum before going into the other deep minimum. A visualization can be found in figure 3.15

A trajectory of length 10,000 was calculated with time step $dt = 0.1$. Again, for simplicity we only considered two different potentials for the temperature Replica exchange and the convolution Replica exchange. For the convolution Replica exchange $\lambda_1 = 0$ and $\lambda_2 = 0.7$ were used. The convolution was approximated by a Monte Carlo approximation as shown in 3.1. The inverse temperature was set to $\beta = 10$. In the temperature Replica exchange $\beta_1 = 10$ and $\beta_2 = 2$ were used. Like in the one-dimensional example the swap was tried in each step. For better comparison the same random numbers were used in the two different approaches.

In this example one also sees a difference in the two different samplings. In the temperature Replica exchange all three metastable sets were visited. We also see like in the one-dimensional example that very unlikely positions were visited. The sampling of the convolution Replica exchange looks much more structured. Nearly no unlikely position was sampled. For the convolution Replica exchange all three

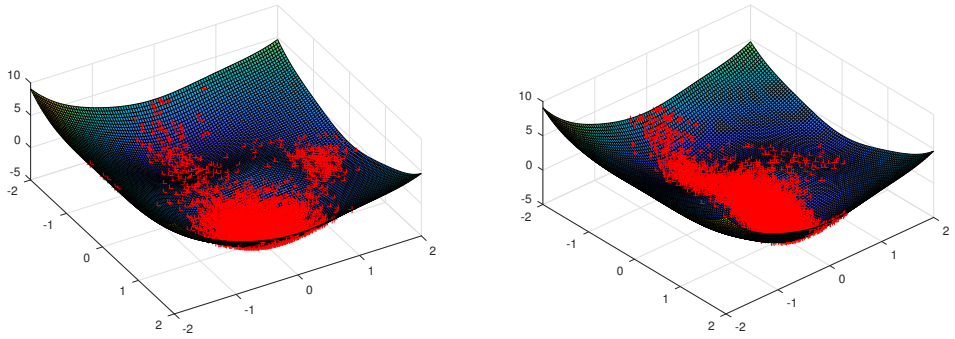


Fig. 3.16: Sampled region for the temperature Replica exchange (left) and the convolution Replica exchange (right).

metastable sets were explored but also the area around the local maximum was explored very well. In the temperature Replica exchange this is not the case. The position has been swapped 54,381 times in the temperature Replica exchange and 58,491 times in the convolution Replica exchange.

3.2.2 Reweighting

In MD and also in Monte Carlo simulations reweighting techniques are used to transfer results to another one referred to a different parameter (e.g. temperature) without additional simulations, for example. These reweighting schemes are often related to importance sampling as presented in the introduction. First, we present the general theory before showing how the convolution approach can be used in a reweighting scheme.
One form of reweighting is based on the Boltzmann distribution (ρ). In case of different temperatures the different Boltzmann distribution can be related easily. Let us consider two different inverse temperatures β and β'. The distributions are related by

$$\rho_{\beta'} \propto \exp(-\beta' V) = C \exp(-(\beta' - \beta)V)\rho_\beta \tag{3.19}$$

where C is a constant depending on β and β' and is often undetermined. The expectation of an observable $\varphi(x)$ with respect to temperature β' can be written as

$$\begin{aligned}
\mathbb{E}_{\beta'}[\varphi] &= \frac{1}{Z_{\beta'}} \int \varphi(x)\rho_{\beta'}(x)dx \\
&= \frac{C}{Z_{\beta'}} \int \varphi(x) \exp(-(\beta' - \beta)V(x))\rho_\beta dx \\
&= \frac{Z_\beta}{Z_{\beta'}} C\mathbb{E}_\beta[\varphi \exp(-(\beta' - \beta)V)]
\end{aligned}$$

where Z_β and $Z_{\beta'}$ are the normalization constants. The ratio of the normalization constants is given by

$$\frac{Z'_\beta}{Z_\beta} = C\mathbb{E}_\beta[\exp(-(\beta' - \beta)V)]. \tag{3.20}$$

The resulting reweigthing scheme for an observable φ is given by

$$\mathbb{E}_{\beta'}[\varphi] = \frac{\mathbb{E}_\beta[\varphi \exp(-(\beta' - \beta)V)]}{\mathbb{E}_\beta[\exp(-(\beta' - \beta)V)]}. \tag{3.21}$$

This reweighting scheme can also be extended to the convolution approach. Since the stationary distribution depends on the potential and the potential is changed by the convolution, the stationary distribution also changes. We are going to denote the distribution of the convoluted potential by ρ_λ and the original distribution by ρ. The two different distributions can be related in a similar way as in the different temperature setting

$$\rho \propto \exp(-\beta V) = C\exp(-\beta(V - V_\lambda))\rho_\lambda \tag{3.22}$$

where C is again a constant.

Let us again consider the expectation of an observable φ

$$\mathbb{E}[\varphi] = \frac{1}{Z}\int \varphi(x)\rho(x)dx \tag{3.23}$$

$$= \frac{C}{Z}\int \varphi(x)\exp(-\beta(V(x) - V_\lambda(x)))\rho_\lambda(x)dx \tag{3.24}$$

$$= \frac{Z_\lambda}{Z}C\mathbb{E}_\lambda[\varphi \exp(-\beta(V(x) - V_\lambda(x)))]. \tag{3.25}$$

The ratio of the two normalization constants is given by

$$\frac{Z}{Z_\lambda} = C\mathbb{E}_\lambda[\exp(-\beta(V(x) - V_\lambda(x))]. \tag{3.26}$$

So the main difference compared to the temperature reweighting is that the difference has to be calculated between the potentials instead of the temperatures.

Example
As a proof of concept we are going to show this approach for the sampling of a stationary distribution of the same three well potential given in (3.18)

For the sampling we use again a Euler-Maruyama discretization. We sampled one trajectory with $10,000$ steps. The temporal discretization was set to $dt = 0.1$. We compared the method with a standard sampling in the unperturbed potential. The inverse temperature was set to $\beta = 5$. The convolution was approximated by a Monte Carlo procedure as shown in 3.1. The smoothing parameter was set to $\lambda = 0.5$.

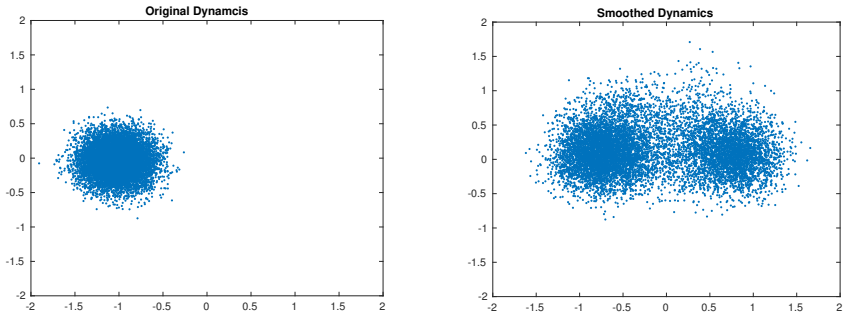


Fig. 3.17: Sampled region in the non-convoluted potential (left) and the sampled region for the convoluted potential (right).

In figure 3.17 we compared the individual trajectories of the different samplings. For a better comparison we used the same random numbers for both trajectories.

One can see that the convolution lowers the barrier and thus the trajectory explores more of the state space. The left figure shows the exploration of the standard sampling. The trajectory in the original potential did not explore the region outside the metastable region in which it started. The right figure shows the exploration in the convoluted potential. The trajectory in the convoluted potential explored the two deep metastable regions and the local maximum. It crossed the barrier to the other metastable regions several times. The two different pathways were explored by the trajectory. Even the less metastable region which is above the two bigger metastable regions was visited. While the standard sampling does not give us any information about the numbers of minima and the different barriers, the sampling in the convoluted potential gives us more state space information. The shown results indicate the decreasing effect of the convolution on the metastability.

In order to test the reweighting scheme presented above we are going to reconstruct the Boltzmann distribution of the unconvoluted potential from the convoluted sampling. In order to reconstruct the correct Boltzmann distribution from the biased sampling we have to approximate the reweighting factor given in (3.25). This can be done during the sampling. We applied the reweighting scheme in the two-dimensional example.

Figure 3.18 shows the original Boltzmann distribution and the reweighted Boltzmann distribution from the sampling shown above. We see that the reconstructed Boltzmann distribution and the original Boltzmann distribution agree very well. So the reweighting scheme in combination with the convolution to lower the barrier works. The less metastable region is not visible in this example because of the chosen inverse temperature.

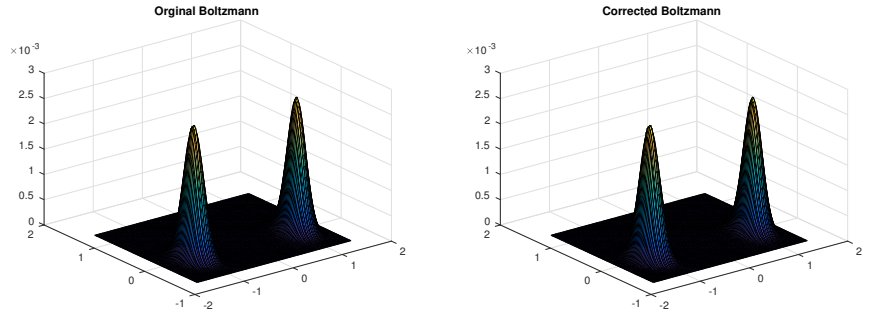


Fig. 3.18: Comparison of the original Boltzmann distribution (left) and the reweighted Boltzmann distribution from the sampling in the convoluted potential.

In this section we have seen how the convolution approach can be used for the sampling of thermodynamic quantities. In various low-dimensional examples these methods have been tested showing quite good results. In the next section we are going to investigate, if the convolution approach can be understood in terms of Linear Response theory. By doing this we can describe the response of the dynamical system on the convolution in a systematic manner.

3.3 Linear response

As we have seen so far the convolution decreases the metastability. But we have also seen that the convolution changes the dynamical system. Therefore, we are going to show in this section how this change can be quantified and that the convolution approach can be understood in Linear Response theory. Linear Response theory is a way to characterize the behaviour of a dynamical system on some small external forcing. The response of the system on this small force can be explicitly quantified. This result is especially useful for thermodynamic quantities because by using the theory we can calculate how the system reacts and thus how the quantity changes by the convolution. Furthermore, this can be used to correct the thermodynamic quantities sampled in a convoluted potential and in this way we can calculate equilibrium quantities with decreased metastability.

So in order to apply Linear Response theory we have to show that the convolution can be interpreted as a small external force. Therefore, it is possible to calculate a response function and thus quantify how the observable changes under the convolution explicitly. It is possible to rewrite this formula such that the convolution can be interpreted as a perturbation of the original potential. Let us consider that the potential $V : \mathbb{R}^n \to \mathbb{R}$ is a polynomial function of the form shown in Theorem 5

$$V(x) = \sum_{i_1, i_2 \ldots i_d} a_{i_1, i_2 \ldots i_d} x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n}. \tag{3.27}$$

Then the convolution is given by

$$V_\lambda(x) = \sum_{i_1,i_2\dots i_n} a_{i_1,i_2\dots i_n}\Big((x_1^{i_1} + \frac{i_1!}{(i_1-2)!}\lambda x^{i_1-2} + G_{i_1}(x_1,\lambda)) \\ +(x_2^{i_2} + \frac{i_2!}{(i_2-2)!}\lambda x^{i_2-2} + G_{i_2}(x_2,\lambda)) \\ \dots \\ +(x_n^{i_n} + \frac{i_n!}{(i_n-2)!}\lambda x^{i_n-2} + G_{i_n}(x_n,\lambda))\Big)$$

where

$$G_n(x,\lambda) = \sum_{k=2}^{\lfloor n/2\rfloor} a_{i_1,i_2\dots i_d}\frac{n!}{k!(n-2k)!}\lambda^k x^{n-2k}.$$

After multiplication and rearranging terms we find

$$V_\lambda(x) = \sum_{i_1,i_2\dots i_n} a_{i_1,i_2\dots i_n}\Big((x_1^{i_1}x_2^{i_2}\dots x_n^{i_n} + \frac{i_1!}{(i_1-2)!}\lambda x^{i_1-2}x_2^{i_2}\dots x_n^{i_n} \\ +x_1^{i_1}\frac{i_2!}{(i_2-2)!}\lambda x^{i_2-2}\dots x_n^{i_n} \\ +\dots + x_1^{i_1}x_2^{i_2}\dots\frac{i_n!}{(i_n-2)!}\lambda x^{i_n-2} + \mathcal{O}(\lambda^2)\Big).$$

This result shows that we can first reconstruct the original function and second we can order the additional terms according to the power of the convolution parameter λ. We are interested in the situation in which λ is close to zero because we only want small perturbations of the dynamical system. Because of the small smoothing parameter we can exclude all terms which include λ with exponent bigger or equal 2 and approximate the convolution by

$$V_\lambda(x) = \sum_{i_1,i_2\dots i_n} a_{i_1,i_2\dots i_n}\Big((x_1^{i_1}x_2^{i_2}\dots x_n^{i_n} + \lambda F(x) + \mathcal{O}(\lambda)^2\Big) \tag{3.28}$$

where $F(x) = \sum_{j=1}^n \frac{i_j!}{(i_j-2)!}x^{i_j-2}\prod_{\substack{k=1\\k\neq j}}^n x_k^{i_k}$.
Considering this a new drift of the SDE we see that the convolution can be seen approximately as a linear perturbation. Thus it is possible to derive a response function for the convoluted potential.
Let us consider a stochastic dynamical system satisfying the SDE (2.16). Furthermore, let us consider that the potential V is polynomial and so is the derivative of the potential. If we now apply the convolution to this potential function, we know that there is a formula given by (3.5). The convoluted SDE is now given by

$$dx_t = -(\nabla V_\lambda(x_t))dt + \sqrt{2\beta^{-1}}dB_t. \tag{3.29}$$

For a small λ we can neglect the higher order terms and write

$$dx_t = (-\nabla \bar{V}_\lambda(x_t))dt + \sqrt{2\beta^{-1}}dB_t \tag{3.30}$$

where $\bar{V}_\lambda(x) = \sum_{i_1,i_2\ldots i_n} a_{i_1,i_2\ldots i_n}\Big((x_1^{i_1}x_2^{i_2}\ldots x_n^{i_n} + \lambda F(x)\Big)$. In order to calculate the deviation of the system from the equilibrium system we assume that the distribution function $f^\lambda(x,t)$ of the convoluted system satisfies the Fokker-Plank equation

$$\frac{\partial f^\lambda}{\partial t} = \mathcal{L}^{\dagger\lambda} f^\lambda \tag{3.31}$$

$$f^\lambda|_{t=t_0} = \rho_{eq}. \tag{3.32}$$

The initial condition represents that the system is at equilibrium and ρ_∞ is the equilibrium distribution. The generator of the process satisfying (3.30) is given by

$$\mathcal{L}^{\dagger\lambda} = \mathcal{L}_0^{\dagger\lambda} + \lambda\mathcal{L}_1^{\dagger\lambda} \tag{3.33}$$

where $\mathcal{L}_0^{\dagger\lambda}$ corresponds to the Fokker-Plank operator of the unperturbed system and $\mathcal{L}_1^{\dagger\lambda}$ is related to the external forcing. Since the convolution can be expressed as a linear perturbation and the generator is also linear we have that

$$\mathcal{L}_1^{\dagger\lambda} = \nabla F \cdot \mathcal{D} \tag{3.34}$$

where $\mathcal{D}$ is some linear differential operator. Furthermore, we know that for the unique equilibrium distribution

$$\mathcal{L}\rho_{eq} = 0. \tag{3.35}$$

The resulting Fokker-Plank operator of the perturbed dynamics is given by

$$\mathcal{L}^{\dagger\lambda} = -\nabla\cdot(\nabla V\cdot) + \frac{1}{2\sigma^2}\nabla^2 - \lambda\nabla F\cdot\nabla \tag{3.36}$$

The solution we are looking for is expressed as a power series expansion in λ

$$f^\lambda = f_0 + \lambda f_1 + +\lambda^2 f_2\ldots \tag{3.37}$$

Substituting this into (3.32) and using the initial condition we find the following system of equations

$$\frac{\partial f_0}{\partial t} = \mathcal{L}_0^{\dagger\lambda} f_0 \quad f_0|_{t=0} = \rho_{eq} \tag{3.38}$$

$$\frac{\partial f_1}{\partial t} = \mathcal{L}_0^{\dagger\lambda} f_1 + \mathcal{L}_1^{\dagger\lambda} f_0 \quad f_1|_{t=0} = 0. \tag{3.39}$$

The only solution to the first equation is the invariant distribution $f_0 = \rho_{eq.}$. Substituting this into the second equation and using the form of $\mathcal{L}_1^{\dagger\lambda}$ we find

$$\frac{\partial f_1}{\partial t} = \mathcal{L}_0^{\dagger\lambda} f_1 + \nabla F \mathcal{D} f_0. \tag{3.40}$$

Using a variation of constants formula to solve the above equation we get

$$f_1(t) = \int_{t_0}^{t} \exp(\mathcal{L}_0^{\dagger\lambda}(t-s)) \nabla F(x) \mathcal{D} \rho_{eq} ds. \tag{3.41}$$

The above calculation can now be used to calculate the deviation of an observable due to the external forcing in expectation. Let $A(\cdot)$ be an observable. Furthermore, we denote with $O(t)$ the deviation of its expectation value from equilibrium

$$\begin{aligned} O(t) &= \langle A(X_t) \rangle - \langle A(X_t) \rangle_{eq} && (3.42)\\ &= \int_{\mathbb{R}^n} A(x)(f^\lambda(x,t) - \rho_{eq}(x)) dx && (3.43)\\ &= \lambda \int_{\mathbb{R}^n} A(x) \Big(\int_{t_0}^{t} \exp(\mathcal{L}_0^{\dagger\lambda}(t-s)) \nabla F(x) \mathcal{D} \rho_{eq} ds \Big) dx && (3.44) \end{aligned}$$

The above formula can be used to calculate observables for a system at equilibrium by sampling the convoluted system. We can simply sample the observable in the convoluted potential and subtract the correction term to get the observable for the unconvoluted potential. In order to apply the correction formula the generator and the stationary distribution have to be known what can be a drawback. But together with a good approximation scheme this would result in an efficient sampling method because the correction term can be calculated on the fly.

So far we have seen that the convolution approach decreases the metastability. We have shown various applications of the approach, e.g. rapid state space exploration and also the efficient sampling of thermodynamic quantities. In the last section of this chapter we are going to apply the convolution approach to the sampling of dynamic quantities.

3.4 Convolution for dynamic quantities

After we have seen that the convolution approach can be applied efficiently for the sampling of thermodynamic quantities we would now like to investigate, if the approach can also be used for the sampling of dynamic quantities. First we are going to consider the problem of sampling exit times for double well potentials. We do this because in this special situation we can use the Arrhenius' law to describe the mean first exit time. The Arrhenius law states that the mean first exit time mainly depends on the barrier. We have already seen that the convolution decreases the barrier and if

we assume that the potential is polynomial or can be approximated by a polynomial, we can quantify the change of the barrier explicitly using the representation formula known from Theorem 5. This enables us to develop an extrapolation algorithm using samplings from convoluted and less metastable dynamical systems.
In the second section we are going to generalize the application of the convolution approach to other dynamic quantities. We do this by building an importance sampling scheme based on Girsanov's theorem. As a biasing potential we are going to use the convoluted potential. In order to apply this method we have to check that the Novikov's condition are satisfied. We conclude with a numerical example.

3.4.1 Convolution for exit times

In this section we consider the application of the convolution approach to the exit time problem in a double well potential. In this situation Arrhenius' or Kramers' law applies which gives us explicit expression for the exit time cf. [5], [8] and [69]. The Arrehnius law states that the mean first exit time mainly depends on the heights of the barrier. We are going to use the convolution to reduce this height such that the exit time is lower and thus easier to sample. So in order to find the exit time for the original potential we are going to develop an extrapolation scheme. For this we assume that the potential is of polynomial form as presented in Theorem 5. So we can make use of the finite sum representation of the convolution and quantify the change of the barrier. Combining this with the formula of Kramer's law we can use this to extrapolate the exit time of the original potential from sampling the less metastable system. In the end we show a one-dimensional application of the extrapolation scheme and compare our results with the Monte Carlo estimator and the exact exit time coming from the PDE formulation of the problem.

Let us first introduce the Eyring-Kramers' law. We consider a stochastic process satisfying (2.38) in a metastable set S be a bounded domain in $\mathbb{R}^n$ with smooth boundary. We denote the exit time to leave the domain S of this stochastic process which started in $x \in S$ by

$$\tau_S = \inf\{t > 0, x_t \notin S\}. \tag{3.45}$$

The quantity we would like to estimate is the mean first exit, which is the expectation of the stopping time $\mathbb{E}^x[\tau_S]$. In order to sample this dynamic quantity we have to sample paths $(x_{0:\tau})$ of the SDE (2.38). So we see that estimating this quantity is a path sampling problem.
From the literature it is well-known that the mean first exit time mainly depends on two things: the barrier height $\Delta E = V(z) - V(y)$ where z is the local maximum and y is the minimum of the metastable set $\mathcal{S}$ and the temperature β of the stochastic

process. Furthermore, we assume that the stochastic fluctuations are weak in comparison to the barrier which means

$$\frac{1}{\beta \Delta E} \ll 1, \tag{3.46}$$

then the mean first exit time is a rare event. In accordance with large deviations theory the relevant time scale of the exit event to happen scales exponentially in β

$$\mathbb{E}^x[\tau_S] \simeq C \exp[\beta(V(z) - V(y))] \tag{3.47}$$

where y is the minimum of the set S and z is the local maximum and C is some constant cf. [5] or [8]. Based on this formula it is also possible to define a so-called hopping rate which quantifies the rate at which the metastable set is left cf. [69]. It is given by

$$\kappa \sim C^{-1} \exp\left(-\beta(V(z) - V(y))\right). \tag{3.48}$$

It was possible to quantify the constant C for different situations leading to the Eyring-Kramers' law cf. [28, 50]. For the one-dimensional case the Eyring-Kramers' formula satisfies

$$\mathbb{E}^x[\tau_S] \simeq \frac{2\pi}{\sqrt{V''(y)|V''(z)|}} \exp[\beta(V(z) - V(y))].$$

From this explicit expression of the constant C we see that its dependency on the curvature of the potential. We can also conclude that a smaller curvature is leading to a longer transition time cf. [5]. The multidimensional version for $n \geq 2$ satisfies

$$\mathbb{E}^x[\tau_S] \simeq \frac{2\pi}{|\alpha_1(z)|} \sqrt{\frac{|\det \nabla^2 V(z)|}{|\det \nabla^2 V(y)|}} \exp[\beta^{-1}(V(z) - V(y))].$$

where $\alpha_1(z)$ is the single negative eigenvalue of the Hessian $|\det \nabla^2 V(z)|$ at the local maximum. Due to the exponential dependence on the barrier heights the sampling of the hopping rate or the mean first exit time is difficult to sample. In general a lower barrier leads to a smaller exit time and thus it is easier to sample. We have seen that the convolution decreases the barrier heights. So we can build an extrapolation scheme by combining the convolution formula of Theorem 5 and the Eyring-Kramers' formula.

Derivation of the Extrapolation formula

The overall idea is to smooth out the barrier by convoluting the potential in order to make the sampling easier. Furthermore, we would like to use the biased sampling to calculate the mean first exit time for the unconvoluted potential. By combining Eyring-Kramers' law and the convolution approach we can build an extrapolation

method. With this method we can use the biased sampling data to extrapolate for the original mean first exit time. We are using the Arrhenius formulation for simplicity.

We have already seen that the mean first exit time mainly depends on the barrier

$$\mathbb{E}^x[\tau_S] \simeq C \exp[\beta(V(z) - V(y))]. \tag{3.49}$$

We assume that the potential is a polynomial such that we can use the representation formula given in (3.2). Plugging this into the Arrhenius' law (3.49) and rearranging terms we find

$$\beta^{-1} \log(1/C(\mathbb{E}^x[\tau_S(\lambda)])) \simeq V_\lambda(z) - V_\lambda(y) \tag{3.50}$$

where $V_\lambda(\cdot)$ is the convoluted potential evaluated at the corresponding minima y or the local maxima z. We see that the log mean first exit time of the convoluted dynamical system changes in a polynomial fashion in λ by using Theorem 5 (this will get clearer in the one-dimensional example). So using this connection we can sample the mean first exit time for different convoluted potentials and use a logarithmic regression to extrapolate the mean first exit time for the unconvoluted potential. The result can be made even more accurate if the change of the potential is taken into account in the constant C. Since the convolution is changing the whole potential, it will also affect the constant. But in many different applications the constant cannot be calculated and as we will see in the one-dimensional example the method gives quite good results without taking this change into account. Let us state an algorithm for the extrapolation in the next section.

Algorithm

The main idea of this approach is to sample the exit time in different smoothed potentials and use this data to extrapolate for the original exit time. We state the algorithm in pseudocode next.

Result: Quantity of Interest
initialisation: $x_0 = x$; decreasing sequence $\lambda_i,\ i = 1, \dots N$
for *i=1:N* **do**
  calculate the $V(x, \lambda_i)$ by solving the convolution (3.2);
  sample the quantity of interest in $V_{\lambda_i}(x)$;
  save quantitiy of interest (λ_i);
end
Extrapolate the quantity of interest for the unconvoluted potential by (3.50)

Algorithm 1: Extrapolation Scheme

In order to investigate the dependency of the smoothing parameter and the mean first exit time we can also use the boundary value problem presented in the example 2.2.1. This will give us some theoretical insights but does not give additional information on the extrapolation formula.
Let us start with the boundary value problem. Due to the Feynman-Kac formula the

mean first exit time $u(x) = \mathbb{E}^x[\tau_S]$ can also be expressed as the following boundary value problem

$$\mathcal{L}u(x) = -1 \quad x \in S \tag{3.51}$$

$$u(x) = 0 \quad x \in \delta S \tag{3.52}$$

where $\mathcal{L}$ is the infinitesimal generator of the stochastic process. Since we are going to use the convolution to lower the barrier, we can use the boundary value problem to calculate the explicit dependency of the exit time on the smoothing parameter. Convoluting the potential introduces a second parameter such that the mean first exit time also depends on the smoothing parameter. In order to explore the λ dependency on the mean first exit time we just calculate the derivative of the above stated boundary value problem.

$$\begin{aligned}
&\frac{\partial}{\partial\lambda}\mathcal{L}^\lambda(x)u_\lambda(x) = -1 \\
\Leftrightarrow &-\frac{\partial}{\partial\lambda}\nabla_x V_\lambda(x)\nabla_x u_\lambda(x) + \frac{\partial}{\partial\lambda}\frac{1}{2}\Delta_x u_\lambda(x) = 0 \\
\Leftrightarrow &-\nabla_x\frac{\partial}{\partial\lambda}V_\lambda(x)\nabla_x u_\lambda(x) - \nabla_x V_\lambda(x)\frac{\partial}{\partial\lambda}\nabla_x u_\lambda(x) + \frac{1}{2}\Delta_x\frac{\partial}{\partial\lambda}u_\lambda(x) = 0 \\
\Leftrightarrow &\mathcal{L}^\lambda(x)\frac{\partial}{\partial\lambda}u_\lambda(x) = \nabla_x\frac{\partial}{\partial\lambda}V_\lambda(x)\nabla_x u_\lambda(x) \\
\Leftrightarrow &\mathcal{L}^\lambda(x)\frac{\partial}{\partial\lambda}u_\lambda(x) = \nabla_x\Delta_x V_\lambda(x)\nabla_x u_\lambda(x)
\end{aligned}$$

This result shows that the mean first exit time in case of smoothing the potential is changing mostly in the direction of the curvature. This observation fits the observation of Eyring-Kramers' formula that a smaller curvature in the stable direction decreases the mean first exit time cf. [5]. But apart from this we do not know to use the above result to correct the biased sampling or extrapolate to get the right result. To solve the last equation is actually very difficult because the resulting system of equations depend on each other.

Example

As a proof of concept we will give a one-dimensional example of a particle moving in a metastable potential. We are interested in the mean first exit time of the particle leaving a specific well and the hopping rate of this event.

The position of the particle $x_t \in \mathbb{R}$ at time t is described by an SDE satisfying (2.38). We consider the same asymmetric bistable potential as in the first one-dimensional example

$$V(x) = 8x^4 - 44/3x^3 + 2x^2 + 11/3x + 1. \tag{3.54}$$

A visualization can be found in figure 3.1. According to the representation formula (3.5) the convolution of the potential is given by

$$V_\lambda(x) = V(x) + 48\lambda^4 + (2 - 44x + 48x^2)\lambda^2. \tag{3.55}$$

To predict the mean first exit time for a stochastic process stating in the left well we are going to generate three different potentials each convoluted with a different λ. In each of these convoluted potentials we sample the exit time. Due to the reduced metastability the sampling effort is reduced. To extrapolate the exit time for the unconvoluted potential based on the sampling in the convoluted potentials we use the log regression based on equation (3.50). In this one-dimensional example the formula simplifies to

$$\begin{aligned}\beta^{-1}\log(1/C(\mathbb{E}^x[\tau_S(\lambda)])) &\simeq V_\lambda(z) - V_\lambda(y)\\ &= \sum_{k=0}^{\lfloor n/2\rfloor} \frac{n!}{k!(n-2k)!}\lambda^k z^{n-2k} - \sum_{k=0}^{\lfloor n/2\rfloor} \frac{n!}{k!(n-2k)!}\lambda^k y^{n-2k}\\ &= \sum_{k=0}^{\lfloor n/2\rfloor} \frac{n!}{k!(n-2k)!}(z-y)^{n-2k}\lambda^k. \end{aligned}\tag{3.56}$$

In order to show that the extrapolation scheme produces good results we calculate the exact exit time by the PDE formulation. Furthermore, we compare the extrapolated result with a standard Monte Carlo sampling for the exit time in the unconvoluted potential.

In this example we want to sample the mean first exit time of an SDE starting in the left well. So the starting point was set to $x = -0.25$ for all simulations. The temperature was set to $\beta = 3$. We used three different smoothing parameters $\lambda_1 = 0.15, \lambda_2 = 0.2, \lambda_3 = 0.25$. The sampling was done with an Euler-Mayurama algorithm with temporal discretization $dt = 0.001$. The exit time was sampled 1000 times for each individual sampling. The Monte Carlo approximation was also done with 1000 samplings. The exact exit time was calculated with a finite difference scheme based on the formulation (3.52).

The example shows that the extrapolated solution and the real solution agree very well. Comparing the absolute errors of the extrapolation scheme and the PDE solution ($error = time_{PDE} - time_{ext} = 1.9$) and the MC and the PDE solution ($error = time_{PDE} - time_{mc} = 3.7$) we see that the extrapolation scheme performs better than the MC estimator. In total, the sampling time for the exit time in the three convoluted potentials together is also lower than the sampling time of the Monte Carlo estimator due to the reduced barrier heights.

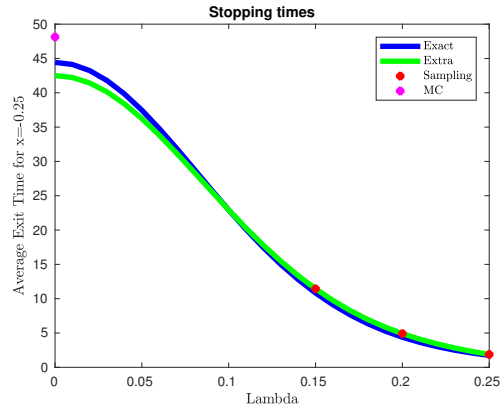


Fig. 3.19: PDE solution (blue) for different parameters λ in [0,0.25] at x=-0.25 and the extrapolated solution (green) which is calculated of the sampled data from the smoothed dynamics. In pink the Monte Carlo estimator for the exit rate is shown. The red crosses show sampled exit time in the convoluted potential which was used for the extrapolation.

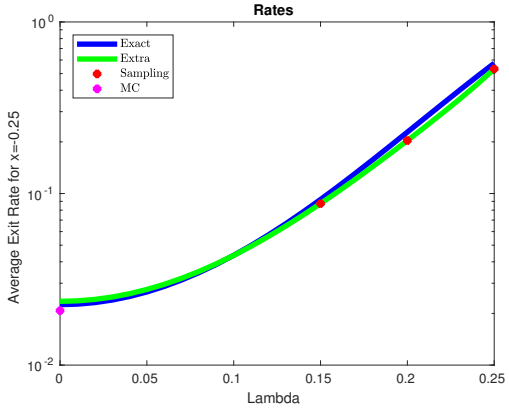


Fig. 3.20: PDE solution (blue) for different parameters λ in [0,0.25] and extrapolated solution (green) with the three calculated data points for $\lambda_1 = 0.15$, $\lambda_2 = 0.2$ and $\lambda_3 = 0.25$.

So we have seen that the convolution approach can be used for the sampling of mean first exit times by combining it with the Eyring-Kramers' formula. In the next section we develop an importance sampling scheme for a more general formulation of dynamic quantities.

3.4.2 Generalization for dynamical quantities

As we have already seen in Chapter 2 dynamic quantities of SDE can be expressed in general as

$$\mathbb{E}^x_{\mathbb{P}}\left[\exp\left(\int_0^{\hat{\tau}} f(x_s)ds + g(x_{\hat{\tau}})\right)\right] \tag{3.57}$$

where $\tau = \hat{\min}(\tau, T_N)$ and T_N is a finite sampling time. Since the convolution decreases the metastability, we can use the convoluted potential as a bias for the

sampling of dynamical quantities. So in order to develop an importance sampling scheme we combine the convolution approach with the reweighting strategy based on Girsanov's theorem as presented in Chapter 2 3. Girsanov's theorem gives us a direct formula how the path measure changes, if the drift of the SDE is changed. In order to apply the theorem we have to make sure that Novikov's condition is satisfied. Furthermore, we have also seen that the resulting estimator is unbiased. We are going to state a Lemma which assures that Girsanov's theorem can be applied and shows a one-dimensional example as a proof of concept of the method.

We are interested in sampling dynamical quantities from a metastable dynamical system satisfying equation (2.16). But in order to reduce the sampling effort and the variance of the estimator we would like to sample the less metastable dynamics given by

$$dy_t = -\nabla V_\lambda(y_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad y_0 = x. \tag{3.58}$$

In order to compensate the change of drift we use the reweighting formula from Girsanov's theorem. In this case it satisfies

$$\begin{aligned} M_t = \exp\Big(&- \frac{1}{\sqrt{2\beta^{-1}}}\int_0^t (\nabla V(x_s) - \nabla V_\lambda(x_s))dB_s \\ &- \frac{1}{4\beta^{-1}}\int_0^t (\nabla V(x_s) - \nabla V_\lambda(x_s))^2 ds\Big) \end{aligned} \tag{3.59}$$

such that we can use

$$\mathbb{E}_{\mathbb{P}}[f(x_{0:\tau})] = \mathbb{E}_{\mathbb{Q}}[f(y_{0:\tau})] = \mathbb{E}_{\mathbb{P}}[M_\tau f(y_{0:\tau})]$$

to calculate dynamic quantities for the original dynamics by sampling the less metastable dynamics given by equation (3.58). Let us next state a Lemma proving Novikov's condition.

Lemma 2. *We assume the potential V to be a continuous differential function for which ∇V_λ exits and is again a continuous function. Furthermore, let us assume that the time $\tau = \hat{\min}(\tau, T_N)$ is finite and the events we are interested in can be sampled on a closed and bounded set $\mathcal{D} \subset \mathbb{R}^n$. Then Novikov's condition holds and we can use Girsanov's theorem to calculate path-dependent quantities from non-equilibrium sampling for the equilibrium dynamics.*

Proof. We have to verify Novikov's condition for the difference $\nabla V(\cdot) - \nabla V_\lambda(\cdot)$ which states that

$$\mathbb{E}\Big[\exp\Big(\frac{1}{2}\int_0^{\hat{\tau}} \Big|\frac{\nabla V(y_t) - \nabla V_\lambda(y_t)}{\sqrt{2\beta^{-1}}}\Big|^2 dt\Big)\Big] < \infty. \tag{3.60}$$

We know that the difference of the continuous function is again a continuous function. So we can calculate

$$\begin{aligned}\frac{1}{2}\int_0^{\hat{\tau}}\left|\frac{\nabla V(y_t)-\nabla V_\lambda(y_t)}{\sqrt{2\beta^{-1}}}\right|^2 dt &= \frac{\beta}{4}\int_0^{\hat{\tau}}\left|\nabla V(y_t)-\nabla V_\lambda(y_t)\right|^2 dt \\ &\leq \frac{\beta}{4}\hat{\tau}\left|\nabla V(y)-\nabla V_\lambda(y)\right|^2 \\ &\leq \frac{\beta}{4}\hat{\tau}\sup_{y\in\mathcal{D}}\left|\nabla V(y)-\nabla V_\lambda(y)\right|^2 < \infty\end{aligned}$$

Since we know that the set $\mathcal{D}$ is closed and bounded, we know by the extreme value theorem that the maximum is attained on $\mathcal{D}$. Due to the fact that the difference of two continuous functions is again continuous we know that the difference is bounded. Furthermore, we assumed the time horizon to be bounded and so it follows that the whole expression is bounded. From this we conclude that (4.17) is satisfied. Therefore, Novikov's condition holds. □

After we have seen that Girsanov's theorem can be applied let us consider an example.

Example

Let us consider a one-dimensional example as a proof of concept. We want to sample the moment generating function of the stopping time

$$\tau = \inf\{t > 0, x_t > 0.5\} \tag{3.61}$$

of the SDE satisfying (2.38) in this potential

$$V(x) = 8x^4 - 44/3x^3 + 2x^2 + 11/3x + 1. \tag{3.62}$$

Instead of sampling in the original potential we are going to sample in the convoluted potential given by

$$V_\lambda(x) = V(x) + 96(\lambda^2/2)^2 + (4 - 88x + 96x^2)(\lambda^2/2) \tag{3.63}$$

for $\lambda = 0.2$. The two different potentials are visualized in figure 3.21.

We simulated 1000 trajectories with a standard Euler-Maruyama algorithm. The time step was set to $dt = 0.0001$ and the temperature was set to $\beta = 3$. The starting point was set to $x_0 = -0.25$. In order to reweight the expectation we used the standard Girsanov formula as shown in equation (3.59).

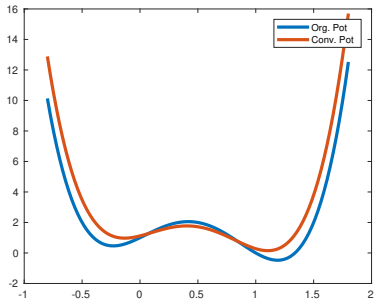


Fig. 3.21: Original potential (blue) and convoluted potential (orange)

	MC	IS
$\mathbb{E}[e^{-\beta\tau}]$	1.958×10^{-3}	1.782×10^{-3}
Var	3.937×10^{-3}	1.773×10^{-4}
$R(I)$	32.04	7.46
$\mathbb{E}[\tau]$	27.06	3.64
$Var[\tau]$	699.65	10.86

Tab. 3.1: Comparison of the importance sampling estimator (IS) in the convoluted potential and the Monte Carlo estimator for the moment generating function and the mean hitting time (non reweighted) and its variance.

The example shows that with the convolution as bias a variance reduction for dynamical quantities can be achieved. As one can see in table 3.1 even for a small number of samples the variance can be reduced by one order of magnitude. The estimator of moment generating function also show good agreement. The only disadvantage of this approach is that the potential has to be evaluated twice to calculate the difference. For high-dimensional problems this can cause a lengthy calculation and a loss in efficiency. Furthermore, we show the non-reweighted estimators of the stopping time and its variance for the convoluted potential and the original potential. The comparison of the two estimators shows that the convolution decreases the stopping time. From this we can again conclude that the convoluted dynamical system is less metastable. So the convolution can be used to accelerate the sampling. The also shown variance of the mean first stopping time shows that the barrier heights has an influence on the variance of the estimator.

3.5 Summary and Discussion

In this chapter we presented the convolution approach to decrease the metastability and thus simplify the sampling of different quantities. We showed that this approach can be used to decrease the metastability without apriori knowledge of the exact location of the metastability in different one- and two-dimensional examples. The application of the approach to Butane showed that the approach can also be used in a high-dimensional setting. We combined the approach with an importance sampling

strategy for the sampling of thermodynamic quantities. After this we could show that the convolution can be seen as a small external force acting on the potential and so it is possible to quantify the response of the dynamical system in terms of Linear Response theory. In the last part of the chapter we applied the convolution approach to the sampling problem of path dependent quantities. We derived an extrapolation scheme based on the Eyring-Kramers' formula for mean first exit times in double well potentials and also showed that we can use Girsanov's theorem in order to build an importance sampling scheme for general dynamic quantities.

During the many numerical simulations which were performed in this chapter the choice of the convolution parameter has always been a very critical point of the approach. It was never clear from the beginning how large the influence of the smoothing parameter on the dynamical system would be. Therefore, to find a good smoothing parameter different numerical simulations have been run and analysed. From a theoretical viewpoint one can say that the smoothing parameter has to be chosen such that the spectral gap becomes smaller but can still be detected. But the actual parameter λ which does that is quite difficult to find. One possibility to investigate the influence of the convolution on the metastability is to investigate the behaviour of the eigenvalues of the transfer operator under the convolution. A possible way in order to do this could be the application of Kato theory; see [44]. Kato theory was developed for the investigation of linear operators under perturbation. Since the infinitesimal generator is a linear operator which is perturbed by the convolution, an analysis could give more theoretical information.
For the importance sampling scheme for thermodynamic quantities the calculation of the reweighting factor for the reweighting method may also become very unhandy for high-dimensional problems. The method proved to be very accurate in the low-dimensional setting but the approximation of the reweighting factor in many dimensions can become very difficult.
It was possible to derive a linear response function. In order to use this response function to correct observables coming from a biased sampling the stationary distribution and the infinitesimal generator of the stochastic process has to be known or approximated. This can be very challenging for high-dimensional problems. So to use this scheme efficiently for numerical simulations a good approximation for the integral has to be developed in future work.
Even though the examples for the application of the methods for path sampling problems showed good results difficulties could occur. The extrapolation scheme for the mean first exit time can be extended to high-dimensional problems because here also a closed formula is known. But the high-dimensional formula requires knowledge about the eigenvalues of the Hessian at the local minimum. This is a drawback because the convolution will also influence these eigenvalues. How large the influence is and if the change of the eigenvalues may be neglected has to be found out in future research.

In the case of general dynamic quantities the convolution of the potential does not steer the dynamical system into a specific direction. This is one drawback of the convolution approach for dynamic quantities. It can be used for quantities like exit times. But for transitions in a potential with many different minima a more directed forcing will be much more efficient. Another disadvantage is that for the evaluation of the Girsanov weight the potential has to be evaluated twice. This can be very costly in higher dimensions and reduce the efficiency of the method. This is why we are going to develop an importance sampling approach for dynamical quantities based on a local bias in the next chapter. In order to build this local bias we are going to transfer the experience from the method developed for thermodynamic quantities to the sampling of dynamic quantities.

4 Adaptive importance sampling

The sampling of dynamic quantities is important to characterize its behaviour and as we have seen in the last section of chapter 3.4 the convolution approach can be used to reduce the variance of the sampling. But the convolution approach can be too naive in many situations. For example, for the sampling of transitions the convolution lowers the barriers but it does not really steer the dynamic system into the right direction in a controlled way. Therefore, we are going to extend a certain class of enhanced sampling algorithms which were developed for the effective sampling of thermodynamic quantities to the sampling of dynamic quantities. In the computational physics community many algorithms have been developed for effective sampling of thermodynamic quantities as we have already seen in chapter 2.5. But we have also seen that the sampling of dynamic quantities is a different sampling problem and this is why it is not possible to directly use these methods for the path sampling problem. The main difference of the two sampling problems are the random variables and the different probability measures which are considered as already presented in the introduction. This also results in a different Radon-Nikodym derivative which has to be used to compensate the biasing. So in order to build an adaptive importance sampling scheme for dynamic quantities of metastable complex dynamical questions we are going to combine the assimilated well-known enhanced sampling methods coming from MD with a theorem from stochastic analysis, Girsanov's theorem.

The class of enhanced sampling algorithms which we are going to use is the class of biasing algorithms. The main idea of these algorithms is to introduce an artificial bias (on the potential or on the force) to decrease the metastability of the dynamical system. Enhanced sampling algorithms have been designed for sampling stationary distributions and thermodynamic quantities. By assimilating these algorithms they can also be used for sampling dynamic quantities and variance reduction.

A very similar method for importance sampling has been proposed by [95]. The main difference to our approach is that we are not interested in finding the optimal bias because it is computationally too expensive. Instead, we use the unbiasedness of the estimator and construct a suboptimal bias which will lead to a variance reduction. Moreover, while the authors of [95] considered a very general approach of variance reduction for solving PDEs by path integrals of SDEs, we only look at dynamical systems which are metastable. In order to construct the bias we use Metadynamics. For this we need an additional sampling but we do not have to know the stationary distribution.

In the introduction we have seen that there exits a optimal bias which would give a zero variance estimator. But the optimal bias depends on the quantity itself and it is not computable without solving the sampling problem. Furthermore, we have also seen in Chapter 2 that the unbiasedness of the estimator is independent from the used bias. Since the main goal of importance sampling is the variance reduction we try to design a bias which decreases the variance sufficiently instead of calculating the optimal one. To construct a good bias we would like to benefit from the enhanced sampling algorithms for thermodynamic quantities. In order to do this these methods have to be assimilated. Combining the assimilated algorithms with an effective reweighting scheme we have built an adaptive importance sampling algorithms for sampling dynamic quantities. In this chapter we are going to use Metadynamics as an example for the class of enhanced sampling techniques. But the introduced framework is not restricted to Metadynamics. We are going to show this by applying Metadynamics directly on the force of the dynamical system and also proving Novikov's condition under general assumptions. In principle any other adaptive importance sampling scheme like Adaptive Biasing Force, Hyperdynamics etc. can be used, as long as Novikov's condition is satisfied. The proposed algorithm has two advantages compared to a standard estimator of dynamic quantities: firstly, it is possible to produce estimators with a lower variance and, secondly, the sampling is speeded up. The proposed method can be seen as a method that creates a non-equilibrium dynamics which is used to sample the equilibrium quantities.
The chapter is structured as followed. First, we are going to introduce additional theory of importance sampling and briefly comment on Girsanov's theorem. Then, we are presenting Metadynamics as an example of the used enhanced sampling techniques. Next, some theoretical aspects of the methods are presented before different numerical examples are shown. We conclude the chapter with a short summary and a discussion of the numerical results.
Parts of this chapter have been published in [73]. In this chapter further analysis of the resulting algorithm has been added. Additional one-dimensional examples have been run in order to show different aspects of the algorithm. A two-dimensional application can be found in [73].

4.1 Importance sampling for dynamic quantities

In this section we present the two main ingredients of the algorithm. In the first part we briefly review the main idea behind importance sampling and show how the variance affects the statistical error. The second section is a short introduction into Metadynamics and how the algorithm is assimilated.

We are going to consider the diffusion process $x_t \in \mathbb{R}^n$ governed by the SDE as the root model for the motion of a molecule. The SDE is given

$$dx_t = -\nabla V(x_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad x_0 = x \tag{4.1}$$

where x_t is the state of the system at time $t \geq 0$, $V : \mathbb{R}^n \to \mathbb{R}$ is a sufficiently smooth (e.g., $\mathcal{C}^\infty$) potential, $\beta^{-1} > 0$ is an arbitrary scaling factor for the noise, often called the inverse temperature, and B_t is a standard n-dimensional Brownian motion with respect to the probability measure $\mathbb{P}$ on the probability space $(\Omega, \mathbb{P}, \mathcal{F})$.
Moreover, we assume that the process is trapped in a metastable region $\mathcal{S} \subset \mathbb{R}^n$ which is an open and bounded set with a smooth boundary. Furthermore, we define a target set $\mathcal{T}$ that is an open and bounded set with a smooth boundary as well. Finally, we define the stopping time $\tau = \inf\{t > 0 : x_t \in \mathcal{T}\}$ to be the first time that the process (2.16) hits the target set $\mathcal{T}$, e. g. when a dihedral angle of a biomolecule reaches a certain value. The presented theory here can be generalized to a state-dependent diffusion constant, cf. [67]. But here we will consider the constant case only.
We are interested in expectations of the form

$$\mathbb{E}[\exp\left(-\beta g(x_{0:T})\right)] \tag{4.2}$$

where $x_{0:T}$ is a trajectory of (2.16) until some finite time T and g is some functional on $\mathcal{C}([0,T] : \mathbb{R}^n)$. We consider this type of quantities because they give us information in terms of the temperature of the system. However, a generalization to other quantities expressed as expectations is possible. As pointed out by [88] an interesting case of this quantity arises when $g = 0$ for $x_{0:T} \in \mathcal{A} \subset \mathcal{C}([0,T], \mathbb{R}^n)$ and $g = \infty$ otherwise. Then, (4.2) becomes

$$\mathbb{P}[x_{0:T} \in \mathcal{A}]. \tag{4.3}$$

Expectations like (4.2) are integrals over the entire state space and cannot be calculated analytically. But, given an ensemble of paths, they can be approximated by an unbiased MC estimator

$$I = \mathbb{E}[\exp\left(-\beta g(x_{0:T})\right)] = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\beta g(x^i_{0:T})\right) \tag{4.4}$$

where $x^i_{0:T}$, $i \in [1, \ldots, N]$ are independent paths of length T, all starting at the same point $x_0 = x \in \mathbb{R}^n$, produced, for example, by numerical integration of (2.16). This estimator is unbiased. Its variance is given by

$$\text{Var(I)} = \frac{1}{N}(\mathbb{E}[\exp\left(-2\beta g(x_{0:T})\right)] - \mathbb{E}[\exp\left(-\beta g(x_{0:T})\right)]^2). \tag{4.5}$$

The relative error is defined by

$$r(I) = \frac{\sqrt{(\mathrm{Var}(I))}}{\mathbb{E}[I]} = \frac{1}{\sqrt{N}} \sqrt{\frac{\mathbb{E}[\exp\left(-2\beta g(x_{0:T})\right)]}{\mathbb{E}[\exp\left(-\beta g(x_{0:T})\right)]^2} - 1} \tag{4.6}$$

cf. [88].

To build an importance sampling scheme for a metastable diffusion process one has to decrease the depth of the minima which cause the metastable behaviour. Since the time evolution of the SDE (2.16) with a low temperature (i. e. large β) is a negative gradient descent perturbed by some Brownian motion, the process x_t will stay in the region around the minimum of V. By filling the metastable region in $V(\cdot)$ we change the metastable behaviour and thus the sampling of the desired quantity of interest gets easier. But this perturbation changes the underlying path distribution as well. To compensate for this perturbation we use Girsanov's theorem to reweight (or correct) the estimators. Another interpretation of this theorem is that it offers a way to sample equilibrium quantities of some dynamics by sampling the dynamics out of equilibrium. We can construct a bias, which influences the multimodality of the stationary distribution in such a way that low probability regions are more probable or high barriers are easier to cross. The main advantage of Girsanov's theorem is that it is not necessary to know the stationary distribution a priori. So we would like to sample a less metastable SDE

$$dy_t = -(\nabla V(y_t) + \nabla V_{bias}(y_t; c, w, \lambda)dt + \sqrt{2\beta^{-1}}dB_t, \quad y_0 = x \tag{4.7}$$

where $\nabla V_{bias}(y_t; c, w, \lambda)$ is now a local perturbation reducing the metastability. Similar to what we have seen in chapter 3 we can use the weight given by Girsanov's theorem satisfying

$$\begin{aligned} M_t = \exp\Big(&\frac{1}{\sqrt{2\beta^{-1}}} \int_0^t (\nabla V_{bias}(y_t; c, w, \lambda))dB_s \\ &- \frac{1}{4\beta^{-1}} \int_0^t (\nabla V_{bias}(y_t; c, w, \lambda))^2 ds\Big). \end{aligned} \tag{4.8}$$

The importance sampling estimator for dynamical quantities of interest is then given by

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(-\beta g(y_{0:T})\Big) \\ &\quad \exp\Big(\frac{1}{\sqrt{2\beta^{-1}}} \int_0^T (\nabla V_{bias}(y_t; c, w, \lambda))dB_s - \frac{1}{4\beta^{-1}} \int_0^T (\nabla V_{bias}(y_t; c, w, \lambda))^2 ds\Big)\Big] \end{aligned} \tag{4.9}$$

Furthermore, as we have seen in chapter 2 we can derive a different formula of the weight if the bias is of gradient structure. The corresponding importance sampling estimator satisfies

$$\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(-\beta g(y_{0:T})+\frac{1}{2\beta^{-1}}\Big(V_{bias}(y_T;c,w,\lambda)-V_{bias}(y_0;c,w,\lambda)\Big)\Big)$$
$$\exp\Big(\frac{1}{2\beta^{-1}}\int_0^T\Big(\nabla V(y_s)\nabla V_{bias}(y_s;c,w,\lambda)+\frac{1}{2}|\nabla V_{bias}(y_s;c,w,\lambda)|^2$$
$$\beta^{-1}\Delta V_{bias}(y_s;c,w,\lambda)ds\Big)\Big]. \tag{4.10}$$

Independent from the used expression of the Girsanov reweighting the expectation can be again approximated by a finite summation similar to (4.4) as

$$\hat{I}=\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\beta g(y^i_{0:\tau})\right)M^i_{0:\tau} \tag{4.11}$$

where $y^i_{0:\tau}$ and $M^i_{0:\tau}$ are independent samples from (4.7) and (2.40). For $a(y_t)$ satisfying Novikov's condition and a bounded stopping time, M_τ is a continuous bounded martingale which yields $\mathbb{E}[M_t]=1,\ t\in[0,\tau]$. Then, the importance sampling estimator is an unbiased estimator with expectation

$$\mathbb{E}[\hat{I}]=\mathbb{E}[\exp\left(-\beta g(x_{0:\tau})\right)] \tag{4.12}$$

cf. [58]. Following [88], we know that the relative error of this estimator is

$$r(\hat{I}):=\frac{1}{\sqrt{N}}\sqrt{\frac{\mathbb{E}[\exp\left(-2\beta g(y_{0:\tau})\right)(M_{0:\tau})^2]}{\mathbb{E}[\exp\left(-\beta g(x_{0:\tau})\right)]^2}-1}. \tag{4.13}$$

In order to control the relative error we have to control the ratio

$$R(\hat{I}):=\sqrt{\frac{\mathbb{E}[\exp\left(-2\beta g(y_{0:\tau})\right)(M_{0:\tau})^2]}{\mathbb{E}[\exp\left(-\beta g(x_{0:\tau})\right)]^2}}. \tag{4.14}$$

To apply proposition 2 to dynamic quantities like exit times we have to guarantee the fulfillment of Novikov's condition. This can be achieved by making an assumption on the stopping time.

Condition 1. *In order to guarantee the applicability of proposition 2 it is to be assumed that the stopping time is bounded for the specific problem. This assumption is by far non-trivial and can only be shown analytically in very few situations. Anyhow, from a numerical viewpoint it is impossible to simulate trajectories which have infinite length. One has to stop the simulation after a finite number of steps. The quantity of interest can be approximated by the quantity of interest conditioned on the event happening in a finite simulation time. This assumption can be formalized by considering the stopping time* $\hat{\tau}=\min(\tau,T_N)$ *where* T_N *is the length of the numerical simulation.*

Then, Novikov's condition follows for a reasonable function $u(\cdot)$. This treatment has been suggested in [58].

Condition 1 means that the sampling of the quantity of interest has to be finite in time. If the sampling is too long $(t > T_N)$, the simulation is stopped.

In conclusion, Proposition 2 gives us an option to sample the dynamic quantity of interest from a different dynamical system without knowing the stationary distribution a priori. The different dynamical system can be changed in such a way that the quantity of interest is observed more often. The main difficulty of applying this strategy to a metastable system is to determine the metastable regions to change it accordingly. For this, we are going to use Metadynamics. This algorithm is used in MD to sample the free energy surface and can be seen as an adaptive biasing method. In order to use this algorithm for the effective sampling of dynamic quantities we are going to assimilate the algorithm slightly.

4.1.1 Metadynamics

The method Metadynamics was first proposed in [41] called *local elevation*. It was reintroduced as Metadynamics in [53]. It is an adaptive method for sampling the free energy surface (FES) of high-dimensional molecular systems. The main purpose of this method is the efficient sampling of the FES and the construction of the corresponding stationary distribution. The method combines dynamics in reaction coordinates with adaptive bias potentials. The idea of this approach is to perturb the energy landscape when the simulation is trapped in a metastable region. This is done by locally adding Gaussian functions along a reaction coordinate which fill up the minima in which the simulation is trapped. In this way it is possible to explore the energy landscape in a rather short time compared to the standard sampling approach. The convergence of some certain variants of Metadynamics was proved in [16].

In order to apply the method it is assumed that the high-dimensional system can be projected onto a few relevant collective coordinates. One possible way to find these collective variables for stochastic dynamics is to average out the fast degrees of freedom; see [54] or [94] for example. A more general overview can be found in [15] and the references therein. In general this projection can be written as $s : \mathbb{R}^n \to \mathbb{R}^d$ with $d \ll n$. Only the dependence of these parameters on the free energy $\mathcal{F}(s(x))$ is considered. The exploration of the FES is guided by the forces $F_i^t = -\partial \mathcal{F}(s_i(x))/\partial s_i^t$. But again metastability is a problem of the effective sampling which is not solved by the projection into reaction coordinates. So in order to sample the FES more efficiently a bias is added to the system whenever the simulation is stuck in such a minimum. With Metadynamics one constructs a bias potential

$V_{bias} : \mathbb{R}^d \to \mathbb{R}$ which is composed of $K \in \mathbb{N}$ Gaussian functions. The complete bias potential is

$$V_{bias}(x) = \sum_{i=1}^{K} \frac{w_i}{\sqrt{2\pi\lambda_i^2}} \exp\left(- \frac{(s(x) - c_i)^2}{2\lambda_i^2} \right) \tag{4.15}$$

where $w_i \in \mathbb{R}$ is a weight, $c_i \in \mathbb{R}^d$ is the centre of the Gaussian, and $\lambda_i \in \mathbb{R}$ is the width. These functions are placed along the trajectory to allow for an easy escape from the metastable sets using the derivatives as an artificial force. The method can be parallelized easily since the bias depends on the history of the individual trajectory only. This makes the method extremely efficient. Additionally, the bias also prevents the trajectory from going back to the visited states.

For simplicity we assume that all considered functions and variables are in the low-dimensional collective variable space and stick with the old notation. Of course, Girsanov's theorem is not restricted to the collective variable space.

4.1.2 Assimilation of Metadynamics

We are going to assimilate the Metadynamics algorithm to the sampling of dynamical quantities of interest. For our framework we do not have to calculate the complete FES. We only need a bias which makes sure the trajectory does not get trapped in the metastable region. This is the reason why we add an additional sampling before we start sampling the quantity of interest to build a bias. In order to build a bias which decreases the metastability, we use Metadynamics in the metastable region only.

The bias is built in the following way: When the trajectory is trapped in a metastable region we start a Metadynamics simulation until the trajectory has left the metastable region. In every kth step we add a Gaussian function to the potential such that the metastability is reduced. The force is then changed with the gradient of these Gaussian functions. When the trajectory hits the target set $\mathcal{T}$ for the first time, we save the bias and stop the Metadynamics simulation. The bias consists of $\#steps\ needed/k$ bias functions. The choice of k is a compromise between adding as few bias functions as necessary getting a small hitting time τ and not perturbing the potential too much. Depending on the choice of the parameters w and λ a certain number of bias functions is needed. It is obvious that the simulation of Metadynamics gets more expensive the more bias functions are added due to the increasing number of function evaluations. That is why all parameters should be adapted to the problem such that the computation does not get too costly.

After having built the bias potential the sampling of the original trajectory is continued with the bias potential. To correct the quantity of interest at the end of the

calculation we must sample the weights (2.40) as well. This can be done on the fly.

4.2 The algorithm

Now we present the algorithm in pseudocode. We will use Metadynamics to build a bias in the metastable regions of the potential. Then, we sample the quantity of interest in this biased potential N times and reweight the sampling with the weight given by (2.40).

Data: dynamics x_t, y_t, starting set $\mathcal{S}$, target set $\mathcal{T}$
initialisation: $x_0 = y_0 = x$; w_i, λ_i
Step 1: Build bias potential
while *Transition has not occurred* **do**
  sample the dynamics x_t given in eq. (4.7)
  every kth steps: add a new bias function to $u(\cdot)$
end
save the bias potential;
Step 2: Sample the quantity of interest
for *N* **do**
  sample the quantity of interest with the additional bias according to eq. (4.7)
  sample the weights according to eq. (2.40)
end
reweight according to eq. (2.42)
return estimator

Algorithm 2: Adaptive importance sampling

4.3 Properties of the method

In this section we are going to explore different properties of the algorithm. First we show under which assumptions Novikov's conditions are satisfied. Then, we conclude that the method preserves ergodicity. In the end we will give a couple of remarks concerning different aspects of the method.

4.3.1 Proof of Novikov's condition

To apply Girsanov's theorem one has to make sure that the Novikov's condition is satisfied. The proof is very similar to what we have seen in chapter 3.

Lemma 3. *Let $\hat{\tau}$ be the stopping time as given in Condition 1. Further let the bias potential V_{bias} be a continuous differential function for which ∇V_{bias} is bounded. We also assume that the events we are interested in can be sampled on a closed and bounded*

set $\mathcal{D} \subset \mathbb{R}^n$. Then the Novikov condition holds and we can use Proposition 2 to calculate path-dependent quantities from nonequilibrium sampling for the equilibrium dynamics.

Proof. Since the bias function is added to the potential V, the resulting SDE is given by

$$dy_t = (-\nabla V(y_t) + \nabla V_{bias}(y_t; c, w, \lambda)dt + \sqrt{2\beta^{-1}}dB_t, \quad y_0 = x. \tag{4.16}$$

We have to verify Novikov's condition for $\nabla V_{bias}(\cdot)$ which states that

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^{\hat{\tau}} \left|\frac{\nabla V_{bias}(y_t; c, w, \lambda)}{\sqrt{2\beta^{-1}}}\right|^2 dt\right)\right] < \infty. \tag{4.17}$$

We are going to show that the time integral is bounded from which we can then conclude that Novikov's condition holds. We can calculate

$$\begin{aligned}
\frac{1}{2}\int_0^{\hat{\tau}} \left|\frac{\nabla V_{bias}(y_t; c, w, \lambda)}{\sqrt{2\beta^{-1}}}\right|^2 dt &= \frac{\beta}{4}\int_0^{\hat{\tau}} \left|\nabla V_{bias}(y_t; c, w, \lambda)\right|^2 dt \\
&\leq \frac{\beta}{4}\hat{\tau}\left|\nabla V_{bias}(y; c, w, \lambda)\right|^2 dt \\
&\leq \frac{\beta}{4}\hat{\tau}\sup_{y\in\mathcal{D}}\left|\nabla V_{bias}(y; c, w, \lambda)\right|^2 < \infty
\end{aligned}$$

Since we know that the set $\mathcal{D}$ is bounded and that ∇V_{bias} is also bounded by assumption, we know by $\sup_{y\in\mathcal{D}}\left|\nabla V_{bias}(y; c, w, \lambda)\right|^2$ is bounded. Furthermore, we have assumed that time horizon finite and so it follows that the whole expression is bounded. From this we conclude that (4.17) is satisfied. Therefore Novikov's condition holds and Girsanov's theorem can be applied. □

From this Lemma follows that we can use the Gaussian function as the biasing potential such that ∇V is perturbed by the derivative of the Gaussian function or we can use directly Gaussian functions to perturb ∇V.

4.3.2 Ergodicity

In this paragraph we show that both variants of the Adaptive Importance Sampling methods preserve ergodicity. For this we are going to show that the resulting potential is still confining.

Lemma 4. *Let us consider a confining potential V. If the bias is constructed with the assimilated version of Metadynamics with Gaussian functions with variance $\lambda > 0$, then the biased potential is still a confining potential and Proposition 1 still applies.*

Proof. The bias is a sum of Gaussian functions. Since the Gaussian functions decrease to zero

$$\lim_{x \to \pm\infty} V_{bias} = 0$$

the biased potential is still confining

$$\lim_{x \to \pm\infty} V + V_{bias} = \infty.$$

□

Even if Metadynamics is directly applied to the force, the biased potential is still confining.

Lemma 5. *Let us consider a confining potential V. If the bias is constructed with the assimilated version of Metadynamics directly working on the derivative, the biased potential is still a confining potential and Proposition 1 still applies.*

Proof. The gradient of the bias potential is a sum of Gaussian functions so the bias itself is a sum of error functions due to linearity of integration. Since we assume the number of biasing functions to be bounded $N < \infty$ and the sum of the weights is also finite, the bias potential is also bounded.

$$\lim_{x \to \pm\infty} V_{bias} = |C|$$

The result follows from the exact calculation as before. □

4.3.3 Remarks

We conclude this section with some final remarks about the presented method.

Remark 1. *The construction of the bias potential depends on the history of the trajectory. Since the simulation to get the bias function is done in an additional step, the potential is not time dependent. Furthermore, the discretization of* (2.16) *always gives a discrete time Markov process because of the time independence of the Brownian motion. The construction of the bias potential itself is not Markovian because it depends on the history of the trajectory. Since the construction of the bias function and the sampling of the quantity of interest are done independently of each other, the bias does not have any influence on the Markovianity of the perturbed SDE* (4.7)*. In general, an extension of*

the proposed method for non-Markovian dynamics should be possible. In this regard one could use the Metadynamics methods proposed in [11] and the general reweighting formula given in [67].

Remark 2. *The method is not restricted to the use of Metadynamics. Any stochastic approximation algorithm (e.g. Adaptive Biasing Force) or even a deterministic algorithm could be used, provided the bias satisfies Novikov's condition. The method is also not restricted to the case in which the drift term ($b(\cdot)$) of the SDE is of gradient form. Since the Girsanov formula does not use the stationary distribution, all calculations are still valid. However, if the bias is not of gradient form the alternative Girsanov formula cannot be applied.*

Remark 3. *In order to create a good bias potential within a reasonable computational cost one can use the history of the trajectory to estimate the parameters of the bias functions. The midpoint c_i can be chosen to be the mean of the average of the last k steps and the λ_i can be chosen to be the maximal distance from the starting point of the last k steps times a constant, $C(\lambda_i = \max(|x_{i*(1:k)} - c_i|)),\ C \in \mathbb{R}$. This can be more efficient as we could see in the example shown above. In the literature one can find many extensions and variants of Metadynamics which could be used as well, e.g. [4]*

4.4 Examples

In the following we study different numerical examples of the method presented above. We assume that the reaction coordinates are given such that we have a low-dimensional representation of the high-dimensional dynamics. In the first section of the example section we are going to present different applications of the assimilated Metadynamics. The first one-dimensional example shows the construction of the bias potential with fixed parameters. In the second example we are going to use the alternative reweighting formula (4.10) to correct the statistics from the nonequilibrium sampling. The third example shows the application of the method for smaller inverse temperatures. In the last example of this part the reverse transition is considered as the quantity of interest. In the second part we are going to show different examples for the application of Metadynamics on the force. The first example is again the construction of the bias potential with fixed parameters. The second example shows how the parameters can be estimated from the sampling of the additional trajectory. In the last example of this section again the reverse transition is considered as the quantity of interest.

For the examples we consider the dynamics given by (2.16) and the potential given by

$$V(x) = \frac{1}{2}(x^2 - 1)^2. \tag{4.18}$$

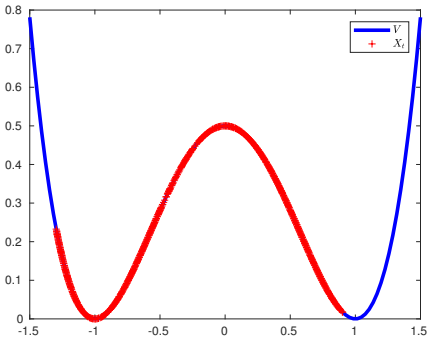


Fig. 4.1: In blue, the potential function (4.18) is shown and in red a realization of (2.16) showing the desired transition is presented.

This potential has two minima at $x = \pm 1$ and a local maximum at $x = 0$. We are going to calculate two different quantities of interest. The first quantity of interest is the probability of all continuous paths which start at a point x in the metastable region $\mathcal{S}$ and reach the target set in time $\hat{\tau} = \min(\tau, T_N)$ as shown in Condition 1. This can be written as $\mathbb{P}(\mathcal{A})$, where $\mathcal{A} = \{x_{0:\hat{\tau}} \in \mathcal{C}([0, \hat{\tau}], \mathbb{R}^n) | x_0 = x (x \in \mathcal{S}), x_{\hat{\tau}} \in \mathcal{T}\}$. For this quantity of interest we choose $g(y_t) = 0$ for $y_t \in \mathcal{S},\ t \in [0, \hat{\tau}]$ and choose $g(y_t) = 1$ for $y_t \in \mathcal{T},\ t \in [0, \hat{\tau}]$. The second quantity of interest is the moment generating function of the stopping $\hat{\tau}$. To sample this we set $g(y_{0:\hat{\tau}}) = \hat{\tau}$. The trajectories $y_{0:\hat{\tau}}$ are realizations of (4.7) with $b(\cdot) = -\nabla V(\cdot)$, $u(\cdot)$ is the bias constructed by the Metadynamics simulation and $\sigma = \sqrt{2\beta^{-1}}$. We compare our method with the results of a standard MC estimator for the different quantities. We will see that in the examples our method achieves the variance reduction for both reweighting formulas given in this thesis. Furthermore, the average sampling time was decreased in the biased simulation. For this, we estimate the mean sampling time (MST) for our experiments. The MST is the average time trajectories need to reach the target set. If the trajectory does not hit the target set the MST is set to T_N.

In all examples 1000 trajectories of (4.7) are calculated by using a standard Euler-Maruyama discretization with a time step $\Delta t = 10^{-4}$ in MATLAB, cf. [39]. Our aim is to investigate the variance of the different estimators and the MST. At maximum, we calculate $T_N = 15.000$ time steps. The random number generator is fixed to have a better comparison within the different examples. The seed of the random number generator is given in the introduction of the example to simplify reproducibility. The examples have been tested with other random number generators which are not shown here, but showing similar results.

4.4.1 Assimilated Metadynamics

In this section different examples for the assimilated version of Metadynamics are presented.

Example 1: Diffusion in a double well
For this example we define the metastable region $\mathcal{S} = [-1.5, 0]$. We choose the starting point of the SDE (2.16) in the metastable region $x_0 = -1$ and set the temperature to $\beta = 3.0$ for all simulations. The stopping time is defined as the first hitting time of the target set $\mathcal{T} = [0.9, 1.1]$. The random number generator was set to rng(1,'twister').

The parameters of the bias function have been set to $w_i = 0.05,\ \lambda_i = 0.8\ \forall i$. The centre c_i of every bias function is chosen as the current value of the trajectory when the new bias function is added. The Girsanov weights to reweight the expectation have been calculated by (4.8).

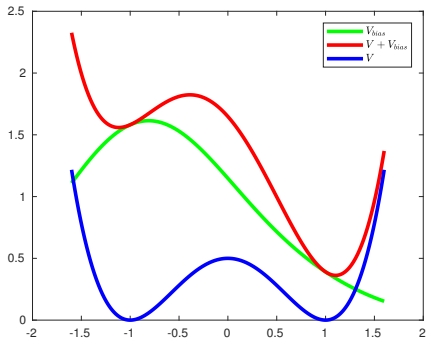


Fig. 4.2: The blue curve shows the original potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics and in red the corresponding biased potential $(V + V_{bias})$ is shown.

	MC	GIR
$P(\mathcal{A})$	4.8470×10^{-2}	4.8323×10^{-2}
Var	4.6121×10^{-2}	1.6404×10^{-2}
$R(I)$	4.4307	2.6504
$\mathbb{E}[\exp(-\beta\tau)]$	2.569×10^{-3}	2.4885×10^{-3}
Var	2.5850×10^{-4}	6.9180×10^{-5}
$R(I)$	6.2561	3.3463
MST	1.4804	1.4425

Tab. 4.1: Comparison of the importance sampling estimators and the Monte Carlo estimators for the simulation with fixed parameters of the biased potential.

In this example 79 bias functions have been used; see figure 4.2 for the calculated bias and the resulting potential. The estimators of the MC and the importance sampling are in good agreement for both cases; see 4.1. The results show that the

variance of the biased estimator is reduced for both quantities of interest using the reweighting approach. The variance for the transition probability is reduced by 65% and for the moment generating function by 76%. Hence, the automatically generated bias potential by the adjusted Metadynamics is actually a good potential in the sense of importance sampling. Additionally, the MST is faster compared to the plain MC approach. This example shows that our method achieves the desired goals of variance reduction and computational speed-up.

Example 2: Alternative Reweighting Formula
In the following example we use the alternative reweighting formula as shown in (4.10). In order to calculate the bias we use the same parameters as in the first example. Since the random number generator was set to rng(1,'twister') , the bias is exactly the same as in the first example; see figure 4.2.

In order to calculate the Girsanov's weights we need V_{bias}, ∇V_{bias} and $\nabla^2 V_{bias}$. For V_{bias} as given in (4.15), the derivatives can be calculated easily.

	MC	AGIR
$P(\mathcal{A})$	4.8470×10^{-2}	4.8329×10^{-2}
Var	4.6121×10^{-2}	1.6407×10^{-2}
$R(I)$	4.4307	2.6504
$\mathbb{E}[\exp(-\beta\tau)]$	2.5690×10^{-3}	2.4856×10^{-3}
Var	2.5850×10^{-4}	6.9170×10^{-5}
$R(I)$	6.2561	3.3459
MST	1.4804	1.4425

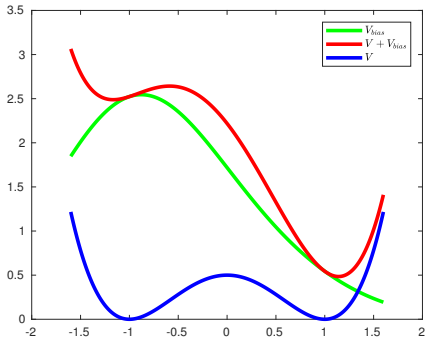
Tab. 4.2: Comparison of the importance sampling estimators and the MC estimators for the simulation with the alternative Girsanov formula.

In this case the MC estimator and the importance sampling estimator agree very well. The variance reduction is very similar to the other reweighting formula. The variance for the transition probability is reduced by 64% and the variance for the moment generating function is reduced by 73%. These examples show that the alternative Girsanov formula can be applied well to correct the biased estimators. The reduction of the MST is the same as in the first example since we used the same seed for the random number generator.

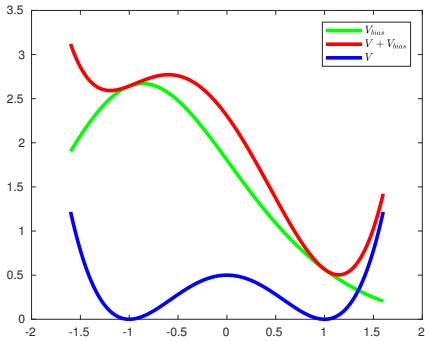
Example 3: Diffusion in a double well with lower temperature

In order to test the method in a low temperature situation we calculate the probability of all continuous paths which start at a point x in the metastable region $\mathcal{S}$ and reach the target set in time $\hat{\tau}$ for $\beta = 7$ and for $\beta = 10$. The random number generator is

set to rng(3,'twister'). The Girsanov weights to reweight the expectation have been calculated by (4.8).



(a) The blue curve shows the original potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics and in red the corresponding biased potential $(V + V_{bias})$ is shown for $\beta = 7$.



(b) The blue curve shows the original potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics and in red the corresponding biased potential $(V + V_{bias})$ is shown for $\beta = 10$.

Fig. 4.3: Resulting bias for a low temperature sampling.

$\beta = 7$	MC	GIR
$P(\mathcal{A})$	1×10^{-3}	1.9×10^{-3}
Var	1×10^{-3}	6.845×10^{-5}
$R(I)$	31.62	4.319
$\beta = 10$	MC	GIR
$P(\mathcal{A})$	0	2.04×10^{-4}
Var	0	1.845×10^{-6}
$R(I)$	∞	6.602

Tab. 4.3: Comparison of the importance sampling estimators and the Monte Carlo estimators for the simulation with fixed parameters of the biased potential.

In the example, for $\beta = 7$, 123 bias functions have been used; see figure 4.3 for the calculated bias and the resulting potential. The estimators of the MC and the importance sampling for the probability leaving the set within a certain time are in good agreement. In the case of the biased sampling 87 trajectories have reached the set within the simulation time. In the unbiased sampling only 1 trajectory has reached the target set in the simulation time.
In this example, for $\beta = 10$, 128 bias functions have been used; see figure 4.3 for the calculated bias and the resulting potential. In the unbiased sampling no trajectory has reached the target set within the simulation time. So the MC estimator is zero. In the biased sampling 49 trajectories have reached the set within the simulation time. So we still can calculate an estimator.
The above examples show that the proposed algorithm also works for low temperature samplings. In both tested cases a variance reduction could be achieved.

Furthermore, the algorithm could still give results when the MC method produces no estimator at all.

Example 4: Reverse Problem

In a second example we test the method for the transition and the mean first exit time for the reverse problem. For this example we define the metastable region $\mathcal{S} = [0, 1.5]$. We choose the starting point of the SDE (4.7) in the metastable region $x_0 = 1$ and fix $\beta = 3.0$ for all simulations. The stopping time is defined as the first hitting time of the target set $\mathcal{T} = [-0.9, -1.1]$. The random number generator was set to rng(3,'twister'). The parameters of the bias potential are the same as in the previous example. The Girsanov weights to reweight the expectation have been calculated by (4.8).

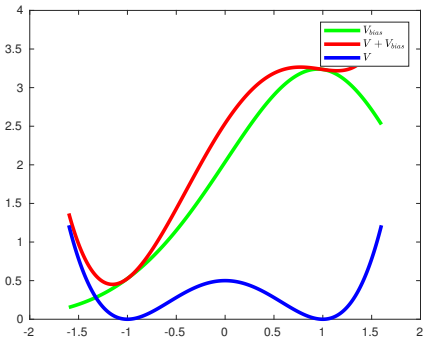


Fig. 4.4: The blue curve shows the original potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics and in red the corresponding biased potential $(V + V_{bias})$ fo the reverse problem is shown.

	MC	GIR
$P(\mathcal{A})$	4.64×10^{-2}	5.138×10^{-2}
Var	4.3927×10^{-2}	8.25×10^{-3}
$R(I)$	4.5563	1.7679
$\mathbb{E}[\exp(-\beta\tau)]$	2.4021×10^{-3}	2.558×10^{-3}
Var	2.42×10^{-4}	1.567×10^{-5}
$R(I)$	6.4837	1.5475
MST	1.4812	1.3132

Tab. 4.4: Comparison of the importance sampling estimators and the Monte Carlo estimators for the simulation with fixed parameters of the biased potential.

Even though the problem is symmetric the algorithm generates different biasing potentials for the reverse problem. In this example 155 bias potentials have been used; see figure 4.4 for the calculated bias and the resulting potential. The example shows again a good agreement of the MC estimators and the importance sampling for both quantity of interest; see table 4.4. The results show that the variance of the biased estimator is reduced for both using the reweighting approach. The variance for the transition probability is reduced by 80% and for the moment generating

function by 95%. This example shows that the method achieves the desired goals of variance reduction and computational speedup, also for different problems.
The difference in the bias potentials can come from the difference of the gradient of the potential. In the left well the gradient is negative while in the right well the gradient is positive. Another explanation is the sampled trajectory. The bias is built with an extra sampling which only depends on one trajectory. So the bias function strongly depends on this extra sampling.

4.4.2 Metadynamics applied on the force

In this example we are going to explore the idea to apply the assimilated version of Metadynamics directly on the gradient of the potential. Since the metastability arises from the structure of the gradient, it makes sense to directly perturb the gradient without taking the detour via the potential. The main reason why we propose this version of adaptive importance sampling scheme are the Gaussian functions used in the previous examples. Since the potential is biased with Gaussian functions, the gradient is perturbed by the gradient of the Gaussian. The derivative of a Gaussians has a positive and a negative part and thus a small minimum and a small maximum are added. It seems that the resulting bias is very rugged which could cause new sampling problems. This is why changing the gradient directly with Gaussian functions should result in smoother bias.
Furthermore, the examples show that the proposed method is not restricted to the usage of Metadynamics. Basically, each algorithm producing a bias can be integrated into the proposed framework.

The biased dynamical system again satisfies

$$dy_t = (\nabla V_{bias}(y_t) - \nabla V(y_t))dt + \sqrt{2\beta^{-1}}dB_t. \tag{4.19}$$

The main difference to the previous approach is that the gradient of the system is biased with a sum of Gaussian functions. So the bias function is defined by

$$\nabla V_{bias}(y_t) = \begin{bmatrix} \frac{dV_{bias}}{dx_1} \\ \vdots \\ \frac{dV_{bias}}{dx_d} \end{bmatrix} := \begin{bmatrix} \sum_{i=1}^{K} b_i(y_t, c_i, w_i, \lambda) \\ \vdots \\ \sum_{i=1}^{K} b_i(y_t, c_i, w_i, \lambda) \end{bmatrix}. \tag{4.20}$$

where

$$b_i(x; c, w, \lambda) = \frac{w_i}{\sqrt{2\pi\lambda_i^2}} \exp\Big(- \frac{(s(x) - c_i)^2}{2\lambda_i^2} \Big), \quad i \in [1, \dots, K] \tag{4.21}$$

where $c_i \in \mathbb{R}^n \ \forall i \in [1, \dots, N]$ is the centre of the bias function, $w_i \in \mathbb{R} \ \forall i \in [1, \dots, K]$ is a weight, and $\lambda_i \in \mathbb{R} \ \forall i \in [1, \dots, N]$ is the width of the bias function.

The bias function is built in the same way as the bias potential in Metadynamics. We again use the assimilated version of Metadynamics in order to generate the biasing force only in the metastable region. From sampling the dynamical system one gets blurred gradient information which can be used to approximate the gradient such that the metastability is removed or decreased.

Example 5: Diffusion in a double well

The quantities of interest are as in the previous examples. As in the first example the metastable region are defined as $\mathcal{S} = [-1.5, 0]$. We choose the starting point of the SDE (4.7) in the metastable region $x_0 = -1$ and fix $\beta = 3.0$ for all simulations. The stopping time is defined as the first hitting time of the target set $\mathcal{T} = [0.9, 1.1]$. The random number generator was set to rng(1,'twister') .The Girsanov weights to reweight the expectation have been calculated by (4.8).

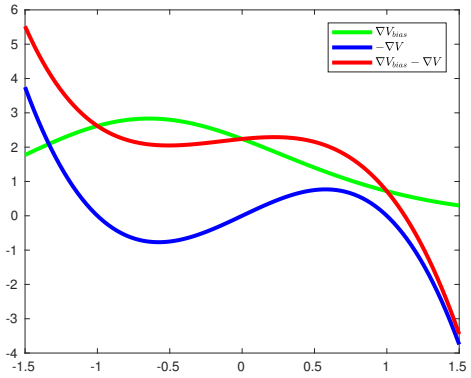


Fig. 4.5: The blue curve shows the negative gradient of the original potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics applied directly on the gradient and in red the corresponding negative biased gradient is shown.

	MC	GIR
$P(\mathcal{A})$	4.4×10^{-2}	4.5379×10^{-2}
Var	4.2106×10^{-2}	6.1682×10^{-3}
$R(I)$	4.6635	1.7306
$\mathbb{E}[\exp(-\beta\tau)]$	2.5320×10^{-3}	2.5315×10^{-3}
Var	2.2788×10^{-4}	3.44×10^{-6}
$R(I)$	5.9617	0.7327
MST	1.4801	0.84695

Tab. 4.5: Comparison of the importance sampling estimators and the Monte Carlo estimators for the simulation with the standard Girsanov formula where a biasing force is calculated.

We can also see a very good agreement of the MC estimator and the importance sampling estimator for both quantities of interest. Furthermore, a variance reduction could be achieved for both estimators. The variance for the transition probability is

reduced by 85% and the variance for the moment generating function is reduced by 98%. These examples show that the direct application of Metadynamics on the force can be used for building good biasing forces. Again a reduction of the MST could be achieved as well.
Comparing the results with the results from the biasing potential approach one sees that the estimators have lower variance. The results can be interpreted as the direct biasing on the force.

Example 6: Estimated parameter example
In this example the parameters for the constructed bias have been estimated from the sampling. The parameter λ_i is calculated from the history of the trajectory as described in Remark 3 with the constant $C = 10$. The centre of the bias function was calculated by the average over the last 200 steps of the trajectory. The random number generator was set to rng(1,'twister'). All other parameters are kept as in the previous example ($\beta = 3.0$). The Girsanov weights to reweight the expectation have been calculated by (4.8).

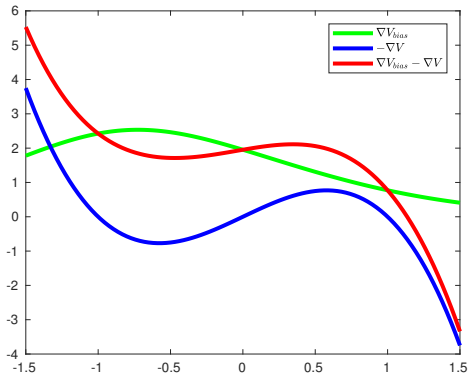


Fig. 4.6: The blue curve shows the negative gradient of the original potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics directly applied on the force with estimated parameters and in red the corresponding biased negative gradient is shown.

	MC	GIR
$P(\mathcal{A})$	4.8470×10^{-2}	4.7408×10^{-2}
Var	4.6121×10^{-2}	6.9288×10^{-3}
$R(I)$	4.4307	1.7557
$\mathbb{E}[\exp(-\beta\tau)]$	2.569×10^{-3}	2.5278×10^{-3}
Var	2.5850×10^{-4}	4.41×10^{-6}
$R(I)$	6.2561	0.8312
MST	1.4804	0.9614

Tab. 4.6: Comparison of the importance sampling estimators and the Monte Carlo estimators for the simulation with the standard Girsanov formula.

This example shows a very good agreement of the MC estimator and the importance sampling estimator for both quantities of interest. So the estimation of the parameters for the biasing force work. The variance reduction could be achieved for both estimators. The variance for the transition probability is reduced by 85% and the variance for the moment generating function is reduced by 98%. These examples show that the parameters on the biasing force have an impact on the variance reduction. Comparing the results with the results from the biasing potential approach one sees that the estimators also have a lower variance. We also have a bigger reduction of the MST. So one sees that a carefully designed biasing function can lead to a significant variance reduction and a reduction of the sampling time.

Example 7: Reverse example

A naive application of the method on the reverse problem shows that the direct force biasing needs more information about the dynamical system. In the proposed method the bias is always positive. But a positive force will only push the trajectory into one direction. In order to be flexible the bias has to be adapted to the structure of the gradient. Let us consider the above example 4.4. First we have considered the problem of going from the left well into the right well. The negative gradient is negative in the metastable area. To decrease the influence of the potential on the trajectory the gradient has to be raised. Consider now the reverse problem going from the right to the left well. The negative gradient is positive in the metastable set. So a negative bias has to be used to lower the gradient in the metastable region. A naive application of the method with only positive bias will not be able to generate the necessary bias. The potential version of the method does not have such problems because the resulting force is positive and negative. So in order to apply this method directly on the gradient one needs additional information on the system.

To show that the force method with additional information can be used to build biases for the reverse problem we consider the reverse problem again. As in the previous example we test the methods for the transition and the mean first exit time for the reverse problem. For this example we define the metastable region $\mathcal{S} = [0, 1.5]$. We choose the starting point of the SDE (4.7) in the metastable region $x_0 = 1$ and fix $\beta = 3.0$ for all simulations. The stopping time is defined as the first hitting time of the target set $\mathcal{T} = [-0.9, -1.1]$. The random number generator was set to rng(3,'twister') in order to have a better comparison within the different simulations. The parameters of the bias potential are the same as in the previous example. For this computation we chose $w_i = -0.1,\ \lambda_i = 0.8$ for all bias functions. The c_i of every bias function is chosen as the current value of the trajectory when the new bias function is added. The Girsanov weights to reweight the expectation have been calculated by (4.8).

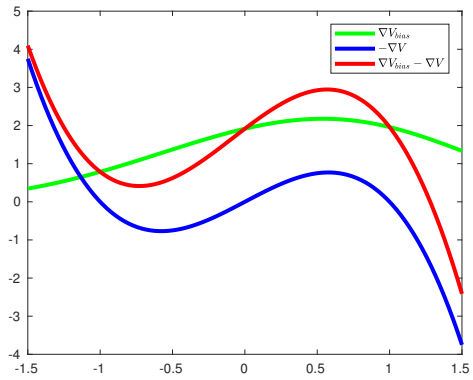


Fig. 4.7: The blue curve shows the negative gradient of the potential (4.18). The green curve shows the bias V_{bias} produced by Metadynamics applied directly on the force and in red the corresponding biased negative gradient is shown.

	MC	GIR
$P(\mathcal{A})$	4.6×10^{-2}	5.799×10^{-2}
Var	4.3927×10^{-2}	5.5×10^{-3}
$R(I)$	4.5563	1.2790
$\mathbb{E}[\exp(-\beta\tau)]$	2.4021×10^{-3}	3.0043×10^{-3}
Var	2.4258×10^{-4}	2.5×10^{-6}
$R(I)$	6.4837	0.52680
MST	1.4812	1.0007

Tab. 4.7: Comparison of the importance sampling estimators and the Monte Carlo estimators for the simulation with the standard Girsanov formula.

We see that the method with the additional information achieves the variance reduction. Moreover, the MC estimator and the importance sampling estimator agree very well. The variance reduction is in the same order of magnitude as in the previous examples. The variance for the transition probability is reduced by 87% and the variance for the moment generating function is reduced by 98%. The reduction of the MST could be achieved in this example.

4.5 Summary and Discussion

In this chapter we have developed an algorithm for the adaptive importance sampling of dynamical quantities in metastable dynamic systems. The main idea for this approach was to combine well-known MD enhanced sampling techniques with Girsanov's theorem. For the adaptive importance sampling the enhanced sampling techniques have been assimilated. We could show under which conditions the method can be applied and that the method preserves ergodicity. The method has been tested for many different applications. All of these different applications show

that the method achieves the variance reduction and the reduction of the sampling time. The examples also show that the bias depends on the additional sampling. So for different realizations of the additional sampling different biases are generated. Furthermore, one can see that variance reduction depends on the bias. Since in all different numerical examples a variance reduction has been achieved, it seems that the general framework is working. From the last example one can learn that a naive construction of the bias can also have a negative impact on the sampling. So in order to construct a good bias much information of the system has to be used. Similar results have been shown in [23].
We have seen that the method achieved a larger variance reduction if a larger β (a smaller temperature) is considered (see example 4.4.1). A possible explanation for this could be that we use a stationary bias. The constructed bias is generated by an additional realization of the dynamical system and is then used for the sampling of the quantity of interest. So the construction of the bias only considers one possible realization of many. It does not cover all possible movements of the trajectories. For a smaller β the trajectories are more affected by the diffusion constant and so the stationary bias does not reflect this. So in order to overcome this difficulty one could try to model a bias which reacts on some special behaviour. For example, one could supervise the sampling of the quantity of interest and if a certain behaviour (e.g. a movement in some direction) is shown, the bias is changed according to this. Also a more sophisticated way of building the bias could be used, e.g. reinforcement learning [43].
Comparing the variance reduction for the application of the bias on the potential and on the derivative of the potential we see that the application on the derivative achieves a larger variance reduction. A possible reason for this could be that the resulting bias is smoother compared to the application of the method on the potential. The construction of the bias by Metadynamics implies that the dynamical system is perturbed by derivative of the Gaussian functions. The derivatives have a negative and a positive part and thus the resulting bias is less smooth compared to the direct application of the method on the derivative. But we have also seen that the method needs additional information such that the perturbation is working accordingly. How the application of another algorithm to construct the bias like Adaptive Biasing Force impacts the variance of the estimator is a question for future research.

One of the main disadvantages of Adaptive Importance Sampling based on Metadynamics is the restriction on effective coordinates. Since Metadynamics only works in effective coordinates, our algorithm is also restricted to this. But other enhanced sampling techniques are also based on the assumption that the high-dimensional molecular system can be approximated by some low-dimensional model. Our algorithm has been applied in a two dimensional example; see [73]. Theoretically, the algorithm can also be applied in high-dimensional examples. But different problems arise in a high-dimensional set-up.

The first problem is how a bias can be built. For high-dimensional problems it is hard to understand which part of the potential causes the metastable behaviour. So the location of the metastable behaviour for high-dimensional problems is difficult. The accurate biasing of the potential is challenging on its own. Much information of the system has to be known a priori to force a certain behaviour of the dynamical system. From this perspective it does make sense to use the method which is working with an assimilated version of Metadynamics. Metadynamics has been applied for different high-dimensional problems very efficiently, see e.g. [87], and by using this method we can benefit from this experience.
Another problem is that the variance of the Girsanov weight scales with the dimension. It is well-known in the literature that the Radon-Nikodym is affected by the dimensionality of the problem [68] and this will definitely have an negative impact on the variance reduction. A possible solution for this could be a different way of sampling the Girsanov weight, e.g. by using the SDE formulation of the martingale [67] or considering a low-dimensional representation of the molecule, e.g. by effective coordinates. Owen states that one has to make sure that the variances decrease sharply if the dimension increases [68]. So by using a reduced model with only the relevant direction we can reduce the dimensionality of the model and thus reduce the variance in irrelevant directions.
The last critical point of the algorithm is the constructed bias. As we have seen in the many examples the bias highly depends on the additional trajectory. Furthermore, in many cases the constructed bias only destabilizes the metastability. As we have seen in the examples this works for different quantities of interest very well. For more sophisticated problems the constructed bias might be too naive. In importance sampling the bias has a significant impact on the variance reduction. There is also no guarantee that there is always a variance reduction and it is also possible that a bias designed in a wrong way can result in a variance increase. So for more complicated problems more information has to be introduced into the design of the bias. This could be done, for example, by combining the adaptive importance sampling method with the string method similar to [94].

From all of these observations the proposed algorithm seems to work best if combined with a proper dimension reduction technique in order to achieve a significant variance reduction.

5 Gradient estimators and non-parametric representation

In this chapter we are presenting approaches to optimize the bias in order to find the bias which gives the best variance reduction. On the one hand we have seen in the introduction that unbiasedness of the estimator does not depend on the bias. We have used this in the previous chapter to build an importance sampling scheme. On the other hand we have also seen that the bias has an impact on the variance reduction and that there theoretically exits a bias for which the variance of the estimator is zero. So instead of using any bias as done in the previous chapter we are optimizing the bias in this chapter. The main idea of this optimization approach is a formulation of finding the optimal bias as a optimization problem. In order to derive the optimization problem the bias is approximated by parametric ansatz functions. The approximation procedure can be seen as a Galerkin projection and has been proposed by C. Hartmann and Ch. Schütte in [38]. Based on this formulation different algorithmic methods have been suggested to solve the optimization problem, e.g. the gradient descent method [38] or the cross entropy method [93]. An overview on the current algorithmic developments can be found in [12].
In this chapter we are going to extend the proposed methods. For the stochastic gradient descent method we derive different gradient estimators. In the original paper [38] the authors derived a gradient estimator from the discretized problem. In the literature other approaches can be found which we are going to assimilate for the problem formulation presented in [38]. First we are going to derive a gradient estimator based on ideas presented in [27]. But the resulting formulas are only applicable if the bias can be represented with one ansatz function. This is why we derive two different gradient estimators based on ideas presented in [31]. The resulting formulas can be used to optimize an approximation with many ansatz functions. Numerical examples are presented to show different aspects of the method.
In the second part of the chapter we derive a non-parametric representation of the bias by kernelizing the Cross-Entropy method. This reformulation has the advantage that the control is expressed in terms of kernel functions and the resulting representation formula of the control is a non-parametric. Furthermore, the control depends on the data coming from the sampling. This approach can be seen as a data-driven approximation of the control.
First we give a brief introduction into the common approaches for gradient estimation

of expectations. Then we derive the optimization problem. After this we derive the one-dimensional gradient estimator based on the Malliavin gradient descent approach. Two other gradient estimators are derived by the likelihood approach. These gradient estimators are tested in a numerical example. In the last part of the chapter the non-parametric representation formula is calculated. Numerical tests of the non-parametric representation formula support our findings. A discussion closes the chapter.

5.1 Derivation of the optimization problem

Let us consider the SDE given by

$$dx_t = -\nabla V(x_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad x_0 = x \tag{5.1}$$

We are interested in quantities of interest related to these metastable sets. These quantities of interest $F(x) : \mathbb{R}^n \to \mathbb{R}$ are formulated as a moment generating function

$$F(x) = -\beta \log \mathbb{E}_{\mathbb{P}}\Big[\exp\Big(-\frac{1}{\beta}W(x_{0:\hat{\tau}})\Big)|x_0 = x\Big] \tag{5.2}$$

with

$$W(x_{0:\hat{\tau}}) = \int_0^{\hat{\tau}} f(x_s)ds \tag{5.3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a sufficiently smooth and bounded function. The quantity of interest depends on the whole path of the trajectory. For the special case in which f is the indicator function of the target set the function W is only the hitting time of this set. The quantity of interest is in this case the moment generating function of the stopping time. In the following $\mathbb{E}[\cdot|x_0 = x]$ is denoted by $\mathbb{E}^x[\cdot]$ To now overcome the metastability of the dynamics the drift of the dynamics is changed. In this way a perturbed dynamic system is generated which satisfies

$$dx_t^u = (u(x_t^u) - \nabla V(x_t^u))dt + \sqrt{2\beta^{-1}}dB_t, \quad x_0^u = x \tag{5.4}$$

where $u : \mathbb{R}^n \to \mathbb{R}^n$ has to be a measurable function which satisfies Novikov's condition; see proposition 3.
Then we can use Girsanov's theorem to reweight the quantity of interest and write

$$F(x) = -\beta \log \mathbb{E}_{\mathbb{Q}}^x\Big[\exp\Big(-\frac{1}{\beta}W(x_{0:\hat{\tau}}^u) + \int_0^{\hat{\tau}} \gamma(x_s^u)dB_s + \frac{1}{2}\int_0^{\hat{\tau}} |\gamma(x_s^u)|^2 ds\Big)\Big] \tag{5.5}$$

where $\sqrt{2\beta^{-1}}\gamma(x_s^u) = u(x_s^u)$. Due to the convexity of the negative logarithm we can apply Jensen's inequality to get

$$F(x) \leq \mathbb{E}_{\mathbb{Q}}^x\Big[W(x_{0:\hat{\tau}}^u) - \beta\int_0^{\hat{\tau}} \gamma(x_s^u)dB_s - \frac{\beta}{2}\int_0^{\hat{\tau}} |\gamma(x_s^u)|^2 ds\Big]. \tag{5.6}$$

Substituting the Brownian motion B_s now into a Brownian motion under the new probability measure $\mathbb{Q}$ by using $B_t^{\mathbb{Q}} = B_t + \int_0^{\hat{\tau}} |\gamma(x_s^u)|^2 ds$ we find

$$F(x) \leq \mathbb{E}_{\mathbb{Q}}^x \Big[W(x_{0:\hat{\tau}}^u) + \frac{\beta}{2} \int_0^{\hat{\tau}} |\gamma(x_s^u)|^2 ds - \beta \int_0^{\hat{\tau}} \gamma(x_s^u) dB_s^{\mathbb{Q}} \Big]. \tag{5.7}$$

Now using the linearity of the expectation we see that one of the terms is just an expectation over a stochastic integral. Since the expectation of a stochastic integral is zero, we can further simplify

$$F(x) \leq \mathbb{E}_{\mathbb{Q}}^x \Big[W(x_{0:\hat{\tau}}^u) + \frac{\beta}{2} \int_0^{\hat{\tau}} |\gamma(x_s^u)|^2 ds \Big]. \tag{5.8}$$

This is an upper bound for the quantity of interest. The best perturbation can now be found by minimizing the functional

$$F(x) = \inf_{\gamma} \{ \mathbb{E}_{\mathbb{Q}}^x \Big[W(x_{0:\hat{\tau}}^u) + \frac{\beta}{2} \int_0^{\hat{\tau}} |\gamma(x_s^u)|^2 ds \Big] \}. \tag{5.9}$$

Since there is a priori no information about the function space in which γ has to be in, this problem is not solvable numerically. In order to solve it numerically the change of drift is approximated by a sum over $N \in \mathbb{N}$ ansatz functions $u(x) \approx \sum_{i=1}^N a_i b_i(x)$ where $a_i \in \mathbb{R}$ and the $b_i : \mathbb{R}^n \to \mathbb{R}^n$ are bounded continuous functions similar to the approach presented in the previous chapter. The optimization problem can now be approximated by

$$F(x) \approx \min_{a \in \mathbb{R}^N} \{ \mathbb{E}_{\mathbb{Q}}^x \Big[W(x_{0:\hat{\tau}}^a) + \frac{\beta^2}{4} \int_0^{\hat{\tau}} | \sum_{i=1}^N a_i b_i(x_t^a)|^2 ds \Big] \} \tag{5.10}$$

where the dynamic x_t^a is given by

$$dx_t^a = (\sum_{i=1}^N a_i b_i(x_t^a) - \nabla V(x_t^a))dt + \sqrt{2\beta^{-1}} dB_t, \quad x_0^a = x. \tag{5.11}$$

On possible method to solve this stochastic optimization problem is a gradient descent. In the literature different approaches have been proposed how gradients of expectations can be computed efficiently. We summarize these approaches very briefly in the next section and give a short introduction into the gradient descent method.

Gradient descent

The gradient descent method is a first order iterative optimization method for solving optimization problems. In order to find a minimum of a function F steps along the

direction of the negative gradient of this function ($-\nabla F$) are taken (gradient decent methods). The new iterate of the gradient descent method x_{new} is calculated by

$$x_{new} = x_{old} - \alpha \nabla F(x_{old}); \quad n > 0$$

where x_{old} is the old iterate, α is a so-called step size and ∇F is the gradient of the function evaluated for the old iterate x_{old}. If the gradient cannot be calculated analytically, different approximation methods can be used; see e.g. [66]. Maximization problems can also be solved in a similar way by stepping into the direction of the positive gradient direction (gradient ascent method). The algorithm is iterated until some stopping condition, usually something like $\|\nabla F(x_{new})\|_2 \leq \epsilon$ for ϵ small, is satisfied.

Step size

For deterministic optimization problems different sophisticated step length algorithms following either the Wolf condition or the Goldstein condition are available [66]. These step length algorithms can help to speed up the convergence of the method and help to prevent zigzagging. But to our knowledge these algorithms cannot be applied for stochastic optimization problems.

In stochastic optimization the step size sequence which is often used is either a constant or satisfies

$$\sum_{n=0}^{\infty} \alpha_n = \infty, \ \alpha_n \geq 0, \ \alpha_n \to 0 \text{ for n} \geq 0.$$

With these step sizes and some further assumptions one can show convergence of the stochastic gradient descent method, e.g. [51].

Gradient estimators

If the objective function is an expectation, the derivative cannot be calculated analytically. We summarize different approaches how the approximation of derivatives of expectations can be calculated.

1. The *finite difference approach* or *resampling approach*, e.g. [3]. In this approach the derivative of the expectations is approximated by a finite difference

$$\nabla_a \mathbb{E}[f(x^a_{0:T})] \approx \frac{\mathbb{E}[f(x^{a+\epsilon}_{0:T})] - \mathbb{E}[f(x^a_{0:T})]}{\epsilon}$$

for a small ϵ. The resulting estimator is a biased estimator [3]. Moreover, this approach is extremely costly because the process x has to be sampled for every perturbation of the parameter sufficiently often. The approach gets even worse if the dimension of the parameter is large. There are several variants of this finite difference approach to overcome the resampling, for example Simultane-

ous Perturbation Stochastic Approximation (SPSA) introduced by Spall [81]. Here all entries of the parameter space are perturbed at once. This reduces the resampling effort for a high-dimensional parameter space. To reduce the variance of the estimator different numerical extensions were proposed. For example, common random numbers make this estimator extremely efficient in terms of variance [46].

2. The *pathwise approach* proposed by Yang and Kushner [92]. The idea of this approach is to interchange the expectation and the derivative. The resulting expectation then involves $\nabla_a f$ and $\nabla_a x_{0:T}$. This expression can be again written as an expectation and evaluated by Monte Carlo methods. The only restriction for this method is that f has to be a smooth function to calculate $\nabla_a f$. The derivative of the process with respect to the parameter $\nabla_a x_{0:T}$ can be calculated explicitly.

3. The *likelihood approach* or *score method* introduced by [35] [34] and [75]. This approach uses the fact that the gradient can be written as $\mathbb{E}[f(x_{0:T})H]$ with some random variable H. This representation is not unique since any random variable H which is orthogonal to $x_{0:T}$ can be added cf. [36]. Normally, H is equal to the gradient of the log-likelihood ($\nabla_a \log(p(a, x_{0:T}))$) with $p(a, \cdot)$ being the density with respect to the Lebesques measure of the law of x_T. But in many applications the density is not known. If the parameter is only in the drift term and the diffusion term is elliptic, it is possible to calculate H explicitly using Girsanov's theorem cf. [92]. More general situations in which the parameter is also in the diffusion term have been investigated by Gobet and Munos [36].

5.2 Malliavin gradient descent one-dimensional

In the following section a one-dimensional derivative estimator is derived. This estimator is motivated by [27]. In the paper O. Ewald introduces a method which is called Malliavin gradient descent. The main idea of the method is to combine gradient descent methods with Malliavin Calculus in order to solve stochastic optimization problems. The method is used for the calibration of stochastic systems. In this paragraph we are going to apply this approach to the parametric optimization problem. The main advantage of the Malliavin gradient descent is that the objective function does not have to be continuous at all. Even functions which are discontinuous or have singularities can be considered. This is, for example, the case if one is interested in the sampling hitting times. If this estimator can be generalized to a high-dimensional application, is not quite clear yet. This is why we only present the derivation without any further numerical testing.

Before deriving the gradient formula we will summarize the presented theory. At the end of this paragraph we apply this theory to the parametric optimization problem.

Let us consider the probability space $(\Omega, \mathbb{P}, \mathcal{F})$ with a Brownian motion (B_t). Furthermore, we consider the following SDE

$$dx_t = \beta(x_t, u)dt + \sigma(x_t, v)dB_t \tag{5.12}$$

where u respectively v are parameters from some subset $U \in \mathbb{R}$ and $V \in \mathbb{R}$. The functions $\beta : \mathbb{R} \times U \to \mathbb{R}$ and $\sigma : \mathbb{R} \times V \to \mathbb{R}$ are two times continuously differentiable with bounded derivatives up to order two. Then, it follows that there exists a family $\{(x_t(x, u, v))|x \in \mathbb{R}, u \in U, v \in V\}$ of stochastic processes such that the process $(x_t(x, u, v))$ satisfies (5.12) for any choice of x, u, v with initial condition $x_0(x, u, v) = x$ $\mathbb{P}$-almost sure and for $\mathbb{P}$-almost any $\omega \in \Omega$ and any time t the map

$$(x, u, v) \to x_t(x, u, v)(\omega) \tag{5.13}$$

is continuous differentiable [72]. In general, the condition on the bounded derivatives can be relaxed but then the family $(x_t(x, u, v))$ might only exist until an explosion time.
For simplification we drop the explicit dependence of x_t on (x, u, v) but keep it in the drift term β and the diffusion term σ. The derivatives are denoted as $\frac{\partial}{\partial x}x_t$, $\frac{\partial}{\partial u}x_t$ and $\frac{\partial}{\partial v}x_t$. The following proposition shows how the different derivatives can be calculated.

Proposition 3. *Assume that $\beta : \mathbb{R} \times U \to \mathbb{R}$ and $\sigma : \mathbb{R} \times V \to \mathbb{R}$ are two times continuously differentiable with bounded derivatives up to order two and x_t satisfies the SDE given by* (5.12)*. Then the derivatives with respect to the different parameters are given by*

$$\begin{aligned}
\frac{\partial}{\partial x}x_t &= \exp\left(\int_0^t \frac{\partial}{\partial x}\beta(x_s, u)ds\right)\exp\left(\int_0^t \frac{\partial}{\partial x}\sigma(x_s, v)dB_s - \frac{1}{2}\frac{\partial}{\partial x}\sigma(x_s, v)^2 ds\right)\\
\frac{\partial}{\partial u}x_t &= \left(\frac{\partial}{\partial x}x_t\right)\left(\int_0^t \left(\frac{\partial}{\partial x}x_t\right)^{-1}\frac{\partial}{\partial u}\beta(x_s, u)ds\right)\\
\frac{\partial}{\partial v}x_t &= \left(\frac{\partial}{\partial x}x_t\right)\left(\int_0^t \left(\frac{\partial}{\partial x}x_t\right)^{-1}\frac{\partial}{\partial v}\sigma(x_s, v)dB_s\right)
\end{aligned}$$

$$- \int_0^t \left(\frac{\partial}{\partial x} x_t \right)^{-1} \frac{\partial}{\partial x} \sigma(x_t, v) \frac{\partial}{\partial v} \sigma(x_t, v) ds \Bigg)$$

Proof. A short version of the proof can be found in [27]. A more general proof of this proposition can be found in [72]. The proposition is only a special case of the presented theorem in [72]. □

With this formula for the derivatives we can now calculate the derivative of an expectation with respect to the different parameters.

Proposition 4. *Given a measurable function $h : \mathbb{R} \to \mathbb{R}_+$, a time $\hat{\tau}$, $\sigma > 0$ and assume that $h(x_{\hat{\tau}}) \in L(\Omega)$. Consider the function in u and v*

$$(u, v) \to \mathbb{E}[h(x_t(x, u, v))]. \tag{5.14}$$

Define the following weight functions

$$\begin{aligned}
\mathcal{X}_t &:= \int_0^t \sigma(x_s, v)^{-1} \frac{\partial}{\partial x} x_s dB_s \\
\mathcal{U}_t &:= \int_0^t \left(\frac{\partial}{\partial x} x_s \right)^{-1} \frac{\partial}{\partial u} \beta(x_s, u) ds \\
\mathcal{V}_t &:= \int_0^t \left(\frac{\partial}{\partial x} x_s \right)^{-1} \frac{\partial}{\partial v} \sigma(x_s, v) dB_s - \int_0^t \left(\frac{\partial}{\partial x} x_s \right)^{-1} \frac{\partial}{\partial x} \sigma(x_s, v) \frac{\partial}{\partial v} \sigma(x_s, v) ds \\
\mathcal{S}_t &:= \sigma(x_s, v)^{-1} \frac{\partial}{\partial x} x_t
\end{aligned}$$

If $h \in L^2_{loc}(\mathbb{R})$ with at most linear growth at infinity, then the following formulas hold

$$\begin{aligned}
\frac{\partial}{\partial u} \mathbb{E}[h(x_{\hat{\tau}})] &= \mathbb{E}\left[\frac{1}{\hat{\tau}} h(x_{\hat{\tau}}) \mathcal{X}_{\hat{\tau}} \mathcal{U}_{\hat{\tau}} \right] - \mathbb{E}\left[\frac{1}{\hat{\tau}} h(x_{\hat{\tau}}) \int_0^{\hat{\tau}} (D_t \mathcal{U}_{\hat{\tau}}) \mathcal{S}_t dt \right] \\
\frac{\partial}{\partial v} \mathbb{E}[h(x_{\hat{\tau}})] &= \mathbb{E}\left[\frac{1}{\hat{\tau}} h(x_{\hat{\tau}}) \mathcal{X}_{\hat{\tau}} \mathcal{V}_{\hat{\tau}} \right] - \mathbb{E}\left[\frac{1}{\hat{\tau}} h(x_{\hat{\tau}}) \int_0^{\hat{\tau}} (D_t \mathcal{V}_{\hat{\tau}}) \mathcal{S}_t dt \right]
\end{aligned}$$

where $D_t \mathcal{U}_{\hat{\tau}}$ and $D_t \mathcal{V}_{\hat{\tau}}$ are the Malliavin derivatives.

Proof. A proof of this proposition can be found in [27]. □

Application

In the following we are going to apply the presented theory for the stochastic optimization problem (5.10). In the following we are going to denote the controlled SDE by

$$dx_t^a = (ab(x_t^a) - \frac{\partial}{\partial x} V(x_t^a))dt + \sqrt{2\beta^{-1}} dB_t. \tag{5.15}$$

The function V is, as before, a metastable potential. The function $ab(x_t^a)$ is the parametric approximation of the optimal control. We want to calculate the derivative of the expectation with respect to the parameters of the control. To use the same notation as in the presented theory we choose $\beta(x,a) = (ab(x_t^a) - \frac{\partial}{\partial x}V(x_t^a))$ and $\sigma = \sqrt{2\beta^{-1}}$. The derivative are given by

$$\frac{\partial}{\partial x}\beta(x,a) = a\frac{\partial}{\partial x}b(x_t^a) - \frac{\partial}{\partial xx}V(x), \quad \frac{\partial}{\partial a}\beta(x,a) = b(x), \quad \frac{\partial}{\partial x}\sigma = 0, \quad \frac{\partial}{\partial v}\sigma = 0. \tag{5.16}$$

Due to proposition 3 $\frac{\partial}{\partial x}x_t^a$ is given by

$$\frac{\partial}{\partial x}x_t^a = \exp\Big(\int_0^t a\frac{\partial}{\partial x}b(x_t^a)ds\Big). \tag{5.17}$$

Since we are only interested in the derivative with respect to a, we are going to use $\mathcal{A}_t$ instead of $\mathcal{U}_t$. I also drop the terms which are related to the derivative with respect to v. The other stochastic processes which are needed to calculate the derivative of the expectation are due to Proposition 4

$$\begin{aligned}
\mathcal{X}_t &:= \int_0^t \frac{1}{\sqrt{2\beta^{-1}}}\exp\Big(\int_0^t a\frac{\partial}{\partial x}b(x_t^a)ds\Big)dB_s \\
\mathcal{A}_t &:= \int_0^t \exp\Big(-\int_0^t a\frac{\partial}{\partial x}b(x_t^a)ds\Big)f(x_s^a)ds \\
\mathcal{S}_t &:= \frac{1}{\sqrt{2\beta^{-1}}}\exp\Big(\int_0^t a\frac{\partial}{\partial x}b(x_t^a)ds\Big).
\end{aligned}$$

The Malliavin derivative of $\mathcal{A}_t$ which is denoted by $D_t\mathcal{A}_{\hat{\tau}}$ is given by

$$\begin{aligned}
D_t\mathcal{A}_{\hat{\tau}} &= D_t\int_0^{\hat{\tau}}\exp\Big(-\int_0^t a\frac{\partial}{\partial x}b(x_s^a)ds\Big)f(x_s^a)ds \\
&= \int_t^{\hat{\tau}} D_t[\exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_s^a)dq\Big)b(x_s^a)]ds \\
&= \int_t^{\hat{\tau}} D_t(-\int_0^s a\frac{\partial}{\partial x}b(x_q^a)dq)\exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_q^a)dq\Big)b(x_s^a) \\
&+ \exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_s^a)dq\Big)D_tb(x_s^a)ds \\
&= \int_t^{\hat{\tau}} (-\int_t^s D_ta\frac{\partial}{\partial x}b(x_q^a)dq)\exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_q^a)dq\Big)b(x_s^a) \\
&+ \exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_q^a)dq\Big)\frac{\partial}{\partial x}b(x_s^a)\frac{\partial}{\partial x}x_s^a\Big(\frac{\partial}{\partial x}x_s^a\Big)^{-1}\frac{1}{\sqrt{2\beta^{-1}}}ds \\
&= \int_t^{\hat{\tau}} (-\int_t^s \frac{\partial}{\partial x}a\frac{\partial}{\partial x}b(x_q^a)\frac{\partial}{\partial x}x_t^a\Big(\frac{\partial}{\partial x}x_s^a\Big)^{-1}dq)\frac{1}{\sqrt{2\beta^{-1}}}\exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_q^a)dq\Big)b(x_s^a)
\end{aligned}$$

$$+\exp\Big(-\int_0^s a\frac{\partial}{\partial x}b(x_t^a)dq\Big)\frac{\partial}{\partial x}b(x_s^a)\frac{\partial}{\partial x}x_s^a\Big(\frac{\partial}{\partial x}x_s^a\Big)^{-1}\frac{1}{\sqrt{2\beta^{-1}}}ds$$

where $\frac{\partial}{\partial x}x_t^a\Big(\frac{\partial}{\partial x}x_s^a\Big)^{-1}$ is given by

$$\frac{\partial}{\partial x}x_t^a\Big(\frac{\partial}{\partial x}x_s^a\Big)^{-1}=\exp\Big(\int_s^t a\frac{\partial}{\partial x}b(x_q^a)dq\Big).$$

The derivative is then given by

$$\begin{aligned}\frac{\partial}{\partial a}\mathbb{E}[\phi(x_t^a,a)] = \mathbb{E}[&\frac{1}{\hat{\tau}}\phi(x_{\hat{\tau}}^a,a)\Big(\int_0^t\frac{1}{\sqrt{2\beta^{-1}}}\exp\Big(\int_0^t a\frac{\partial}{\partial x}b(x_s^a)ds\Big)dB_s\Big)\\&\Big(\int_0^t\exp\Big(-\int_0^t a\frac{\partial}{\partial x}b(x_s^a)ds\Big)b(x_s^a)ds\Big)]\\&-\mathbb{E}[\frac{1}{\hat{\tau}}\phi(x_{\hat{\tau}}^a,a)\int_0^{\hat{\tau}}D_tA_{\hat{\tau}}\frac{1}{\sqrt{2\beta^{-1}}}\exp\Big(\int_0^{\hat{\tau}}a\frac{\partial}{\partial x}b(x_s^a)ds\Big)ds].\end{aligned}$$

This gradient estimator does not require the differentiability of ϕ. This is why it can be extremely useful to solve optimal control problems with non-continuous cost functions. In this case the PDE formulation cannot be solved because there is no valid formulation of the problem. But still the gradient estimator is quite complex due to the nested integrals which could limit the application. In the next section we are going to calculate a different gradient estimator which is easier to implement.

5.3 Likelihood approach to parametric optimal control

In this section we present the likelihood approach to estimate the gradient of the objective function. The main idea of this section is motivated by [31]. The authors of this paper used Malliavin Calculus to calculate derivatives for objective functions which can again be formulated as expectations. In this paragraph we adapted this approach to the special objective function of the optimization problem stated above.

In order to perform a gradient descent we have to calculate the partial derivatives with respect to the parameter a_i of the expectation

$$\varphi^a(x)=\mathbb{E}_{\mathbb{Q}}[\phi(x_{0:\hat{\tau}}^a,a)|x_0^a=x]. \tag{5.18}$$

In the following we assume that ϕ is a path functional depending on the whole path of the process and a continuous differential function in the second argument. Furthermore, we assume that

$$\mathbb{E}_{\mathbb{Q}}^x[\phi(x_{0:\hat{\tau}}^a,\xi)^2]<\infty \tag{5.19}$$

where the controlled dynamical system is given as 5.11 and ξ is the maximum in the second argument.The following result gives an expression for the partial derivative.

Theorem 6. *Assuming that ϕ is a continuous differentiable function for which the boundedness condition as given in* (5.19) *holds and a finite time $T < \infty$ (or a finite stopping time) then the partial derivative of the expectation is given by*

$$\frac{\partial}{\partial a_i}\varphi^a(x) = \mathbb{E}^x_{\mathbb{Q}}\Big[\frac{\partial \phi}{\partial a_i}(x^a_{0:\hat{\tau}}, a) + \phi(x^a_{0:\hat{\tau}}, a)\int_0^{\hat{\tau}} \frac{\partial}{\partial a_i}\sum_{i=1}^{N} a_i b_i(x^a_t) dB_t\Big]. \tag{5.20}$$

Proof. To show that the above result is a derivative we show that (5.20) converges against the finite difference. For this we define an auxiliary process x^ϵ_t

$$dx^\epsilon_t = (\sum_{i=1}^{N}(a_i + \epsilon_i)b_i(x^\epsilon_t) - \nabla V(x^\epsilon_t))dt + \sqrt{2\beta^{-1}}dB_t, \quad x^\epsilon_0 = x \tag{5.21}$$

where ϵ is a small perturbation in one of the parameters ($\epsilon = [0, 0, \dots \epsilon_i, \dots, 0] \in \mathbb{R}^N$). The probability measure which is introduced by the SDE (5.11) will be denoted by $\mathbb{Q}$ and the probability measure introduced by (5.21) will be denoted by $\mathbb{Q}^\epsilon$. Let us assume that the two probability measures are absolutely continuous which respect to each other. Now applying Girsanov's theorem a new random variable can be defined

$$M_{\hat{\tau}} = \exp\Big(-\epsilon_i \int_0^{\hat{\tau}} \frac{b_i(x^a_t)}{\sqrt{2\beta^{-1}}} dB_t - \frac{\epsilon_i^2}{2}\int_0^{\hat{\tau}} \Big|\frac{b_i(x^a_t)}{\sqrt{2\beta^{-1}}}\Big|^2 dt\Big).$$

Due to the boundedness of the function b_i we have that $\mathbb{E}_{\mathbb{Q}}[M_{\hat{\tau}}] = 1$ for any $\epsilon_i \geq 0$ because Novikov's condition is satisfied. By assumption the two considered measures are absolutely continuous. This is why the objective function can be rewritten as

$$\varphi^{a+\epsilon}(x) = \mathbb{E}_{\mathbb{Q}^\epsilon}[\bar{M}_{\hat{\tau}}\phi(x^\epsilon_{0:\hat{\tau}}, a)|x^\epsilon = x], \tag{5.22}$$

where

$$\bar{M}_{\hat{\tau}} = \exp\Big(-\epsilon_1 \int_0^{\hat{\tau}} \frac{(b_i(x^\epsilon_t))}{\sqrt{2\beta^{-1}}} dB^\epsilon_t - \frac{\epsilon_i^2}{2}\int_0^{\hat{\tau}} \Big|\frac{(b_i(x^\epsilon_t))}{\sqrt{2\beta^{-1}}}\Big|^2 dt\Big). \tag{5.23}$$

The Brownian motion B^ϵ_t is a Brownian motion under the measure $\mathbb{Q}^\epsilon$ defined by $B^\epsilon_t = B_t + \epsilon_i \int_0^t \frac{(b_i(x_t{}^\epsilon)}{\sqrt{2\beta^{-1}}} dt$. The joint distribution of $(x^\epsilon_\cdot, B^\epsilon_\cdot)$ under $\mathbb{Q}^\epsilon$ coincides with the joint distribution of $(x_\cdot, B_\cdot)$ under $\mathbb{Q}$. So the objective function can be expressed as

$$\varphi^{a+\epsilon}(x) = \mathbb{E}^x_{\mathbb{Q}}[M_{\hat{\tau}}\phi(x^\epsilon_{0:\hat{\tau}}, a)]. \tag{5.24}$$

Using Itô's formula $M_{\hat{\tau}}$ can be expressed as

$$M_{\hat{\tau}} = 1 + \int_0^{\hat{\tau}} M_t \frac{(\epsilon_i b_i(x^a_t))}{\sqrt{2\beta^{-1}}} dB_t. \tag{5.25}$$

Rearranging and dividing both sides by ϵ_i one gets

$$\frac{1}{\epsilon_i}(M_{\hat{\tau}}-1)=\int_0^{\hat{\tau}} M_t\frac{(b_i(x_t^a))}{\sqrt{2\beta^{-1}}}dB_t. \tag{5.26}$$

Since M_T is a martingale, it is bounded and so by the dominated convergence theorem we get for $\epsilon_i\to 0$

$$\lim_{\epsilon_i\to 0}\frac{1}{\epsilon_i}(M_{\hat{\tau}}-1)=\int_0^{\hat{\tau}}\frac{b_i(x_t^a)}{\sqrt{2\beta^{-1}}}dB_t=\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}}\frac{\partial}{\partial a_i}\sum_{i=1}^N a_ib_i(x_t^a)dB_t \tag{5.27}$$

in L^2. Therefore we can calculate

$$\begin{aligned}
&\left|\frac{1}{\epsilon}(\varphi^{a+\epsilon}(x)-\varphi^a(x))-\frac{\partial}{\partial a_i}\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a)\right]\right|\\
=&\left|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}^\epsilon}^x\left[\phi(x_{0:\hat{\tau}}^{a+\epsilon},a+\epsilon)\right]-\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a)\right])-\frac{\partial}{\partial a_i}\mathbb{E}_{\mathbb{Q}}^x[\phi(x_{0:\hat{\tau}}^a,a)]\right|\\
=&\left|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a+\epsilon)M_{\hat{\tau}}\right]-\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a)\right])-\frac{\partial}{\partial a_i}\mathbb{E}_{\mathbb{Q}}\left[\phi(x_{0:\hat{\tau}}^a,a)\right]\right|\\
=&\Bigg|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\Big[\phi(x_{0:\hat{\tau}}^a,a+\epsilon)M_{\hat{\tau}}-\phi(x_{0:\hat{\tau}}^a,a+\epsilon)\\
&+\phi(x_{0:\hat{\tau}}^a,a+\epsilon)-\phi(x_{0:\hat{\tau}}^a,a)\Big])-\frac{\partial}{\partial a_i}\mathbb{E}_{\mathbb{Q}}\left[\phi(x_{0:\hat{\tau}}^a,a)\right]\Bigg|\\
=&\left|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\left[(\phi(x_{0:\hat{\tau}}^a,a+\epsilon)-\phi(x_{0:\hat{\tau}}^a,a))+\phi(x_\cdot^a,a+\epsilon)(M_{\hat{\tau}}-1)\right])-\frac{\partial}{\partial a_i}\mathbb{E}_{\mathbb{Q}}[\phi(x_{0:\hat{\tau}}^a,a)\right]\right|\\
=&\Bigg|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\left[(\phi(x_{0:\hat{\tau}}^a,a+\epsilon)-\phi(x_{0:\hat{\tau}}^a,a))-\frac{\partial\phi}{\partial a_i}(x_{0:\hat{\tau}}^a,a)\right]\\
&+\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a+\epsilon)(M_{\hat{\tau}}-1)-\phi(x_{0:\hat{\tau}}^a,a)\int_0^{\hat{\tau}}\frac{\partial}{\partial a_i}\sum_{i=1}^N a_ib_i(x_t^a)dB_t\right]\Bigg|\\
\leq&\left|\mathbb{E}_{\mathbb{Q}}^x\left[(\frac{\phi(x_{0:\hat{\tau}}^a,a+\epsilon)-\phi(x_{0:\hat{\tau}}^a,a)}{\epsilon})-\frac{\partial\phi}{\partial a_i}(x_{0:\hat{\tau}}^a,a)\right]\right|\\
&+\left|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a+\epsilon)(M_{\hat{\tau}}-1)-\phi(x_{0:\hat{\tau}}^a,a)\int_0^{\hat{\tau}}\frac{\partial}{\partial a_i}\sum_{i=1}^N a_ib_i(x_t^a)dB_t\right]\right|.
\end{aligned}$$

The first part of the inequality converges to zero because ϕ is a square integrable function cf. [47]. Applying the Cauchy-Schwarz inequality we find

$$\begin{aligned}
&\left|\frac{1}{\epsilon}(\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,a+\epsilon)(M_{\hat{\tau}}-1)-\phi(x_{0:\hat{\tau}}^a,a)\int_0^{\hat{\tau}}\frac{\partial}{\partial a_i}\sum_{i=1}^N a_ib_i(x_t^a)dB_t\right]\right|\\
\leq&\mathbb{E}_{\mathbb{Q}}^x\left[\phi(x_{0:\hat{\tau}}^a,\xi)^2\right]\mathbb{E}_{\mathbb{Q}}^x\left[\left(\frac{1}{\epsilon}(M_{\hat{\tau}}-1)-\int_0^{\hat{\tau}}\frac{\partial}{\partial a_i}\sum_{i=1}^N a_ib_i(x_t^a)dB_t\right)^2\right].
\end{aligned}$$

By (5.27) the second term converges to zero and this gives the required result. □

The resulting expression for the gradient of the expectation is still an expectation. This is very useful in the sampling context because it can be approximated by a

standard Monte Carlo procedure. In the next paragraph we derive a second gradient estimator based on some assumptions on the biasing potential.

5.3.1 Gradient estimator of the alternative Girsanov formula

As we have already seen in the derivation of the zero variance property there exists another representation formula of the Girsanov weight if the biasing potential is of gradient form. This different representation formula can also be used for deriving another gradient estimator.

As we have seen in the previous calculation the gradient descent includes a stochastic integral

$$\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} \frac{\partial}{\partial a_i}\sum_{i=1}^{N} a_i b_i(x_t^a)dB_s. \tag{5.28}$$

Let us now consider that the bias function is of gradient form such that the perturbed SDE can be written as

$$dx_s^a = -(\nabla V(x_s^a) - \nabla V_{bias}(x_s^a))ds + \sqrt{2\beta^{-1}}dB_s \tag{5.29}$$

where $\nabla V_{bias}(x) = \sum_{i=1}^{N} a_i \nabla b_i(x)$ where $b_i(x) : \mathbb{R}^n \to \mathbb{R}$. One can derive an alternative gradient formula using Itô calculus similar to the alternative reweighting formula shown in Chapter 2. Applying Itô calculus one gets

$$\begin{aligned} V_{bias}(x_t^a) - V_{bias}(x_0^a) = &\int_0^t (-\nabla V \cdot \nabla V_{bias} + |\nabla V_{bias}|^2 + \beta^{-1}\nabla^2 V_{bias})(x_s^a)ds \\ &+ \sqrt{2\beta^{-1}}\int_0^t \nabla V_{bias}(x_s^a)dB_s. \end{aligned}$$

Using again the parametric basis functions as approximation of the control we need the integral $V_{bias}(x) = \sum_i a_i \int b_i(x)$ and the derivative $\nabla^2 V(x) = \sum_i a_i \nabla b_i(x)$ to use alternative representation formula. Now after rearranging terms one finds

$$\begin{aligned} \frac{\partial}{\partial a_i}\varphi^{\epsilon}(x) = \mathbb{E}\Big[\frac{\partial \phi}{\partial a_i}(x_\cdot^a, a) + \phi(x_\cdot^a, a)\frac{1}{\sqrt{2\beta^{-1}}}\Big[\int_0^{\hat{\tau}} \frac{\partial}{\partial a_i}(\nabla V \cdot \nabla V_{bias} \\ - |\nabla V_{bias}|^2 - \beta^{-1}\nabla^2 V_{bias})(x_s^a)ds + \frac{\partial \phi}{\partial a_i}\Big(V_{bias}(x_{\hat{\tau}}^a) - V_{bias}(x_0^a)\Big)\Big]\Big]. \end{aligned}$$

Since the alternative expression does not include a stochastic integral one could suppose that the alternative gradient estimator has a lower variance. We are going to investigate this in our numerical examples.

5.3.2 Examples

We tested the different gradient estimators for a one-dimensional bistable system. Let us consider the controlled dynamics

$$dx_s^u = (-\frac{\partial}{\partial x}V(x_s) + u(x_s))ds + \sqrt{2\beta^{-1}}dB_s \quad x_0 = x. \tag{5.30}$$

The asymmetric bistable potential satisfies

$$V(x) = (x^2 - 1)^2 - 0.2x + 0.3.$$

A visualization can be found in figure 5.1(a). This potential has two minima at $x_0 \approx -0.973994$ and $x_2 \approx 1.02412$ while the right well is deeper than the left well. Furthermore, the potential has a local maximum at $x_1 \approx -0.0501259$. We define the stopping time

$$\hat{\tau} = \inf\{s > 0 : |x_s - x_2| \leq 0.1\}. \tag{5.31}$$

The goal of the considered control problem is to effectively sample the transition into the metastable set around x_2 starting in the metastable set x_0. For this the objective function is expressed as: Minimize

$$\mathbb{E}_{\mathbb{Q}}^x[\hat{\tau} + \int_0^{\hat{\tau}} |u(x_s)|^2 ds]. \tag{5.32}$$

This expectation can be understood as a regularized expectation for the stopping time. The stopping time should be reduced without controlling the dynamical system too much.

For all simulations the inverse temperature was set to $\beta = 4$. The initial condition is chosen to be $x_0 = 1$ for all realizations. The trajectories are sampled by using a Euler-Maruyama scheme with a time discretization $dt = 0.001$.

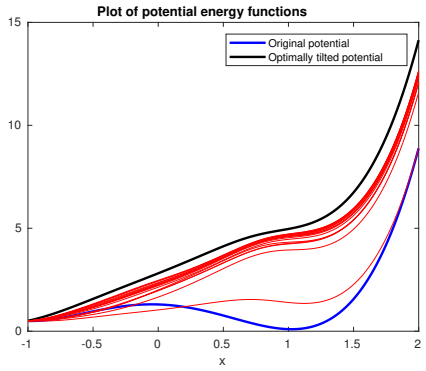
Representation and optimization of control
The control is represented by 20 normalized Gaussians of the form

$$u(x) = \sum_{k=1}^{20} \frac{a_k}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - m_k)^2}{2\sigma^2}\right) \tag{5.33}$$

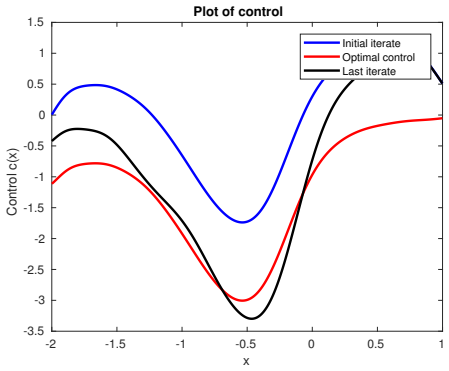
are used with $\sigma = 0.1579$. The basis functions are place along the path of the trajectory $m_k = -1.1 + (k-1)\sigma \quad k = 1\ldots, 20$. The chosen representation is independent of time because the time dependence of the optimal control is relatively weak for this problem and thus left out for simplification.

Stochastic gradient
In this example we test the stochastic gradient descent (5.20). The gradient of

the objective function was approximated by a standard Monte Carlo estimator. To evaluate the gradient estimator 200 trajectories have been calculated for each evaluation. The decreasing step size is chosen to be $0.4/(\text{iteration}+10)$. In total 15 optimization steps were calculated. As a starting point for the optimization we used an perturbed approximation of the optimal solution calculated by the related PDE.



(a) In blue the original potential is shown. In black the optimal tilted potential is shown calculated by the PDE formulation of the problem. In red the result after every gradient descent step is shown.



(b) In blue the starting control is shown. In red the optimal control calculated by the PDE formulation is shown. In black the last iterate after 15 optimization steps is shown.

Fig. 5.1: Resulting potential and control for a gradient descent done with the stochastic gradient estimator with 200 trajectories.

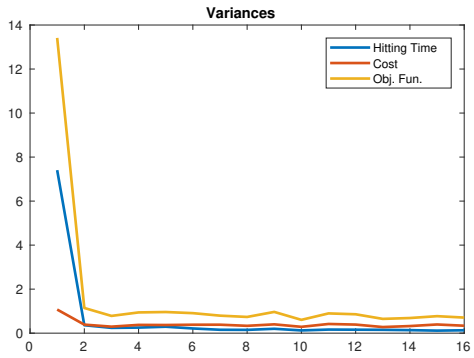


Fig. 5.2: The different variances of the objective function (yellow), hitting time (blue), cost function (orange) are shown for every evaluation of the gradient descent method.

The example shows that this gradient estimator works. The controlled potential is converging into the direction of the optimal tilted potential. Furthermore, the control of the last iterate shows a quite good fit of the optimal control in the relevant region. The variance of the objective function decreases mostly in the first 3 optimization steps. After 12 steps the best variance is reached and afterwards the variance is slightly decreasing. This can be a sign of zigzagging behaviour of the gradient methods.

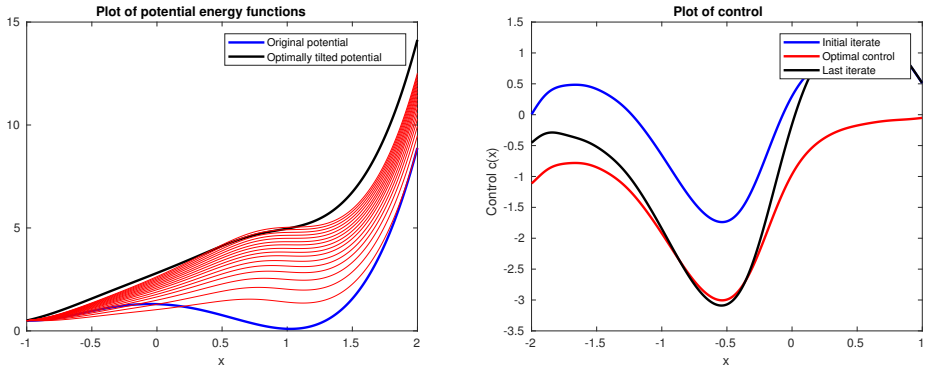
In different numerical examples (not shown here) one could observe that the performance of the stochastic gradient descent highly depends on the choice of the step

size. Furthermore, the step size will influence the convergence of the method. If the step size is chosen too small, the gradient descent will not move very far from the starting point. If the step size is chosen too big, the method needs many iterations to find the optimum.

The approximation quality also depends on the number of chosen basis functions. The intuitive conjecture, saying the more basis functions the better the approximation, is not complied.

Alternative formula

In this example we test the alternative estimator of the gradient estimator. The gradient of the objective function was approximated by a standard Monte Carlo estimator. The decreasing step size is chosen to be $0.01/(\text{iteration}+10)$. In total 15 optimization steps were calculated. As an starting point for the optimization we used a perturbed approximation of the optimal solution calculated by the related PDE. To evaluate the gradient estimator 200 trajectories were calculated.



(a) In blue the original potential is shown. In black the optimal tilted potential is shown calculated by the PDE formulation of the problem. In red the result after every gradient descent step is shown.

(b) In blue the starting control is shown. In red the optimal control calculated by the PDE formulation is shown. In black the last iterate after 15 optimization steps is shown.

Fig. 5.3: Resulting potential and control for a gradient descent done with the alternative stochastic gradient estimator with 200 trajectories.

The example shows that the alternative gradient estimator works. The controlled potential is converging into the direction of the optimal tilted potential. Similar to the previous example the convergence of the method depends very much on the chosen step size. Furthermore, the control of the last iterate shows again a quite good fit of the optimal control in the relevant region. Compared to the stochastic gradient estimator it seems that the stochastic behaviour is reduced. The gradient estimates are less affected by the individual behaviour of the different trajectories. The variance decreases in every gradient descent iteration. The gradient descent

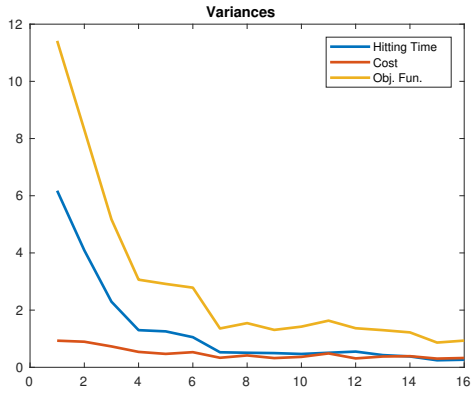
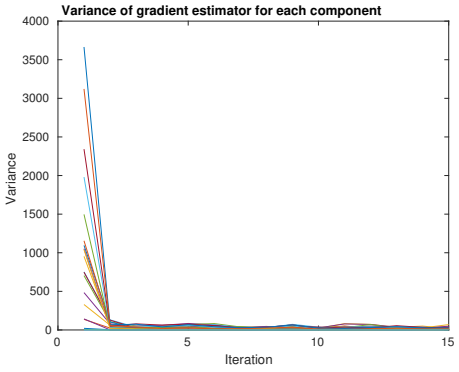


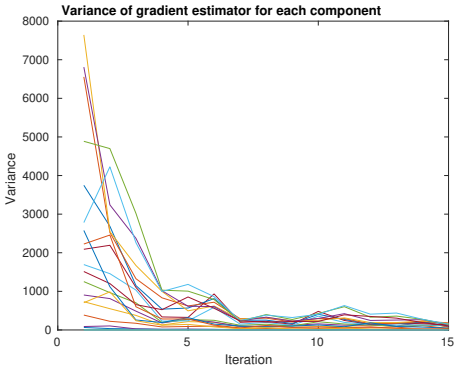
Fig. 5.4: The different variances of the objective function (yellow), hitting time (blue), cost function (orange) are shown for each evaluation of the gradient descent method.

methods with the alternative gradient estimator seems to be even more sensitive to the step size.

Comparing the variances of the different gradient estimators we see that the variance of the stochastic gradient estimator decays much quicker than the variance of the estimator based on the alternative formula. Looking at the two different formulas one could suppose that the variance of the alternative formula is much better because there is no stochastic integral which has to be calculated. But the shown numerical example clearly contradicts this expectation see 5.5. In the last calculated iteration the variance of the stochastic estimator is also lower. It seems that the stochastic integral in the gradient estimator stabilizes the calculation. But this has to be evaluated in future research.



(a) Variance of each component of the gradient estimator for the stochastic gradient descent shown in each iteration.



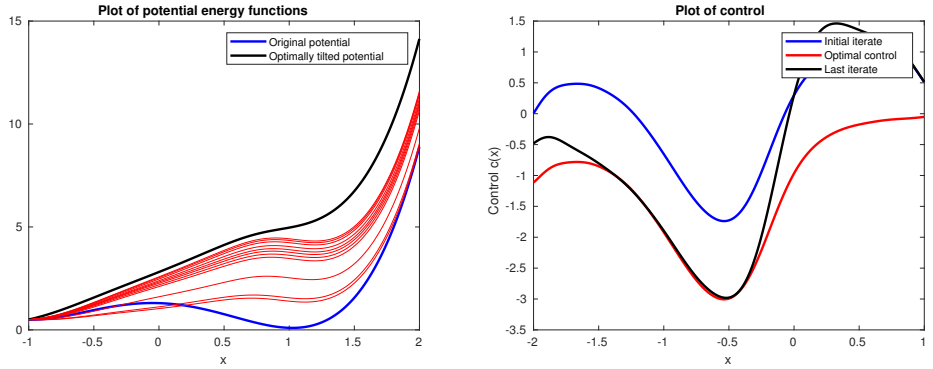
(b) Variance of each component of the gradient estimator for the stochastic gradient descent with the alternative formula shown in each iteration.

Fig. 5.5: Decay of the variance of the different gradient estimators.

Alternative formula with one realization

In the last example we test the alternative estimator with only one evaluation of

the gradient. Even though the variance of the estimator based on the alternative formula is worse compared to the estimator, including the stochastic integral, the first experiment with the alternative formula showed a very stable behaviour of the estimator; see 5.3.2. This is why we want to see if the number of evaluations could be decreased to decrease the computational effort of the gradient evaluations. The decreasing step size is chosen to be 0.007/iteration. In total 10 optimization steps were calculated. As starting point for the optimization we used a perturbed approximation of the optimal solution calculated by the related PDE. To evaluate the gradient estimator 1 trajectory was calculated.



(a) In blue the original potential is shown. In black the optimal tilted potential is shown calculated by the PDE formulation of the problem. In red the result after every gradient descent step is shown.

(b) In blue the starting control is shown. In red the optimal control calculated by the PDE formulation is shown. In black the last iterate after 10 optimization steps is shown.

Fig. 5.6: Resulting potential and control for a gradient descent done with the stochastic gradient estimator with 1 trajectory.

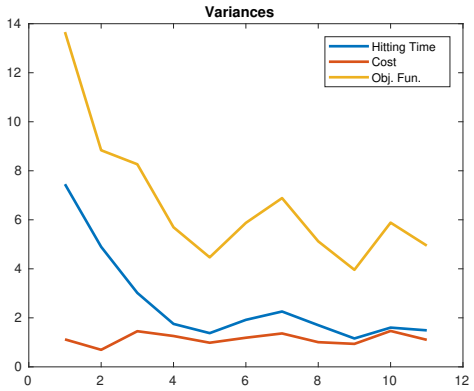


Fig. 5.7: The different variances of the objective function (yellow), hitting time (blue), cost function (orange) are shown for every evaluation of the gradient descent method.

The example shows that the alternative gradient estimator with only one evaluation of the dynamical system works. The controlled potential is converging into to of the optimally tilted potential. Furthermore, the control of the last iterate shows a quite good fit of the optimal control in the relevant region. The variance decreases

in every gradient descent iteration. But the variance decay is much slower compared to the gradient estimator with averaging. So in order to reduce the computational cost for the gradient evaluation is does make sense to combine both techniques. In order to get close to a possible minimum without computational effort one could start the optimization with only a few evaluations of the dynamical system and after a number of iterations the number of function evaluations is increased to get a better gradient estimator.

After we have developed and tested different gradient estimators for the optimization problem of finding the optimal bias we are now going to take a closer look at another method which was proposed to solve the optimization problem, namely the Cross-Entropy method. We are going to show that the Cross-Entropy method can be kernelized and that the methods can be seen as a Gaussian process approach finding the optimal bias.

5.4 Non-parametric representation

In this section we develop a non-parametric expression of the control based on the Cross-Entropy method. The parametric Cross-Entropy for importance sampling of diffusions was first introduced by [93]. We extend this approach by introducing a kernel function and in this way we can derive a non-parametric representation formula of the optimal bias. The derived formula can be used to express the optimal bias on the trajectories and also to predict the bias. The prediction also depends on the trajectories from the sampling. Because of this dependency on the observed trajectories the kernelized Cross-Entropy method can be seen as a data driven approximation of the optimal bias. Furthermore, the kernelization of the parametric Cross-Entropy method turns the linear method into a non-linear method and thus gives the approach more flexibility to express the optimal bias.
Kernel methods are well-known in the context of machine learning, e.g. support vector machines or Gaussian processes [65]. The main advantage of these methods is that they can operate in a high-dimensional space by using so-called kernel functions without the explicit calculating of the coordinates in this space. This is done by expressing these high-dimensional calculations in terms of inner products which is often called the kernel trick. Kernel functions are weighted sums of integrals which are used in this application to express similarity of observed data points (e.g. time points in a trajectory) and thus encode spacial information cf. [65].

The section is structured as follows: We introduce kernel functions very briefly before presenting the parametric Cross-Entropy method. Then, we derive the kernelized version of the Cross-Entropy method and close this section by giving a numerical example of the derived method.

5.4.1 Kernel functions

In machine learning kernel functions are used as a measure of similarity because it is assumed that two inputs which are very close to each other will give a similar output. The term kernel arises in the theory of integral operators. A function $k : \mathcal{D} \in \mathbb{R}^n \times \mathcal{D} \in \mathbb{R}^n \to \mathbb{R}$ of two arguments mapping into the $\mathbb{R}$ is called a kernel. With this kernel we can now define an integral transformation T_k.

Definition 2. *Consider a function $f \in L^2$ and $k \in L^2(\mu)$. Then the integral operator $T_k : L^2(\mu) \to L^2(\mu)$ is defined by*

$$T_k(f) = \int_{\mathcal{D}} k(y, x) f(x) \mu(dx)$$

where μ denotes a measure on $\mathcal{D}$.

A kernel is said to be symmetric if $k(x, y) = k(y, x)$ holds. In machine learning only positive semidefinite kernels are considered. A kernel is positive semidefinite if

$$\int k(y, x) f(x) f(y) \mu(dx) \mu(dy) \geq 0$$

holds for all $f \in L^2(\mu)$ and $x, y \in \mathcal{D}$.

Famous examples of symmetric positive semidefinite kernels defined on $\mathbb{R}^n$ are the linear kernel

$$k(x, y) = x^T y \quad x, y \in \mathbb{R}^n$$

or the Gaussian kernel (RBF kernel)

$$k(x, y) = \exp\left(- \frac{\|x - y\|_2^2}{2\sigma^2} \right) \quad x, y \in \mathbb{R}^n, \sigma \in \mathbb{R}_+ \tag{5.34}$$

In many applications only stationary kernel functions are considered but one can also find non-stationary kernel functions in the literature [74].
At least we present Mercer's theorem which states that a kernel can be represented in terms of eigenvalues and eigenfunctions. This representation will be useful for the derivation of the non-parametric representation formula.

Theorem 7 (Mercer's theorem [74])**.** *Let $(\mathcal{D}, \mu)$ be a finite measurable space and $k \in L^\infty(\mu \times \mu)$ be a kernel such that $T_k : L^2(\mu) \to L^2(\mu)$ is positive definite. Let $\phi_i \in L^2(\mu)$ be normalized eigenfunctions of T_k associated with the eigenvalue $\lambda_i > 0$. Then*

- *the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are absolutely summable*

- $$k(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y) \tag{5.35}$$

  holds $\mu \times \mu$ almost everywhere and the series converges absolutely and uniformly $\mu \times \mu$ almost everywhere.

Proof. A proof of this representation theorem can be found in [52]. □

After this brief presentation of kernel functions we are going to present the derivation of the Cross-Entropy method and how it can be used to solve the optimization problem stated in equation (5.10).

5.4.2 Parametric Cross-Entropy method

In order to derive the Cross-Entropy method for the minimization problem

$$\begin{aligned} \min_u J(u) &= \mathbb{E}_{\mathbb{Q}}^x \left[W(x^u_{0:\hat{\tau}}) + \frac{\beta}{2} \int_0^{\hat{\tau}} |\gamma(x^u_s)|^2 ds \right] \\ \text{s.t. } dx^u_t &= -(\nabla V(x^u_t) + u(x^u_t))dt + \sqrt{2\beta^{-1}} dB_t, \quad x^u_0 = x \end{aligned}$$

we note that we can rewrite the objective function J based on a entropy representation

$$J(u) = J(u^*) + D(\mathbb{Q}|\mathbb{Q}^*) \tag{5.36}$$

where u^* is the optimal control, $D(\mathbb{Q}|\mathbb{Q}^*)$ is the Kullback-Leibler divergence and $\mathbb{Q} = \mathbb{Q}(u)$ and $\mathbb{Q}^* = \mathbb{Q}(u^*)$ are the measures corresponding to the different controls. Due to the Galerkin procedure the optimization problem turns into the problem of minimizing

$$\bar{H}(a) = D(\mathbb{Q}(u(a))|\mathbb{Q}^*) \tag{5.37}$$

over $a \in \mathbb{R}^N$, such that the measure $\mathbb{Q}(u(a))$ is absolutely continuous with respect to $\mathbb{Q}^*$. However, minimizing D is not easy possible because the function may have several minima or the optimal measure $\mathbb{Q}^*$ is unknown. The problem can be turned into a feasible minimization problem by flipping the arguments

$$H(a) = D(\mathbb{Q}^*|\mathbb{Q}(u(a))). \tag{5.38}$$

We lose equality in (5.36) since D is not symmetric. But it is well-known that

$$\bar{H}(a) \geq 0, H(a) \geq 0 \text{ and } \bar{H}(a) = 0 \text{ if and only if } H(a) = 0 \tag{5.39}$$

holds where the minimum is attained if and only if $\mathbb{Q}^* = \mathbb{Q}(u(a))$. Let us denote the path functional $W(x^u_{0:\hat{\tau}}) = \int_0^{\hat{\tau}} f(x^u_s)ds + g(x^u_{\hat{\tau}})$. It is known from the literature

[12] that the minimization of (5.38) is equivalent to the minimization of the cross entropy functional

$$CE(a) = -\mathbb{E}[\log M_{\hat{\tau}}(a)) \exp(-W(x^u_{0:\hat{\tau}}))] \tag{5.40}$$

where the likelihood ration $M_{\hat{\tau}}(a) = (\frac{d\mathbb{Q}^*}{d\mathbb{Q}(u(a))})$ between the controlled and the uncontrolled dynamic is quadratic in the parameter a and can be expressed via a Girsanov transformation. Many different methods have been proposed to solve this optimization problem; see e.g. [38, 93, 94].

5.4.3 Non-parametric Cross-Entropy method

The aim of the Cross Entropy method is to estimate a good approximation of the bias. In [93] a Galerkin projection of the bias was chosen such that the bias is represented by some weighted ansatz functions

$$u(x) \approx \sum_{i=1}^{M} a_i b_i(x).$$

In order to determine the weights a_i the Cross-Entropy functional (5.40) has to be minimized. In this linear approach the ansatz functions have to be chosen and placed. The choice and the placing influence the convergence and the variance behaviour of the quantity of interest. Especially, the sufficient placing of the ansatz function can be very challenging. This is why we are interested in a non-parametric representation of the bias. Furthermore, a non-parametric representation introduces more flexibility into the Cross-Entropy method. This non-parametric representation can be seen as allowing for an infinite set of ansatz functions b_i and from this point of view we can make use of Mercer's theorem to represent the kernel in terms of eigenfunctions as we will see later.
Since one only has a finite number of samples for the estimation of $u(x)$, we need a regularization by introducing a penalty term. We choose a regularization of quadratic form $r = -\frac{1}{2}\sum_j \frac{a_j^2}{\lambda_j}$ where $\lambda_j \in \mathbb{R}$ are the so called hyper-parameters. We will see later that the hyper-parameters can be seen as the eigenvalues of the kernel.

Let us recall Girsanov's theorem. We consider the two SDEs

$$\begin{aligned} dx_t &= -\nabla V(x_t)dt + \sqrt{2\beta^{-1}}dB_t \quad x_0 = x \\ dx^u_t &= -(\nabla V(x^u_t) + u(x^u_t))dt + \sqrt{2\beta^{-1}}dB_t, \quad x^u_0 = x. \end{aligned}$$

Then, we know from Girsanov's theorem that the likelihood $M_{\hat{\tau}}$ is given by

$$M_{\hat{\tau}} = \exp\Big(\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} u(x_t)dB_t - \frac{1}{4\beta^{-1}}\int_0^{\hat{\tau}}(u(x_t))^2dt\Big). \tag{5.41}$$

The discrete regularized Cross-Entropy functional is then given by

$$CEr(a) = -\frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\Big(\log M_{\hat{\tau}}(a))\Big) - \frac{1}{2}\sum_j \frac{a_j^2}{\lambda_j} \tag{5.42}$$

where $M_{\hat{\tau}}(a)$ is (5.41) with $u(x) = \sum_{j=1}^{M} a_j b_j(x)$. In order to minimize the Cross-Entropy functional to find the optimal weights a^* we are taking the gradient of (5.42) with respect to a and set it to zero. Since only the likelihood and the regularization depend on the parameter a, the derivative is given by

$$\begin{aligned}\frac{\partial}{\partial a_j}CEr(a) = -\frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\Big(&\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} b_j(x_t)dB_t \\ &- \frac{1}{2\beta^{-1}}\int_0^{\hat{\tau}}(\sum_j a_j b_j(x_t))b_j(x_t)dt\Big) - \frac{a_j}{\lambda_j}.\end{aligned}$$

In order to kernelize the above equation we multiply the above equation such that we can use Mercer's theorem. We multiply by λ_j and $b_j(x)$. Then, by introducing a summation over the ansatz functions we see that $(\sum_j \lambda_j b_j(x)b_j(y))$ is the representation of a kernel given by Mercer's theorem. The λ_j can be interpreted as the eigenvalues and $b_j(x)$ as the eigenfunctions of the corresponding integral transformation.
The derivative of the regularized Cross-Entropy functional is given by

$$\begin{aligned}0 = -\frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\Big(&\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} k(x_t^i, x)dB_t \\ &- \frac{1}{2\beta^{-1}}\int_0^{\hat{\tau}}(\sum_j a_j b_j(x_t)k(x_t^i, x)dt\Big) - \sum_j a_j b_j(x).\end{aligned}$$

Rewriting again the approximation $\sum_j a_j b_j(x)$ as $u(x)$ we have found an integral equation for the function $u(x)$

$$\begin{aligned}0 = -\frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\Big(&\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} k(x_t^i, x)dB_t \\ &- \frac{1}{2\beta^{-1}}\int_0^{\hat{\tau}}(u(x_t)k(x_t^i, x)dt\Big) - u(x).\end{aligned}$$

We now assume that we have observations of the dynamical system consisting of N paths each of individual length T_i $i = 1, \dots, N$ (we call the collection of all time points of all trajectories observations). Let us denote the size of the observation is N_O. Using these we can discretize the above integral equation to get a system of

linear equations. Solving this gives us an estimate of the control on the observed trajectory. We discretize the integrals in time at points $x^i_{t_k}$

$$A = \frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\Big(-\frac{1}{2\beta^{-1}}\sum_{t_k=0}^{M_i}(u(x^i_{t_k})k(x^i_{t_k},x')dt_k\Big) + I_{N_O\times N_O}$$
$$b = -\frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\frac{1}{\sqrt{2\beta^{-1}}}\sum_{t_k=0}^{M_i}k(x^i_{t_k},x')dB_{t_k}$$

where the kernel k is evaluated on all sampled time points of all trajectories. Based on the estimate of the observations we can use the estimate of $u(x)$ to predict the control for every point. This is done in a second step by rearranging the above formula and use the solution of the linear system

$$u(x) = \frac{1}{N}\sum_{i=1}^{N} e^{-W(x^i_{0:\hat{\tau}})}\Big(-\frac{1}{\sqrt{2\beta^{-1}}}\sum_{t_k=0}^{M_i}k(x^i_{t_k},x)dB_{t_k} + \frac{1}{2\beta^{-1}}\sum_{t_k=0}^{M_i}u(x^i_{t_k})k(x^i_{t_k},x)dt_k\Big)$$

where the kernel k is evaluated on the old observations and the new point x.
Furthermore, it is not necessary to start from the unperturbed dynamical system. By assimilating the path function using again the weights from Girsanov's theorem we can start with any perturbed dynamical system to find the optimal control. The resulting formulas look very similar to the equations derived above, only the path function changes. We only present the regularized Cross-Entropy functional since the derivation does not change

$$CEr(a) = -\frac{1}{N}\sum_{i=1}^{N} e^{-W(x^v_{0:\hat{\tau}})M^v_{\hat{\tau}}}\Big(\log M_{\hat{\tau}}(a))\Big) - \frac{1}{2}\sum_j \frac{a_j^2}{\lambda_j}$$

where

$$dx^v_t = -(\nabla V(x^v_t) + v(x^v_t))dt + \sqrt{2\beta^{-1}}dB_t, \quad x^v_0 = x$$

and

$$M^v_{\hat{\tau}} = \exp\Big(-\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} v(x^v_t)dB_t + \frac{1}{4\beta^{-1}}\int_0^{\hat{\tau}}(v(x^v_t))^2dt\Big).$$

By doing this it is possible to start the optimization problem with an already perturbed stochastic process. This can be very helpful if the stochastic process is metastable. Furthermore, the here presented scheme can be used to iterate over the predicted bias in order to optimize it.
Let us briefly comment on the connection to Gaussian processes before presenting a one-dimensional example of the presented approach. The regularized Cross-Entropy

functional can also be interpreted from a pseudo-Bayesian viewpoint. For this we rewrite the functional

$$CEr(a) = -\mathbb{E}\Big[\log\Big(\exp\Big(C\Big(\frac{1}{\sqrt{2\beta^{-1}}}\int_0^{\hat{\tau}} u(x_t)dB_t - \frac{1}{4\beta^{-1}}\int_0^{\hat{\tau}}(u(x_t))^2dt\Big)\Big) \exp\Big(-\frac{1}{2}\sum_j \frac{a_j^2}{\lambda_j}\Big)\Big)\Big].$$

where $C = \exp(-W(x_{0:\hat{\tau}}))$. We can interpret the first term as a weighted likelihood (weighted by the path functional) and $\exp\Big(-\frac{1}{2}\sum_j \frac{a_j^2}{\lambda_j}\Big)$ as a Gaussian prior distribution over the parameter a_j. From this point of view the regularized Cross-Entropy functional can be understood as a Gaussian process model for the function u. This interpretation of the method could be very useful because there are many very efficient algorithms for Gaussian processes in the literature which might help to develop fast algorithms for high-dimensional generalization of the presented method; see e.g. [74].

5.4.4 Examples

In the following we show the application of our method in a simple one-dimensional example. Let us consider a dynamical system satisfying (2.16) where V is a symmetric bistable potential

$$V(x) = \frac{1}{2}(x^2 - 1)^2. \tag{5.43}$$

The potential has two minima which are separated by a local maximum. The dynamical system following the SDE is metastable. A visualization can be found in 3.1. We are interested in sampling the moment generating function for the stopping time reaching the local maximum when starting in the local minimum $x = -1$. The resulting path functional is given by

$$\mathbb{E}[\exp(-\frac{1}{\beta}\tau)] \tag{5.44}$$

where $\tau = \inf\{t > 0, x_t > 0\}$. In this example 20 trajectories of (2.16) are calculated by using a standard Euler-Maruyama discretization with a time step $\Delta t = 10^{-2}$ and $\beta = 2$ in MATLAB. In order to sample the quantity of interest we sampled until all trajectories reached the goal. We used a Gaussian kernel as a measure of similarity given by

$$k(x, y) = \alpha \exp\Big(-\frac{(x-y)^2}{2\sigma^2}\Big) \tag{5.45}$$

with $\alpha = 5$ and $\sigma = 0.5$. In total 10 optimization steps have been performed to find the optimal bias.

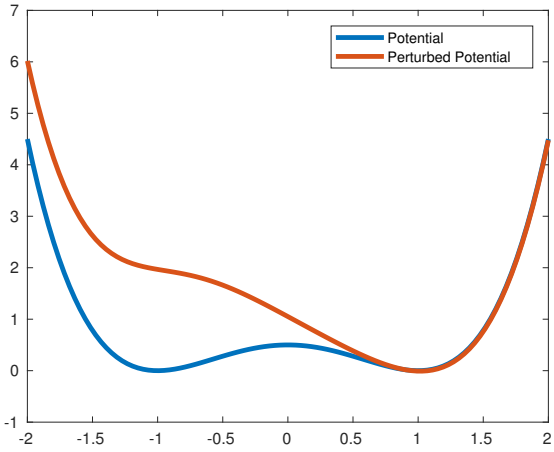


Fig. 5.8: Original potential (blue) and perturbed potential (red)

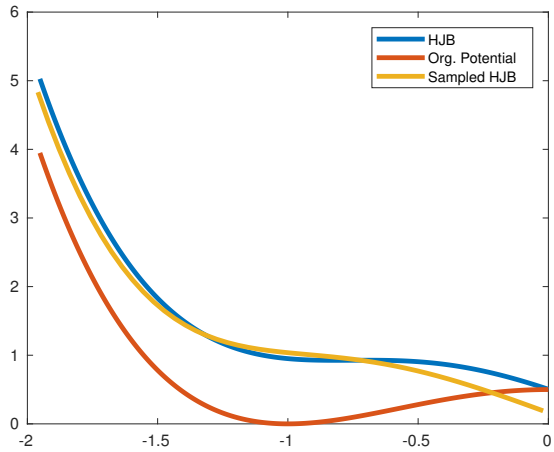


Fig. 5.9: Comparison of the optimal biasing potential of the sampling approach (yellow) and the solution of the corresponding HJB equation (blue), the original potential is shown in (red)

	MC	IS
$\mathbb{E}[\tau]$	2.81	1.85
$\mathrm{Var}[\tau]$	4.087	2.020
$\mathbb{E}[\exp(-\frac{1}{\beta}\tau)]$	0.3419	0.3754
$\mathrm{Var}[\exp(-\frac{1}{\beta}\tau)]$	0.0461	0.0218

Tab. 5.1: Comparison of the importance sampling estimator (IS) and the Monte Carlo estimator after 10 optimization steps.

We see that the estimator of the MC and the importance sampling are in good agreement for the quantity of interest; see 5.1. Comparing the reweighted estimator with a MC estimator with 10,000 trajectories $\mathbb{E}[\exp(-\frac{1}{\beta}\tau)] = 0.3714$ we see that the IS estimator is very close. So the IS estimator with 20 trajectories is better than the MC estimator with 20 trajectories. Furthermore, our approach could achieve

a variance reduction. The variance of the moment generating function could be reduced by 50%. The sampling time could also be reduced with results in a shorter computation time. In figure 5.9 the bias computed by the regularized Cross-Entropy approach is compared to the solution of the corresponding HJB equation. The results show a good agreement.

One limitation of the presented method is that the kernel function has to be evaluated on all sampled time points of all trajectories. This leads to very huge linear systems which can become difficult to solve for many observations. To overcome this problem a sparsity approach has to be developed. One idea is to use so-called inducing points (z) and to express the optimal bias by a linear combination of different kernel functions ($k(x, z_j) = \sum_j \alpha_j K(x, z_j)$). In this way the kernel functions only have the dimension of the inducing points and the size of the kernel matrix can be reduced. There are also other ideas to deal with high-dimensional data sets; see e.g [74].

5.5 Summary and Discussion

In the first part of this chapter different gradient estimators for the optimization problem after projecting the optimal control into a weighted ansatz function space have been derived. The main idea is motivated by differentiating the rewritten path functional. Here different formulation of the path functional can be used. In a first approach we derived a gradient formula based on ideas presented in [27]. The formula has the main advantage that the objective function of the optimization problem does not have to be continuous. But the derived formula relies on the fact that the approximation of the bias only depends on one parameter. If and how the presented method can be extended to high-dimensional problems has to be dealt with in future research. Another gradient estimator was derived by using an idea presented in [31]. We could show that the resulting derivative converges to the finite difference approximation and thus is a derivative of the path function. Furthermore, a second gradient estimator could be derived if the bias is of gradient structure based on ideas presented in [58]. These different gradient estimators have been tested for one-dimensional examples. The examples show that the gradient estimators work and that by applying a gradient descent a good approximation of the optimal control can be found. Comparing the variances of the gradient estimators we could clearly see that the variance of the estimator based on the alternative formula decays much slower than the variance of the estimator including the stochastic integral. So it seems that the stochastic integral stabilizes the gradient estimation if averaging is used. The second gradient estimator showed a very stable behaviour even though it is a stochastic expression. This is why it was tested with only one trajectory sampling. The example showed that even in this case the gradient descent was working. Comparing this example with the example of the estimator using averaging

the behaviour was not very stable but still it performed quite well. The advantage of this stable estimator is that the computational cost of the evaluation can be reduced drastically if the amount of sampling is reduced. If this behaviour of the gradient estimator generalizes to other applications and in which way it is affected by the step size has to be dealt with in future research.

The numerical examples show that the convergence of the gradient descent method is very sensitive to the starting point and the chosen step size. The starting point also influences how strong the metastability is influenced and how much the sampling effort can be reduced. So in order to have a fast converging gradient descent it is necessary to have a good starting point. For this the optimization method and the adaptive importance sampling method presented in Chapter 4 could be combined in future work. The adaptive importance sampling method could be used to build a suboptimal bias which is then, in a second step optimized, by the gradient descent method.

To determine a good step size is an even harder problem. If the step size is too small, the convergence of the gradient descent method is very slow. If the step size is too big, the algorithm will need a very long time to find the optimum. Moreover, this can lead to biases which hinder the trajectory from reaching the sampling goal. This can then also lead to a zigzagging behaviour of the gradient descent.

So finding the optimal bias with the gradient descent method is a difficult task. There are many parameters to be tuned and there is also a discretization error from the projection of the control and also from the Euler-Maruyama discretization of the SDE. This is why a zero variance estimator might be impossible to be found. But as we have seen in the numerical experiments a significant variance reduction can be achieved by performing only a few optimization steps and using the resulting bias as a suboptimal bias in combination with a reweighting scheme.

In the second part of the chapter we have derived a non-parametric representation of the optimal bias. For this we used the Cross-Entropy method and developed a regularized estimator. The non-parametric estimator is derived by expressing the derivative of the ansatz function as a kernel function based on Mercer's theorem. The kernel is evaluated on the sampled trajectories and encodes the similarity information. The estimator can be assimilated such that a already perturbed process can be used to calculate the optimal bias. For this only the path functional has to be adapted by Girsanov's theorem. In this way the approach can be used to optimize the already used. We also showed how the kernelized Cross-Entropy method can be viewed as a Gaussian process model. At the end of the section the approach was tested in a numerical example. In this example a variance reduction was achieved and also a good approximation of the HJB equation was found.

The here presented approach can be easily extended to high-dimensional problems. In order to do this one has to express the biasing potential as a kernel function. The optimal bias is then expressed as the derivative of the kernel. So in order to calculate the optimal bias a linear system for the derivative can be derived and with this the optimal bias can be calculated.

6 Summary and Outlook

In this thesis methods have been developed how the sampling of different quantities of metastable dynamical systems can be improved. The main idea of all of these approaches is to sample the quantity of interest in a biased potential. The bias reduces the metastability and thus the sampling can be speeded up. Furthermore, if the bias is designed in a good way the variance of the sampling can be reduced.

In Chapter 3 a global perturbation was used to bias the metastable system. The method can be used to decrease a metastability in a structural way without a priori knowledge where the metastability is located. The approach is based on a convolution approach with was first used by Scheraga for global optimization of molecular systems. In preliminary numerical tests we could show that the convolution approach has a decreasing impact on the metastability for low- and high-dimensional examples. Furthermore, different schemes for the approximation of the convolution were summarized. We showed how the convolution approach can be integrated into different well-known MD algorithms like replica exchange and developed a reweighting technique for thermodynamic quantities. We could also show that the convolution approach can be interpreted as a small external force acting on the dynamical system. In this way it was possible to use Linear Response theory to understand the behaviour of the dynamical system on the convolution. Furthermore, we showed how the convolution approach can also be used for the sampling of dynamic quantities. For this we first combined the convolution approach with the Eyring-Kramers' formula to develop an extrapolation scheme for mean first exit times and exit rates. In a second approach we combined the convolution approach with Girsanov's theorem to develop an importance sampling scheme for general dynamic quantities.
In future research the convolution approach could be used to build a multilevel Monte Carlo estimator. Different potentials can be generated with different smoothing parameters and the quantity of interest is then sampled in each individual potential. These samples have to be combined then in order to find the real estimator. In Multilevel Monte Carlo this is done by combining the sampled estimators from different levels in a telescope sum. If this can also work for convolution approaches, has not been tested or proved yet. As we have seen the convolution approach reduces the metastability which also reduces the sampling time dramatically. If the convolution approach can be integrated into the Multilevel Monte Carlo framework, this could result in a fast low variance estimator for metastable systems.

In order to analyse the convolution approach analytically Kato theory could be used. Kato theory was invented to investigate the behaviour of linear operators under perturbation. The convolution can be seen as a perturbation of the original dynamical system especially in the polynomial case. It would be of great value to investigate the behaviour of the eigenvalues under the convolution in order to get a better understanding how the spectral gap and thus the metastability of the system is changed.
Another interesting question is if and how ergodicity is influenced under convolution. We have seen in the numerical tests that the position of the metastable states is also changed by the convolution. For small smoothing parameters this might be negligible. But since the convolution is a global transformation, the behaviour of the potential at infinity will also change. How this influences the ergodicity is still unclear and should be dealt with in future work.
All in all it was possible to show that the convolution approach can be used to decrease the metastability without exact knowledge of the location. The approach offers many interesting possibilities to be integrated in well-known frameworks. Also analytical investigation seems to be possible for special situations. All this makes the convolution approach an interesting method to be investigated further.

In Chapter 4 we extended well-known algorithms from MD, namely Metadynamics which have been designed for the sampling of stationary distributions, to the sampling of dynamic quantities. In order to achieve this we adapted the existing algorithm in a way that it only constructs a bias which reduces the metastability locally leaving the rest of the dynamical system unchanged. The quantity of interest is then sampled in the bias potential and the expectation is reweighted. The weights are calculated according to Girsanov's theorem. The resulting estimator is an unbiased estimator. It was also possible to derive a second unbiased estimator if the bias has a gradient structure. Under relatively mild assumptions we could show that Novikov's condition holds and thus Girsanov's theorem can be applied. A further analysis showed that the bias which is constructed by our algorithm preserves ergodicity of the dynamical system. Different low-dimensional examples show that a suboptimal control can achieve a variance reduction and a reasonable reduction of the sampling effort. In general, the method is not restricted to the usage of Metadynamics and other methods can be used to construct the bias. This was indicated by applying a Metadynamics-like algorithm to construct a bias directly on the force.
For future research different strategies could be tested. There is a third way how the Girsanov's weights can be expressed. Based on the Martingale Representation theorem the Girsanov weight can be expressed as a SDE c.f. [67]. Since the resulting SDE is of very simple form and the diffusion term is not space-dependent a more sophisticated discretization scheme can be used [77]. This could lead to a reduced variance of the Girsanov weight in high-dimensional problems. If this has an impact on the variance of the Girsanov weight and the resulting estimator has to be evalu-

ated in future research.
The Adaptive Importance Sampling algorithm can also be combined with the optimization procedure presented in Chapter 5. The bias constructed by the assimilated Metadynamics algorithm can be used as a starting point for the optimization. The combination of these two methods would improve the optimization procedure because it is often very hard to find a good starting point. Especially in the case of a metastable dynamical system it is very important to have a good bias which improves the evaluation of the expectation and thus reduces the sampling effort. If the bias is chosen in a bad way, the evaluation of the expectation can be very slow.

In Chapter 5 we dealt with the optimization formulation of finding the optimal bias. In the first part new gradient estimators for the gradient descent algorithm proposed in [38] were developed. At first we derived a gradient estimator for non-continuous objective functions based on an idea presented by [27]. The estimator was developed in a low-dimensional framework. If this estimator can be extended to high-dimensional problems has to be verified in the future. If this is possible it would be interesting to compare this gradient estimator with the others developed in this thesis to see which performs better for the given problem. In a second approach we used an idea from financial mathematics and adapted it to the control problem at hand. Based on a different representation of the likelihood ratio another derivative estimator was derived. The different estimators were tested numerically showing quite good results. But the numerical test also showed that the convergence of the approach heavily depends on the starting point and the chosen step size. For future research other gradient estimators could be tested. There were many interesting ideas presented in the literature, e.g. [36] or [2]. Especially the ideas presented in [36] have also been tested for high-dimensional problems such that it could be very interesting to see if they can be applied to the optimal control problem. The authors interpret the expectation as a solution of a PDE and differentiate the PDE to calculate the derivative of the expectation. In a second step the solution of the differentiated PDE is reformulated as an expectation. The gradient estimators presented in this work are also applied in high-dimensional problems and show a very stable behaviour. So since the expression for the gradient is not unique, many other formulas can be used.
In the second part of the chapter non-parametric estimators for the optimal bias were developed. The key idea was a reformulation of the Cross Entropy approach proposed by [93]. For the derivation a regularization was introduced and so the ansatz functions can be rewritten in terms of kernel functions. The resulting formula includes all the sampled points and this changes the point of view on the optimization problem. The optimization problem turned into a data-driven problem taking much more dynamical information into account. The shown interpretation in term of Gaussian processes also strengthens this point of view. The first low-dimensional examples show good first results in terms of variance reduction and approximation

of the optimal bias.
A future research direction for the non-parametric representation formula could focus on the connection of the Cross-Entropy method and Gaussian processes. As we have seen the regularized Cross-Entropy functional can be interpreted as a likelihood times a prior which leads to the Gaussian process interpretation. The main advantage of this interpretation is that many algorithmic achievements from Gaussian processes could also be used for the optimization problem, for example, the sparsity approach. The sparsity approach offers the application of the method for high-dimensional problems with many observations. Another interesting question is to understand how different kernel functions influence the variance reduction and the approximation error.

A high-dimensional application of this methodology is very challenging. Many of the enhanced sampling algorithms which could be used for constructing a bias only work in low- dimensional reaction coordinates. Many of the here developed methods can be applied at least in theory for high-dimensional problems. But if we do not use the projection into reaction coordinates, we need to find what causes the metastability in the dynamical system. Since metastability is a phenomenon which involves many particles, it is not quite clear how these particles can be detected. If the relevant particles causing the metastability are detected, the interaction potentials could be changed in order to influence it. If this leads to an optimal bias or a suboptimal bias, is unclear. In principle, this could also lead to a complete different behaviour of the dynamical system which would be undesirable. Furthermore, it is also well-known that the Radon-Nikodym derivative is affected by the dimension. So in order to prevent this dimension effect the variance has to decrease sharply as the dimension increases [68]. This can only be achieved if information about the system is available. From this point of view we think the best option to achieve a significant variance reduction for dynamic systems is to combine the importance sampling approach with a dimension reduction method.

All in all, the variance of the estimator is reduced if the metastability is decreased. For the low-dimensional examples the metastable set is easy to find. In this case the dynamical system is described only in some relevant coordinates and the related HJB equation must only be solved in a low-dimensional space. For very low-dimensional problems the HJB equation can then be solved numerically which will lead to a very good variance reduction and a massive speed-up in sampling.

6.1 Future work

In this section we are going to summarize different directions which could be explored in future research.

Machine learning

We have seen in this thesis that there are many connections of the problem and subjects close to machine learning like stochastic optimization or Gaussian processes. Another different connection of optimal control and machine learning can be found in [43]. In this paper H. Kappen suggests to use a machine learning approach to solve the optimal control problem called reinforcement learning. The main idea of this approach is that the a control is learned by an algorithm based on the state space exploration. Since in molecular dynamics the equations of motion are given, reinforcement learning (RL) could be used to find the optimal control. In RL a system or agent can try an action to reach a certain goal. After some time the action is evaluated. If the action helps achieving the goal, a reward is given. The goal is to maximize the reward. One example of RL is Q-learning. Here the system can choose the action freely at the beginning of the learning process. Due to the performance of the different actions they are rewarded differently. A so-called learning rate determines how the new information gained by the exploration overrides the old information. A decreasing learning rate ensures that the best action of the past is chosen more often in the future. In this way the optimal action can be learned by exploring the system. In order to apply RL in MD possible actions have to be found. The actions can be explored automatically by different simulations. To determine a good set of actions a priori knowledge about the system must be available.

Analytical expression of the variance decay

In order to build a sufficient importance sampling scheme it would be very interesting to get a better understanding how the bias or a suboptimal control influences the variance. In the literature it was proved that for finite time problems the second moment of the estimator can be characterized as a solution to a PDE, cf. [82]. The theorem uses an equivalent representation of the second moment. For this representation the PDE can be derived by using Itô formula and Feynman-Kac formula. To study the variance decay a perturbation approach is used. This approach might be used to investigate much smaller problems in which the PDE can be solved analytically or at least numerically. This would give the possibility to understand the decay of the variance under a suboptimal control much better and how an effective bias has to be designed to achieve a significant variance reduction.

7 Appendix

Example: Girsanov discrete version

In this section we are going to consider a derivation of Girsanov's theorem. The example is taken from [56].

Let us consider a stochastic process x_t. We are going to shift the mean of the stochastic process. The process satisfies

$$dx_t = -h(t) + dB_t \tag{7.1}$$

where $h \in \mathcal{C}^2 : \mathbb{R}_+ \to \mathbb{R}$ is a deterministic function with $h(0) = 0$ and dB_t is a standard Brownian motion. Furthermore, let $t > 0$, $n \in \mathbb{N}$ is a constant and $0 = t_0 < t_1 < t_2 \ldots < t_n = t$ is a partition of the interval $[0, t]$. We have to show that the density of the process (7.1) is given by

$$\prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi(t_{i+1} - t_i)}} \exp\Big(-\sum_{i=0}^{n-1} \frac{(x_{i+1} - x_i + h(t_{i+1}) - h(t_i))^2}{2(t_{i+1} - t_i)} \Big). \tag{7.2}$$

We note that the discrete process satisfies $x_{t_1} = -h(t_1) + W_{t_1}, x_{t_2} = -h(t_2) + h(t_1) + W_{t_2} - W_{t_1}, \ldots, x_{t_n} = -h(t_n) + h(t_{n-1}) + W_{t_n} - W_{t_{n-1}}$. Since the increment of the Brownian motion is an independent Gaussian random variable with variance $t_1, t_2 - t_1, \ldots, t_n - t_{n-1}$, it follows that $x_{t_1}, x_{t_2} - x_{t_1}, \ldots x_{t_n} - x_{t_n - 1}$ are again independent Gaussian random variables with mean $-h(t_1), -h(t_2) + h(t_1), \ldots, -h(t_n) + h(t_{n-1})$ and variance $t_1, t_2 - t_1, \ldots, t_n - t_{n-1}$. So we can calculate for a test function $\varphi : \mathbb{R}^n \to \mathbb{R}$

$$\mathbb{E}[\varphi(x_{t_1}, x_{t_2} - x_{t_1}, \ldots, x_{t_n} - x_{t_n - 1})] = \int \varphi(y_1, \ldots, y_n) q(y_1, \ldots, y_n) dy_1 \ldots dy_n \tag{7.3}$$

with density

$$q(y_1, \ldots, y_n) = \prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi(t_{i+1} - t_i)}} \exp\Big(-\sum_{i=0}^{n-1} \frac{(y_i + h(t_{i+1}) - h(t_i))^2}{2(t_{i+1} - t_i)} \Big). \tag{7.4}$$

Using a transformation of variables $\psi : \mathbb{R}^n \to \mathbb{R}^n$

$$\psi = (y_1, y_2, \ldots, y_n) \to (y_1, y_1 + y_2, \ldots, y_1 + y_2 + \ldots y_n). \tag{7.5}$$

This is a $\mathcal{C}^1$ diffeomorphism with $\det(Jacobi) = 1$. One sees that $\psi(x_1, x_2 - x_1, \ldots, x_n - x_{n-1}) = (x_1, \ldots, x_n)$ has the inverse $\psi^{-1}(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n) = (x_1, x_2 - x_1, \ldots, x_n - x_{n-1})$. So we can apply the transformation of variables on a test function $\bar{\varphi} : \mathbb{R}^n \to \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[\bar{\varphi}(x_{t_1}, x_{t_2}, \ldots, x_{t_n})] &= \mathbb{E}[\bar{\varphi} \circ \psi(x_{t_1}, x_{t_2} - x_{t_1}, \ldots, x_{t_n} - x_{t_n - 1})] \\
&= \int \bar{\varphi}(y_1, \ldots, y_n) \circ \psi(y_1, \ldots, y_n) q(y_1, \ldots, y_n) dy_1 \ldots dy_n \\
&= \int \bar{\varphi}(x_1, \ldots, x_n) q \circ \psi^{-1}(x_1, \ldots, x_n) Jac(\psi^{-1})(x_1, \ldots, x_n) dx_1 \ldots dx_n \\
&= \int \bar{\varphi}(x_1, \ldots, x_n) q(x_1, x_2 - x_1 \ldots, x_n - x_{n-1}) dx_1 \ldots dx_n
\end{aligned}$$

which gives the desired result.

Zusammenfassung

Das Verhalten von Molekülen wird bestimmt von seltenen Ereignissen. So kann zum Beispiel eine Konformationsänderung dazu führen, dass sich die Funktionalität eines Moleküls komplett ändert. Darüberhinaus haben diese seltenen Ereignisse auch einen großen Einfluss auf numerische Simulationen von Molekülen. Darum ist es wichtig effektive und zuverlässige numerische Methoden zu haben, um diese seltenen Ereignisse vorherzusagen.

Die Probleme, die durch seltene Ereignisse hervorgerufen werden, werden hauptsächlich durch das stochastische Verhalten des dynamischen Systems und einem daraus resultierenden Phänomen, welches Metastabilität genannt wird, verursacht. Metastabilität heißt, dass das dynamische System für lange Zeit in einem bestimmten metastabilen Zustand verweilt, bevor es sehr schnell in einen anderen metastabilen Zustand übergeht. Deshalb ist Metastabilität eines der größten Probleme für die effektive Schätzung der unterschiedlichen Größen. In der Molekulardynamik gibt es zwei unterschiedliche Größen und die Schätzung von beiden wird durch seltene Ereignisse beeinflusst. Für thermodynamsche Größen sind viele unterschiedliche Methoden entwickelt worden, die sich nicht ohne Weiteres auf die Schätzung von dynamischen Größen übertragen lassen.

Diese Arbeit beschäftigt sich mit der Verbesserung von Schätzmethoden dieser Größen. Die zugrundeliegende Idee ist, die Metastabilität des Systems zu beeinflussen, um den Simulationsaufwand zu verringern und eine Varianzreduktion des Schätzers zu bekommen. Nach einer Einführung und einer Zusammenfassung der relevanten Theorie beschäftigt sich das 3. Kapitel mit einer Idee aus der globalen Optimierung, um die Metastabilität zu reduzieren. Wir zeigen, dass der Ansatz sowohl für thermodynamische Größen als auch für dynamische Größen genutzt werden kann.

Im 4. Kapitel werden lokale Ansätze genutzt, um ein Importance-Sampling-Schema für dynamische Größen zu entwickeln. Wir nutzen die Expertise gut etablierter MD-Methoden, um eine gute lokale Perturbation zu erstellen. Für das Importance-Sampling-Schema müssen diese Algorithmen angepasst und mit Ergebnissen aus der stochastischen Analysis verbunden werden. Die Methode wird an unterschiedlichen Beispielen getestet.

Das letzte Kapitel beschäftigt sich mit zwei Methoden, die eine optimale Perturbation im Sinne der Varianz finden können (Gradientenabstieg und Cross-Entropy-Methode). Für den Gradientenabstieg werden unterschiedliche Schätzer des Gradienten entwickelt und mithilfe der Cross-Entropy-Methode wird eine nicht parametrische Approximation der optimalen Pertubration hergeleitet. Am Ende der Arbeit werden die Ergebnisse zusammengefasst, diskutiert und weiterführende Ideen präsentiert.

Bibliography

[1]R. J. Allen, P. B. Warren, and P. R. Ten Wolde. „Sampling rare switching events in biochemical networks“. In: *Physical Review Letters* 94.1 (2005) (cit. on p. 8).

[2]G. Arampatzis, M. Katsoulakis, and L. Rey-Belett. „Efficient estimators for likelihood ratio sensitivity indices of complex stochastic dynamics“. In: *Journal of Chemical Physics* 144 (2016) (cit. on p. 129).

[3]S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer Verlag New York, 2007 (cit. on pp. 33, 102).

[4]A. Barducci, G. Bussi, and M. Parrinello. „Well Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method“. In: *Phys. Rev. Lett.* 100 (2 2008) (cit. on p. 85).

[5]N. Berglund. „Kramers' law: Validity, derivations and generalisations“. In: *Markov Process. Related Fields* (2013) (cit. on pp. 63, 64, 66).

[6]R.C. Bernardi, M.C.R Melo, and K. Schulten. „Enhanced sampling techniques in molecular dynamics simulations of biological systems“. In: *Bio. et Biophy. Acta (BBA)* 5 (2015) (cit. on pp. 34, 53).

[7]M. Born and R. Oppenheimer. „Zur Quantentheorie der Molekülen“. In: *Annalen der Physik.* 389.20 (1927), pp. 457–484 (cit. on p. 13).

[8]A. Bovier and F. den Hollander. *Metastability: A Potential-Theoretic Approach*. Springer International Publishing Switzerland: Springer International Publishing, 2015 (cit. on pp. 32, 63, 64).

[9]L. Brust. *Molekulardynamische Simulation via Faltungsansatz im Potential*. Tech. rep. 2017 (cit. on p. 48).

[10]J. A. Bucklew. *Introduction into rare event sampling*. Springer, 2004 (cit. on p. 7).

[11]G. Bussi, A. Laio, and M. Parrinello. „Equilibrium Free Energies from Nonequilibrium Metadynamics“. In: *Phys. Rev. Lett.* 96 (9 2006), p. 090601 (cit. on p. 85).

[12]C. C. Hartmann, L. Richter, Ch. Schütte, and W. Zhang. „Variational characterization of free energy: theory and algorithms“. In: *Entropy* 19 (2017), pp. 626–653 (cit. on pp. 29, 99, 119).

[13]F. Cerou and A. Guyader. „Adaptive multilevel splitting for rare event analysis“. In: *Stochastic Analysis and Application* (2006) (cit. on pp. 8, 36).

[14]P. Chaudhari, A. Oberman, S. Osher S. Soatto, and G. Carlier. *Deep Relaxation: partial differential equations for optimizing deep neural networks*. Tech. rep. 2017 (cit. on pp. 38, 46).

[15]C. Chipot and A. Pohorille. *Free Energy Calculations*. Springer Series in Chemical Physics. Berlin Heidelberg: Springer, 2007 (cit. on pp. 34, 80).

[16]J. F. Dama, M. Parrinello, and G. A. Voth. „Well-Tempered Metadynamics Converges Asymptotically". In: *Physical Review Letters* (2014) (cit. on p. 80).

[17]E. Darve and A. Porohille. „Calculating free energy using average forces". In: *Journal of Chemical Physics* (2001) (cit. on p. 7).

[18]E. Darve and A. Porohille. „Calculating free energy using average forces". In: *Journal for Chemical Physics* (2001) (cit. on p. 34).

[19]E.B Davies. „Spectral properties of metastable Markov semigroups". In: *Journal of Functional Analysis* 52.3 (1983), pp. 315 –329 (cit. on p. 32).

[20]P. Dupuis and H. Wang. „Importance Sampling, Large Deviations, and Differential Games". In: *Stochastics and Stochastic Reports* 76.6 (2004), pp. 481–508 (cit. on pp. 8, 9, 33).

[21]P. Dupuis and H. Wang. „Subsolutions of an Isaacs Equation and Efficient Schemes for Importance Sampling". In: *Mathematics of Operations Research* 32.3 (2007), pp. 723–757 (cit. on pp. 9, 33).

[22]P. Dupuis, A. D. Sezer, and H. Wang. „Dynamic importance sampling for queueing networks". In: *The Annals of Applied Probability* 17.4 (2007), pp. 1306–1346 (cit. on pp. 9, 33).

[23]P. Dupuis, K. Spiliopoulos, and X. Zhou. „Escaping from an Attractor: Importance Sampling and Rest Points I". In: *Annals of Applied Probability* 25.5 (2015), pp. 2909–2958 (cit. on pp. 9, 33, 96).

[24]R. Elber and A. West. „Atomically detailed simulation of the recovery stroke in myosin by Milestoning". In: *Proceedings of the National Academy of Science of the United States of America* (2010) (cit. on p. 36).

[25]J. Elstrodt. *Maß und Integrationstheorie*. 7th ed. Springer, 2010 (cit. on p. 27).

[26]L. C. Evans. *Partial Differential Equations*. Vol. 2. American Mathematical Society, 2010 (cit. on p. 40).

[27]Ch. O. Ewald. „The Malliavin gradient method for the calibration of stochastic dynamical models". In: *Statistics and Probability Letters* (2006) (cit. on pp. 99, 103, 105, 124, 129).

[28]H. Eyring. „The activated complex in chemical reactions". In: *Journal of Chemical Physics* 3 (1935) (cit. on p. 64).

[29]A. K. Faradjian and R. Elber. „Computing time scales from reaction coordinates by milestoning". In: *The Journal of Chemical Physics* 120.23 (2004) (cit. on p. 8).

[30]W. Fleming and H.M. Soner. *Controlled Markov Processes and Viscosity Solutions*. 2nd ed. Springer, 2005 (cit. on p. 31).

[31]E. Fournie, J. Lasry, J. Lebuchoux, P. Lions, and N. Tozi. „Application of Malliavin Calculus to Monte-Carlo methods in finance“. In: *Finance and Stochastics* (1999) (cit. on pp. 99, 107, 124).

[32]M. Freidlin and A. Wentzell. „Random Perturbations of Dynamical Systems“. In: (2012) (cit. on p. 32).

[33]D Frenkel and B. Smit. *Understanding Molecular Simulation*. 2nd. Academic Press, Inc., 2001 (cit. on p. 18).

[34]P. W. Glynn. „Likelihood Ratio Gradient Estimation: An Overview.“ In: *Proceedings of the 1987 Winter Simulation Conference*. ACM, 1987 (cit. on p. 103).

[35]P. W. Glynn. „Stochastic Approximation for Monte Carlo Optimization“. In: *Proceedings of the 18th Conference on Winter Simulation*. ACM, 1986 (cit. on p. 103).

[36]E. Gobet and R. Munos. „Sensitivity analysis using Itô Malliavin calculus and martingales, and Application to stochastic optimal control“. In: *SIAM Journal on Control and Optimization* (2005) (cit. on pp. 103, 129).

[37]R. van Handel. *Stochastic Calculus, Filtering and Stochastic Control*. `https://web.math.princeton.edu/~rvan/acm217/ACM217.pdf`. Lecture Notes. 2007 (cit. on p. 31).

[38]C. Hartmann and Ch. Schütte. „Efficient Rare Event Simulation by Optimal Nonequilibrium Forcing“. In: *Journal Statistical Mechanics Theory and Experiments* 2012 (2012) (cit. on pp. 9, 99, 119, 129).

[39]D.J Higham. „An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations“. In: *SIAM Review* 43.03 (2001) (cit. on pp. 26, 86).

[40]C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni. „Mori–Zwanzig formalism as a practical computational tool“. In: *Faraday Discussions* 144 (2010), pp. 301–322 (cit. on p. 16).

[41]T. Huber, A. E. Torda, and W. F. van Gunsteren. „Local elevation: A method for improving the searching properties of molecular dynamics simulation“. In: *Journal of Computer-Aided Molecular Design* 8 (1994), pp. 695–708 (cit. on p. 80).

[42]W. Huisinga and B. Schmidt. „Metastability and Dominant Eigenvalues of Transfer Operators“. In: *Leimkuhler B. et al. (eds) New Algorithms for Macromolecular Simulation. Lecture Notes in Computational Science and Engineering* 49 (2006) (cit. on p. 47).

[43]H. Kappen. „An introduction to stochastic control theory, path integrals and reinforcement learning“. In: *AIP Conference Proceedings* 887.1 (2007), pp. 149–181 (cit. on pp. 96, 131).

[44]T. Kato. *Perturbation Theory for Linear Operators*. Springer, 1995 (cit. on p. 72).

[45]S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. „Optimization by Simulated Annealing“. In: *Science* 220 (1983), pp. 671–680 (cit. on p. 35).

[46]N. L. Kleinman, J. C. Spall, and D. Q. Naiman. „Simulation-Based Optimization with Stochastic Approximation Using Common Random Numbers“. In: *Manage. Sci.* 45.11 (1999) (cit. on p. 103).

[47]A. Klenke. *Wahrscheinlichkeitstheorie*. Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, 2008 (cit. on p. 109).

[48]P. Kloeden and A. Platen. *Numerical Solution of Stochastic Differential Equations*. Vol. 1. Springer-Verlag Berlin Heidelberg, 1992 (cit. on pp. 25, 26).

[49]J. Kostrowicki and L. Piela. „Diffusion equation method of global minimization: Performance for standard test functions". In: *Journal of Optimization Theory and Applications* (1991) (cit. on pp. 38, 40, 41).

[50]H. A. Kramers. „Brownian motion in a field of force and the diffusion model of chemical reactions". In: *Physica* 7 (1940) (cit. on p. 64).

[51]H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Vol. 2. Springer, 2003 (cit. on p. 102).

[52]H. König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser Basel, 1986 (cit. on p. 118).

[53]A. Laio and M. Parrinello. „Escaping free- energy minima". In: *PNAS* 20.10 (2002), pp. 12562–12566 (cit. on pp. 7, 34, 80).

[54]F. Legoll, T. Lelièvre, and S. Olla. „Pathwise estimates for an effective dynamics". In: *Stochastic Processes and their Applications* 127.9 (2017), pp. 2841 –2863 (cit. on p. 80).

[55]B. Leimkuhler and C. Matthews. *Molecular Dynamics*. Springer, 2015 (cit. on pp. 1, 13–15, 17, 26).

[56]T. Lelièvre. *Stochastic Calculus*. Lecture Notes. 2016 (cit. on p. 133).

[57]T. Lelièvre. „Two mathematical tools to analyze metastable stochastic processes". In: *Numerical Mathematics and Advanced Applications* (2011) (cit. on pp. 32, 33, 39).

[58]T. Lelièvre and G. Stoltz. „Partial differential equations and stochastic methods in molecular dynamics". In: *Acta Numerica* 25 (2016), pp. 681–880 (cit. on pp. 3, 14, 19, 21, 25, 29, 31, 79, 80, 124).

[59]T. Lelièvre, M. Rousset, and G. Stoltz. *Free energy computations: a mathematical perspective*. London, Hackensack (N.J.), Singapore: Imperial College Press, 2010 (cit. on p. 34).

[60]H. Lie and J. Quer. „Some connections between importance sampling and enhanced sampling methods in molecular dynamics". In: *The Journal of Chemical Physics* 147 (2017) (cit. on p. 34).

[61]J. Lu and E. Vanden-Eijnden. „Infinite swapping replica exchange molecular dynamics leads to a simple simulation patch using mixture potentials". In: *The Journal of chemical physics* 138 (2013) (cit. on p. 35).

[62]J. Lu and E. Vanden-Eijnden. „Infinite swapping replica exchange molecular dynamics leads to a simple simulation patch using mixture potentials". In: *The Journal of Chemical Physics* 138.8 (2013), p. 084105 (cit. on p. 53).

[63]G. N. Milstein. *Numerical integration of stochastic differential equations*. Vol. 313. Springer Science & Business Media, 1994 (cit. on pp. 8, 26, 29).

[64]J. Moré and Z. Wu. „Global Continuation For Distance Geometry Problems". In: *Siam Journal of Optimization* (1995) (cit. on p. 38).

[65]K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012 (cit. on p. 116).

[66]J. Nocedal and S. Wright. *Numerical Optimization*. 2nd ed. Springer, 2006 (cit. on p. 102).

[67]B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Hochschultext / Universitext. Berlin: Springer, 2003 (cit. on pp. 19, 22, 23, 25, 77, 85, 97, 128).

[68]A. Owen. *Monte-Carlo theory, methods and examples*. `http://statweb.stanford.edu/~owen/mc/`. Book preprint. 2013 (cit. on pp. 6, 7, 27, 28, 97, 130).

[69]G. Pavliotis. *Stochastic Processes and Applications*. Vol. 1. American Mathematical Society, 2014 (cit. on pp. 15, 16, 20, 23, 63, 64).

[70]G. Pavliotis and A. Stuart. *Multiscale Methods Averages and Homogenization*. Vol. 1. Springer Science+Business Media, LLC, 2008 (cit. on pp. 45, 46).

[71]L. Piela, J. Kostrowicki, and H. Scheraga. „On the multiple-minima problem in the conformational analysis of molecules: deformation of the potential energy hypersurface by the diffusion equation method". In: *Journal Physical Chemistry* (1989) (cit. on pp. 39, 40).

[72]P. E. Protter. *Introduction to Stochastic Integration and Differential Equations*. 2nd ed. Springer, 2004 (cit. on pp. 104, 105).

[73]J. Quer, L. Donati, B.G. Keller, and M. Weber. „An Automatic Adaptive Importance Sampling Algorithm for Molecular Dynamics in Reaction Coordinates". In: *SIAM Journal on Scientific Computing* 40.2 (2018), pp. 653–670 (cit. on pp. 76, 96).

[74]C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006 (cit. on pp. 117, 122, 124).

[75]M. I. Reiman and A. Weiss. „Sensitivity Analysis for Simulations via Likelihood Ratios". In: *Oper. Res.* 37.5 (1989) (cit. on p. 103).

[76]L. Rey-Belett. „Open classical systems". In: *Open quantum systems volume 1881 of Lecture Notes in Math* (2006) (cit. on p. 15).

[77]A. Rössler. „Runge–Kutta Methods for the Strong Approximation of Solutions of Stochastic Differential Equations". In: *SIAM Journal on Numerical Analysis* 48.3 (2010), pp. 922–952 (cit. on pp. 26, 128).

[78]T. Schlick. *Molecular Modeling and Simulation*. Springer Berlin-Heidelberg, 2002 (cit. on pp. 1, 18).

[79]Ch. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics*. American Mathematical Society, 2013 (cit. on pp. 14, 47).

[80]V. Spahn, G. Del Vecchio, D. Labuz, et al. „A nontoxic pain killer designed by modeling of pathological receptor conformations". In: *Science* 355.6328 (2017), pp. 966 –969 (cit. on p. 1).

[81]J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. 1st ed. Wiley, 2003 (cit. on p. 103).

[82]K. Spiliopoulos. „Non-asymptotic performance analysis of importance sampling schemes for small noise diffusions". In: *Journal of Applied Probability* 52 (2015), pp. 1–14 (cit. on pp. 33, 131).

[83]R.H. Swendsen and J.S. Wang. „Replica Monte Carlo simulation of spin glasses". In: *Physical Review Letters* (1986) (cit. on pp. 8, 35).

[84]C. Tsallis and D. A. Stariolo. „Generalized simulated annealing“. In: *Physica A: Statistical Mechanics and its Applications* (1996) (cit. on p. 8).

[85]M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford, 2010 (cit. on p. 18).

[86]O. Valsson and M. Parrinello. „Variational approach to enhanced sampling and free energy calculations“. In: *Physical Review Letters* 113.9 (2014) (cit. on pp. 7, 34).

[87]O. Valsson, P. Tiwary, and M. Parrinello. „Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint“. In: *Annual Review of Physical Chemistry* 67.1 (2016), pp. 159–184 (cit. on pp. 34, 97).

[88]E. Vanden-Eijden and J. Weare. „Rare Event Simulation of Small Noise Diffusions“. In: *Communications on Pure and Applied Mathematics* 65 (2012), pp. 1770 –1803 (cit. on pp. 33, 77–79).

[89]A. Voter. „A method for accelerating the molecular dynamics simulation of infrequent events“. In: *The Journal of Chemical Physics* (1997) (cit. on p. 7).

[90]A. Voter. „Paral replica method for dynamics of infrequent events“. In: *Physical Reviews* (1998) (cit. on p. 8).

[91]F. Wang and D. P. Landau. „Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram“. In: *Physical Review E* 64 (5 2001) (cit. on pp. 7, 34).

[92]J. Yang and H. J. Kushner. „A Monte Carlo Method for Sensitivity Analysis and Parametric Optimization of Nonlinear Stochastic Systems“. In: *SIAM J. Control Optim.* 29.5 (1991) (cit. on p. 103).

[93]W. Zhang, H. Wang, C. Hartmann, M. Weber, and Ch. Schütte. „Applications of the cross-entropy method to importance sampling and optimal control of diffusions“. In: *Siam Journal of Scientific Computing* (2014) (cit. on pp. 99, 116, 119, 129).

[94]W. Zhang, C. Hartmann, and Ch. Schütte. „Effective dynamics along given reaction coordinates, and reaction rate theory“. In: *Farraday Discussions* (2016) (cit. on pp. 80, 97, 119).

[95]G. Zou and R. D. Skeel. „Robust variance reduction for random walk methods“. In: *SIAM Journal on Scientific Computing* 25.6 (2004) (cit. on pp. 8, 9, 75).

[96]R. Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, New York, 2001 (cit. on p. 15).

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download of the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich alle Hilfsmittel und Hilfen angegeben habe und dass ich die Arbeit auf dieser Grundlage selbstständig verfasst habe. Ich versichere, dass die Arbeit nicht schon einmal in einem früheren Promotionsverfahren eingereicht wurde.

Berlin, Juli 2018

Jannes Quer