Convergence of Multilevel MCMC methods on path spaces

Dissertation

zur

Erlangung des Doktorsgrades (Dr. rer. nat)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universitt Bonn

vorgelegt von

Daniel Gruhlke

aus Dinslaken

Bonn 2013

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich–Wilhelms-Universität Bonn

Erstreferent:	Prof. Dr. Andreas Eberle
Zweitreferent:	Prof. Dr. Karl–Theodor Sturm

Tag der Promotion: 12.05.2013
Erscheinungsjahr: 2014

Acknowledgements

I would not have finished this thesis without the help of many people and I want to take the opportunity to thank them.
First of all, I want to thank Andreas Eberle for the trust he put into me when accepting me as his student as well as the support during these years. He gave me the freedom to pursue my own ideas but always offered help when I got stucked.
It was a pleasure to work with everyone in the groups of Anton Bovier, Patrik Ferrari, Karl–Theodor Sturm and Stefan Ankirchner. This page is to small to mention everyone, therefore I would like to highlight just a few: Martin Slovik, Shi–Dong Wang and Hendrik Weber for many, many kilometers of mathematical arguments as well as Nikolaus Schweizer with whom I shared the office for most of the time at the institute. The discussions with him were always very fruitful and I could not expect a better atmosphere – a big thank for this, and also to Martina Baar for being a worthy successor. Mei–Ling Wang deserves a special compliment for all the passion she puts into the probability group, including me.

My non–mathematical work at the institute was greatly supported by Casten Rezny, Roland Schregle, Sergio Conti and Patrik Müller. They all do a great job in keeping the infrastructure together and I learned a lot from all of them.

The final words go to my family and especially to Noreen. Without your support and patience there would not exist anything that these acknowledgements could preface.

Abstract

Within this work, the efficiency of Markov Chain Monte Carlo methods on infinite dimensional spaces, such as function spaces, is analyzed. We study two aspects in this respect:
The first aspect is a Multilevel Markov Chain Monte Carlo algorithm. It extends a Multilevel Monte Carlo method introduced by Giles to Markov Chains, and overcomes the need for a trade–off between discretization error and Monte Carlo error. We develop the Multilevel algorithm, state and prove its order of convergence and show results of numerical simulations.
The second part of this work deals with the analysis of the speed of convergence of the Metropolis Adjusted Langevin Algorithm (MALA). Controlling the speed of convergence is an important tool for bounding the error of Markov Chain Monte Carlo methods. It is also a crucial ingredient for bounding the order of convergence of the Multilevel algorithm. We apply a method of Eberle to the Hilbert space case and obtain a subexponential bound on the distance of the distribution of the MALA–process to its invariant measure.
Both aspects are illustrated by an application from molecular dynamics called Transition Path Sampling. In this example, Markov Chain Monte Carlo methods on path spaces are used to simulate the properties of transitions from one metastable state of a molecule to another. We present this application and apply the results on the Multilevel estimator and the MALA–process in this context.

Contents

Chapter 1

Introduction

Since their introduction by Metropolis et al. [35] and Hastings [25], Markov Chain Monte Carlo methods have been applied in a wide range of fields from biology to economics. The main idea of these methods is the approximation of an integral $\mu(f)$ of a function f with respect to a measure μ by generating a Markov Chain $(X_k)_{k\in\mathbb{N}}$ with ergodic measure μ, and using the ergodic average $\frac{1}{N}\sum_{k=1}^{N} f(X_k)$ as an estimator for $\mu(f)$.

This thesis deals with the efficiency of Markov Chain Monte Carlo methods on high– and infinite–dimensional continuous state spaces with an invariant measure μ which is absolutely continuous to a Gaussian measure ν. This setting often arises in different applications. Withhin this thesis, the main application is Transition Path Sampling, which is explained in a later part of this introduction. We analyze the question of efficiency by two means:

The first approach is the development and analysis of a Multilevel Markov Chain Monte Carlo algorithm. This Multilevel algorithm extends an idea of Giles [18] from the Monte Carlo setting to Markov Chains. In case it is applicable, the Multilevel algorithm can leat to a significant speed–up in computational time compared to the classical Markov Chain Monte Carlo.

The second approach is the analysis of the process of the Metropolis Adjusted Langevin Algorithm (MALA) in possibly infinite–dimensional state spaces. This is partially motivated by the analysis of the Multilevel algorithm, as exponentially fast convergence with uniform constants is needed for the Markov Chains used in the algorithm to prove its efficiency. The methods applied here could be tools to proof such bounds. But of course, results on the speed of convergence of Markov Chain Monte Carlo processes are of interest on their

own.

In the next two sections, we introduce Multilevel Markov Chain Monte Carlo and the MALA–process, state our main results, and give references to prior work. The third section describes the application Transition Path Sampling, which is used as running example during this thesis.

Multilevel Markov Chain Monte Carlo

The Multilevel Monte Carlo method of Giles

For solving quadrature problems of stochastic differential equations, the Multilevel Monte Carlo method was introduced in 2008 by Giles [18]. Giles was interested in approximating expectation values with respect to the distribution μ of a stochastic differential equation with Lipschitz coefficients:

$$\mu(f) := \int_E f(x)\mu(\mathrm{d}x).$$

The standard way to approach this kind of problems is via discretization and Monte Carlo simulations. We discretize the function f and the stochastic differential equation, for example with the Euler–Maruyama scheme, to obtain discretizations μ_M, f_M, generate i.i.d. samples $(X_i^M)_{1\leq i\leq N}$ from μ_M and approximate

$$\mu_M(f) \approx \frac{1}{N}\sum_{k=1}^{N} f_M(X_k^M).$$

By applying this procedure, we induce two kinds of errors. The discretization error $|\mu_M(f_M) - \mu(f)|$, and the Monte Carlo error $\left|\mu_M(f_M) - \frac{1}{N}\sum_{k=1}^{N} f_M(X_k^M)\right|$. For a given discretization method, the discetization error is reduced by increasing the dimension of the approximation. The Monte Carlo error is reduced by increasing the sample size N. At this point, we face a trade–off, as sampling usually is more computational expensive in higher dimension, increasing the dimension to reduce the discretization error decreases the number of samples that can be computed in given time, thus it increases the Monte Carlo error.
Giles proposed a Multilevel algorithm to improve the performance of this scheme. The idea

is to use a high–dimensional approximation, but to shift calculations to low–dimensional spaces where the sampling can be performed at low costs. To this end, μ and f are approximated on a sequence of spaces E_i with increasing dimension, and $\mu_M(f_M)$ is decomposed in

$$\mu_M(f_M) = \sum_{i=1}^{M} (\mu_i(f_i) - \mu_{i-1}(f_{i-1})) + \mu_0(f_0).$$

The crucial step in the construction of the Multilevel algorithm is to find low–variance estimators for each $\theta_i := (\mu_i(f_i) - \mu_{i-1}(f_{i-1}))$ for given $i \in \mathbb{N}$. In the case of [18], this is solved via coupling. A coupling $(X_k^i, \tilde{X}_k^i)_{1\leq k\leq N_i}$ is constructed, such that $(X_k^i)_{1\leq k\leq N_i}$ are i.i.d. μ_i—distributed variables and $(\tilde{X}_k^i)_{1\leq k\leq N_i}$ are i.i.d. μ_{i-1}–distributed variables. Furthermore, $\left\|X_k^i - \tilde{X}_k^i\right\|_{E_i}$ is small for large i. Then

$$\mu_i(f_i) - \mu_{i-1}(f_{i-1}) \approx \frac{1}{N_i} \sum_{k=1}^{N_i} f_i(X_k^i) - f_{i-1}(\tilde{X}_k^i) =: \hat{\theta}_i$$

is an unbiased estimator, and by the Central Limit Theorem the mean square error is asymptotically given by

$$\mathbb{E}\left[\left|\theta_i - \hat{\theta}_i\right|^2\right] \approx \frac{1}{N_i} \mathrm{var}\left(f_i(X_1^i) - f_{i-1}(\tilde{X}_1^i)\right).$$

The small distance $\left\|X_i - \tilde{X}_i\right\|_{E_i}$ for large i implies that for Lipschitz–continuous $f_i : E_i \to \mathbb{R}$ the variance is small, too. Therefore, fewer steps N_i are required on the higher levels. Exploiting this mechanism, Giles Multilevel Monte Carlo method manages to improve the order of convergence to $\frac{1}{2}$ with logarithmic corrections.

Multilevel Markov Chain Monte Carlo

We apply the Multilevel idea of Giles to the Markov Chain Monte Carlo setting. The aim is to construct an efficient method to approximate integrals of the form

$$\mu(f) := \int_E f(x)\mu(\mathrm{d}x)$$

for functions $f : E \to \mathbb{R}$ and probability measures μ which are absolutely continuous to a Gaussian measure ν on E of the form

$$\mu(\mathrm{d}x) = \frac{1}{Z} \exp\left(-V(x)\right) \nu(\mathrm{d}x),$$

where Z is an unknown normalization constant. In contrast to the setting of Giles described above, we assume that we can not sample from the distributions μ_i directly. Instead, we are running Markov Chains $(X_k^i)_{k\in\mathbb{N}}$ to approximate these distributions:

$$\mu_i(f) \approx \frac{1}{N_i}\sum_{i=1}^{N_i} f(X_k^i) =: \Theta_i^S(N_i)$$

for large N_i. Here, we face the same trade–off as in the Monte Carlo setting. On the one hand, we need to minimize $|\mu(f) - \mu_M(f_M)|$, which requires a high–dimensional approximation. On the other hand, evaluation of f_M and sampling of $(X_i^M)_{0\leq i\leq N}$ is often more computational expensive in higher dimension, so that for a given amount of resources, fewer steps of the Markov Chain can be simulated in higher dimensions. This leads to a worse approximation of $\left|\mu_M(f_M) - \hat{\Theta}_M^S(N_M)\right|$.

In the remaining part of this section of the work, we develop a Multilevel Markov Chain Monte Carlo method based on the Multilevel Monte Carlo method of Giles, which overcomes this conflict by performing a significant part of the calculations for a high–dimensional approximation in low–dimensional spaces.

As in [18], we decompose the expectation value $\mu_M(f_M)$ into

$$\mu_M(f_M) = \sum_{i=1}^{M}(\mu_i(f_i) - \mu_{i-1}(f_{i-1})) + \mu_0(f_0)$$

It is now the crucial part to find a low variance estimator

$$\hat{\theta}_i \approx \mu_i(f_i) - \mu_{i-1}(f_{i-1})$$

that works in the Markov Chain Monte Carlo setting.

Results

To describe the computational complexity, we introduce the notation for cost, denoted by $\mathrm{cost}(X)$. For a random variable X, $\mathrm{cost}(X)$ models the cost needed for sampling X. In a specific application, this could for example be the computational time required for sampling X. For our analysis, we assume we are in the following situation:

Assumption 1.1.

1. *The space E is approximated by a sequence of subspaces E_i with dimension $d_i = 2^i$. There exist projections $\Pi_i : E \to E_i$ satisfying $\Pi_i \circ \Pi_j = \Pi_i$ for $j \geq i$.*

2. *φ is approximated by a sequence of functions $\varphi_i : E_i \to \mathbb{R}$. Assume $\varphi_0 \equiv 1$. μ_i is defined as $\mu_i(\mathrm{d}x) := \frac{1}{Z_i}\varphi_i(x)\nu_i(\mathrm{d}x)$ where ν_i is the image measure of ν under the projection Π_i: $\nu_i := \nu \circ \Pi_i^{-1}$ and Z_i are normalization constants.*

3. *f is approximated by a sequence of functions $f_i : E_i \to \mathbb{R}$. Assume $f_0 \equiv 0$.*

We construct extensions of f_i and φ_i to functions $f_i : E \to \mathbb{R}$ and $\varphi_i : E \to \mathbb{R}$ by

$$f_i := f_i \circ \Pi_i$$
$$\varphi_i := \varphi_i \circ \Pi_i.$$

In the following, we use the same symbol for the functions f_i, φ_i and their extensions. We also define

$$\tilde{\mu}_i(\mathrm{d}x) := \frac{1}{\tilde{Z}_i}\varphi_{i-1}(x)\nu_i(\mathrm{d}x).$$

We now present the Multilevel Markov Chain Monte Carlo method: First, we define $h_i : E \times E \to \mathbb{R}$ by

$$h_i(x, y) := f_i(x) - f_{i-1}(x)\frac{\varphi_{i-1}(x)}{\varphi_i(x)}\frac{\varphi_i(y)}{\varphi_{i-1}(y)}. \tag{1.1}$$

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and for each $i = 1 \dots M$, let X_0^i, Y_0^i be independent random variables with distribution ν_i on $(\Omega, \mathcal{F}, \mathrm{P})$. Furthermore, let $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}^i$ be two independent Markov Chains on $(\Omega, \mathcal{F}, \mathrm{P})$, starting in X_0^i, Y_0^i, with unique ergodic measure μ_i, $\tilde{\mu}_i$ respectively. Then for N_i, $n_i \in \mathbb{N}$, define the estimators

$$\hat{\theta}_i(n_i, N_i) := \frac{1}{N_i}\sum_{k=0}^{N_i} h_i(X_{n_i+k}^i, Y_{n_i+k}^i),$$
$$\hat{\Theta}_M := \sum_{i=1}^{M} \hat{\theta}_i(n_i, N_i). \tag{1.2}$$

We include two parameters n_i and N_i for each estimator $\hat{\theta}_i$ that can improve the estimation, an increase of N_i takes the average over more states of the Markov Chain, while increasing n_i gives the chain some time to converge to its invariant measure before the averaging is started. Although $\hat{\Theta}_M$ depends on $(N_i)_{i\in\{1,\dots,M\}}$ and $(n_i)_{i\in\{1,\dots,M\}}$, we omit this dependencies in the

notation for convenience.

The Multilevel algorithm relies on two observations: The function h_i is constructed in such a way that its expectation value with respect to $\mu_i \otimes \tilde{\mu}_i$ equals θ_i:

$$\int_{E_i} \int_{E_i} h_i(x,y)\mu_i(\mathrm{d}x)\tilde{\mu}_i(\mathrm{d}y) = \mu_i(f_i) - \mu_{i-1}(f_{i-1}),$$

As the chain $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$ is assumed to be ergodic, with unique ergodic measure $(\mu_i \otimes \tilde{\mu}_i)$, this implies that the estimator $\hat{\theta}_i(n_i, N_i)$ converges to $\mu(f_i) - \mu_{i-1}(f_{i-1})$ P–almost surely as $N_i \to \infty$. Under additional assumptions outline below, the variance of h_i decreases exponentially in i:

$$\mathrm{var}_{\mu_i\otimes\tilde{\mu}_i}(h_i) \lesssim \frac{1}{2^i}.$$

Here and in the following, the notation $a_i \lesssim b_i$ means that for given sequences $(a_i)_{i\in\mathbb{N}}$ and $(b_i)_{i\in N}$, there exists a constant C, such that for all $i \in \mathbb{N}$

$$a_i \leq C b_i.$$

The variance is an important factor in the error estimates for Markov Chain Monte Carlo methods. When the variance decreases like 2^{-i}, the number of steps that a sufficiently well mixing chain requires achieving a given error scales like 2^{-i} in i. This will counteract the fact that operations on higher dimensional spaces are more expensive. In particular, we can assume that the costs for operations on the space E_i can scale linearly with the dimension, and still a given error for approximations of $\mu_i(f_i) - \mu_{i-1}(f_{i-1})$ can be achieved with constant costs independent of the discretization level i. This is crucial for the efficiency of the Multilevel algorithm.

We now present the main theorem of the Multilevel section. It states that under assumptions on the approximations of the density φ and the integrand f, uniform conditions on the speed of convergence to equilibrium of the Markov Chains $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$, the order of convergence of the Multilevel scheme is $\frac{1}{2}$ up to logarithmic corrections.

Theorem 1.1. *Under Assumption 1.1, as well as Assumptions 2.1 – 2.4 (see below), the following statements hold: For given η, $\varepsilon > 0$, there exists $M(\eta,\varepsilon)$, $N_i(\eta,\varepsilon)$, $n_i(\eta,\varepsilon)$ and C, $\eta_0 > 0$ such that for $\eta \leq \eta_0$,*

$$\mathrm{P}\left[|\hat{\Theta}_{M(\eta,\varepsilon)} - \mu(f)| > \eta\right] < \varepsilon,$$

and

$$\operatorname{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) \leq \frac{C}{\eta^2\varepsilon}\log^4\left(\frac{1}{\eta\varepsilon}\right).$$

The full version of this theorem including the explicit form of $M(\eta,\varepsilon)$, $N_i(\eta,\varepsilon)$ and $n_i(\eta,\varepsilon)$ will be given in Chapter 2.

Comparison of the Multilevel MCMC to the Multilevel Monte Carlo approach of Giles

The estimator for $\mu_i(f_i) - \mu_{i-1}(f_{i-1})$ used in the Multilevel Monte Carlo setting is constructed via coupling. We construct a sequence of i.i.d couplings $(X_k^i, \tilde{X}_k^i)_{k\in\mathbb{N}}$, where X_k^i is μ_i–distributed and $\tilde{X}_k^i$ is μ_{i-1} distributed. The coupling is chosen in such a way that the distance between X^i and $\tilde{X}^i$ is small, which implies that the random variable

$$h_i(X_k^i, \tilde{X}_k^i) := f_i(X_k^i) - f_{i-1}(\tilde{X}_k^i)$$

has a small variance for large i.

The main difference compared to our approach is the following: While we want to transfer the Multilevel Monte Carlo method to the Markov Chain Monte Carlo setting, we do not use a coupling approach to construct low–variance estimators for the terms $\mu_i(f_i) - \mu_{i-1}(f_{i-1})$. This has the reason that we did not succeed in doing so. A direct transfer of the coupling idea to the Markov Chain Monte Carlo setting would either require to construct couplings $\pi_i(\mathrm{d}x, \mathrm{d}y)$ of the measures $\mu_i(\mathrm{d}x) = \frac{1}{Z_i}\varphi_i(x)\nu_i(\mathrm{d}x)$ and $\mu_{i-1}(\mathrm{d}y) = \frac{1}{Z_{i-1}}\varphi_{i-1}(y)\nu_{i-1}(\mathrm{d}y)$ which is concentrated near the diagonal, and to construct processes which are reversible with respect to π_i. Or we at least need to construct a processes $(X_k^i, \tilde{X}_k^i)_{k\in\mathbb{N}}$ such that their marginals $(X_k^i)_{i\in\mathbb{N}}$, $(\tilde{X}_k^i)_{i\in\mathbb{N}}$ are reversible with respect to μ_i, μ_{i-1} respectively, and $d(X_k^i, \tilde{X}_k^i)$ is small for an appropriate metric d.

In neither of these approaches, we managed to construct a coupling whose variance decreases fast enough to be useful for a Multilevel estimator. Instead, we use the importance sampling technique as seen in (1.1). For this purpose, the expectation value of f_{i-1} with respect to μ_{i-1} is expressed as an expectation value with respect to μ_i by

$$\begin{aligned}\mu_{i-1}(f_{i-1}) &:= \frac{Z_i}{Z_{i-1}}\mu_i\left(f_{i-1}\frac{\varphi_{i-1}}{\varphi_i}\right)\\ &= \tilde{\mu}_i\left(\frac{\varphi_i}{\varphi_{i-1}}\right)\mu_i\left(f_{i-1}\frac{\varphi_{i-1}}{\varphi_i}\right).\end{aligned}$$

Therefore, ignoring the normalization constants for now, we can use the Markov Chain $(X_k^i)_{k\in\mathbb{N}}$ to estimate both $\mu_i(f_i)$ and $\mu_{i-1}(f_{i-1})$. The chain $(\tilde{X}_k^i)_{k\in\mathbb{N}}$ accounts for the quotient of the normalization constants and is chosen to be independent of $(X_k^i)_{k\in\mathbb{N}}$.
While this approach fulfills the desired low variance characteristics, it has some disadvantages compared to the coupling method. In order for the Multilevel method to be efficient, the function h_i has to have a small variance with respect to $\mu_i \otimes \tilde{\mu}_i$, which requires the term $\frac{\varphi_i}{\varphi_{i-1}}$ to be close to 1. This implies that $(\mu_i)_{i\in\mathbb{N}}$ has to converge in total variation sufficiently fast. In contrast, for the coupling approach, we only need $\left|f_i(X^i) - f_{i-1}(\tilde{X}^i)\right|$ to be small, which can be fulfilled if $\left\|X^i - \tilde{X}^i\right\|_E \to 0$ sufficiently fast for $i \to \infty$ on some space E, and the functions f_i converge with respect to the same norm.

Overview on further work on Multilevel Monte Carlo methods

The Multilevel Monte Carlo method has been widely applied. Even before the work of Giles, Heinreich [26] used a Multilevel to compute solution of integral equations, and Heinreich and Sindambiwe [28] for parametric integration. An overview on these applications can be found in Heinreich [27].
In [10], Creutzig, Dereich, Müller–Gronbach and Ritter analyzed the integration of Gaussian measures on Banach spaces, and the Multilevel Monte Carlo method in particular, from the complexity theoretical point–of view an derived lower bounds on the complexity. This is also surveyed by Müller–Gronbach and Ritter in [36]. By using the Milstein scheme to discretize the stochastic differential equation, Giles [19] proposed a modification of his Multilevel algorithm for stochastic differential equations, which improved the order of convergence to $\frac{1}{2}$ without logarithmic corrections.
Hutzenthaler, Jentzen, Kloeden [30] analyzed Giles' Multilevel scheme for stochastic differential equations with non–Lipschitz coefficients and realized that the scheme is divergent in this case. The effect is related to the fact that the Euler scheme with fixed time discretization is eventually divergent if the considered time–horizon is large enough. The Multilevel algorithms uses large sample sizes on low–dimensional discretization levels. Therefore, eventually one of these samples will diverge. Hutzenthaler, Jentzen, Kloeden showed that this happens at such a high rate that the algorithm is divergent. They proposed a modification

to the algorithm by truncating the discretization of the drift term. This modified Multilevel algorithm also works for stochastic differential equations with non–Lipschitz coefficients.
There have been other approaches for an extension of the Multilevel idea of Giles to the Markov Chain Monte Carlo setting, like in the paper of of Hoang, Schwab and Stuart [29]. They work in a setting where a coupling of the Markov Processes can be performed. Basically, the decomposition of $\mu_M(f_M)$ chosen is

$$\mu_M(f_M) \approx \mu_{\frac{M}{2}}(f_{\frac{M}{2}}) \sum_{i=\frac{M}{2}}^{M} \mu_i(f_i) - \tilde{\mu}_i(f_{i-1}), \tag{1.3}$$

where $\tilde{\mu}_i$ is the projection of μ_i to E_{i-1}. This allows to construct a coupling $(X_k^i, \tilde{X}_k^i)_{k\in\mathbb{N}}$ reversible with respect to $(\mu_i, \tilde{\mu}_i)$ by constructing $(X_k^i)_{k\in\mathbb{N}}$ reversible with respect to μ_i, and obtaining $(\tilde{X}_k^i)_{k\in\mathbb{N}}$ as the projection of $(X_k^i)_{k\in\mathbb{N}}$ to E_{i-1}. This procedure introduces a new error $\mu_i(f_i) - \tilde{\mu}_i(f_i)$ on each level. In the setting of [29], this error decreases with order 2^{-2i} for approximations on spaces E_i whose dimension increase like 2^i. As the decomposition is started on level $\frac{M}{2}$, the total error introduced by not having exactly the same measures in the telescopic sum in (1.3) is negligible. For our Multilevel algorithm, we only assume that the error decrease with a factor of $2^{-\frac{i}{2}}$ on 2^i–dimensional spaces, and thus we can not follow this procedure.
Finally, we want to note that there is a long history of Multilevel methods in the Markov Chain Monte Carlo literature, starting with Goodman and Sokal [20], who developed a Multilevel method for sampling spin systems. Simulated Tempering and Parallel Tempering [17, 33] can also be considered Multilevel algorithms, although their method is not a reduction of the dimension but a modification of the density in order to overcome bottlenecks of the distribution. All these methods improve the rate of convergence of a Markov chain Monte Carlo scheme by using multiple levels, in [20] by reducing the dimensionality, exploiting that the lower dimensional chains have better mixing properties, and in Simulated Tempering or Parallel Tempering by varying the measure which leads to faster converging chains. While all these methods also uses multiple levels of approximations, this is done to improve the speed of convergence of the Markov Chain to its invariant measure. Our goal is a different one, we already are in a situation where the chain is rapidly mixing, but nevertheless the mentioned conflict between low discretization error and high computational

costs for high dimensional approximation arises. We use the different levels to mitigate this conflict and improve the order of convergence of the estimator.

Speed of Convergence of the MALA–process

One central assumption that is required for the proof of Theorem 1.1, that shows the efficiency of the Multilevel algorithm, is a uniform spectral gap of the applied Markov Processes. This is needed to guarantee an exponentially fast speed of convergence of the used chains. The second part of this thesis analyzes the speed of convergence of Markov Chain Monte Carlo processes in high–dimensional state spaces as they arise when constructing the Multilevel Markov Chain Monte Carlo estimator.

The speed of convergence of Markov Chains on high– and even infinite–dimensional state spaces has been attracted attention since several years. For finite–dimensional state spaces, a good overview can be found in Roberts, Rosenthal [38]. The first steps to the infinite–dimensional state spaces were the works of Roberts, Gelman and Gilks [37] and Roberts, Rosenthal [39]. They studied the Random Walk Metropolis (RWM) and the Metropolis adjusted Langevin algorithm (MALA) with product measure targets, and developed optimal scaling results as the dimension increases to infinity.

Roberts, Gelman and Gilks [37] and Roberts, Rosenthal [39] analyzed the speed of the diffusion limits of one–dimensional marginals of the RWM– and MALA–process on $\mathbb{R}^d$. The result was that, in order to have non–zero speed, the step–sizes of the RWM– and MALA–process have to be scaled with order $\mathcal{O}(d^{-1})$ and $\mathcal{O}(d^{-\frac{1}{3}})$ respectively. In addition, they characterized the optimal acceptance rate of these processes to be 0.574 and 0.234 respectively in the scaling limit for product measure targets. Afterwards, these acceptance rates were used as heuristics to optimize the speed of Markov Chains, see Bédard, Rosenthal [2]

In more recent years, these scaling results have been extended to non–product measures. The focus here was shifted to measures that are absolutely continuous with respect to Gaussian measures. The scaling limit results from the product case where extended by Beskos, Roberts, Stuart [3]. These methods also require the step size of the process to converge to zero as the dimension increases, in order to obtain a non–zero acceptance probability

in the limit. By using a semi–implicit discretization of the Langevin s.d.e. as proposal for the Markov chain, Beskos, Stuart, Roberts, Voss [8] and Beskos, Stuart [4] constructed non–degenerating processes, which allow strictly positive step–sizes as the dimension of the state space converges to infinity. This allowed Mattingly, Pillai, Stuart [34] to define the corresponding MALA–process directly on the infinite–dimensional Hilbert space. For step–sizes h converging to 0, they showed weak convergence of the MALA–process to the corresponding stochastic differential equation.

Analysis of the speed of convergence of discrete time processes on high– or infinite–dimensional state spaces has been carried out by Bou–Rabée, Hairer, Vanden–Eijnden [9], who proved exponentially fast convergence to equilibrium of the MALA–process by comparing it to the continuous time diffusion limit and using established convergence results for that limit. However, the result depends on the dimension of the state space and thus does not scale to the infinite–dimensional setting. An infinite–dimensional result was achieved by Hairer, Vollmer, Stuart [22]. They analyzed the RWM–algorithm and obtained exponential convergence even without assuming log–concavity of the measure in this case. However, the result is rather non–quantitative with non–explicit constants in the bounds on the rate of convergence.

There are lots of works studying the speed of convergence of Langevin diffusion to its equilibrium measure, see Roberts, Tweedie [40], Roberst, Rosenthal [39] or da Prato, Zabczyk [13]. In general, one can expect exponentially fast convergence if the equilibrium measure is log–concave. However, the comparison of the Markov Chain Monte Carlo methods and their diffusion limits must be handled with care. Convergence of the process to a diffusion is usually only known for fixed time horizons, while results for the speed of convergence of Markov Chain Monte Carlo methods are asymptotic results for time to infinity. Furthermore, the Markov Chain Monte Carlo process converge to the diffusion limit only as the step–size converges to 0. For efficient sampling, one prefers larger step–sizes to improve the mixing properties of the chain. Thus, the heuristic implied from results on the diffusion limits could be misleading when being directly applied to Markov Chain Monte Carlo methods, and explicit analysis of this methods is needed when one wants to state results on their efficiency.

In this work, we study the MALA–process on a Hilbert space W in the semi–

implicit form. This process is constructed to have invariant measure μ given by

$$\mu(\mathrm{d}x) := \frac{1}{Z} \exp\left(-V(x)\right) \nu(\mathrm{d}x),$$

where ν is a Gaussian measure on W with mean 0 and covariance operator $\mathcal{C}$ and the potential V is a function $V : E \to \mathbb{R}$. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and $(N_i)_{i\in\mathbb{N}}$ be a i.i.d. sequence of ν–distributed random variables, and $(U_i)_{i\in\mathbb{N}}$ be an i.i.d sequence of uniform distributed variables on $[0,1]$ independent of $(N_i)_{i\in\mathbb{N}}$ on $(\Omega, \mathcal{F}, \mathrm{P})$. Define the scalar product $\langle\cdot,\cdot\rangle_S$ as

$$\langle x, y\rangle_S = \left\langle \mathcal{C}^{-\frac{1}{2}} x, \mathcal{C}^{-\frac{1}{2}} y \right\rangle_W,$$

and let S be the Hilbert space

$$S = \{x \in W | \|x\|_S < \infty\}.$$

Given $x \in W$ and $h \in (0,2)$, the proposal $Y_{h,n}(x)$ is given by

$$Y_{h,n}(x) := \left(1 - \frac{h}{2}\right) x - \frac{h}{2} \nabla_S V(x) + \sqrt{\tilde{h}} N_{n+1}.$$

Here $\tilde{h} := h - \frac{h^2}{4}$. Let $X_0 \in W$ be a random variable, and let P_x be the probability measure $\mathrm{P}[\cdot | X_0 = x]$ for $x \in W$. The MALA–process is now constructed as follows: At time $n+1$, the state $Y_{h,n}(X_n)$ is proposed. It is accepted as new state of the process with acceptance probability $a(X_n, Y_{h,n}(X_n))$; if it is rejected, the state does not change:

$$X_{n+1} := \begin{cases} Y_{h,n}(X_n) & \text{if } U_{n+1} < a(X_n, Y_{h,n}(X_n)), \\ X_n & \text{otherwise.} \end{cases} \tag{1.4}$$

The acceptance probability is chosen is such a way that $(X_n)_{n\in\mathbb{N}}$ is reversible with respect to ν. The exact form is derived in Chapter 3.

The choice of the proposal corresponds to the ones in Beskos and Stuart [4], and Beskos, Roberts, Stuart and Voss [8]. It has the advantage that the acceptance probability remains positive for fixed $h > 0$ even in the infinite–dimensional limit. This is mainly due to the fact that in the Gaussian case $V \equiv 0$, the Gaussian measure ν is reversible with respect to the kernel induced by the proposal $Y_{h,n}(x)$. This is not true for other proposals like in the original RWM or MALA, whose proposals are singular with respect to ν in infinite

dimensions.

As a Markov Process, $(X_i)_{i\in\mathbb{N}}$ can be described by its kernel $q_h : W \times \mathcal{B}(W) \to [0,1]$, where $\mathcal{B}(W)$ denotes the Borel sets on W. It is defined by

$$q_h(x)(A) := \mathrm{P}_x\left[X_1 \in A\right] \qquad \text{for } x \in W,\ A \in \mathcal{B}(W).$$

Our aim of this thesis is to bound the speed of convergence of the MALA–process to its invariant measure μ, if said measure fulfills log–concavity. The distance of the distribution to its invariant measure is measured in the Wasserstein distance. Given a metric $d : W \times W \to [0,\infty]$, the Wasserstein distance $\mathcal{W}_d : \mathcal{P}(W) \times \mathcal{P}(W) \to [0,\infty)$ is defined by

$$\mathcal{W}_d(\eta,\tilde{\eta}) := \inf_{\pi} \int_{W\times W} d(x,\tilde{x})\mathrm{d}\pi(x,\tilde{x}),$$

where the infimum is taken over all couplings π of η and $\tilde{\eta}$. We will apply this to the metrics d_R given by

$$d_R(x,y) = \|x-y\|_W \wedge R \qquad \text{for } x,y \in W,\ R \in [0,\infty).$$

We use two assumptions on the potential V to prove our result. First, we need a fixed bound on the second derivative of V.

Assumption 1.2. *There exists a constant $0 \le L < 1$ such that*

$$\|\nabla_S V(x) - \nabla_S V(y)\|_W \le L\|x-y\|_W. \tag{1.5}$$

Furthermore, we need polynomial bounds on the first four derivatives of the potential.

Assumption 1.3. *The potential V is four times differentiable with respect to W, and constants C_n, $p_n \in [0,\infty)$ exist, such that the derivatives, as operators from $W^{\otimes n}$ to $\mathbb{R}$ are bounded by a polynomial:*

$$|\mathrm{D}^n V(x)(\xi_1,\ldots,\xi_n)| \le C_n \max\{1,\|x\|_W\}^{p_n},$$

for all $x \in W$, $\|\xi_1\|_W = \ldots = \|\xi_n\|_W = 1$, and $n \in \{2,3,4\}$.

Our main result states a subexponential bound on the distance of the distribution of the MALA–process to its invariant measure μ, if Assumptions 1.2 and 1.3 are satisfied.

Theorem 1.2. *Let q_h be the kernel of the MALA–process with step-size $h \in (0,2)$. Let Assumptions 1.2 and 1.3 be satisfied. Then there exists $C > 0$, $r > 0$ and $n_0 > 0$ such that for given $n \geq n_0$, there exists $h(n) > 0$ such that*

$$\mathcal{W}_{d_1}(\nu q^n_{h(n)}, \mu) \leq \exp\left(-cn^{\frac{2}{1+r}}\right)(\mathcal{W}_{d_\infty}(\mu,\nu) + C).$$

The constant r depends on the degree of the polynomial bounds of the derivatives of the potential V, and is given by (3.33).

The idea of the proof is based on the construction of a coupling of the MALA–process. We will show that it is contracting with constant $\gamma(R) < 1$ on a ball with radius R with respect to a suitable metric. The contraction property is proven by showing that the proposal of the MALA–process is contracting. The contraction rate of the MALA–process itself is then controlled via an asymptotic analysis of the rejection probabilities.
As outlined in Eberle [16], this leads to an estimate of the Wasserstein distance of the distribution of the MALA–process to its invariant measure of the form

$$\mathcal{W}_d(\mu q^n, \nu q^n) \leq \gamma^n \mathcal{W}_d(\mu,\nu) + \frac{R}{1-\gamma}(C_n(B_R,\mu) + C_n(B_R,\nu)), \tag{1.6}$$

where $C_n(B_R,\mu)$ is the highest probability that the process, started with distribution μ, is not in the ball with radius R after i steps, and i runs from 1 to n. In Section 3.3.3, we show that the second term is exponentially small in R, such that the theorem can be proven by bound (1.6), and the correct choice of $R(n)$ and $h(n)$.

The chapter 3 is a generalization of the work of Eberle [16], who proved a convergence result for the MALA–process via coupling methods on finite–dimensional state spaces. However, the result is independent of the dimension, and the proof can be extended to infinite–dimensional state spaces quite directly, so that the main ideas and calculations were adopted.
Note that the bound we obtain here is only sub–exponential and thus does not suffice to prove the exponential bound needed for the Multilevel algorithm. This is also expected for the MALA–process, as already for the one–dimensional setting it is known that MALA–process in a polynomial potential does not converge exponentially fast, see Roberts and Tweedie [40]. However, the analysis in this work can be seen as a step towards a better understanding of Markov Chain Monte Carlo settings, and might, applied to another Markov

Process eventually leads to an exponential bound.

Transition Path Sampling

As announced, we use Transition Path Sampling throughout this work as visualization. In this chapter, we introduce the general definitions of this setting. This thesis deals with efficient integration with respect to and sampling from a measure μ of the type

$$\mu(\mathrm{d}x) = \frac{1}{Z}\exp(-V(x))\nu(\mathrm{d}x),$$

where ν is a Gaussian measure and the potential V is a function $V : E \to \mathbb{R}$. There are plenty of applications where one is interested in this type of measures, including models in molecular dynamics, signal processing and data assimilation. See e.g. Beskos and Stuart [5] for an overview.

Our main motivation arises from Transition Path Sampling. Transition Path Sampling is a technique used in biological chemistry, especially in molecular dynamics. Within this area, the transition of a molecule from one metastable state to another is analyzed. The main idea behind Transition Path Sampling is the following: Consider the dynamics of a molecule modelled by a stochastic differential equation in $\mathbb{R}^d$

$$\mathrm{d}X_t = g(X_t)\,\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t, \tag{1.7}$$

$$X_0 = x_0 \in \mathbb{R}^d, \tag{1.8}$$

where B_t is a d–dimensional Brownian Motion, $g : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ are functions controlling the dynamic. We are interested in transitions of the molecule from the state x_0, or some metastable region $A_0 \ni x_0$, to the state x_1, or the metastable region $A_1 \ni x_1$. Typically, these transitions occur very rarely, and happen very fast when they do. This means that very small time–steps are needed within a simulation of the solution of the stochastic differential equation, to resolve the transition we are interested in. Moreover, the process would spent most of the time in the meta–stable regions, and very little time in the transition regions we are interested in. So most of the computational time is wasted, as only a tiny fraction of the path captures the relevant behaviour.

To find a workaround, we consider paths which are conditioned on performing the wanted transition in a given time frame. Therefore, the solution of the stochastic differential equation as presented above is conditioned on the event $\{X_T = x_T\}$. Here, $x_0 \in \mathbb{R}^d$ represents one metastable configuration and $x_T \in \mathbb{R}^d$ represents the other one. This strategy was first proposed by Dellago, Bolhuis, Chandler [14]. Since then, the technique is widely used, see Dellago, Bolhuis [15] for a survey on the topic.

Mathematically, the easiest case to analyze is the so called "gradient case" when $g = -\nabla f$ for some smooth function $f : \mathbb{R}^d \to \mathbb{R}$, and $\sigma \equiv 1$. This is a strong limitation, as many important examples, including equations arising from Newton's law of motion, are not captured in this setting. However, the gradient case is often assumed, as the density of the measure is smooth in this case.
The first rigorous mathematical study of the processes arising in the (gradient case) Transition Path Sampling setting was done by Hairer, Stuart, Voss, Wiberg [23], [24]. They showed that μ is absolute continuous with respect to a Brownian Bridge on $\mathbb{R}^d$ with density

$$\varphi(x) := \frac{1}{Z} \exp\left(- \int_0^1 \Phi(x_s) \mathrm{d}s \right),$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\Phi(z) := \frac{1}{2} \left(\Delta f(z) + |\nabla f(z)| \right).$$

Additionally, they constructed solutions to the stochastic partial differential equations

$$\begin{aligned} \partial_t z &= \partial_u^2 z - \nabla \Phi(z) + \sqrt{2} \partial_t w \\ z(t,0) &= x_0, \qquad z(t,1) = x_1, \\ z(0,t) &= z_0(t) \end{aligned} \tag{1.9}$$

and

$$\begin{aligned} \partial_t z(t,u) &= -z(t,u) + y(t,u) + \sqrt{2} \partial_t \tilde{w}(t,u) \\ \partial_u^2 y &= \nabla \Phi(z) \\ y(t,0) = x_0 \quad y(t,1) &= x_1, \\ z(0,u) &= z_0(u). \end{aligned} \tag{1.10}$$

and showed that μ is the unique ergodic measure of the solutions of (1.9) and (1.10). The first equation is called the "non–preconditioned equation", the second one the "preconditioned equation" in these papers. The MALA–process as analyzed in this thesis can be considered as a time–discretization of equation (1.10) which preserves the invariant measure. We briefly discuss the implication of using discretizations of other equations like (1.9) or intermediate equations in Section 3.5.

The non–gradient case is significantly more difficult to analyze mathematically, because the density is much more irregular in this case. Hairer, Stuart, Voss [22] studied the fourth–order stochastic partial differential equation

$$\partial_t z(t,u) = (\partial_u^2 - m^2\partial_u^4)z(t,u) + N(z)(t,u) + \sqrt{2}\partial_t w(t,u).$$

The invariant measure of its solution is characterized as

$$m\ddot{x}(t) = f(x(t)) - \dot{x}(t) + \dot{w}(t) \tag{1.11}$$

conditioned on $x(0) = x_0$ and $x(1) = x_1$, when $N(z)(t,u)$ is chosen in the right way. Here, the additional smoothing of the fourth order differential operator guarantees the existence of the solution. Formally, the stochastic differential equation conditioned on its endpoint is given by the limit $m \to 0$ in (1.11). This relation is analyzed in Hairer [21]. They showed that for this particular equations, the limit of the solution of (1.11), as m converges to 0, is indeed a solution of the conditioned stochastic differential equation.

This measure is our main motivating example, and it will reoccur during the following chapters. In particular, the results on the Multilevel algorithm and the speed of convergence of the MALA–process are applied in the context of Transition Path Sampling.

We are going to apply the results on the Multilevel algorithm and on the speed of convergence of the MALA–process to the Transition Path Sampling setting.

Results related to Transition Path Sampling

In Chapter 2, we apply the Multilevel algorithm to the Transition Path Sampling setting. Under conditions on the drift term g in equation (1.7), we show that the order of the Multilevel algorithm convergence, in the sense of Theorem 1.1, is $\frac{1}{2}$.

Theorem 1.3. *Let μ, Φ and $(X_k^i)_{k\in\mathbb{N}}$, $(Y_k^i)_{k\in\mathbb{N}}$ be defined as constructed above. Let $f : C_0([0,T],\mathbb{R}^d) \to \mathbb{R}$ be given. Assume that for constants $c, L > 0$,*

$$\begin{aligned} |f(x) - f(x)| &\le L\|x-y\|_{L^q([0,T],\mathbb{R}^d)} && \text{for all } x, y \in C([0,T],\mathbb{R}^d) \\ \operatorname{cost}(f_i(\xi)) &\lesssim 2^i + \operatorname{cost}(\xi), \\ |\Phi(u) - \Phi(v)| &\le L\|u-v\|_{\mathbb{R}^d} && \text{for all } u, v \in \mathbb{R}^d \\ c^{-1} &\le \Phi(u) \le c && \text{for all } u \in \mathbb{R}^d. \end{aligned}$$

Then the Multilevel estimator $\hat{\Theta}_{M(\eta,\varepsilon)}$ constructed in (1.2) satisfies

$$\mathrm{P}\left[|\hat{\Theta}_{M(\eta,\varepsilon)} - \mu(f)| > \eta\right] < \varepsilon,$$

and

$$\operatorname{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) \le \frac{C}{\eta^2\varepsilon}\log^4\left(\frac{1}{\eta\varepsilon}\right).$$

In Chapter 3, we analyze the MALA–process in the Transition Path Sampling context. We show that its speed of convergence is bounded under growth conditions on the derivatives of Φ. We assume that Φ and its derivatives are bounded by a polynomial.

Assumption 1.4. *For all $\eta^1, \dots, \eta^n \in \mathbb{R}^d$ with $\|\eta^i\|_{\mathbb{R}^d} = 1$ for $i \in \{1, \dots, n\}$,*

$$\left|\mathrm{D}^n\Phi(z)(\eta^1, \dots, \eta^n)\right| \le C_n \left(\max\{1, \|z\|_{\mathbb{R}^d}\}\right)^{p_n} \tag{1.12}$$

for $n = 1, \dots, 4$ and constants C_n and p_n.

The second assumption deals with a uniform bound on the second derivative on Φ.

Assumption 1.5. *The second derivative of Φ is uniformly bounded by $L_\Phi < \frac{\pi}{\sqrt{2}}$: For all $z \in \mathbb{R}^d$, and all η^1, η^2 in $\mathbb{R}^d$*

$$\left|\mathrm{D}^2\Phi(z)(\eta^1, \eta^2)\right| \le L_\Phi \|\eta^1\|_{\mathbb{R}^d}\|\eta^2\|_{\mathbb{R}^d}.$$

Theorem 1.4. *Let q_h be the kernel of the process $(X_n)_{n\in\mathbb{N}}$ with step–size h as constructed in (1.4) for the measure μ. Let Assumptions 1.4 and 1.5 be satisfied. Then for given $n \in \mathbb{N}$, there exist $h(n) > 0$ and constants c, C and r such that*

$$\mathcal{W}_{d_1}(\nu q_{h(n)}^n, \mu) \le \exp\left(-cn^{\frac{1}{1+r}}\right)\left(\mathcal{W}_{d_\infty}(\mu,\nu) + C\right),$$

We also show a uniform bound on the speed of convergence of the MALA–process for a sequence of finite–dimensional discretizations of μ. Define $d_N := 2^{N-1}$, let E_N be the piece–wise linear functions on the partition $\left\{0, \frac{1}{d_N}, \dots, 1\right\}$, and $\Pi_N : E \to E_N$ be the projection to E_N. Consider the measures μ_N given by

$$\mu_N(\mathrm{d}x) := \frac{1}{Z_N} \exp\left(\frac{1}{d_N} \sum_{k=1}^{d_N} \Phi\left(x_{\frac{k}{d_N}}\right)\right) \nu_N(\mathrm{d}x), \tag{1.13}$$

where ν_N is the image measure of ν under Π_N. Let $(X_i^N)_{i\in\mathbb{N}}$ be the MALA–process constructed according to (1.4) and $q_{N,h}$ its kernel.
Applying the results on the MALA–process on Hilbert spaces, we obtain the following uniform result on distance to equilibrium of the distribution of the MALA–process:

Theorem 1.5. *Let $q_{N,h}$ be the kernel of the process $(X_n^N)_{n\in\mathbb{N}}$ with step–size h for the measure μ_N given by (1.13). Let Assumptions 1.4 and 1.5 be satisfied. Additionally assume $L_\Phi \leq \frac{\pi}{3}$. Then for given $n \in \mathbb{N}$, there exists $h(n) > 0$ and constants c, C and r such that*

$$\mathcal{W}_{d_1}(\nu(q_{N,h(n)}^n), \mu) \leq \exp\left(-cn^{\frac{1}{1+r}}\right) (\mathcal{W}_{d_\infty}(\mu, \nu) + C),$$

The constants are independent of N.

Again, we will specify the constants in later chapters.

Organization of the thesis

The remaining parts of this thesis are organized in the following way.

Chapter 2 is devoted to the Multilevel Markov Chain Monte Carlo estimator. In Section 2.1 the estimator is introduced, and theorem 1.1 as well as the assumptions needed to prove it are stated. Within Chapter 2.2, basic results that are implied by the assumptions are proven. These are used in Section 2.3 to prove Theorem 1.1 on the order of convergence of the Multilevel estimator. In Section 2.4, the Multilevel estimator in the context of Transition Path Sampling is analyzed and Theorem 1.3 is proven. In Section 2.5, numerical examples are presented which compare the performance of the Multilevel Markov Chain Monte Carlo method to classical Markov Chain Monte Carlo alogrithms.

Chapter 3 deals with the speed of convergence of the MALA–process on high– or infinite–dimensional spaces. In Section 3.1, the Hilbert–space valued MALA–process is constructed. The Transition Path Sampling version of this process is covered as an example in Section 3.2. Within Section 3.3 a bound for the speed of convergence of the process is derived and Theorem 1.2 is proven. In Section 3.4, this theorem is applied to the Transition Path Sampling setting and Theorems 1.4 and 1.5 are proven. Finally, in Section 3.5, we analyze different possible choices of Metropolis Chains that are reversible with respect to μ, and show that these choices would not lead to a contracting process.

Chapter 2

Multilevel Markov Chain Monte Carlo

2.1 Setting

Let $\mu : \mathcal{B}(E) \to [0,1]$ be a probability measure on a seperable Banach space E with Borel sets $\mathcal{B}(E)$. Assume μ is absolutely continuous with respect to a Gaussian reference measure ν on E with density φ known up to a normalization constant Z:

$$\mu(\mathrm{d}x) = \frac{1}{Z}\varphi(x)\nu(\mathrm{d}x).$$

Let $f \in L^1(E,\mu)$ be a integrable function. We are interested in approximations of the integral

$$\mu(f) := \int_E f(x)\mu(\mathrm{d}x).$$

In 2008, Giles [18] introduced the Multilevel Monte Carlo method to improve the order of convergence of Monte Carlo estimators in the infinite–dimensional setting, for example when μ is the distribution of the solution of a stochastic differential equation. A survey by Müller–Gronbach and Ritter on Multilevel Monte Carlo can be found in [36]. Giles was interested in expectation values with respect to distributions of solutions of stochastic differential equations. He noticed that in the common approach, in which we discretize the s.d.e. with (say) the Euler–Maruyama scheme and generate independent samples of this discretized equation, we face two opposing effects: While higher–dimensional approximations lead to a smaller discretization error, sampling of the distribution gets more

expensive. The Multilevel method of Giles overcomes this issue by transforming parts of the calculations to low–dimensional spaces. This improves the order of convergence from $\frac{1}{3}$ to $\frac{1}{2}$ with logarithmic corrections. However, the method relies on the possibility to draw i.i.d. samples of the measures μ_n to approximate μ.
In many application, it is not possible to generate i.i.d. samples, and Markov Chain Monte Carlo methods are used. A wide range of application with infinite–dimensional target measures is presented e.g. in [5], including problems in signal processing, geophyics and molecular dynamics. In principle, MCMC methods for infinite–dimensional targets face the same effect as Monte Carlo methods: Higher–dimensional approximations improve the discretization error, but sampling of the Markov Chain is more expensive in higher dimensions. This raises the question if, and how, the Multilevel method can be modified to work in the Markov Chain Monte Carlo setting. This work gives a first approach to answer this question.

We assume that we have a sequence of approximating spaces $E_i \subset E$ as well as approximation $\varphi_i : E_i \to \mathbb{R}$, $f_i : E_i \to \mathbb{R}$, and ν_i on E_i for $i \in \mathbb{N}$. We further assume that $f_0 \equiv 0$ and that ν_i is given as image measure of ν under a sequence of projections: There exist continuous projection operators $\Pi_i : E \to E_i$ from E to E_i, that is for $j \geq i$, $\Pi_i \circ \Pi_j = \Pi_i$, as well as $\Pi_i x_i = x_i$ for $x_i \in E_i$, and ν_i is given by

$$\nu_i(A) := \nu(\Pi_i^{-1}(A)) \qquad \text{for } A \in \mathcal{B}(E_i).$$

To simplify the notation later on, we extend functions $g_i : E_i \to \mathbb{R}$ to functions on E by defining

$$g_i(x) := g_i(\Pi_i(x)) \qquad \text{for } x \in E.$$

We use the same symbol for the function and its extension. We define the probability measure $\tilde{\mu}_i$ by

$$\tilde{\mu}_i(\mathrm{d}x) := \frac{1}{\tilde{Z}_i}\varphi_{i-1}(x)\nu_i(\mathrm{d}x).$$

where $\tilde{Z}_i$ is the normalization constant. This measure is chosen in such a way that its marginal on E_{i-1} is μ_{i-1}.
Finally, for $i \in \mathbb{N}$, let $(X_k^i)_{k\in\mathbb{N}}$ be a Markov Chain on E_i that is reversible with respect to μ_i.

As in the Monte Carlo setting of Giles introduced in Chapter 1, we decompose the expectation $\mu_M(f_M)$ in

$$\begin{aligned}\mu_M(f_M) &= \sum_{i=1}^{M} (\mu_i(f_i) - \mu_{i-1}(f_{i-1})) + \mu_0(f_0) \\ &= \sum_{i=1}^{M} (\mu_i(f_i) - \mu_{i-1}(f_{i-1}))\end{aligned}$$

as $f_0 \equiv 0$. We will construct estimators $\hat{\theta}_i$ for $\theta_i := \mu_i(f_i) - \mu_{i-1}(f_{i-1})$. Then we define

$$\hat{\Theta}_M := \sum_{i=1}^{M} \hat{\theta}_i$$

as an estimator for $\mu_M(f_M)$. An important point in the Multilevel approach is the construction of an estimator with low variance on high–dimensional spaces E_i. When the variance decreases like 2^{-i}, the number of required steps of the chain to achieve a given error also scales like 2^{-i} in i. This allows us to find estimators with a given error with identical costs compared to the low–dimensional spaces, despite the fact that sampling is more expensive in high dimensions. In this work, this is achieved by performing fewer steps of the Markov Chain on higher–dimensional spaces. Nevertheless, we can obtain a small error, if we have an integrand with small variance. To obtain this, we define $h_i : E \times E \to \mathbb{R}$ by

$$h_i(x, y) := f_i(x) - f_{i-1}(x) \frac{\varphi_{i-1}(x)}{\varphi_i(x)} \frac{\varphi_i(y)}{\varphi_{i-1}(y)}. \tag{2.1}$$

The function h_i is constructed in a way that its expectation value with respect to $\mu_i \otimes \tilde{\mu}_i$ equals θ_i:

Lemma 2.1. *For $h_i : E_i \times E_i \to \mathbb{R}$ defined in (2.1),*

$$\int_{E_i} \int_{E_i} h_i(x, y) \mu_i(\mathrm{d}x) \tilde{\mu}_i(\mathrm{d}y) = \mu_i(f_i) - \mu_{i-1}(f_{i-1})$$

holds.

Additionally, under assumptions outlined below, the variance of h_i decreases exponentially:

Lemma 2.2. *Under Assumptions 2.2 and 2.3, for μ_i, $\tilde{\mu}_i$ and h_i defined above, there exists $C < \infty$ independent of i such that*

$$\mathrm{var}_{\mu_i \otimes \tilde{\mu}_i}(h_i) \leq C \frac{1}{2^i}.$$

Finally, we define the estimators for θ_i and Θ_M: For N_i, $n_i \in \mathbb{N}$:

$$\hat{\theta}_i := \frac{1}{N_i} \sum_{k=0}^{N_i} h_i(X^i_{n_i+k}, Y^i_{n_i+k}),$$
$$\hat{\Theta}_M := \sum_{i=1}^{M} \hat{\theta}_i. \tag{2.2}$$

We will now present the assumptions we need to bound the order of convergence of the Multilevel Markov Chain Monte Carlo estimator, followed by the corresponding Theorem which bounds the order of convergence.

First, we introduce the following notation: For two sequences $(a_i)_{i\in\mathbb{N}}$ and $(b_i)_{i\in\mathbb{N}}$ we will write $a_i \lesssim b_i$ if there exists $C < \infty$ such that for all $i \in \mathbb{N}$ $a_i \leq Cb_i$ holds.

Now, we introduce a cost model to measure the efficiency of algorithms. For a random variable X, we model the algorithmic costs needed to sample this variable by $\text{cost}(X)$. To prove the efficiency result on the Multilevel algorithm, we need the following assumption on the costs of the Singlelevel estimators.

Assumption 2.1. *For $i, M \in \mathbb{N}$, the costs for sampling the estimators $\hat{\theta}_i$ and $\hat{\Theta}_M$ are bounded by*

$$\begin{aligned}\text{cost}(\hat{\Theta}_M) &\lesssim \sum_{i=1}^{M} \text{cost}(\hat{\theta}_i) \\ &\lesssim \sum_{i=1}^{M} (N_i + \text{cost}((h_i(X^i_k, Y^i_k))_{0\leq k\leq n_i+N_i})).\end{aligned}$$

For $i \in \mathbb{N}$, the costs for evaluating $(h_i(X^i_k, Y^i_k)_{0\leq k\leq n_i+N_i}$ is bounded by

$$\text{cost}\left((h_i(X^i_k, Y^i_k)_{0\leq k\leq n_i+N_i}\right) \lesssim 2^i(n_i + N_i).$$

The first part of the assumption represents the requirement that in a reasonable cost model, the cost for calculating the sum or the product of random variables is roughly the sum of the costs for calculating each summand or factor. The second part bounds the cost for sampling the variable $(h_i(X^i_k, Y^i_k))_{0\leq k\leq n_i+N_i}$. This can for example be satisfied if the costs of the evaluation of f_i and φ_i which are needed to calculate h_i and one step of the Markov Chains $(X_k)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ are of order 2^i.

f_i, φ_i and φ_i^{-1} satisfy the following uniform integrability–bounds:

Assumption 2.2. *There exists $Z_* < \infty$, such that for all $i \in \mathbb{N}$,*

$$\begin{aligned} \|\varphi_i\|_{L^{32}(E_i,\nu_i)} &\leq Z_* \\ \left\|\varphi_i^{-1}\right\|_{L^4(E_i,\nu_i)} &\leq Z_* \\ \|f_i\|_{L^8(E_i,\nu_i)} &\leq Z_*. \end{aligned}$$

Furthermore, we assume that the approximations f_i, φ_i converge sufficiently fast

Assumption 2.3.

$$\lim_{M\to\infty} \left| \int_{E_M} f_M(x)\mu_M(\mathrm{d}x) - \int_E f(x)\mu(\mathrm{d}x) \right| = 0,$$

and for $i \in \mathbb{N}$,

$$\begin{aligned} \|f_i - f_{i-1}\|_{L^4(E_i,\nu_i)} &\lesssim 2^{-\frac{i}{2}} \\ \|\varphi_i - \varphi_{i-1}\|_{L^{32}(E_i,\nu_i)} &\lesssim 2^{-\frac{i}{2}}. \end{aligned}$$

Remark 2.3. *Assumption* 2.2 *guarantees (by Jensen's inequality) that $(Z_i)_{i\in\mathbb{N}}$ is uniformly bounded away from* 0 *and* ∞ *by*

$$0 < Z_*^{-1} \leq Z_i \leq Z_* < \infty \qquad \textit{for all } i \in \mathbb{N}.$$

We now define two sequences of Markov Chains $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k^i)_{k\in\mathbb{N}}$, which are used in the Multilevel Monte Carlo estimator. Let $(\Omega, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathrm{P})$ be a filtered probability space. For each $i = 1 \dots M$, let X_0^i, Y_0^i be independent $\mathcal{F}_0$–measurable random variables on $(\Omega, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathrm{P})$ with distribution ν_i. Furthermore, let $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k^i)_{k\in\mathbb{N}}$ be two independent Markov Chains on $(\Omega, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathrm{P})$, starting in X_0^i, Y_0^i, adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and with unique invariant measure μ_i, $\tilde{\mu}_i$ respectively.

Our last assumption is related to the speed of convergence of the chains $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k^i)_{k\in\mathbb{N}}$. Denote by $(p_n^i)_{n\in\mathbb{N}}$ the semigroup generated by $(X_k^i)_{k\in\mathbb{N}}$ and q_n^i the semigroup generated by $(Y_k^i)_{k\in\mathbb{N}}$, defined by

$$\begin{aligned} p_n^i(g)(x) &:= \mathbb{E}_{x,i}[g(X_n^i)] && \text{for } g \in L^2(E_i, \mu_i), x \in E_i \\ q_n^i(g)(x) &:= \mathbb{E}_{x,i}[g(Y_n^i)] && \text{for } g \in L^2(E_i, \tilde{\mu}_i), x \in E_i, \end{aligned}$$

where $\mathbb{E}_{x,i}$ denotes the expectation value with respect to P conditioned on $\{X_0^i = x, Y_0^i = x\}$.

We assume that the chains posses a uniform spectral gap:

Assumption 2.4. *$(p_n^i)_{n\in\mathbb{N}}$ and (q_n^i)_$n \in \mathbb{N}$ have uniform spectral gaps with constant ρ: There exists $\rho > 0$ such that for all $i \in \mathbb{N}$ and for all $g : E_i \to \mathbb{R}$ with $\int_{E_i} g(x)\mu_i(\mathrm{d}x) = 0$*

$$\int_{E_i} g(x)\,(p_1^i g)(x)\mu_i(\mathrm{d}x) \leq (1-\rho)\int_{E_i} g(x)^2\mu_i(\mathrm{d}x),$$

and for all g with $\int_{E_i} g(x)\tilde{\mu}_i(\mathrm{d}x) = 0$

$$\int_{E_i} g(x)\,(q_1^i g)(x)\tilde{\mu}_i(\mathrm{d}x) \leq (1-\rho)\int_{E_i} g(x)^2\tilde{\mu}_i(\mathrm{d}x).$$

Furthermore, the cost for sampling $(X_k^i)_{k\leq N}$ is bounded by

$$\mathrm{cost}((X_k^i)_{k\leq N}) \lesssim 2^i N.$$

We can now present the main theorem of this chapter. It states that the Multilevel estimator can achieve an error of η for costs that scale like $\frac{1}{\eta^2}$ with logarithmic corrections.

Theorem 2.1. *Under Assumptions 2.1 - 2.4, the following statements hold: There exists $M(\eta,\varepsilon)$, $N_i(\eta,\varepsilon)$, $n_i(\eta,\varepsilon)$ and $\eta_0 > 0$ such that for $\eta \leq \eta_0$ and $\varepsilon > 0$*

$$\mathrm{P}\left[|\hat{\Theta}_{M(\eta,\varepsilon)} - \mu(f)| > \eta\right] < \varepsilon.$$

Furthermore, there exists $C > 0$, such that the cost of the evaluation of (2.2) is bounded by

$$\mathrm{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) \lesssim \frac{1}{\eta^2\varepsilon}\log^4\left(\frac{1}{\eta\varepsilon}\right).$$

$M(\eta,\varepsilon)$, $N_i(\eta,\varepsilon)$ and $n_i(\eta,\varepsilon)$ are known explicitly and will be stated in Section 2.3.

In this context, for functions $a, b : [0,1]\times[0,1] \to \mathbb{R}$, the notation $a(\eta,\varepsilon) \lesssim b(\eta,\varepsilon)$ denotes that there exists $C < \infty$, such that for all $0 < \varepsilon, \eta < \frac{1}{2}$, $a(\eta,\varepsilon) \leq Cb(\eta,\varepsilon)$ holds. The proofs of Lemmas 2.1 and 2.2 and Theorem 2.1 will be presented in Sections 2.2 and 2.3.

2.1.1 Example: Transition Path Sampling

As an example we consider the Transition Path Sampling setting. Here, we are interested in the distribution μ of the solution of the equation

$$\mathrm{d}X_t = -V(X_t)\,\mathrm{d}t + \varepsilon\mathrm{d}B_t \tag{2.3}$$

$$X_0 = x_0 \tag{2.4}$$

conditioned on the event $\{X_1 = x_1\}$, for $t \in [0,1]$. Here $x_0, x_1 \in \mathbb{R}^d$, $V : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth vector field and B_t is a d-dimensional Brownian Motion. In the case where V is a gradient ∇U of a function $U : \mathbb{R}^d \to \mathbb{R}$, μ is absolutely continuous with respect to a Brownian Bridge with density proportional to

$$\varphi(x) = \exp\left(-\int_0^1 \Phi_\varepsilon(x_s)\,\mathrm{d}s\right). \tag{2.5}$$

The function $\Phi_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\Phi_\varepsilon(z) = \frac{1}{2}\left(\Delta U(z) + \frac{1}{\varepsilon^2}|\nabla U(z)|^2\right),$$

see e.g. [24].

We consider the linear interpolations on a equidistant partition as approximations. For each level i, we take the partition $0 =: l_0^i < \ldots < l_{2^i}^i := 1$ of the interval $[0,1]$ with 2^i sub-intervals where

$$l_k^i := \frac{k}{2^i} \qquad 0 \le k \le 2^i, \tag{2.6}$$

and construct finite–dimensional approximations of E by the piece-wise linear functions on this partition

$$E_i := \left\{(f^1, \ldots, f^d) \in E \,\middle|\, \exists z_1^j, \ldots, z_{2^i}^j \in \mathbb{R}, \forall t \in [l_k^i, l_{k-1}^i] : f^j(t) = L(z_{k-1}^j, z_k^j, l_{k-1}^i, l_k^i; t)\right\}.$$

Here, $L(x, y, v, w; t)$ is the line spanned by (v, x) and (w, y), given by

$$L(x, y, v, w; t) := x\frac{t-w}{v-w} + y\frac{t-v}{w-v}.$$

The projections $\Pi_i(x)$ are defined as the linear interpolations of the values of $(x(l_k^i))_{0 \le k \le 2^i}$. For $i \le j$ the partition $\{l_k^i\}_k$ is a subset of $\{l_k^j\}_k$, so the projections are consistent: $\Pi_i \circ \Pi_j = \Pi_i$.

An with respect to $\langle \cdot, \cdot \rangle_S$ orthonormal basis of E_i is given by $\{e_{k,m,j}\}$ for $m \in \{1, \ldots, i\}$, $k \in \{1, \ldots 2^{m-1}\}$ and $j \in \{1, \ldots, d\}$.

$$e_{k,m,j}(t) := \begin{cases} 2^{-\frac{m}{2}-1}(x - 2^{-m}k)\,\mathbf{e}_j & 2^{-m}(k-1) \le t \le 2^{-m}(k-\frac{1}{2}) \\ -2^{-\frac{m}{2}-1}(x - 2^{-m}(k+1))\,\mathbf{e}_j & \text{if} \quad 2^{-m}(k-\frac{1}{2}) \le t \le 2^{-m}k \\ 0 & \text{otherwise.} \end{cases}$$

where $\{\mathbf{e}_j\}_{j=1,\dots,d}$ are the unit–vectors in $\mathbb{R}^d$. As a first step in the definition of the Markov Chains $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$, we construct ν_i–distributed random variables: For $j \in \{1,\dots,d\}$, $m \in \{1,\dots,i\}$ and $k \in \{1,\dots,2^{m-1}\}$, let $\xi_{m,k}^j$ be i.i.d. Gaussian random variables on $\mathbb{R}$ with mean 0 and variance 1. For $j = 1,\dots,d$, the one–dimensional Brownian Bridge N^j is now constructed iteratively by

$$N^j(0) := 0\,\mathbf{e}_j,$$
$$N^j(1) := 1\,\mathbf{e}_j$$

and for $l \in \{1,\dots,i\}$, $m \in \{1,\dots,i\}$ and $k \in \{1,\dots,2^{m-1}\}$

$$N^j\left(l_{2k-1}^m\right) := \frac{1}{2}\left(N^j\left(l_{k-1}^{m-1}\right) + N^j\left(l_k^{m-1}\right)\right) + 2^{-\frac{m}{2}-\frac{1}{2}}\xi_{m,k}^j\,\mathbf{e}_j. \tag{2.7}$$

For points $s \notin \{l_1^i,\dots,l_{2^i}^i\}$, $N^j(s)$ is given by linear interpolation.
This construction implies that

$$N^j(s) = \sum_{l=1}^{i}\sum_{k=1}^{2^{l-1}} \xi_{l,k}\mathbf{e}_k^l.$$

Therefore, $N^j(s)$ is a one–dimensional Brownian Bridge. This follows from [41, Theorem 6.1], where a similar construction is given for the Brownian Motion, and the fact that for a Brownian Motion $(\tilde{B}_s)_{s\geq 0}$,

$$B_s := \tilde{B}_s - s\tilde{B}_1$$

is a Brownian Bridge. We now set $N := \left(N^1,\dots,N^d\right)$, which then is ν_i–distributed.
As reversible Markov Chains, we construct the Random Walk Metropolis algorithm with respect to μ_i. Given a sequence of independent ν_i–distributed random variables $(N_k^i)_{k\in\mathbb{N}}$, the discrete Ornstein–Uhlenbeck process

$$\tilde{Z}_{k+1} := \sqrt{1-h^2}\tilde{Z}_k + hN_k^i$$

is reversible with respect to ν_i for each $0 < h \leq 1$. We construct a chain reversible with respect to μ_i by adding a Metropolis rejection step: Given a sequence $(U_k^i)_{k\in\mathbb{N}}$ of i.i.d. uniformly distributed variables on $[0,1]$, we define the acceptance function $a_i : E_i \times E_i \to [0,1]$ by

$$a_i(x,y) := \min\left(1, \frac{\varphi_i(y)}{\varphi_i(x)}\right)$$

and set

$$\tilde{Z}^i_{k+1} := \sqrt{1-h^2}Z^i_k + hN^i_k$$
$$Z^i_{k+1} := \begin{cases} \tilde{Z}^i_{k+1} & \text{if} \qquad U^i_k < a_i(Z^i_k, \tilde{Z}^i_{k+1}) \\ Z^i_k & \text{otherwise.} \end{cases}$$

The process $(Z^i_k)_{k\in\mathbb{N}}$ is reversible with respect to μ_i, see e.g. [8], [38].

In Section 2.4, we check Assumptions 2.1 – 2.4 to apply Theorem 2.1 in the Transition Path Sampling setting.

2.2 Basic Lemmas

We now prove some basic lemmas that are implied by the assumptions.
As ν_i is defined as the image–measure of ν under Π_i, the following lemma holds:

Lemma 2.4. *For $g : E \to \mathbb{R}$,*

$$\int_{E_i} g(\Pi_{i-1}(x))\nu_i(\mathrm{d}x) = \int_{E_{i-1}} g(x)\nu_{i-1}(\mathrm{d}x).$$

In particular,

$$\int_{E_i} g(\Pi_{i-1}(x))\tilde{\mu}_i(\mathrm{d}x) = \int_{E_{i-1}} g(x)\mu_{i-1}(\mathrm{d}x)$$

and

$$\tilde{Z}_i = Z_{i-1}.$$

Proof. We know that ν_{i-1} is the image–measure of ν_i under Π_{i-1}, as $\Pi_{i-1} = \Pi_{i-1} \circ \Pi_i$. Consequently,

$$\int_{E_i} g(\Pi_{i-1}(x))\nu_i(\mathrm{d}x) = \int_{E_{i-1}} g(x)\nu_{i-1}(\mathrm{d}x).$$

The third statement follows from the first one applied to the function $\varphi_{i-1}(x) = \varphi(\Pi_{i-1}(x))$. For the second statement note that

$$\begin{aligned}\int_{E_i} g(\Pi_{i-1}(x))\tilde{\mu}_i(\mathrm{d}x) &= \frac{1}{\tilde{Z}_i}\int_{E_i} (g\varphi)(\Pi_{i-1}(x))\nu_i(\mathrm{d}x) \\ &= \frac{1}{Z_i}\int_{E_{i-1}} g(x)\varphi(x)\nu_{i-1}(\mathrm{d}x).\end{aligned}$$

□

We now restate and prove Lemma 2.1: .1

Lemma 2.2. *For $h_i : E_i \times E_i \to \mathbb{R}$ defined in (2.1),*

$$\int_{E_i}\int_{E_i} h_i(x,y)\mu_i(\mathrm{d}x)\tilde{\mu}_i(\mathrm{d}y) = \mu_i(f_i) - \mu_{i-1}(f_{i-1}).$$

Proof. We have

$$\begin{aligned}
&\int_{E_i}\int_{E_i} h_i(x,y)\,\mu_i(\mathrm{d}x)\tilde{\mu}_i(\mathrm{d}y) \\
&= \int_{E_i} f_i(x)\mu_i(\mathrm{d}x) - \int_{E_i}\int_{E_i} f_{i-1}(x)\frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\mu_i(\mathrm{d}x)\tilde{\mu}_i(\mathrm{d}y)
\end{aligned}$$

Using Lemma 2.4, we get for the second term

$$\begin{aligned}
&\int_{E_i}\int_{E_i} f_{i-1}(x)\frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\mu_i(\mathrm{d}x)\tilde{\mu}_i(\mathrm{d}y) \\
&= \int_{E_i} f_{i-1}(x)\frac{\varphi_{i-1}}{\varphi_i}(x)\mu_i(\mathrm{d}x)\int_{E_i}\frac{\varphi_i}{\varphi_{i-1}}(y)\tilde{\mu}_i(\mathrm{d}y) \\
&= \frac{Z_{i-1}}{Z_i}\int_{E_i} f_{i-1}(x)\tilde{\mu}_i(\mathrm{d}x)\,\frac{1}{Z_{i-1}}\int_{E_i}\varphi_i(y)\nu_i(\mathrm{d}y) \\
&= \int_{E_i} f_{i-1}(x)\tilde{\mu}_i(\mathrm{d}x) \\
&= \int_{E_{i-1}} f_{i-1}(x)\mu_{i-1}(\mathrm{d}x),
\end{aligned}$$

which shows

$$\int_{E_i}\int_{E_i} h_i(x,y)\mu_i(\mathrm{d}x)\tilde{\mu}_i(\mathrm{d}y) = \mu_i(f_i) - \mu_{i-1}(f_{i-1}).$$

□

Assumption 2.2 allows us to bound the variance of the relative density $\frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_i}$. The variance controls the distance of the starting measure ν_i of the chains $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k^i)_{k\in\mathbb{N}}$ with respect to the target measure μ_i. It is important that this distance is uniformly bounded for all levels i to ensure uniform bounds on the errors of the estimators $\hat{\theta}_i$.

Lemma 2.5. *Under Assumption 2.2, the variance of the relative density,* $\mathrm{var}_{\mu_i}\left(\frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_i}\right)$ *can be bounded uniformly in i:*

$$\sup_{i\in\mathbb{N}} \mathrm{var}_{\mu_i}\left(\frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_i}\right) < \infty.$$

We set

$$V_{\sup} := \sup_{i\in\mathbb{N}} \mathrm{var}_{\mu_i}\left(\frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_i}\right)^{\frac{1}{2}}. \tag{2.8}$$

Proof. The lemma follows from:

$$\begin{aligned}\operatorname{var}_{\mu_i}\left(\frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_i}\right) &= \frac{1}{Z_i}\int_{E_i}\left(Z_i\varphi_i^{-1}(x)-1\right)^2\varphi_i(x)\nu_i(\mathrm{d}x)\\ &\leq Z_i\int_{E_i}\varphi_i^{-1}(x)\nu_i(\mathrm{d}x)+1\\ &\leq Z_*^2+1.\end{aligned}$$

□

The assumptions we have made so far allows us to estimate the discretization error.

Lemma 2.6. *Under Assumption 2.2 and 2.3,*

$$|\mu(f)-\mu_M(f_M)|\lesssim 2^{-\frac{M}{2}}.$$

In particular, there exists $c_0>0$ such that for $M(\eta,\varepsilon):=2\log_2\left(\frac{1}{\eta}\right)+c_0$,

$$\left|\mu(f)-\mu_{M(\eta,\varepsilon)}(f_{M(\eta,\varepsilon)})\right|\leq\frac{\eta}{2}.$$

Proof. We apply the triangular inequality and Hölder's inequality and get:

$$\begin{aligned}&\left|\int_E f(x)\mu(\mathrm{d}x)-\int_E f_M(x)\mu_M(\mathrm{d}x)\right|\\ &\leq\sum_{i=M}^{\infty}\left|\int_{E_{i+1}}f_{i+1}(x)\mu_{i+1}(\mathrm{d}x)-\int_{E_i}f_i(x)\mu_i(\mathrm{d}x)\right|\\ &\leq\sum_{i=M}^{\infty}\left|\int_{E_{i+1}}f_{i+1}(x)\left(\frac{1}{Z_{i+1}}\varphi_{i+1}(x)-\frac{1}{Z_i}\varphi_i(x)\right)\nu_{i+1}(\mathrm{d}x)\right|\\ &\quad+\sum_{i=M}^{\infty}\left|\int_{E_{i+1}}\frac{1}{Z_i}\varphi_i(x)(f_{i+1}(x)-f_i(x))\nu_{i+1}(\mathrm{d}x)\right|\\ &\leq\sum_{i=M}^{\infty}\left(\int_{E_{i+1}}f_{i+1}^2(x)\nu_{i+1}(\mathrm{d}x)\int_{E_{i+1}}\left(\frac{\varphi_{i+1}}{Z_{i+1}}(x)-\frac{\varphi_i}{Z_i}(x)\right)^2\nu_{i+1}(\mathrm{d}x)\right)^{\frac{1}{2}}\\ &\quad+\sum_{i=M}^{\infty}\left(\int_{E_i}\frac{\varphi_i^2}{Z_i^2}(x)\nu_i(\mathrm{d}x)\int_{E_{i+1}}(f_{i+1}(x)-f_i(x))^2\nu_{i+1}(\mathrm{d}x)\right)^{\frac{1}{2}}\end{aligned}$$

Using Remark 2.3 and Assumption 2.3, we can find an upper bound:

$$\begin{aligned}
&\int_{E_{i+1}} \left(\frac{\varphi_{i+1}}{Z_{i+1}}(x) - \frac{\varphi_i}{Z_i}(x)\right)^2 \nu_{i+1}(\mathrm{d}x) \\
&= Z_*^2 \int_{E_{i+1}} (Z_i\varphi_{i+1}(x) - Z_{i+1}\varphi_i(x))^2 \nu_{i+1}(\mathrm{d}x) \\
&\leq 2Z_*^4 \int_{E_{i+1}} (\varphi_{i+1}(x) - \varphi_i(x))^2 \nu_{i+1}(\mathrm{d}x) + 2Z_*^2 \int_{E_i} ((Z_i - Z_{i+1})\varphi_i(x))^2 \nu_i(\mathrm{d}x) \\
&\leq 2Z_*^4 \int_{E_{i+1}} (\varphi_{i+1}(x) - \varphi_i(x))^2 \nu_{i+1}(\mathrm{d}x) + 2Z_*^4 \left(\int_{E_{i+1}} (\varphi_i(x) - \varphi_{i+1}(x))\nu_{i+1}(\mathrm{d}x)\right)^2 \\
&\lesssim \int_{E_{i+1}} (\varphi_{i+1}(x) - \varphi_i(x))^2 \nu_{i+1}(\mathrm{d}x) \\
&\lesssim 2^{-i}.
\end{aligned}$$

Consequently, with Assumption 2.2 and 2.3 this leads to

$$\begin{aligned}
&\left|\int_E f(x)\mu(\mathrm{d}x) - \int_E f_M(x)\mu_M(\mathrm{d}x)\right| \\
&\lesssim \sum_{i=M}^{\infty} \left(\int_{E_{i+1}} f_{i+1}^2(x)\nu_{i+1}(\mathrm{d}x) \int_{E_{i+1}} (\varphi_{i+1}(x) - \varphi_i(x))^2 \nu_{i+1}(\mathrm{d}x)\right)^{\frac{1}{2}} \\
&\quad + \sum_{i=M}^{\infty} \left(\int_{E_i} \varphi_i^2(x)\nu_i(\mathrm{d}x) \int_{E_{i+1}} (f_{i+1}(x) - f_i(x))^2 \nu_{i+1}(\mathrm{d}x)\right)^{\frac{1}{2}}. \\
&\lesssim \sum_{i=M}^{\infty} 2^{-\frac{i}{2}} \\
&\lesssim 2^{-\frac{M}{2}},
\end{aligned}$$

which proves the first statement. For the second one, set $M(\eta, \varepsilon) := 2\log_2\left(\frac{1}{\eta}\right) + c_0$. Then for a given C, there exists c_0 such that

$$\begin{aligned}
|\mu(f) - \mu_M(f_M)| &\leq C\, 2^{-\frac{M(\eta,\varepsilon)}{2}} \\
&\leq \eta\, C\, 2^{-\frac{c_0}{2}} \\
&\leq \frac{\eta}{2}.
\end{aligned}$$

□

To measure the distance of a Markov Process to its invariant measure we use the total variation distance. For two probability measures $\eta_1, \eta_2 : \mathcal{B}(E) \to [0, 1]$, it is defined by

$$\|\eta_1 - \eta_2\|_{TV} := \frac{1}{2} \sup_{|f|\leq 1} (\eta_1 f - \eta_2 f),$$

where the supremum is taken over all measurable functions $f : E \to \mathbb{R}$ such that $|f(x)| \leq 1$ for all $x \in E$.

We now define the mixing time $t_{\text{mix}}(\varepsilon; X, (\nu_i, \nu_i))$ of a Markov Process $(X_k)_{k\in\mathbb{N}}$. It is the first time at which the total variation distance of the process $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$ started in (ν_i, ν_i) to its invariant measure decreases below ε. The mixing time will be used to control the burn–in n_i in the definition of the estimator $\hat{\theta}_i$. For times later than the mixing time, the chain is guaranteed to be sufficiently close to its invariant measure, which is needed to control the error of $\hat{\theta}_i$.

Definition 2.1. *Let $(X_k)_{k\in\mathbb{N}}$ be a Markov Process on E with invariant measure π and semigroup r on $L^2(E, \pi)$. Let η be a probability measure on E. We define*

$$t_{\text{mix}}(\varepsilon; X, \eta) := \min\{t \in \mathbb{N} : \|r_t\eta - \pi\|_{TV} \leq \varepsilon\}$$

We denote the mixing times of the process $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$ starting in $\nu_i \otimes \nu_i$ by t_{mix}^i.

$$t_{\text{mix}}^i(\varepsilon) := t_{\text{mix}}(\varepsilon; (X^i, Y^i), \nu_i \otimes \nu_i) \leq \min\left\{t \in \mathbb{N} : \left\|p_t^i\nu_i - \mu_i\right\|_{TV} + \left\|q_t^i\nu_i - \tilde{\mu}_i\right\|_{TV} \leq \varepsilon\right\}$$

It is well–known that a spectral gap implies exponentially fast convergence of the chain to its equilibrium. This can be used to bound the mixing time.

Lemma 2.7. *Assume r is a reversible semigroup with spectral gap σ on $L^2(E, \mu)$. Then*

$$\|\nu r_n - \mu\|_{TV} \leq (1-\sigma)^n \operatorname{var}_\mu\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^{\frac{1}{2}}.$$

In particular, given Assumption 2.4 and

$$\sup_{i\in\mathbb{N}} \operatorname{var}_{\mu_i}\left(\frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_i}\right)^{\frac{1}{2}} \leq V_{\sup},$$

the mixing time t_{mix}^i is bounded by

$$t_{\text{mix}}^i(\varepsilon) \leq \frac{1}{\log((1-\rho)^{-1})} \log\left(\frac{2V_{\sup}}{\varepsilon}\right).$$

uniformly in i.

Proof. The total variation distance is bounded by the L^2-norm of the relative density:

$$\begin{aligned}\|\nu r_n - \mu\|_{TV} &= \int_E \left|\frac{\mathrm{d}\nu r_n}{\mathrm{d}\mu} - 1\right| \mathrm{d}\mu \\ &\leq \left(\int_E \left(\frac{\mathrm{d}\nu r_n}{\mathrm{d}\mu} - 1\right)^2 \mathrm{d}\mu\right)^{\frac{1}{2}} \\ &= \left(\int_E \left(r_n\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) - 1\right)^2 \mathrm{d}\mu\right)^{\frac{1}{2}} \\ &\leq (1-\sigma)^n \left(\int_E \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu} - 1\right)^2 \mathrm{d}\mu\right)^{\frac{1}{2}}.\end{aligned}$$

Applying this for $r = p$, as well as Assumption 2.4, we get

$$\left\|p_t^i \nu_i - \mu_i\right\|_{TV} \leq \frac{\varepsilon}{2}.$$

for $t > \log_{1-\rho}\left(\frac{\varepsilon}{2V_{\sup}}\right)$. The second summand $\left\|q_t^i \nu_i - \tilde{\mu}_i\right\|_{TV}$ can be bounded similarly by $\frac{\varepsilon}{2}$, leading to $t_{\mathrm{mix}} \leq \frac{1}{\log((1-\rho)^{-1})} \log\left(\frac{2V_{\sup}}{\varepsilon}\right)$. □

The next lemma bounds the probability that the difference between the ergodic average of a Markov Chain and its expectation value is larger than a given $\eta > 0$. It is a modification of a similar result for finite state spaces as shown in [32, Theorem 12.19] to continuous state spaces of, and is the main tool in our proof of Theorem 2.1.

Lemma 2.8. *Let $(X_k)_{k\in\mathbb{N}}$ be a Markov Chain with reversible measure μ on E, starting in X_0 with distribution ν under P_ν. Let Assumptions 2.2 and 2.4 be satisfied. If $n > t_{\mathrm{mix}}(\frac{\varepsilon}{2})$ and $N \geq 4\frac{\mathrm{var}_\mu(f)}{\eta^2 \varepsilon \rho}$, then*

$$\mathrm{P}_\nu\left[\left|\frac{1}{N}\sum_{k=1}^N f(X_{n+k}) - \mu(f)\right| > \eta\right] < \varepsilon.$$

To proof Lemma 2.8, we need the following error bound of the ergodic average of a Markov Chain started in its invariant measure:

Lemma 2.9. *Let $(X_k)_{k\in\mathbb{N}}$ be a Markov Chain with semigroup p and reversible measure μ satisfying a spectral gap with constant ρ. Assume $(X_k)_{k\in\mathbb{N}}$ is starting with the reversible distribution μ. Then for each $g \in L^2(E,\mu)$,*

$$\mathrm{E}_\mu\left[\left(\frac{1}{N}\sum_{k=1}^N (g(X_k) - \mu(g))\right)^2\right] \leq 4\frac{\mathrm{var}_\mu(g)}{N\rho}.$$

The proof given here is a modification of the proof of [32, Lemma 12.20] to non–finite state–spaces.

Proof. Without loss of generality we demand $\mu(g) = 0$, otherwise consider $\tilde{g}(x) := g(x) - \mu(g)$. We apply the following bound:

$$\begin{aligned}\mathrm{E}_\mu\left[\left(\frac{1}{N}\sum_{k=1}^{N} g(X_k)\right)^2\right] &= \frac{1}{N^2}\mathrm{E}_\mu\left[\left(\sum_{k=1}^{N} g(X_k)\right)^2\right] \\ &\leq \frac{2}{N^2}\sum_{k=1}^{N}\sum_{l=k}^{N}\mathrm{E}_\mu[g(X_k)g(X_l)].\end{aligned}$$

By the Markov property, we have for $l \geq k$

$$\begin{aligned}\mathrm{E}_\mu[g(X_k)g(X_l)] &\leq \mathrm{E}_\mu[g(X_k)(p_{l-k}g)(X_k)] \\ &= \mathrm{E}_\mu\left[g(X_k)\int_{-\infty}^{\infty}\lambda^{l-k}\,\mathrm{d}E_\lambda(g(X_k))\right],\end{aligned}$$

where E_λ is the spectral family of the positive, self–adjoint operator p_1. Moreover,

$$\begin{aligned}\mathrm{E}_\mu\left[g(X_k)\int_E \lambda^{l-k}\,\mathrm{d}E_\lambda(g(X_k))\right] &= \int_{-\infty}^{\infty}\lambda^{l-k}\,\mathrm{d}\langle E_\lambda g(X_k), g(X_k)\rangle_{L^2(\mu)} \\ &= \int_0^{1-\rho}\lambda^{l-k}\,\mathrm{d}\langle E_\lambda g(X_k), g(X_k)\rangle_{L^2(\mu)} \\ &\leq (1-\rho)^{l-k}\int_0^{\rho} 1\,\mathrm{d}\langle E_\lambda g(X_k), g(X_k)\rangle_{L^2(\mu)},\end{aligned}$$

as the spectrum of p_1 on the subspace $\{g \in E : \mu(g) = 0\}$ is concentrated on $[0, 1-\rho]$ and $\mathrm{d}\langle E_\lambda g(X_k), g(X_k)\rangle_{L^2(\mu)}$ is a positive measure. Finally the invariance of X with respect to μ leads to the following transformation:

$$\begin{aligned}(1-\rho)^{l-k}\int_0^{1-\rho} 1\,\mathrm{d}\langle E_\lambda g(X_k), g(X_k)\rangle_{L^2(\mu)} &= (1-\rho)^{l-k}\|g(X_k)\|_{L^2(\mu)} \\ &= (1-\rho)^{l-k}\mathrm{var}_\mu(g).\end{aligned}$$

We thus have $\mathrm{E}_\mu[g(X_k)g(X_l)] \leq (1-\rho)^{l-k}\mathrm{var}_\mu(g)$ and conclude

$$\begin{aligned}
\mathrm{E}_\mu\left[\left(\frac{1}{N}\sum_{k=1}^{N} g(X_k)\right)^2\right] &\leq \frac{2}{N^2}\sum_{k=1}^{N}\sum_{l=k}^{N}\mathrm{E}_\mu[g(X_k)g(X_l)] \\
&\leq \frac{2}{N^2}\sum_{k=1}^{N}\sum_{l=k}^{N}(1-\rho)^{l-k}\mathrm{var}_\mu(g) \\
&\leq \frac{2}{N^2}\sum_{k=1}^{N} N\,(1-\rho)^{k}\mathrm{var}_\mu(g) \\
&\leq \frac{2}{N\rho}\mathrm{var}_\mu(g).
\end{aligned}$$

□

We can now prove Lemma 2.8

Proof. (Lemma 2.8)
We adopt the proof of Theorem [32, Theorem 12.19] for non-finite state spaces: Let p be the kernel of X, and let π_n be the optimal coupling of νp_n and μ, that is $\pi_n : \mathcal{B}(E)\times\mathcal{B}(E)$ is the probability measure such that for every $A \in \mathcal{B}(E)$,

$$\begin{aligned}
\pi_n(A, E) &= \int_E p_n(x, A)\nu(\mathrm{d}x), \\
\pi_n(E, A) &= \mu(A), \\
\int_E \pi_n(\mathrm{d}x, \mathrm{d}x) &= 1 - \|\nu p_n - \mu\|_{TV}.
\end{aligned}$$

Define a Markov Process $(V_k, W_k)_{k\in\mathbb{N}}$ on $E\times E$ by the kernel

$$Q((x,y),\mathrm{d}(v,w)) = \begin{cases} p(x,\mathrm{d}v) & \text{if } x = y \text{ and } v = w \\ p(x,\mathrm{d}v)p(y,\mathrm{d}w) & \text{if } x \neq y \\ 0 & \text{otherwise,} \end{cases}$$

and let the distribution of (V_0, W_0) under P_{π_n} be π_n. So $(V_k)_{k\in\mathbb{N}}$ and $(W_k)_{k\in\mathbb{N}}$ start with the optimal coupling of $p_n(x,\mathrm{d}y)$ and $\mu(\mathrm{d}y)$ and move independently until they meet. Afterwards, they move together. Because of the Markov property, $(X_{n+k})_{k\in\mathbb{N}}$ has the same

distribution under P_ν as $(V_k)_{k\in\mathbb{N}}$ has under P_{π_n}. We can now write

$$\begin{aligned}
&\mathrm{P}_\nu\left[\left|\frac{1}{N}\sum_{k=1}^{N} f(X_{n+k}) - \mu(f)\right| > \eta\right] \\
&= \mathrm{P}_{\pi_n}\left[\left|\frac{1}{N}\sum_{k=1}^{N} f(V_k) - \mu(f)\right| > \eta\right] \\
&\leq \mathrm{P}_{\pi_n}\left[V_0 \neq W_0\right] + \mathrm{P}_{\pi_n}\left[\left|\frac{1}{N}\sum_{k=1}^{N} f(W_k) - \mu(f)\right| > \eta\right].
\end{aligned}$$

The construction of π_n gives us

$$P_{\pi_n}[V_0 \neq W_0] = \|\nu p_n(x, \mathrm{d}y) - \mu(\mathrm{d}y)\|_{TV}.$$

Therefore, the first summand is bounded by $\frac{\varepsilon}{2}$ due to the definition of $t_{\mathrm{mix}}(\frac{\varepsilon}{2})$. As W_0 is distributed according to the reversible measure μ, we can apply Lemma 2.9 to the second summand and in connection with Tschebycheff inequality, we conclude

$$\begin{aligned}
\mathrm{P}_{\pi_n}\left[\left|\frac{1}{N}\sum_{k=1}^{N} f(W_n) - \mu(f)\right| > \eta\right] &\leq \frac{1}{\eta^2}\mathrm{var}_\mu\left(\frac{1}{N}\sum_{k=1}^{N} f(W_n)\right) \\
&\leq \frac{2}{N\eta^2\rho}\mathrm{var}_\mu(f).
\end{aligned}$$

By choosing $N \geq 4\frac{\mathrm{var}_\mu(f)}{\eta^2\varepsilon\rho}$ we can bound this term by $\frac{\varepsilon}{2}$ which proves the result. □

2.3 Order of convergence of the Multilevel estimator

We restate the main results and give the explicit form of the constants M, N_i and n_i appearing in the theorem:

Theorem 2.1. *Under Assumptions 2.1 - 2.4, the following statements hold: For given $\eta, \varepsilon > 0$, choose $M(\eta,\varepsilon) = 2\log_2\left(\frac{1}{\eta}\right) + c_0$, $N_i(\eta,\varepsilon) = 16\, M(\eta,\varepsilon)^3 \frac{\mathrm{var}_{\mu_i\otimes\tilde{\mu}_i}(h_i)}{\eta^2\varepsilon\rho}$, and $n_i = \frac{1}{\log((1-\rho)^{-1})}\log\left(\frac{8M(\eta,\varepsilon)V_{\sup}}{\varepsilon}\right)$, where $V_{\sup}$ is given by (2.8) and c_0 is introduced in Lemma 2.6. Then*

$$\mathrm{P}\left[|\hat{\Theta}_{M(\eta,\varepsilon)} - \mu(f)| > \eta\right] < \varepsilon.$$

Furthermore, the cost of the evaluation of $\hat{\Theta}_M$ is bounded by

$$\mathrm{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) \lesssim \frac{1}{\eta^2\varepsilon}\log^4\left(\frac{1}{\eta\varepsilon}\right).$$

Before we prove the theorem, we present the outstanding proof of Lemma 2.2, which is important for bounding the costs of the Multilevel estimator. .2

Lemma 2.2. *Under the Assumptions 2.2 - 2.3,*

$$\mathrm{var}_{\mu_i\otimes\tilde{\mu}_i}(h_i) \lesssim \frac{1}{2^i}.$$

Proof. Applying Hölder's inequality and using Assumption 2.2, we get

$$\begin{aligned}\mathrm{var}_{\mu_i\otimes\tilde{\mu}_i}(h_i) &\leq \int_{E_i}\int_{E_i}\left(f_i(x) - f_{i-1}(x)\frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\right)^2 \frac{\varphi_i}{Z_i}(x)\frac{\varphi_{i-1}}{Z_{i-1}}(y)\nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\\ &\lesssim \left(\int_{E_i}\int_{E_i}\left(f_i(x) - f_{i-1}(x)\frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\right)^4 \nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\right)^{\frac{1}{2}}.\end{aligned}$$

Using the triangular inequality, Hölder's inequality and Assumptions 2.2 and 2.3, we get

$$\begin{aligned}&\int_{E_i}\int_{E_i}\left(f_i(x) - f_{i-1}(x)\frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\right)^4 \nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\\ &\lesssim \int_{E_i}(f_i(x) - f_{i-1}(x))^4\nu_i(\mathrm{d}x)\\ &\quad + \int_{E_i}\int_{E_i} f_{i-1}^4(x)\left(1 - \frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\right)^4 \nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\\ &\lesssim \frac{1}{2^{2i}} + \left(\int_{E_i} f_{i-1}^8(x)\nu_i(\mathrm{d}x)\right)^{\frac{1}{2}}\left(\int_{E_i}\int_{E_i}\left(1 - \frac{\varphi_{i-1}}{\varphi_i}(x)\frac{\varphi_i}{\varphi_{i-1}}(y)\right)^8 \nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\right)^{\frac{1}{2}}.\end{aligned}$$

$\int_{E_i} f_{i-1}^8(x)\nu_i(\mathrm{d}x)$ is uniformly bounded because of Assumption 2.2, and

$$\begin{aligned}&\int_{E_i}\int_{E_i}\left(1 - \frac{\varphi_i(x)}{\varphi_{i-1}(x)}\frac{\varphi_{i-1}(y)}{\varphi_i(y)}\right)^8 \nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\\ &\leq \left(\int_{E_i}\frac{1}{\varphi_i(x)^{16}}\nu_i(\mathrm{d}x)\int_{E_i}\frac{1}{\varphi_{i-1}(y)^{16}}\nu_i(\mathrm{d}y)\right)^{\frac{1}{2}}\\ &\qquad\cdot\left(\int_{E_i}\int_{E_i}\left(\varphi_i(x)\varphi_{i-1}(y) - \varphi_{i-1}(x)\varphi_i(y)\right)^{16}\nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y)\right)^{\frac{1}{2}}.\end{aligned}$$

By Assumption 2.2, $\int_{E_i}\frac{1}{\varphi_i(x)^{16}}\nu_i(\mathrm{d}x)$ and $\int_{E_i}\frac{1}{\varphi_{i-1}(y)^{16}}\nu_i(\mathrm{d}y)$ are also uniformly bounded.

The last term can be bounded by

$$\begin{aligned}
&\int_{E_i}\int_{E_i} \left(\varphi_i(x)\varphi_{i-1}(y) - \varphi_{i-1}(x)\varphi_i(y)\right)^{16} \nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y) \\
\lesssim &\int_{E_i}\int_{E_i} \varphi_i(x)^{16}(\varphi_{i-1}(y) - \varphi_i(y))^{16}\nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y) \\
&\quad + \int_{E_i}\int_{E_i} \varphi_i(y)^{16}(\varphi_i(x) - \varphi_{i-1}(x))^{16}\nu_i(\mathrm{d}x)\nu_i(\mathrm{d}y) \\
\leq &\left(\int_{E_i} \varphi_i(x)^{32}\nu_i(\mathrm{d}x) \int_{E_i} (\varphi_{i-1}(y) - \varphi_i(y))^{32}\right)^{\frac{1}{2}} \\
&\quad + \left(\int_{E_i} \varphi_i(y)^{32}\nu_i(\mathrm{d}y) \int_{E_i} (\varphi_i(x) - \varphi_{i-1}(x))^{32}\nu_i(\mathrm{d}x)\right)^{\frac{1}{2}} \\
\leq &\, 2^{-8i}
\end{aligned}$$

due to Assumptions 2.2 and 2.3. Inserting this into the inequalities above yields

$$\mathrm{var}_{\mu_i\otimes\tilde{\mu}_i}(h_i) \lesssim 2^{-i}.$$

□

We now introduce two lemmas to prove the two statements of the theorem.

Lemma 2.10. *Under the assumptions of Theorem* 2.1*, the following holds*

$$\mathrm{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) \lesssim \frac{1}{\eta^2\varepsilon}\log^4\left(\frac{1}{\eta\varepsilon}\right). \tag{2.9}$$

Proof. By Assumption 2.1, the cost for evaluating $\hat{\Theta}_{M(\eta,\varepsilon)}$ is bounded by

$$\begin{aligned}
\mathrm{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) &\lesssim \sum_{i=1}^{M(\eta,\varepsilon)} \mathrm{cost}(\hat{\theta}_i) \\
&\lesssim \sum_{i=1}^{M(\eta,\varepsilon)} 2^i(N_i(\eta,\varepsilon) + n_i(\eta,\varepsilon)).
\end{aligned}$$

With our choices of N_i and n_i, and since i is bounded by M, we get

$$\begin{aligned}
&2^i(N_i(\eta,\varepsilon) + n_i(\eta,\varepsilon)) \\
\lesssim\; &2^i M(\eta,\varepsilon)^3 \frac{\mathrm{var}_{\mu_i\otimes\tilde{\mu}_i}(h_i)}{\eta^2\varepsilon} + 2^i\left(\log\left(\frac{M(\eta,\varepsilon)}{\varepsilon\eta}\right)\right)
\end{aligned}$$

Applying Lemma 2.2, we can bound the variance

$$\begin{aligned}
&2^i M(\eta,\varepsilon)^3 \left(\frac{\operatorname{var}_{(\mu_i\otimes\tilde{\mu}_i)}(h_i)}{\eta^2\varepsilon}\right)\\
\lesssim\ & 2^i \frac{1}{\eta^2\varepsilon}\log^3\left(\frac{1}{\eta}\right) 2^{-i}\\
=\ & \frac{1}{\eta^2\varepsilon}\log^3\left(\frac{1}{\eta}\right).
\end{aligned}$$

The second summand is bounded by

$$\begin{aligned}
&2^i\left(c_0+\log\left(\frac{M(\eta,\varepsilon)}{\varepsilon}\right)\right)\\
\lesssim\ & 2^{M(\eta,\varepsilon)}\log\left(\log_2(\eta^{-2})\frac{1}{\varepsilon}\right)\\
\lesssim\ & \frac{1}{\eta^2}\log\left(\frac{1}{\varepsilon\eta}\right).
\end{aligned}$$

This bound is now used within the sum, and by choosing $M(\eta,\varepsilon) = 2\log_2\left(\frac{1}{\eta}\right) + c_0$, we get

$$\begin{aligned}
\operatorname{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) &\lesssim \sum_{i=1}^{M(\eta,\varepsilon)} \operatorname{cost}(\hat{\theta}_i)\\
&\lesssim \frac{1}{\eta^2\varepsilon}\log^4\left(\frac{1}{\eta\varepsilon}\right).
\end{aligned}$$

□

Lemma 2.11. *Under the assumptions of Theorem* 2.1*, the following holds*

$$\mathrm{P}\left[|\hat{\Theta}_{M(\eta,\varepsilon)} - \mu(f)| > \eta\right] < \varepsilon.$$

Proof. This result follows from the error estimates we have established for the Markov Chains $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$ on each level i in Lemma 2.8. Applying the triangular inequality, we can bound

$$\begin{aligned}
&\mathrm{P}\left[\left|\sum_{i=1}^{M(\eta,\varepsilon)} \frac{1}{N_i(\eta,\varepsilon)} \sum_{k=0}^{N_i(\eta,\varepsilon)} h_i(X^i_{n_i(\eta,\varepsilon)+k}, Y^i_{n_i(\eta,\varepsilon)+k}) - \int_E f(x)\mu(\mathrm{d}x)\right| > \eta\right]\\
&\le \mathrm{P}\left[\left|\int_E f(x)\mu(\mathrm{d}x) - \int_{E_{M(\eta,\varepsilon)}} f_{M(\eta,\varepsilon)}(x)\mu_{M(\eta,\varepsilon)}(\mathrm{d}x)\right| > \frac{\eta}{2}\right]\\
&\quad + \mathrm{P}\left[\left|\int_{E_{M(\eta,\varepsilon)}} f_{M(\eta,\varepsilon)}(x)\mu_{M(\eta,\varepsilon)}(\mathrm{d}x) - \sum_{i=1}^{M(\eta,\varepsilon)} \frac{1}{N_i(\eta,\varepsilon)} \sum_{k=0}^{N_i(\eta,\varepsilon)} h_i(X^i_{n_i(\eta,\varepsilon)+k}, Y^i_{n_i(\eta,\varepsilon)+k})\right| > \frac{\eta}{2}\right].
\end{aligned}$$

The first term equals 0 if $\left|\int_E f(x)\mu(\mathrm{d}x) - \int_{E_{M(\eta,\varepsilon)}} f_{M(\eta,\varepsilon)}(x)\mu_{M((\eta,\varepsilon)}\mathrm{d}x)\right| < \frac{\eta}{2}$, which is follows from Lemma 2.6. For the second term, the error of the Multilevel estimator can be bounded by the sum of the errors at each level i, and therefore we get with $\theta_i := \left(\int_{E_i} f_i(x)\mu_i(\mathrm{d}x) - \int_{E_i} f_{i-1}(x)\mu_{i-1}(\mathrm{d}x)\right)$:

$$\begin{aligned}
&\mathrm{P}\left[\left|\sum_{i=1}^{M(\eta,\varepsilon)} \frac{1}{N_i(\eta,\varepsilon)} \sum_{k=0}^{N_i(\eta,\varepsilon)} h_i(X^i_{n_i(\eta,\varepsilon)+k}, Y^i_{n_i(\eta,\varepsilon)+k}) - \int_{E_{M(\eta,\varepsilon)}} f_{M(\eta,\varepsilon)}(x)\mu_{M(\eta,\varepsilon)}(\mathrm{d}x)\right| > \frac{\eta}{2}\right] \\
&\leq \sum_{i=1}^{M(\eta,\varepsilon)} \mathrm{P}\left[\left|\frac{1}{N_i(\eta,\varepsilon)} \sum_{k=0}^{N_i(\eta,\varepsilon)} h_i(X^i_{n_i(\eta,\varepsilon)+k}, Y^i_{n_i(\eta,\varepsilon)+k}) - \theta_i\right| > \frac{\eta}{2M(\eta,\varepsilon)}\right].
\end{aligned}$$

We apply Lemma 2.8 which states

$$\begin{aligned}
&\mathrm{P}\left[\left|\frac{1}{N_i(\eta,\varepsilon)} \sum_{k=0}^{N_i(\eta,\varepsilon)} h_i(X^i_{n_i(\eta,\varepsilon)+k}, Y^i_{n_i(\eta,\varepsilon)+k}) - \theta_i\right| > \frac{\eta}{2M(\eta,\varepsilon)}\right] \\
&< \frac{\varepsilon}{M(\eta,\varepsilon)},
\end{aligned}$$

as $n_i(\eta,\varepsilon) = \frac{1}{\log((1-\rho)^{-1})} \log\left(\frac{8M(\eta,\varepsilon)V_{\sup}}{\varepsilon}\right) \geq t^i_{\mathrm{mix}}\left(\frac{\varepsilon}{2M(\eta,\varepsilon)}\right)$ by Lemma 2.7 and Assumption 2.4. Therefore, we get

$$\mathrm{P}\left[\left|\sum_{i=1}^{M(\eta,\varepsilon)} \frac{1}{N_i(\eta,\varepsilon)} \sum_{k=0}^{N_i(\eta,\varepsilon)} h_i(X^i_{n_i(\eta,\varepsilon)+k}, Y^i_{n_i(\eta,\varepsilon)+k}) - \int_E f(x)\mu(\mathrm{d}x)\right| > \eta\right] < \varepsilon,$$

which proves the lemma. □

The two previous lemmas imply Theorem 2.1:

Proof. (Theorem 2.1) Combining Lemma 2.10 and 2.11 proves the Theorem. □

2.4 Application to Transition Path Sampling

As an example, we apply the results of Theorem 2.1 the Transition Path Sampling setting, where we are interested in the distribution μ of the solution of the equation

$$\mathrm{d}X_t = -V(X_t)\,\mathrm{d}t + \varepsilon \mathrm{d}B_t \tag{2.10}$$

$$X_0 = x_0 \tag{2.11}$$

conditioned on the event $\{X_1 = x_1\}$. Here $x_0, x_1 \in \mathbb{R}^d$, $V : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth vector field and B_t is a d-dimensional Brownian Motion. In the case where V is a gradient ∇U of a function $U : \mathbb{R}^d \to \mathbb{R}$, μ is absolutely continuous with respect to a Brownian Bridge with density proportional to

$$\varphi(x) = \exp\left(-\int_0^1 \Phi_\varepsilon(x_s)\,\mathrm{d}s\right). \tag{2.12}$$

The function $\Phi_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\Phi_\varepsilon(z) = \frac{1}{2}\left(\Delta U(z) + \frac{1}{\varepsilon^2}|\nabla U(z)|^2\right),$$

see e.g. [24]. In this setting, direct Monte Carlo simulations of μ (or its approximations) are often not possible and Markov Chain Monte Carlo methods are used. Analysis of MCMC–method in the Transition Path Sampling setting can be found in [8]. We give a discretization of the space E and conditions on Φ_ε and f such that the Assumptions 2.2 - 2.3 of the previous sections hold, construct chains $(X_k^i, Y_k^i)_{k\in\mathbb{N}}$ that satisfy Assumption 2.4 and introduce a cost model that satisfies Assumption 2.1

We assume that Φ is positive and Lipschitz–continuous. For each level i, we generate an equidistant partition $0 = l_0^i < \ldots < l_{2^i}^i = T$ of the interval $[0, 1]$ with 2^i subintervals where

$$l_k^i := \frac{k}{2^i} \qquad 0 \le k \le 2^i, \tag{2.13}$$

and construct finite–dimensional approximations of E by the piece-wise linear functions on this partition,

$$E_i := \left\{(f^1, \ldots, f^d) \in E \,\middle|\, \exists z_1^j, \ldots, z_{2^i}^j \in \mathbb{R}, \forall t \in [l_{k-1}^i, l_k^i] : f^j(t) = L(z_{k-1}^j, z_k^j, l_{k-1}^i, l_k^i; t)\right\},$$

where L is given by

$$L(x, y, v, w; t) := x\frac{t-w}{v-w} + y\frac{t-v}{w-v}.$$

The projections $\Pi_i(x)$ are defined as the linear interpolations of the values at $(x(l_k^i))_{0\le k\le 2^i}$. For $i \le j$ the partition $\{l_k^i\}_{0\le k\le 2^i}$ is a subset of $\{l_k^j\}_{0\le k\le 2^j}$, so the projections are consistent: $\Pi_i \circ \Pi_j = \Pi_i$.

The approximation $\varphi_i : E_i \to \mathbb{R}$ are defined using the Riemann-sum approximation of the integral:

$$\varphi_i(x) = \frac{1}{Z_i} \exp\left(-\frac{1}{d_i} \sum_{k=1}^{d_i - 1} \Phi(x_{l_k^i}) \right), \tag{2.14}$$

where $d_i := 2^i$. The boundary terms $\Phi(x_{l_0^i})$ and $\Phi(x_{l_{2^i}^i})$ can be neglected as they are fixed by the boundary conditions, and therefore just appear in the normalization constant Z_i.

To measure the computational complexity, we use the following cost model: We define $\text{cost}\,(X) := 1$, if

- X is a uniform distributed random variable on $[0, 1]$, or
- for $k \leq d$, X is a Gaussian random variable on $\mathbb{R}^k$ with mean $m \in \mathbb{R}^k$ and variance $\sigma \in \mathbb{R}^{d \times d}$, or
- X is a constant.

For other random variables, the costs can be bound recursively by the following rules:
For $k \leq d$, given an injective map $\pi : \{1, \ldots, l\} \to \{1, \ldots, k\}$, and $\Lambda : \mathbb{R}^k \times \ldots \times \mathbb{R}^k \to \mathbb{R}^l$ be one of the following functions:

$$\begin{aligned}
(x_1, \ldots, x_n) &\mapsto \sum_{i=1}^{n} x_i \\
(x_1, \ldots, x_n) &\mapsto \prod_{i=1}^{n} x_i && \text{for } x_1, \ldots, x_n \in \mathbb{R} \\
x &\mapsto \Phi(x) && \text{for } x \in \mathbb{R}^d \\
x &\mapsto x^{-1} && \text{for } x \in \mathbb{R}, x \neq 0 \\
x &\mapsto -x \\
x &\mapsto \exp(x) && \text{for } x \in \mathbb{R}
\end{aligned}$$

Then

$$\text{cost}\,((X_1, \ldots, X_n, \Lambda(X_{\pi_1}, \ldots, X_{\pi_k}))) \leq \text{cost}((X_1, \ldots, X_k)) + k.$$

Furthermore, the cost of a vector is bounded by the sum of the costs of its components: For $k \leq d$ amd $X_1, \ldots, X_n \in \mathbb{R}^k$,

$$\operatorname{cost}(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} \operatorname{cost}(X_i).$$

This is a coarse model that allows basic operations on $\mathbb{R}^d$ for unit costs, and does not measure the exact effort for e.g. sampling a Gaussian random variable. However, this is not required for further analysis since we focus on the asymptotics of the algorithm as the dimension d_N of the approximation converges to infinity. For that, constant factors on the costs of low–dimensional operations are not of interest.

We verify Assumptions 2.1 – 2.4 for our choice of the density and its approximation. Conditions to satisfy Assumptions 2.2 and 2.3 are given in the next theorem:

Theorem 2.2. *Let φ and φ_i be given by (2.12) and (2.14), where $\Phi : \mathbb{R}^d \to \mathbb{R}$ is positive and Lipschitz–continuous. For $f : E \to \mathbb{R}$ let f_i be defined as*

$$f_i := f \circ \Pi_i.$$

Assume f is Lipschitz–continuous with respect to the L^q–norm for some $q \geq 1$:

$$|f(x) - f(y)| \leq L\|x - y\|_{L^q([0,1],\mathbb{R}^d)} \qquad \text{for all } x, y \in C_0([0,1], \mathbb{R}^d).$$

Then Assumptions 2.2 and 2.3 are satisfied.

The proof proceeds in a number of lemmas.

Lemma 2.12. *Under the assumptions of Theorem 2.2, there exists Z_* such that*

$$\begin{aligned} \|\varphi_i\|_{L^{32}(E_i,\nu_i)} &< Z_*, \\ \left\|\varphi_i^{-1}\right\|_{L^4(E_i,\nu_i)} &< Z_*. \end{aligned}$$

Proof. As Φ is positive, $\int_{E_i} \varphi_i^{32}(x)\nu_i(\mathrm{d}x) \leq 1$ holds for all i. Using the Lipschitz–continuity

of f, the inverse moment can be bounded by

$$\begin{aligned}\int_{E_i} \frac{1}{\varphi_i(x)^4}\nu_i(\mathrm{d}x) &= \int_{E_i} \exp\left(\frac{4}{d_i}\sum_{k=1}^{d_i}|\Phi(x_{l_k^i})|\right)\nu_i(\mathrm{d}x)\\ &\leq \int_{E_i} \exp\left(4|x_0| + \frac{4L}{d_i}\sum_{k=1}^{d_i}|x_{l_k^i}|\right)\nu_i(\mathrm{d}x)\\ &\leq \int_{E_i} \exp\left(4|x_0| + 4L\max_{k\in\{1,\dots,d_i\}}|x_{l_k^i}|\right)\nu_i(\mathrm{d}x)\\ &\leq \int_{E} \exp\left(4|x_0| + 4L\max_{s\in[0,1]}|x_s|\right)\nu(\mathrm{d}x),\end{aligned}$$

where we bounded the maximum of the finite dimensional marginal of the Brownian Bridge by the maximum of the Brownian Bridge in the last line. By applying the formula for the distribution of the maximum of a Brownian Bridge (see e.g. [31, Example 3.12]), we get

$$\begin{aligned}&\int_{E} \exp\left(4|x_0| + 4L\max_{s\in[0,1]}|x_s|\right)\nu(\mathrm{d}x)\\ &\leq \exp(4|x_0|)\int_0^\infty 4z\exp\left(4Ldz\right)\exp(-2z^2)\nu(\mathrm{d}z)\\ &< C,\end{aligned}$$

for a constant C independent of i. □

Lemma 2.13. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be positive and Lipschitz–continuous. Let φ_i be given by (2.14). Then for $p \geq 1$,*

$$\|\varphi_i - \varphi_{i-1}\|_{L^{32}(E_i,\nu_i)} \lesssim 2^{-\frac{i}{2}}.$$

Proof. We estimate

$$\begin{aligned}&\int_{E_i} (\varphi_i(x) - \varphi_{i-1}(x))^{32}\nu_i(\mathrm{d}x)\\ &\leq \int_{E_i}\left|\frac{1}{d_{i-1}}\sum_{k=1}^{d_{i-1}}\Phi(x_{l_k^{i-1}}) - \frac{1}{d_i}\sum_{k=1}^{d_i}\Phi(x_{l_k^i})\right|^{32}\nu_i(\mathrm{d}x)\\ &\leq \int_{E_i}\left|\frac{1}{d_i}\sum_{k=1}^{d_{i-1}}\Phi(x_{l_{2k}^i}) - \Phi(x_{l_{2k-1}^i})\right|^{32}\nu_i(\mathrm{d}x)\\ &\leq \int_{E_i}\left(\frac{L}{2}\right)^{32}\frac{1}{d_{i-1}}\sum_{k=1}^{d_{i-1}}\left|x_{l_{2k}^i} - x_{l_{2k-1}^i}\right|^{32}\nu_i(\mathrm{d}x)\\ &= \left(\frac{L}{2}\right)^{32}\frac{1}{d_{i-1}}\int_{E_i}\sum_{k=1}^{d_{i-1}}\left|x_{l_{2k}^i} - x_{l_{2k-1}^i}\right|^{32}\nu_i(\mathrm{d}x).\end{aligned}$$

The mean of the Gaussian random variable $(x_{l^i_{2k}} - x_{l^i_{2k-1}})$ is given by $\frac{1}{d_i}(x_1 - x_0)$, its variance is bounded by $\frac{1}{d_i}$. Consequently, we can bound the 32th moment by

$$\int_{E_i} \left| x_{l^i_{2k}} - x_{l^i_{2k-1}} \right|^{32} \leq C d_i^{-16}.$$

for a constant $C < \infty$. Putting all terms together, we finally get

$$\|\varphi_i - \varphi_{i-1}\|_{L^{32}(E_i,\nu_i)} \lesssim 2^{-\frac{i}{2}}.$$

□

The following lemma provides conditions on f to satisfy the assumptions of Theorem 2.1:

Lemma 2.14. *Let $f : C_0([0,1], \mathbb{R}^d) \to \mathbb{R}$ be Lipschitz–continuous with respect to the L^q–norm for some $q \geq 1$, i.e. there exists $L < \infty$, such that for all $x, y \in C_0([0,1], \mathbb{R}^d)$,*

$$|f(x) - f(y)| \leq L\|x - y\|_{L^q([0,1],\mathbb{R}^d)},$$

and let the approximations $f_i : E_i \to \mathbb{R}$ be given as

$$f_i := f \circ \Pi_i.$$

Then for all $p \geq 1$,

$$\|f_i - f_{i-1}\|_{L^p(E_i,\nu_i)} \lesssim 2^{-\frac{i}{2}}.$$

Furthermore, there exists Z_* such that

$$\|f_i\|_{L^8(E_i,\nu_i)} < Z_*$$

uniformly in i.

Proof. The Lipschitz–continuity of f implies

$$\begin{aligned}\int_{E_i} (f_i(x) - f_{i-1}(x))^p \nu_i(\mathrm{d}x) &= \int_E (f(\Pi_i(x)) - f(\Pi_{i-1}(x)))^p \nu(\mathrm{d}x) \\ &\leq L \int_E \|\Pi_i(x) - \Pi_{i-1}(x)\|^p_{L^q([0,1],\mathbb{R}^d)} \nu(\mathrm{d}x).\end{aligned}$$

Considering the Schauder decomposition of the Brownian Bridge, we see that

$$\int_E \|\Pi_i(x) - \Pi_{i-1}(x)\|_{L^q([0,1],\mathbb{R}^d)} \nu(\mathrm{d}x) \leq \mathbb{E}\left[\sum_{k=1}^{2^{i-1}} \left\| \mathbf{e}_k^{i-1} \right\|_{L^q([0,1],\mathbb{R}^d)} |\xi_k^i|^p \right],$$

where for each i, ξ_k^i are independent Gaussian random variables with mean 0 and variance 2^{-i}, and $\mathbf{e}_k^i$ is given by

$$\mathbf{e}_k^i(t) := \begin{cases} 2^{i+1}(x - 2^{-i}k) & 2^{-i}(k-1) \le t \le 2^{-i}(k-\frac{1}{2}) \\ -2^{i+1}(x - 2^{-i}(k+1)) & \text{if} \quad 2^{-i}(k-\frac{1}{2}) \le t \le 2^{-i}k \\ 0 & \text{otherwise.} \end{cases}$$

Estimating the p–th moment of a Gaussian random variable with variance 2^{-i}, we get

$$\int_E \|\Pi_i(x) - \Pi_{i-1}(x)\|^p_{L^q([0,1],\mathbb{R}^d)} \nu(\mathrm{d}x) \lesssim 2^{-p\frac{i}{2}}.$$

To prove the second statement, note that

$$\begin{aligned} \int_{E_i} f_i(x)^8 \nu_i(\mathrm{d}x) &\lesssim \int_E f(x)^8 \nu(\mathrm{d}x) + \int_E (f(x) - f_i(x))^8 \nu(\mathrm{d}x) \\ &\lesssim f(0)^8 + \int_E \|x\|^8_{L^q([0,1],\mathbb{R}^d)} \nu(\mathrm{d}x) + \int_E \|x - \Pi_i(x)\|^8_{L^q([0,1],\mathbb{R}^d)} \nu(\mathrm{d}x). \end{aligned}$$

Using the Schauder decomposition to represent x and $(x - \Pi_i(x))$ we can easily bound these terms independently of i. □

We now construct the sequence of Markov Chains for the Multilevel algorithm: On a fixed level i, a Metropolis chain $(Z_n^i)_{n\in\mathbb{N}}$ with invariant measure μ_i can be constructed the following way: Given a sequence of independent ν_i–distributed random variables $(N_k^i)_{k\in\mathbb{N}}$, the discrete Ornstein–Uhlenbeck process

$$\tilde{Z}_{k+1} := \sqrt{1-h^2}\tilde{Z}_k + hN_k^i$$

is reversible with respect to ν_i for each $0 < h \le 1$. The process becomes reversible with respect to μ_i by adding a Metropolis rejection step: Given a sequence $(U_k^i)_{k\in\mathbb{N}}$ of i.i.d. uniformly distributed variables on $[0,1]$, and a starting point $z_0 \in E_i$, we define the acceptance function $a_i : E_i \times E_i \to [0,1]$ by

$$a_i(x,y) := \min\left(1, \frac{\varphi_i(y)}{\varphi_i(x)}\right).$$

We set $Z_0 := z_0$, and for $k \in \mathbb{N}$,

$$\begin{aligned} \tilde{Z}_{k+1}^i &:= \sqrt{1-h^2}Z_k^i + hN_k^i \\ Z_{k+1}^i &:= \begin{cases} \tilde{Z}_{k+1}^i & \text{if} \quad U_k^i < a_i(Z_k^i, \tilde{Z}_{k+1}^i) \\ Z_k^i & \text{otherwise.} \end{cases} \end{aligned}$$

The process $(Z_k^i)_{k\in\mathbb{N}}$ is reversible with respect to μ_i, see e.g. [8].
For the Multilevel algorithm, we define two independent Metropolis chains $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k^i)_{k\in\mathbb{N}}$ on each level i, X^i being reversible with respect to μ_i, and Y^i being reversible with respect to $\tilde{\mu}_i$. The estimator $\hat{\Theta}_M$ is now set to

$$\hat{\Theta}_M := \sum_{i=1}^{M} \frac{1}{N_i} \sum_{k=0}^{N_i} h_i(X_{n_i+k}^i, Y_{n_i+k}^i), \tag{2.15}$$

where h_i is given by (2.1).

Furthermore, we need to consider the spectral gaps of the processes $(X_k^i)_{k\in\mathbb{N}}$ and $(Y_k^i)_{k\in\mathbb{N}}$. The following lemma provides this result:

Lemma 2.15. *Assume φ_i is given by (2.14) and there exists $C > 0$ such that*

$$c^{-1} \leq \Phi(z) \leq c \qquad \text{for all } z \in \mathbb{R}^d.$$

Then for each $i \in \mathbb{N}$, $(X_k)_{k\in\mathbb{N}}^i$ and $(Y_k)_{k\in\mathbb{N}}^i$ possess a spectral gap of size ρ with

$$\rho \geq -\exp\left(3(c^{-1}-c)\right) \log\left(\sqrt{1-h^2}\right) > 0.$$

Remark 2.16. *Note that if Φ is bounded as in Lemma 2.15, it is possible to use an exact sampling algorithm for Transition Path Sampling as presented in [6, 7]. As one simulates the exact measure with this method, it does not have an approximation error. Given the independent and exact samples $(X_i)_{i\in\mathbb{N}}$ of μ of this method, we can construct the estimator $\hat{\theta}_{ES} := \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ for $\nu(f)$. If also f can be evaluated exactly, its error decreases like $T^{-\frac{1}{2}}$ by the Central Limit Theorem.*
Basically, the Exact Sampler is an Acceptance–Rejection Sampler. It proposes samples of the Brownian Bridge and rejects or accepts them with a rate such that the accepted samples have distribution μ. It works well when the relative density of the target measure with respect to the Brownian Bridge is large for typical realizations of a Brownian Bridge, whereas the acceptance rate and therefore the performance of the algorithm decreases if the density is small. This is not the case for the Multilevel sampler, which is based on a Markov Chain Monte Carlo algorithm, which typically behaves well as long as the state space does not have isolated modes, although a spectral gap is difficult to prove.

Proof. We compare $(X_k)_{k\in\mathbb{N}}^i$ and $(Y_k)_{k\in\mathbb{N}}^i$ with the discrete Ornstein–Uhlenbeck process $(\tilde{Z}_k)_{k\in\mathbb{N}}$ given by

$$\begin{aligned} \tilde{Z}_{k+1} &= \left(1-\frac{h}{2}\right) Z_k + \sqrt{\tilde{h}} N_{n+1} \qquad \text{for } k \in \mathbb{N}, \\ \tilde{Z}_0 &= z_0. \end{aligned}$$

The distribution of $\tilde{Z}_k$ coincides with the distribution of the continuous–time Ornstein–Uhlenbeck process z_t at time $t = -k \log\left(\sqrt{1-h^2}\right)$, where z is given by

$$dz_t = -z_t dt + \sqrt{2} dw_t.$$

Here w_t is a E_N–valued Wiener process with covariance given by $(-\Delta_{0,N})^{-1}$, see e.g. [11, Propositions 8.13, 9.13]. z_t possesses a spectral gap of size 1 [1, Remarque 1.5.8], therefore $\tilde{Z}_k$ possesses a spectral gap of size $\gamma_{OU} := -\log\left(\sqrt{1-h^2}\right)$. As the density φ_i is bounded from above and below, we have for $f \in L^1(E_i, \mu_i)$

$$\int_{E_i} f(x)\mu_i(\mathrm{d}x) = \frac{1}{Z_i}\int_{E_i} f(x)\varphi_i(x)\nu_i(\mathrm{d}x) \leq \exp(-c^{-1}+c)\int_{E_i} f(x)\nu_i(\mathrm{d}x),$$
$$\int_{E_i} f(x)\nu_i(\mathrm{d}x) = Z_i\int_{E_i} f(x)\varphi_i(x)^{-1}\mu_i(\mathrm{d}x) \leq \exp(c-c^{-1})\int_{E_i} f(x)\mu_i(\mathrm{d}x).$$

Furthermore, the acceptance probability is bounded from below by

$$a_i(x,y) \geq \exp(-c+c^{-1}).$$

So if p_i denotes the semigroup of (X_k^i), and q_i denotes the semigroup of the discrete Ornstein–Uhlenbeck process, we can split p_i into q_i and $\tilde{p}_i$ by

$$p_i f(x) = \exp(-c+c^{-1}) q_i f(x) + \left(1-\exp(-c+c^{-1})\right)\tilde{p}_i f(x),$$

where $\tilde{p}_i$ is the semigroup

$$\tilde{p}_i f(x) := \int_{E_i} \tilde{a}_i(x,y) q_i(x,\mathrm{d}y) + \delta_x(\mathrm{d}y)\int_{E_i}(1-\tilde{a}_i(x,y))q_i(x,\mathrm{d}y),$$

for the modified acceptance probability

$$\tilde{a}_i(x,y) = \left(1-\exp(-c+c^{-1})\right)^{-1}\left(a_i(x,y)-\exp(c-c^{-1})\right) \in [0,1].$$

As it is a kernel of a Metropolis chain, $\tilde{p}_i$ is a Markov kernel again, and we can represent the semigroup p_i by

$$\int_{E_i} f(x)p_i f(x)\mu_i(\mathrm{d}x) = \exp(-c+c^{-1})\int_{E_i} f(x)q_i f(x)\mu_i(\mathrm{d}x)$$
$$+\left(1-\exp(c-c^{-1})\right)\int_{E_i} f(x)\tilde{p}_i f(x)\mu_i(\mathrm{d}x).$$

Applying the bound on $\frac{\mu_i(\mathrm{d}x)}{\nu_i(\mathrm{d}x)}$ and using the spectral gap of the Ornstein–Uhlenbeck process, we get

$$\exp(-c+c^{-1})\int_{E_i} f(x)q_i f(x)\mu_i(\mathrm{d}x) \leq \exp(-2(c-c^{-1}))\int_{E_i} f(x)q_i f(x)\nu_i(\mathrm{d}x)$$
$$\leq \exp(-3(c-c^{-1}))\gamma_{OU}\int_{E_i} f(x)^2\mu_i(\mathrm{d}x),$$

leading to

$$\int_{E_i} f(x)p_i f(x)\mu_i(\mathrm{d}x) \leq \left(1 - \gamma_{OU}\exp(-3(c - c^{-1}))\right) \int_{E_i} f(x)^2\mu_i(\mathrm{d}x).$$

The proof for $(Y_k)_{k\in\mathbb{N}}^i$ works analogously when we replace the acceptance rate a_i by a_{i-1}. □

To apply to apply Theorem 2.1, in the Transition Path Sampling setting, it remains to verify Assumption 2.1.

Lemma 2.17. *Let for every random variable ξ on $\mathbb{R}^d$,*

$$\mathrm{cost}(f_i(\xi)) \lesssim 2^i + \mathrm{cost}(\xi).$$

Then the Multilevel Markov Chain Monte Carlo estimator Θ_M as defined in (2.15) satisfies Assumption 2.1.

Proof. Assumption 2.1 consists of 3 substatements: The first is

$$\mathrm{cost}\left(\hat{\Theta}_M\right) \lesssim \sum_{i=1}^{M} \mathrm{cost}\left(\hat{\theta}_i\right).$$

As $\Theta_M := \sum_{i=1}^M \hat{\theta}_i$, we have by the construction of our cost model

$$\begin{aligned}\mathrm{cost}\left(\hat{\Theta}_M\right) &= M + \mathrm{cost}\left((\hat{\theta}_1, \ldots, \hat{\theta}_M)\right) \\ &\leq M + \sum_{i=1}^{M} \mathrm{cost}\left(\hat{\theta}_i\right) \\ &\leq 2\sum_{i=1}^{M} \mathrm{cost}\left(\hat{\theta}_i\right).\end{aligned}$$

The second statement is

$$\mathrm{cost}\left(\hat{\theta}_i\right) \lesssim N_i + \mathrm{cost}\left(h_i(X_k^i, Y_k^i)_{0\leq k\leq n_i+N_i}\right).$$

This follows form the definition $\hat{\theta}_i := \frac{1}{N_i}\sum_{i=n_i}^{N_i} h_i(X_k^i, Y_k^i)$, we have

$$\begin{aligned}\mathrm{cost}\left(\hat{\theta}_i\right) &\leq 2 + \mathrm{cost}\left(\sum_{i=n_i}^{N_i} h_i(X_k^i, Y_k^i)\right) \\ &\lesssim N_i + \mathrm{cost}\left(h_i(X_k^i, Y_k^i)_{0\leq k\leq n_i+N_i}\right).\end{aligned}$$

It remains to show the third part, which is

$$\mathrm{cost}\left(h_i(X_k^i, Y_k^i)_{0\leq k\leq n_i+N_i}\right) \lesssim 2^i(N_i + n_i).$$

By the construction in (2.7), we can construct the Gaussian random variables (N_k^i) with costs bounded by

$$\operatorname{cost}(N_k^i) \lesssim 2^i$$

for $i \in \{1, \dots, M\}$, $k \in \{1, \dots, N_i + n_i\}$. Using this construction, we can construct the values of the Markov Chain (X_k^i, Y_k^i) up to time $N_i + n_i$ by

$$\operatorname{cost}\left((X_k^i, Y_k^i)_{0 \le k \le N_i + n_i}\right) \lesssim 2^i (N_i + n_i),$$

as evaluation of f_i and φ_i can be done for additional costs bounded by a constant factor of 2^i by the assumptions of this lemma and the construction of φ_i. Furthermore, with definition (2.1) we have

$$\operatorname{cost}\left(h_i(X_k^i, Y_k^i)_{0 \le k \le n_i + N_i}\right) \lesssim 2^i + \operatorname{cost}\left((X_k^i, Y_k^i)_{0 \le k \le n_i + N_i}\right).$$

□

Summarizing the previous Lemmas, we obtain the following theorem addressing the order of convergence of the Multilevel algorithm in the Transition Path Sampling setting.

Theorem 2.3. *Let μ, Φ and $(X_k^i)_{k\in\mathbb{N}}$, $(Y_k^i)_{k\in\mathbb{N}}$ as constructed above. Let $f : C_0([0,T], \mathbb{R}^d) \to \mathbb{R}$ be given. Assume that for constants $c, L > 0$, and every random variable ξ on $\mathbb{R}^d$,*

$$\begin{aligned}
|f(x) - f(x)| &\le L \|x - y\|_{L^q([0,T],\mathbb{R}^d)} && \text{for all } x, y \in C([0,T], \mathbb{R}^d),\\
\operatorname{cost}(f_i(\xi)) &\lesssim 2^i + \operatorname{cost}(\xi), &&\\
|\Phi(u) - \Phi(v)| &\le L \|u - v\|_{\mathbb{R}^d} && \text{for all } u, v \in \mathbb{R}^d,\\
c^{-1} \le \Phi(u) &\le c && \text{for all } u \in \mathbb{R}^d.
\end{aligned}$$

Then the Multilevel estimator $\hat{\Theta}_{M(\eta,\varepsilon)}$ defined in (2.2) satisfies

$$\mathrm{P}\left[|\hat{\Theta}_{M(\eta,\varepsilon)} - \mu(f)| > \eta\right] < \varepsilon,$$

and

$$\operatorname{cost}\left(\hat{\Theta}_{M(\eta,\varepsilon)}\right) \le \frac{C}{\eta^2 \varepsilon} \log^4\left(\frac{1}{\eta\varepsilon}\right).$$

Proof. Under the assumptions of this theorem, Lemmas 2.13 and 2.14 imply Assumption 1 and 2. Assumption 3 follows from Lemma 2.12. Finally, Lemma 2.15 shows that Assumption 4 is satisfied, such that we can apply Theorem 2.1 which implies the result. □

2.5 Numerical Results

This section presents the results of a numerical implementation of the Multilevel algorithm to show its behaviour in a concrete example. The results are compared to the implementation of a Singlelevel algorithm defined on only one discretization level. The examples discussed here are not completely covered by the theory of the previous sections, for example the considered functions are not Lipschitz–continuous as assumed in Section 2.4. Nevertheless the Multilevel algorithm is shown to outperform the Singlelevel algorithm.

We analyze the estimation of $\mu(f)$, where μ is a distribution that is absolutely continuous with respect to the distribution ν of the one–dimensional Brownian Bridge starting in $x_0 = 0$ and ending in $x_1 = 0$. The density $\varphi = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is given by

$$\varphi(x) = \frac{1}{Z} \exp\left(-\lambda \left(\max_{s\in[0,1]} x_s \right)^2 \right),$$

where Z is the normalization constant. The function f is given by $f(x) := \max\limits_{s\in[0,1]} x_s$. As the maximum of a Brownian Bridge satisfies

$$\mathrm{P}\left[\max_{t\in[0,1]} B_t \in \mathrm{d}z \right] = 4z \exp\left(-2z^2\right) \qquad \text{for } z \geq 0,$$

see e.g. [31, Proposition 8.1], the exact expectation value is given by

$$\begin{aligned} \mu(f) &= \frac{\int_0^\infty z^2 \exp\left(-(2+\lambda)z^2\right) \mathrm{d}z}{\int_0^\infty z \exp\left(-(2+\lambda)z^2\right) \mathrm{d}z} \\ &= \sqrt{\frac{\pi}{4(\lambda+2)}} \end{aligned}$$

As finite–dimensional approximations of the infinite–dimensional space $C_0([0,1],\mathbb{R})$ we choose piece–wise linear functions on a equidistant partition $\{(l_k^i)_k,\ k \in D_i\}$ with $D_i := \{1,\ldots,2^i\}$ as in Section 2.4, and f and φ are discretized by

$$\begin{aligned} \varphi_i(x) &:= \exp\left(-\lambda \left(\max_{l\in D_i} x_{l_k^i} \right)^2 \right) \\ f_i(x) &:= \max_{l\in D_i} x_{l_k^i}. \end{aligned}$$

Π_i is given by piece-wise interpolations as in Section 2.4, and ν_i as the image–measure of ν under Π_i.

We assess the performance of the Multilevel estimator for different values of λ. Therefore, we calculate 120 samples of the Multilevel estimator $\hat{\Theta}_{M_T}$ defined in Section 2.4,:

$$\hat{\Theta}_{M_T} = \sum_{i=1}^{M_T} \sum_{k=0}^{N_i} h_i(X^i_{n_i+k}, Y^i_{n_i+k}).$$

As we analyze the error of the Multilevel estimator over a long period of time, the optimal number of levels changes over time. The chain is started with $M_0 = 10$ levels, adding an additional level after 2^i minutes, for $i \in \{-5, -4, \ldots\}$.

We take 64 minutes of CPU-time to calculate the estimator, where the number of steps on level i decrease exponentially. More precisely, the chain on level i is calculated for $n_i + N_i = \lfloor \frac{1}{M_T}(2^{i-1} - 1)^{-1}(N_1 + n_1) \rfloor$ steps. Here the factor $(2^{i-1} - 1)$ corresponds to the dimension of the approximation on level i. N_1 is then determined by the limit of CPU–time. The burn-in is chosen as $n_i = 100$ and the step–size as $h = 0.7$.

For comparison, we calculate the ergodic average of Singlelevel chains Z_i reversible with respect to μ_i,

$$\hat{\Theta}^S_i = \frac{1}{N_i} \sum_{k=n_i}^{N_i} f_i(Z^i_k)$$

for $i = 14 \ldots 20$, also with 64 minutes CPU–time each. This is repeated 120 times to produce 120 independent samples of the estimators $\hat{\Theta}_M$ and $\hat{\Theta}^S_i$.

Figure 2.1 shows the mean square error of the estimators for $\lambda = 2$. The Multilevel algorithm's quadratic error (black line) is compared to the Singlelevel's errors (coloured lines). The Multilevel error is always lower than the one of the best instance of the Singlelevel algorithms, for roughly a factor 3 after some seconds, and a factor 30 after 1 hour. Furthermore, it can be observed for the Singlelevel algorithm, that the lowest–dimensional approximation has the smallest error of all instances in the beginning, but after some time its error does not decrease any more and one of the higher–dimensional approximations has the lowest error.

The steps in the Multilevel's graph can be explained by the chosen burn–in and the successive addition of levels. The higher levels enter the calculation only after some time, and when they do, the error drops fast.

The Multilevel estimator is quite sensible with respect to the increase of the density's oscillation. To demonstrate the effect, we set $\lambda = 10$ and repeat the simulations. In

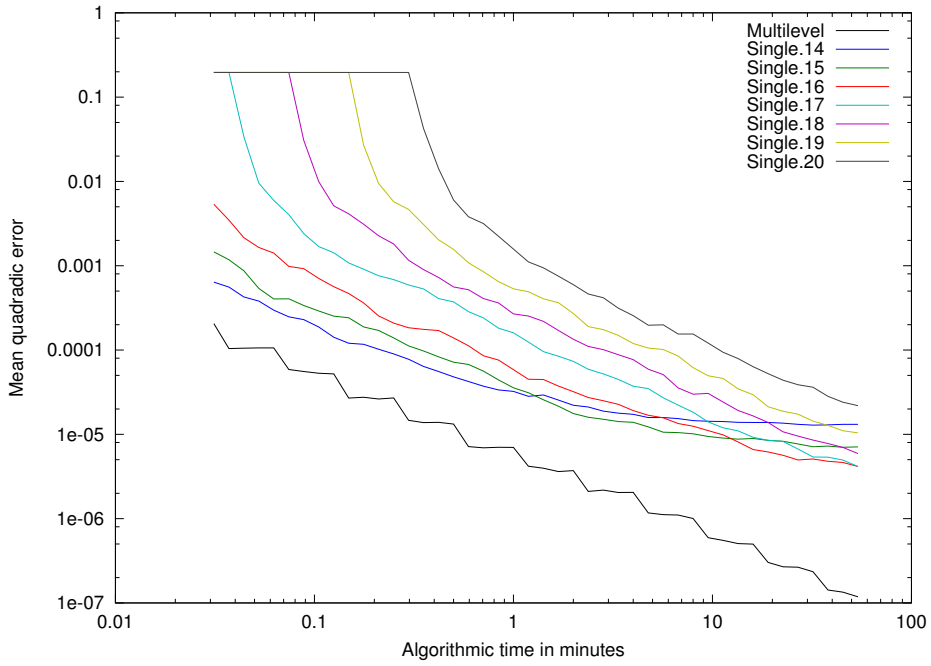

Figure 2.1: Comparison of the Multilevel algorithm to Singlelevel algorithm on different discretization levels for parameters $\lambda = 2$.

Figure 2.2, the mean square error of the estimators are depicted. For $\lambda = 10$, the Multilevel algorithm takes some time to perform better than the best instance of the Singlelevel algorithms, and in the end it is only a factor of about 6 better than the Singlelevel's. Furthermore, we see bumps in the Multilevel's graph where the error temporarily increases. These are caused by the quotient of the densities $\frac{\varphi_i(x)}{\varphi_{i+1}(x)}$ which can get very large for some values of x, leading to large values of the function h_i and to an increasing of the estimator. The likelihood for this scenario depends on the choice of the approximation of φ. In our example the quotient grows exponentially in λ. For $\lambda = 20$, the effect is so strong that several instances of the Singlelevel algorithms outperform the Multilevel algorithm due to these effects. This demonstrates that the performance of the Multilevel estimator crucially depends on a good sequence of approximations for φ, such that the quotients $\frac{\varphi_i}{\varphi_{i+1}}$ and $\frac{\varphi_{i+1}}{\varphi_i}$ can be controlled.

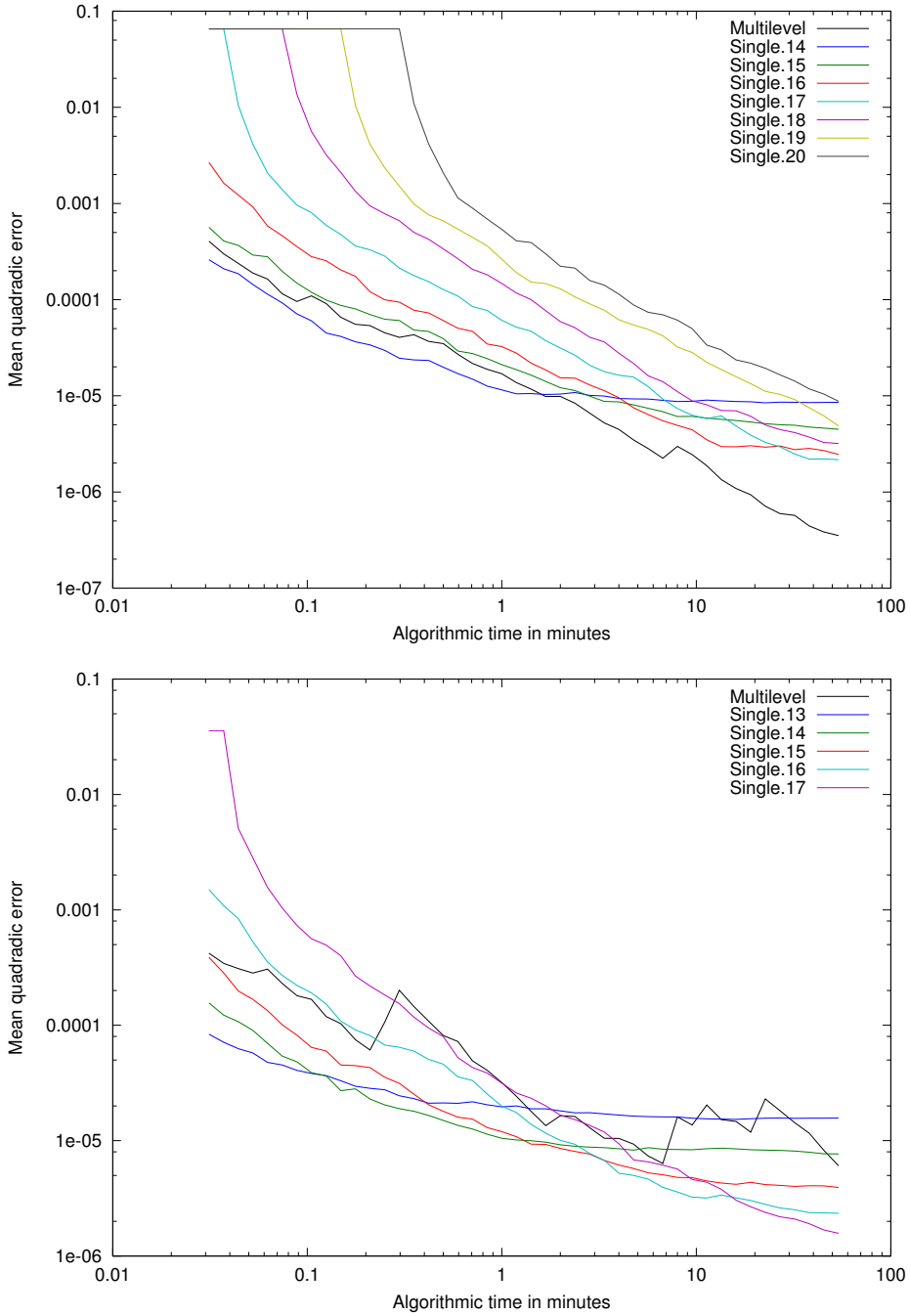

Figure 2.2: Comparison of the Multilevel algorithm to Singlelevel algorithm on different discretization levels for parameters $\lambda = 10$ (left) and $\lambda = 20$ (right).

Chapter 3

Speed of convergence of the MALA–process in infinite dimensions

In this chapter, we analyze the speed of convergence of a Markov Chain Monte Carlo process in a potentially infinite–dimensional state space. This is partly motivated by the results of the previous chapter. For controlling the error of the Multilevel method, we need a control on the speed of convergence of the underlying Markov Chains, c.f. Assumption 2.4. This chapter outlines a method to bound the distance to equilibrium of a particular Markov Chain, called the Langevin Adjusted Metropolis Algorithm (MALA), for log–concave target measures that are absolutely continuous to a Gaussian measure.
Again, the main motivating example are target measures arising in the Transition Path Sampling introduced in Chapter 1, and we apply the results in this setting. The methods applied in this chapter are an application of the approach of Eberle [16]. In that work, the MALA–process with log–concave target measure is analyzed in a finite–dimensional setting, and its distance to equilibrium in an appropriate Wasserstein–metric is bounded using coupling methods. As these techniques are designed to scale well in high–dimensional settings, they carry over quite directly to the infinite–dimensional case.

We are now introducing the setting for the MALA–process before defining it in detail.

Let W be a separable Hilbert space, and ν a Gaussian measure on W with mean 0 and covariance operator $\mathcal{C} : \mathrm{Dom}(\mathcal{C}) \supset W \to W$. We consider the probability measure μ on W given by

$$\mu(\mathrm{d}x) = \frac{1}{Z} \exp(-V(x))\nu(\mathrm{d}x), \tag{3.1}$$

where V is a Borel–measurable function $V : W \to \mathbb{R}$, and $Z > 0$ is the normalization constant such that $\int_W \mu(\mathrm{d}x) = 1$. Let S be the Cameron–Martin space of ν,

$$S := \mathrm{Dom}(\mathcal{C}^{-\frac{1}{2}})$$

equipped with the scalar product

$$\langle x, y \rangle_S := \left\langle \mathcal{C}^{-\frac{1}{2}} x, \mathcal{C}^{-\frac{1}{2}} y \right\rangle_W.$$

We denote with c_π the operator–norm of $\mathcal{C}$ on S, which coincides with the Poincaré constant of $\|\cdot\|_S$ with respect to $\|\cdot\|_W$:

$$\|\mathcal{C}x\|_S \leq c_\pi \|x\|_S \qquad \text{for all } x \in S,$$

which implies

$$\|x\|_W = \|\mathcal{C}x\|_S \leq c_\pi \|x\|_S \qquad \text{for all } x \in S.$$

Given a space $W \supset S$, we denote with W' its topological dual space. As $S \subset W$, W' is continuously embedded in S via Riesz isometry. We identify W' with its embedding in S and also denote it with W'. For $k \in W'$, we can extend the function $\langle k, \cdot \rangle_S : S \to \mathbb{R}$ to a function $\langle k, \cdot \rangle_S : W \to \mathbb{R}$, by defining

$$\langle k, x \rangle_S := k(x) \qquad \forall x \in W. \tag{3.2}$$

We define the function $U : S \to \mathbb{R}$ by

$$U(x) := V(x) + \frac{1}{2}\|x\|_S^2.$$

In finite dimensions, μ can be written as

$$\mu(\mathrm{d}x) \equiv \exp(-U(x))\mathbf{d}x,$$

where $\mathbf{d}x$ is the Lebesgue measure on W. Of course, in infinite dimensions, the Lebesgue measure does not exist, and U is ν–almost surely not defined. Nevertheless, this notation is meaningful in many contexts, for example when considering finite–dimensional approximation or using Girsanov's formula.

We are going to analyze stochastic processes with invariant measure μ as given by (3.1), especially in their speed of convergence to equilibrium. For this purpose, the MALA–process will be appliedin the setting presented above. It was briefly introduced in the Introduction in Chapter 1, a more detailed construction is given in Section 3.1.
Before we move on, we would like to connect the setting above to our running example.
A wide class of distributions which are of type (3.1) are measures on path spaces. In particular, the Transition Path Sampling setting also fits into in this framework. Let $x_0, x_1 \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth potential, and B_t be a $\mathbb{R}^d$–valued Brownian motion, and let μ be the distribution of the solution of the stochastic differential equation

$$\begin{aligned} \mathrm{d}X_t &= -\nabla f(X_t)\mathrm{d}t + \mathrm{d}B_t, \\ X_0 &= x_0, \end{aligned}$$

conditioned on the event $\{X_1 = x_1\}$. This distribution is of the form described above:
Set $E := L^2([0,1], \mathbb{R}^d)$, and let ν be the distribution of the Brownian Bridge. ν is the Gaussian measure with mean 0 and covariance operator $\mathcal{C}_E := (-\Delta_0)^{-1}$ on E, where Δ_0 is the Laplacian on $[0,1]$ with zero boundary conditions. This implies $S := H_0^1([0,1], \mathbb{R}^d)$ with norm $\|x\|_S := \int_0^1 |x_s'|^2 \mathrm{d}s$.
Using Girsanov's formula and integration by parts, it is shown in [24], that μ is absolutely continuous with respect to ν: For

$$\varphi(x) := \exp(-V(x)),$$

μ is given by

$$\mu(\mathrm{d}x) = \frac{1}{Z}\varphi(x)\nu(\mathrm{d}x), \tag{3.3}$$

where Z is a normalization constant and for $x \in E$

$$V(x) := \int_0^1 \Phi(x_s)\mathrm{d}s$$

and

$$\Phi(z) := \frac{1}{2}\left(\Delta f(z) + |\nabla f(z)|^2\right) \qquad \text{for } z \in \mathbb{R}^d.$$

3.1 Construction of the MALA–process

We now give an explicit construction of the MALA–process that we later analyze. The MALA–process goes back to [39], although the version we use here is a slight variation of the original process, that keeps the process stable in the infinite–dimensional limit. The version used here also coincides with the "Preconditioned Implicit Algorithm" in [8] with parameter $\theta = \frac{1}{2}$.

The first step in our construction of the MALA–process is the discrete–time Ornstein–Uhlenbeck process. This process is reversible with respect to the Gaussian measure ν. It can be constructed as follows:
Let $(N_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d ν-distributed random variables on W. For given $h \in (0,2)$, set $(Z_n^h)_{n\in\mathbb{N}}$ as

$$Z_{n+1}^h := \left(1 - \frac{h}{2}\right) Z_n^h + \sqrt{\tilde{h}} N_n. \tag{3.4}$$

Here, and for the rest of this work, $\tilde{h}$ is defined as

$$\tilde{h} := h - \frac{h^2}{4}. \tag{3.5}$$

As $(Z_n^h)_{n\in\mathbb{N}}$ is a time–homogeneous Markov Process, it induces a stochastic kernel $\tilde{q}_h$ by

$$\tilde{q}_h(x, A) := \mathrm{P}\left[Z_{n+1}^h \in A \middle| Z_n^h = x\right] \qquad \text{for } x \in W, A \in \mathcal{B}(W),$$

where $\mathcal{B}(W)$ denotes the Borel sets of W.

We now show that the kernel $\tilde{q}_h$ is reversible with respect to ν:

Proposition 3.1. *The kernel $\tilde{q}_h$ is reversible with respect to ν.*

Proof. We consider the characteristic function of the measure $\nu\tilde{q}_h$. Let $l_1, l_2 \in W$. As ν and $\tilde{q}_h$ are Gaussian measures, we get for the characteristic function

$$\begin{aligned}
&\int_{W\times W} \exp\left(-i\langle (l_1,l_2),(x,y)\rangle_W\right)\nu(\mathrm{d}x)\tilde{q}_h(x,\mathrm{d}y)\\
&= \int_{W\times W} \exp\left(-i\left\langle (l_1,l_2),\left(x,\left(1-\frac{h}{2}\right)x+y\right)\right\rangle_W\right)\nu(\mathrm{d}x)\tilde{q}_h(0,\mathrm{d}y)\\
&= \exp\left(-\frac{1}{2}\left\|l_1+\left(1-\frac{h}{2}\right)l_2\right\|_S^2-\frac{1}{2}\tilde{h}\|l_2\|_S^2\right).
\end{aligned}$$

The characteristic function is symmetric in l_1, l_2 if and only if $\tilde{q}_h$ is reversible with respect to ν. The exponent can be written as

$$\begin{aligned}
&\frac{1}{2}\left\|l_1+\left(1-\frac{h}{2}\right)l_2\right\|_S^2+\frac{1}{2}\tilde{h}\|l_2\|_S^2\\
&= \frac{1}{2}\|l_1\|_S^2+\left\langle l_1,\left(1-\frac{h}{2}\right)l_2\right\rangle_S+\frac{1}{2}\left\|\left(1-\frac{h}{2}\right)l_2\right\|_S^2+\frac{1}{2}\tilde{h}\|l_2\|_S^2
\end{aligned}$$

As $\left\langle l_1, \left(1-\frac{h}{2}\right)l_2\right\rangle_S$ is symmetric and

$$\left(1-\frac{h}{2}\right)^2\|l_2\|_S^2+\tilde{h}\|l_2\|_S^2=\|l_2\|_S^2,$$

the characteristic function is symmetric and $\tilde{q}_h$ is reversible with respect to ν. □

We now construct a discrete–time process which is reversible with respect to μ by a variant of the Metropolis–Hastings scheme, the MALA–process. The MALA–process accounts for the gradient of the potential in the proposal of the Metropolis chain, which asymptotically (for $h \to 0$) leads to an high acceptance probability. This property is needed to get good bounds on the derivatives of the acceptance probability. The bounds are used in the proof of the contraction property of the process.

Let $(N_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d ν–distributed random variables on W and for given $x_0 \in W$, set $X_0 := x_0$. Define the random variable $Y_{h,n}(x)$ by

$$Y_{h,n}(x) := \left(1-\frac{h}{2}\right)x-\frac{h}{2}\nabla_S V(x)+\sqrt{\tilde{h}}N_{n+1}, \tag{3.6}$$

or, in terms of U, by

$$Y_{h,n}(x) = x-\frac{h}{2}\nabla_S U(x)+\sqrt{\tilde{h}}N_{n+1},$$

where $\tilde{h} = h - \frac{h^2}{4}$ as above. $Y_{h,n}(X_n)$ serves as proposal of the Metropolis chain, we denote the kernel generated by $(Y_{h,n})_{n\in\mathbb{N}}$ with q_h.
The proposal is accepted with probability $a_h(X_n, Y_{h,n}(X_n))$, where the acceptance probability $a : W \times W \to [0,1]$ is given by

$$a_h(x,y) := \min\left(1, \frac{\mu(\mathrm{d}y)q_h(y,\mathrm{d}x)}{\mu(\mathrm{d}x)q_h(x,\mathrm{d}y)}\right) \qquad \text{for } x, y \in W. \tag{3.7}$$

The proposals are realized by generating a sequence $(U_n)_{n\in\mathbb{N}}$ of i.i.d. uniformly distributed random variables on $[0,1]$ and set

$$X_{n+1} := \begin{cases} Y_{h,n}(X_n) & \text{if } U_{n+1} < a(X_n, Y_{h,n}(X_n)), \\ X_n & \text{otherwise.} \end{cases}$$

The kernel generated be $(X_n)_{n\in\mathbb{N}}$ is denoted by p_h. It is well–known that is reversible with respect to μ if the process is constructed in the way described above. We will also prove this in Lemma 3.2 for the MALA–process considered here. In the setting outlined above, the acceptance probability satisfies the following equation:

Lemma 3.2. *Let $a_h : W \times W \to [0,1]$ be the acceptance probability defined in (3.7). Then a_h is given by*

$$a_h(x,y) = \min\left(1, \exp(-G_h(x,y))\right) \qquad \textit{for } x, y \in W, \tag{3.8}$$

where

$$\begin{aligned} G_h(x,y) := V(y) - V(x) &- \frac{1}{2}\langle \nabla_S V(x) + \nabla_S V(y), y - x\rangle_S \\ &+ \frac{h}{8-2h}\langle \nabla_S V(y) - \nabla_S V(x), x + y\rangle_S \\ &+ \frac{h}{8-2h}\left(\|\nabla_S V(y)\|_S^2 - \|\nabla_S V(x)\|_S^2\right). \end{aligned} \tag{3.9}$$

Remark 3.3. *Note that as for $z \geq 0$, $\min\{1, \exp(-z)\} \leq 1 - z$, thus*

$$1 - a_h(x,y) \leq \max\{G_h(x,y), 0\} =: G_h(x,y)^+$$

holds for $x, y \in W$.

Proof. (Lemma 3.2)
Let $\tilde{q}_h$ be the kernel induced by Z defined in equation (3.4), and q_h the kernel induced by $Y_{h,n}$ defined in equation (3.6). Due to the Cameron Martin formula (see e.g. [12, Proposition

2.24]) we know that for a centered Gaussian measure η with covariance operator $\mathcal{C}$ and $k \in S$, $\eta^k(\cdot) := \eta(\cdot - k)$ is absolutely continuous with respect to η with density

$$\frac{\eta^k(\mathrm{d}y)}{\eta(\mathrm{d}y)} = \exp\left(\langle y, k\rangle_S - \frac{1}{2}\|k\|_S^2\right).$$

We apply this to the centered Gaussian measure

$$\eta(\mathrm{d}y) := \tilde{q}_h\left(x, \mathrm{d}y + \left(1 - \frac{h}{2}\right)x\right),$$

with covariance operator

$$\mathcal{C}^* := \tilde{h}\mathcal{C}$$

for

$$k := -\frac{h}{2}\nabla_S V(x),$$

where $x \in W$. Note that for this choice of k

$$\eta^k(\mathrm{d}y - Ax) = \eta(\mathrm{d}y - Ax - k) = q_h(x, \mathrm{d}y).$$

Applying the Cameron Martin formula, we see that

$$\begin{aligned}
\frac{q_h(x,\mathrm{d}y)}{\tilde{q}_h(x,\mathrm{d}y)} &= \frac{\eta^k(\mathrm{d}y - Ax)}{\eta(\mathrm{d}y - Ax)} \\
&= \exp\left(\langle y - Ax, k\rangle_S - \frac{1}{2}\|k\|_S^2\right) \\
&= \exp\left(-\frac{1}{\tilde{h}}\left\langle \frac{h}{2}\nabla_S V(x), y - \left(1 - \frac{h}{2}\right)x\right\rangle_S - \frac{h^2}{8\tilde{h}}\|\nabla_S V(x)\|_S^2\right) \\
&= \exp\left(-\frac{2}{4-h}\left\langle \nabla_S V(x), y - \left(1 - \frac{h}{2}\right)x\right\rangle_S - \frac{h}{8-2h}\|\nabla_S V(x)\|_S^2\right).
\end{aligned}$$

We can simplify

$$\begin{aligned}
&\frac{2}{4-h}\left\langle \nabla_S V(x), y - \left(1 - \frac{h}{2}\right)x\right\rangle_S \\
&= \frac{1}{2}\langle \nabla_S V(x), y - x\rangle_S + \frac{h}{8-2h}\langle \nabla_S V(x), y - x\rangle_S + \frac{h}{4-h}\langle \nabla_S V(x), x\rangle_S \\
&= \frac{1}{2}\langle \nabla_S V(x), y - x\rangle_S + \frac{h}{8-2h}\langle \nabla_S V(x), y + x\rangle_S
\end{aligned}$$

which leads to

$$\frac{q_h(x,\mathrm{d}y)}{\tilde{q}_h(x,\mathrm{d}y)} = \exp\left(-\frac{1}{2}\langle \nabla_S V(x), y - x\rangle_S - \frac{h}{8-2h}\left(\langle \nabla_S V(x), y + x\rangle_S + \|\nabla_S V(x)\|_S^2\right)\right). \tag{3.10}$$

We rewrite

$$\begin{aligned}&\frac{\mu(\mathrm{d}y)q_h(y,\mathrm{d}x)}{\mu(\mathrm{d}x)q_h(x,\mathrm{d}y)}\\ =\ &\frac{\varphi(y)}{\varphi(x)}\frac{\nu(\mathrm{d}y)\tilde{q}_h(y,\mathrm{d}x)}{\nu(\mathrm{d}x)\tilde{q}_h(x,\mathrm{d}y)}\frac{q_h(y,\mathrm{d}x)}{\tilde{q}_h(y,\mathrm{d}x)}\frac{\tilde{q}_h(x,\mathrm{d}y)}{q_h(x,\mathrm{d}y)}.\end{aligned}$$

Since by Proposition 3.1, $\tilde{q}_h$ is reversible with respect to ν,

$$\frac{\nu(\mathrm{d}y)\tilde{q}_h(y,\mathrm{d}x)}{\nu(\mathrm{d}x)\tilde{q}_h(x,\mathrm{d}y)} \equiv 1$$

holds. With equation (3.10), we get

$$\begin{aligned}&\frac{\mu(\mathrm{d}y)q_h(y,\mathrm{d}x)}{\mu(\mathrm{d}x)q_h(x,\mathrm{d}y)}\\ =\ &\frac{\varphi(y)}{\varphi(x)}\frac{q_h(y,\mathrm{d}x)}{\tilde{q}_h(y,\mathrm{d}x)}\frac{\tilde{q}_h(x,\mathrm{d}y)}{q_h(x,\mathrm{d}y)}\\ =\ &\exp(-V(y)+V(x))\\ &\cdot\exp\left(-\frac{1}{2}\langle\nabla_S V(y),x-y\rangle_S-\frac{h}{8-2h}\left(\langle\nabla_S V(y),x+y\rangle_S+\|\nabla_S V(y)\|_S^2\right)\right)\\ &\cdot\exp\left(\frac{1}{2}\langle\nabla_S V(x),y-x\rangle_S+\frac{h}{8-2h}\left(\langle\nabla_S V(x),y+x\rangle_S+\|\nabla_S V(x)\|_S^2\right)\right)\\ =\ &\exp(-V(y)+V(x))\\ &\cdot\exp\left(-\frac{1}{2}\langle\nabla_S V(x)+\nabla_S V(y),x-y\rangle_S\right)\\ &\cdot\exp\left(-\frac{h}{8-2h}\langle\nabla_S V(y)-\nabla_S V(x),x+y\rangle_S\right)\\ &\cdot\exp\left(-\frac{h}{8-2h}\left(\|\nabla_S V(y)\|_S^2-\|\nabla_S V(x)\|_S^2\right)\right)\\ =\ &\exp(-G_h(x,y)).\end{aligned}$$

This shows

$$a_h(x,y)=\min\left\{1,\exp\left(-G_h(x,y)\right)\right\}.$$

□

In the following, we also use an alternative representation of the acceptance probability.

Lemma 3.4. *For all $x, y \in W$, $G_h(x,y)$ satisfies*

$$\begin{aligned} G_h(x,y) = V(y) - V(x) - \frac{1}{2}\langle \nabla_S V(y) + \nabla_S V(x), y - x\rangle_S \\ + \frac{h}{8-2h}\langle \nabla_S U(y) + \nabla_S U(x), \nabla_S V(y) - \nabla_S V(x)\rangle_S \end{aligned}$$

Proof. By the definition of U, we have

$$\begin{aligned} &\langle \nabla_S V(y) - \nabla_S V(x), x + y\rangle_S + \|\nabla_S V(y)\|_S^2 - \|\nabla_S V(x)\|_S^2 \\ &= \langle \nabla_S V(y) - \nabla_S V(x), x + y\rangle_S + \langle \nabla_S V(y) - \nabla_S V(x), \nabla_S V(y) + \nabla_S V(x)\rangle_S \\ &= \langle \nabla_S V(y) - \nabla_S V(x), x + \nabla_S V(x) + y + \nabla_S V(y)\rangle_S \\ &= \langle \nabla_S V(y) - \nabla_S V(x), \nabla_S U(x) + \nabla_S U(y)\rangle_S. \end{aligned}$$

Therefore,

$$\begin{aligned} G_h(x,y) &= V(y) - V(x) - \frac{1}{2}\langle \nabla_S V(x) + \nabla_S V(y), y - x\rangle_S \\ &\quad + \frac{h}{8-2h}\langle \nabla_S V(y) - \nabla_S V(x), x + y\rangle_S \\ &\quad + \frac{h}{8-2h}\left(\|\nabla_S V(y)\|_S^2 - \|\nabla_S V(x)\|_S^2\right) \\ &= V(y) - V(x) - \frac{1}{2}\langle \nabla_S V(x) + \nabla_S V(y), y - x\rangle_S \\ &\quad + \frac{h}{8-2h}\langle \nabla_S V(y) - \nabla_S V(x), \nabla_S U(x) + \nabla_S U(y)\rangle_S. \end{aligned}$$

□

3.2 Transition Path Sampling

We now return to the Transition Path Sampling setting and show how the setting can be chosen to fit the abstract frame. We start with the infinite–dimensional distribution on the path space. For simulations on a computer, this distribution will be approximated on finite–dimensional spaces. This case will be handled in the second part.

3.2.1 The infinite–dimensional Transition Path Sampling process

We start with identifying a proper space W on which the process is realized. The straightforward choice $W = E = L^2([0,1], \mathbb{R}^d)$ turns out to suit our needs. Note that for $x \in E$, $V(x)$ might not be defined if $\Phi(x.) \notin L^1([0,1], \mathbb{R}^d)$, e.g. if $\Phi(z)$ growth faster

than quadratically. Assumption 3.1 even requires that derivatives of V up to fourth order are well-defined for all $x \in W$. This will require $\Phi(x.)$ to be an element of $L^5([0,1],\mathbb{R}^d)$. Assuming $\Phi(z)$ is polynomial, we also need $x \in L^q([0,1],\mathbb{R}^d)$ for some sufficient large $q > 5$, depending on the degree of Φ. As we also want W to be a Hilbert space, we rely on fractional Sobolev spaces. More precisely, the spaces W_α constructed below form a slightly different family of spaces. The choice of the norm is motivated by its compatibility with the piece–wise linear approximation we use in Chapter 3.2.2. But they satisfy the same embeddings as the standard fractional Sobolev spaces, namely $W_{\frac{1}{2}-\frac{1}{q}} \subset L^q([0,1],\mathbb{R}^d)$, and the Brownian Bridge is supported on W_α for $\alpha < \frac{1}{2}$.

We now state the main assumptions needed to prove the convergence result. Firstly, we assume that $\Phi(z)$ is indeed bounded by a polynomial.

Assumption 3.1. *For all $\eta^1,\dots,\eta^n \in \mathbb{R}^d$ with $\|\eta^i\|_{\mathbb{R}^d} = 1$,*

$$\left|\mathrm{D}^n\Phi(z)(\eta^1,\dots,\eta^n)\right| \le C_n\left(\max\{1,\|z\|_{\mathbb{R}^d}\}\right)^{p_n} \tag{3.11}$$

for $n = 1,\dots,4$ and constants C_n and p_n. Define

$$p := \max_{i\in\{1,\dots,4\}} p_i.$$

The second assumption we need is a uniform bound on the second derivative on Φ.

Assumption 3.2. *The second derivative of Φ is uniformly bounded by $L_\Phi < \frac{\pi}{\sqrt{2}}$: For all $z \in \mathbb{R}^d$, and all η^1,η^2 in $\mathbb{R}^d$*

$$\left|\mathrm{D}^2\Phi(z)(\eta^1,\eta^2)\right| \le L_\Phi\|\eta^1\|_{\mathbb{R}^d}\|\eta^2\|_{\mathbb{R}^d}.$$

We now construct the spaces W_α. Let $e_{i,k,j}$, $i \in \mathbb{N}$, $k \in \{1,\dots,2^{i-1}\}$, $j \in \{1,\dots,d\}$ be the Schauder basis given by

$$e_{i,k,j}(s) := 2^{-\frac{i}{2}}\phi(2^i s - k + 1)\,\mathbf{e}_j, \qquad \text{for } s \in [0,1], \tag{3.12}$$

where $\mathbf{e}_j$ is the j-th unit vector in $\mathbb{R}^d$ and $\phi : \mathbb{R} \to \mathbb{R}$ is the function

$$\phi(s) := \frac{1}{2}\max\{0, 1 - |1 - s|\}.$$

These vectors form an orthonormal basis of $H_0^1([0,1],\mathbb{R}^d)$, and a basis of $L^2([0,1],\mathbb{R}^d)$. For $x \in L^2([0,1],\mathbb{R}^d)$ of the form

$$x = \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} \sum_{j=1}^{d} x_{i,k,j} e_{i,k,j},$$

define the norm $\|\cdot\|_{W_\alpha}$ by

$$\|x\|_{W_\alpha} := \left(\sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} \sum_{j=1}^{d} 2^{-2(1-\alpha)i} x_{i,k,j}^2 \right)^{\frac{1}{2}} \qquad \text{for } x \in L^2([0,1],\mathbb{R}^d).$$

The space W_α is defined as

$$W_\alpha := \left\{ x \in L^2([0,1],\mathbb{R}^d) \,\middle|\, \|x\|_{W_\alpha} < \infty \right\}.$$

We define the operator $\mathcal{C}_\alpha : \mathcal{W}_\alpha \to S$ by

$$\mathcal{C}_\alpha \left(\sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} \sum_{j=1}^{d} x_{i,k,j} e_{i,k,j} \right) = \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} \sum_{j=1}^{d} 2^{-2(1-\alpha)i} x_{i,k,j} e_{i,k,j}.$$

With this notation, we have for $x \in W_\alpha$,

$$\|x\|_{W_\alpha} = \left\| \mathcal{C}_\alpha^{\frac{1}{2}} x \right\|_S.$$

Depending on the value of α, the support of the measure μ is contained in W_α, and W_α is a subspace of $L^q([0,1],\mathbb{R}^d)$. This is shown by the following two lemmas, This implies that for $\frac{1}{2} - \frac{1}{q} < \alpha < \frac{1}{2}$, both the distribution of the diffusion bridge is supported on the space W_α, and W_α is a subspace of $L^q([0,1],\mathbb{R}^d)$.

Lemma 3.5. *If $0 < \alpha < \frac{1}{2}$, then*

$$\operatorname{supp} \mu \subset W_\alpha.$$

Lemma 3.6. *If $\alpha > \frac{1}{2} - \frac{1}{q}$, then*

$$W_\alpha \subset L^q([0,1],\mathbb{R}^d).$$

Proof. (Lemma 3.5)
The Brownian Bridge can be constructed by the Wiener–Lévy expansion: For $i \in \mathbb{N}$, $k \in$

$\{1,\dots,2^i\}$, $j \in \{1,\dots,d\}$, take one–dimensional Gaussian random variables $Z_{i,k,j}$ with mean 0 and variance 1, and set

$$X := \sum_{j=1}^{d}\sum_{i=1}^{\infty}\sum_{k=1}^{2^{i-1}} Z_{i,k,j}e_{i,k,j}.$$

Then X is a Brownian Bridge. Now assume $\alpha < \frac{1}{2}$, then

$$\begin{aligned}
\mathbb{E}\left[\|X\|_{W_\alpha}^2\right] &= \sum_{j=1}^{d}\sum_{i=1}^{\infty}\sum_{k=1}^{2^{i-1}} 2^{-2(1-\alpha)i}\mathbb{E}\left[Z_{i,k,j}^2\right] \\
&= \sum_{j=1}^{d}\sum_{i=1}^{\infty} 2^{i-1}2^{-2(1-\alpha)i} \\
&= \frac{1}{2}d\sum_{i=1}^{\infty} 2^{-(1-2\alpha)i} \\
&< \infty.
\end{aligned}$$

Thus the Brownian Bridge is almost surely supported on W_α. As μ is absolutely continuous with respect to ν, it is also supported on W_α. □

Proof. (Lemma 3.6) Let $x \in W_\alpha$ with

$$x = \sum_{i=1}^{\infty}\sum_{j=1}^{d}\sum_{k=1}^{2^i} x_{i,k,j}e_{i,k,j}$$

where $e_{i,k,j}$, $i \in \mathbb{N}$, $k \in \{1,\dots,2^{i-1}\}$, $j \in \{1,\dots,d\}$ form the Schauder basis orthonormal in $H_0^1([0,1],\mathbb{R}^d)$. The L^q–norm can be bounded by

$$\begin{aligned}
\|x\|_{L^q} &= \left\|\sum_{i=1}^{\infty}\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} x_{i,k,j}e_{i,k,j}\right\|_{L^q} \\
&\le \sum_{i=1}^{\infty}\left\|\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} x_{i,k,j}e_{i,k,j}\right\|_{L^q}.
\end{aligned}$$

As $e_{i,k,j}$ and $e_{i,\tilde{k},\tilde{j}}$ have disjoint support for $(k,j) \neq (\tilde{k},\tilde{j})$, we get

$$\begin{aligned}\left\|\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} x_{i,k,j}e_{i,k,j}\right\|_{L^q}^q &= \int_0^1 \left|\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} x_{i,k,j}e_{i,k,j}(s)\right|^q \mathrm{d}s \\ &= \sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} |x_{i,k,j}|^q \int_{(k-1)2^{-i}}^{k2^{-i}} e_{i,k,j}^q(s)\mathrm{d}s \\ &\leq 2^{-\frac{iq}{2}} 2^{-i} \sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} |x_{i,k,j}|^q.\end{aligned}$$

This gives

$$\begin{aligned}\left\|\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} x_{i,k,j}e_{i,k,j}\right\|_{L^q} &\leq 2^{-\frac{i}{2}} 2^{-\frac{i}{q}} \left(\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} |x_{i,k,j}|^q\right)^{\frac{1}{q}} \\ &\leq 2^{-\frac{i}{2}} 2^{-\frac{i}{q}} \left(\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} |x_{i,k,j}|^2\right)^{\frac{1}{2}}\end{aligned}$$

resulting in

$$\begin{aligned}\|x\|_{L^q} &\leq \sum_{i=1}^{\infty} \left\|\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} x_{i,k,j}e_{i,k,j}\right\|_{L^q} \\ &\leq \sum_{i=1}^{\infty} 2^{-\frac{i}{2}} 2^{-\frac{i}{q}} \left(\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} |x_{i,k,j}|^2\right)^{\frac{1}{2}}.\end{aligned}$$

We choose a sequence $\varepsilon_i := 2^{i\left(1-2\alpha-\frac{2}{q}\right)}$. Note that $1-2\alpha-\frac{2}{q} < 0$ if and only if $\alpha > \frac{1}{2} - \frac{1}{q}$. For these α, we can show

$$\begin{aligned}\|x\|_{L^q} &\leq \sum_{i=1}^{\infty} 2^{-\frac{i}{2}} 2^{-\frac{i}{q}} \left(\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} |x_{i,k,j}|^2\right)^{\frac{1}{2}} \\ &\leq \left(\sum_{i=1}^{\infty} \varepsilon_i\right)^{\frac{1}{2}} \left(\sum_{i=1}^{\infty}\sum_{j=1}^{d}\sum_{k=1}^{2^{i-1}} \frac{1}{\varepsilon_i} 2^{-\left(1+\frac{2}{q}\right)i} |x_{i,k,j}|^2\right)^{\frac{1}{2}} \\ &\leq C_{\alpha,q} \left(\sum_{j=1}^{d}\sum_{i=1}^{\infty}\sum_{k=1}^{2^{i-1}} 2^{-2(1-\alpha)i} |x_{i,k,j}|^2\right)^{\frac{1}{2}} \\ &\leq C_{\alpha,q} \|x\|_{W_\alpha}\end{aligned}$$

where

$$C_{\alpha,q} := \left(\sum_{i=1}^{\infty} 2^{i\left(1-2\alpha-\frac{2}{q}\right)} \right)^{\frac{1}{2}} < \infty \tag{3.13}$$

for $\alpha > \frac{1}{2} - \frac{1}{q}$. □

We now choose $W := W_\alpha$ for $\alpha := \frac{1}{2} - \frac{1}{6p}$ which implies $W_\alpha \subset L^{rp}([0,1], \mathbb{R}^d)$ for $r < 6$.

For this choice of S and W, the covariance operator of ν on W is given by $\mathcal{C}_\alpha$:

Lemma 3.7. *The covariance operator $\mathcal{C}$ on W of ν is given by*

$$\mathcal{C} = \mathcal{C}_\alpha.$$

Proof. For $i \in \mathbb{N}, k \in \{1, \dots, 2^{i-1}\}$ and $j \in \{1, \dots, d\}$, let $\tilde{e}_{i,k,j}$ be defined by

$$\tilde{e}_{i,k,j} := 2^{(1-\alpha)i}\, e_{i,k,j}.$$

By construction of the scalar product $\langle \cdot, \cdot \rangle_W$, this is an orthonormal basis of W. Let X be a ν-distributed random variables given by

$$X := \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} X_{i,k,j} e_{i,k,j} = \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} X_{i,k,j} 2^{-(1-\alpha)i} \tilde{e}_{i,k,j}$$

where $X_{i,k,j}$ are i.i.d Gaussian random variables on $\mathbb{R}$ with mean 0 and variance 1. The covariance operator $\mathcal{C}$ by definition satisfies for given $h, \tilde{h} \in W$

$$\begin{aligned}
\left\langle h, \mathcal{C}\tilde{h} \right\rangle_W &= \mathbb{E}\left[\langle X, h \rangle_W \left\langle X, \tilde{h} \right\rangle_W \right] \\
&= \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} 2^{-2(1-\alpha)} h_{i,k,j} \tilde{h}_{i,k,j} \mathbb{E}\left[X_{i,k,j} X_{i,k,j} \right] \left\| \tilde{e}_{i,k,j} \right\|_W^2 \\
&= \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} 2^{-2(1-\alpha)} h_{i,k,j} \tilde{h}_{i,k,j} \\
&= \left\langle h, \mathcal{C}_\alpha \tilde{h} \right\rangle_W,
\end{aligned}$$

and thus $\mathcal{C} = \mathcal{C}_\alpha$. □

We now identify the terms arising in the MALA–process, and start with the gradient $\nabla_S V(x)$.

Lemma 3.8. *For $x \in W$, $\nabla_S V(x)$ is given by*

$$(\nabla_S V(x))_t = \left((-\Delta_0^{-1})\nabla\Phi(x.)\right)_t \tag{3.14}$$

$$= -\int_0^t \int_0^s \nabla\Phi(x_u)\mathrm{d}u\mathrm{d}s + t\int_0^1 \int_0^1 \nabla\Phi(x_u)\mathrm{d}u\mathrm{d}s. \tag{3.15}$$

Proof. For $k \in S$ and $x \in W$, the derivative of $V(x)$ in direction k is given by

$$\begin{aligned}\frac{\partial}{\partial k}V(x) &= \int_0^1 \nabla\Phi(x_s)\cdot k_s\mathrm{d}s \\ &= \langle\nabla\Phi(x), k\rangle_E \\ &= \left\langle(-\Delta_0)^{-1}\nabla\Phi(x), k\right\rangle_S\end{aligned}$$

which implies

$$(\nabla_S V(x))_t = \left((-\Delta_0^{-1})\nabla\Phi(x.)\right)_t.$$

A direct calculation yields

$$\left((-\Delta_0^{-1})\nabla\Phi(x.)\right)_t = -\int_0^t \int_0^s \nabla\Phi(x_u)\mathrm{d}u\mathrm{d}s + t\int_0^1 \int_0^1 \nabla\Phi(x_u)\mathrm{d}u\mathrm{d}s.$$

□

We also need to calculate the derivative of V in directions of the larger space W. Under Assumption 3.1, this derivative is given by $\langle\nabla_S V(x), \cdot\rangle_S$, which is a well–definied functional $W \to \mathbb{R}$ in this case.

Lemma 3.9. *Let Assumption 3.1 be satisfied. Then, for $x, y \in W$,*

$$\frac{\partial}{\partial y}V(x) = \langle\nabla_S V(x), y\rangle_S = \int_0^1 \nabla\Phi(x_s)\cdot y_s\mathrm{d}s.$$

Proof. By definition, we have

$$\langle\nabla_S V(x), y\rangle_S = \frac{\partial}{\partial y}V(x).$$

For $x, y \in W$, it holds

$$\frac{\partial}{\partial y}V(x) = \int_0^1 \nabla_S V(x_s)\cdot y_s\mathrm{d}s.$$

Assumption 3.1 guarantees that $\nabla_S V(x_s) \cdot y_s$ is integrable as

$$\begin{aligned}\int_0^1 |\nabla\Phi(x_s) \cdot y_s| \mathrm{d}s &\leq \left(C_2 \int_0^1 (1+|x_s|)^{2p} \mathrm{d}s \int_0^1 |y_s|^2 \mathrm{d}s \right)^{\frac{1}{2}} \\ &\leq C_2^{\frac{1}{2}} \|1+x\|^p_{L^{2p}([0,1],\mathbb{R}^d)} \|y\|_{L^2([0,1],\mathbb{R}^d)} \\ &\leq C_2^{\frac{1}{2}} C_{\alpha,2p} C_{\alpha,2} \|1+x\|^p_W \|y\|_W \\ &\leq C_2^{\frac{1}{2}} C_{\alpha,2p} C_{\alpha,2} (1+\|x\|_W)^p \|y\|_W.\end{aligned}$$

□

As we have identified the gradient of V with respect to S, we can define the MALA–process as presented in Chapter 3.1. The acceptance probability a_h of the MALA–process is given by (3.8) and G_h satisfies:

Lemma 3.10. *Let $x, y \in W$, then*

$$\begin{aligned}G_h(x,y) := &\int_0^1 \Phi(y_s) - \Phi(x_s) \mathrm{d}s - \frac{1}{2} \int_0^1 (\nabla\Phi(x_s) + \nabla\Phi(y_s)) \cdot (y_s - x_s) \mathrm{d}s \\ &+ \frac{h}{8-2h} \int_0^1 (\nabla\Phi(y_s) - \nabla\Phi(x_s)) \cdot (x_s + y_s) \mathrm{d}s \\ &+ \frac{h}{8-2h} \int_0^1 \nabla\Phi(y_s) \cdot \left(-\Delta_0^{-1} \nabla\Phi(y)\right)_s - \nabla\Phi(x_s) \cdot \left(-\Delta_0^{-1} \nabla\Phi(x)\right)_s \mathrm{d}s.\end{aligned}$$

Proof. Let $x, y \in W$. By Lemma 3.2, $G_h(x,y)$ satisfies

$$\begin{aligned}G_h(x,y) := &V(y) - V(x) - \frac{1}{2} \langle \nabla_S V(x) + \nabla_S V(y), y - x \rangle_S \\ &+ \frac{h}{8-2h} \langle \nabla_S V(y) - \nabla_S V(x), x + y \rangle_S \\ &+ \frac{h}{8-2h} \left(\|\nabla_S V(y)\|_S^2 - \|\nabla_S V(x)\|_S^2 \right).\end{aligned}$$

Lemmas 3.8 and 3.9 lead to

$$\begin{aligned}\langle \nabla_S V(x), y \rangle_S &= \int_0^1 \nabla\Phi(x_s) \cdot y_s \mathrm{d}s, \\ (\nabla_S V(x))_t &= \left((-\Delta_0^{-1}) \nabla\Phi(x_\cdot) \right)_t\end{aligned}$$

for $k \in W$. This result in.

$$\begin{aligned}\langle \nabla_S V(x) + \nabla_S V(y), y - x \rangle_S &= \int_0^1 (\nabla\Phi(x_s) + \nabla\Phi(y_s)) \cdot (y_s - x_s) \mathrm{d}s, \\ \langle \nabla_S V(y) - \nabla_S V(x), x + y \rangle_S &= \int_0^1 (\nabla\Phi(y_s) - \nabla\Phi(x_s)) \cdot (x_s + y_s) \mathrm{d}s\end{aligned}$$

and

$$\begin{aligned}&\|\nabla_S V(y)\|_S^2 - \|\nabla_S V(x)\|_S^2\\ &= \int_0^1 \nabla\Phi(y_s)\cdot\left(-\Delta_0^{-1}\nabla\Phi(y)\right)_s - \nabla\Phi(x_s)\cdot\left(-\Delta_0^{-1}\nabla\Phi(x)\right)_s \mathrm{d}s.\end{aligned}$$

□

3.2.2 A finite–dimensional approximation of the Transition Path Sampling setting

We now present a finite–dimensional approximation of the measure μ in the Transition Path Sampling setting from Section 3.2. The infinite–dimensional function spaces are approximated by spaces of piece–wise linear functions. We first introduce the notation

$$\begin{aligned}d_N &:= 2^{N-1}\\ s_i &:= \frac{i}{d_N} \qquad \text{for } i \in \{0,\dots,d_N\}.\end{aligned}$$

For this, recall the Schauder basis defined in (3.12) and set for $N \in \mathbb{N}$

$$\begin{aligned}E_N &:= W_N := S_N\\ &:= \operatorname{span}\left\{e_{i,k,j} \mid i \in \{1,\dots,N\},\, k \in \left\{1,\dots,2^{i-1}\right\}, j \in \{1,\dots,d\}\right\}.\end{aligned}$$

The scalar products on these spaces are defined by

$$\begin{aligned}\langle x,y\rangle_{E_N} &:= \frac{1}{d_N}\sum_{i=1}^{d_N} x_{s_i}\cdot y_{s_i},\\ \langle x,y\rangle_{W_N} &:= \langle x,y\rangle_{W_\alpha},\\ \langle x,y\rangle_{S_N} &:= \langle x,y\rangle_{H_0^1([0,1],\mathbb{R}^d)},\end{aligned}$$

for $x, y \in W_N$. Note that $\langle\cdot,\cdot\rangle_{E_N}$ does not exactly coincide with $\langle\cdot,\cdot\rangle_{L^2}$, because we have for a piece–wise linear function $x \in E_N$

$$\|x\|_{L^2}^2 = \int_0^1 |x_s|^2\mathrm{d}s = \frac{1}{d_N}\sum_{i=1}^{d_N}\frac{2}{3}|x_{s_i}|^2 + \frac{1}{3}x_{s_i}\cdot x_{s_{i-1}}. \tag{3.16}$$

Nevertheless, the norms on E and E_N are equivalent:

Lemma 3.11. *For $x \in E_N$, the inequalities*

$$\|x\|_E^2 \leq \|x\|_{E_N}^2 \leq \frac{3}{2}\|x\|_E^2.$$

hold.

Proof. Equation (3.16) directly implies

$$\frac{3}{2}\|x\|_E^2 \geq \|x\|_{E_N}^2 \qquad \text{for } x \in E_N.$$

As for all $a, b \in \mathbb{R}^d$

$$a \cdot b \leq |a||b| \leq \frac{1}{2}\left(|a|^2 + |b|^2\right)$$

the inequality

$$\|x\|_E^2 \leq \|x\|_{E_N}^2$$

also follows from (3.16). □

The S_N–scalar product has also a point–wise representation given by

$$\begin{aligned}\langle x, y\rangle_{S_N} &= \frac{d_N^2}{d_N}\sum_{i=1}^{d_N}(x_{s_{i+1}} - x_{s_i}) \cdot (y_{s_{i+1}} - y_{s_i}) \\ &= d_N \sum_{i=1}^{d_N-1} (x_{s_{i+1}} - 2x_{s_i} + x_{s_{i-1}}) \cdot y_{s_i}.\end{aligned}$$

Therefore,

$$\langle x, y\rangle_{S_N} = \langle -\Delta_{0,N} x, y\rangle_{E_N}, \tag{3.17}$$

where $\Delta_{0,N}$ is the discrete Laplacian on the partition $\left(\frac{i}{d_N}\right)_{i\in\{0,\dots,d_N\}}$ with zero boundary conditions. Its inverse is given by

$$\left(\Delta_{0,N}^{-1} x\right)_{s_i} = \frac{1}{d_N^2}\sum_{j=1}^{i-1}\sum_{k=1}^{j} x_{s_k} - \frac{1}{d_N^2}\sum_{j=1}^{N-1}\sum_{k=1}^{j} x_{s_k},$$

for $x \in S_N$. While the sets E_N, W_N, S_N coincide, the metric on these sets differ to approximate their infinite–dimensional pendants.

We approximate the distribution of the Brownian Bridge, ν, by its piece–wise linear approximation ν_N, that is the distribution of the random variable

$$Z_N := \frac{1}{\sqrt{2}}\sum_{i=1}^{N}\sum_{k=1}^{2^{i-1}}\sum_{j=1}^{d} N_{i,k,j} e_{i,k,j} \tag{3.18}$$

where $N_{i,k,j}$ are i.i.d. Gaussian random variables with mean 0 and variance 1. The covariance operator $\mathcal{C}_{E_N}$ of ν_N on E_N is given by the inverse discrete Laplacian:

Lemma 3.12. *The covariance operator of ν_N on E_N is given by*

$$\mathcal{C}_{E_N} := (-\Delta_{0,N})^{-1}.$$

Proof. Let Z_N be a ν_N–distributed variable given by (3.18), and define $Z \in E$ by

$$Z := Z_N + \frac{1}{\sqrt{2}} \sum_{i=N+1}^{\infty} \sum_{k=1}^{2^{i-1}} \sum_{j=1}^{d} N_{i,k,j} e_{i,k,j}.$$

It follows that Z is a ν–distributed random variable on E. Let $h, \tilde{h} \in E_N$, then we have

$$\begin{aligned}
\mathbb{E}\left[\langle h, Z_N\rangle_{E_N} \left\langle \tilde{h}, Z_N \right\rangle\right] &= \mathbb{E}\left[\langle h, Z\rangle_E \left\langle \tilde{h}, Z \right\rangle_E\right] \\
&= \left\langle h, (-\Delta_0)^{-1}\tilde{h} \right\rangle_E \\
&= \left\langle h, (-\Delta_{0,N})^{-1}\tilde{h} \right\rangle_{E_N}
\end{aligned}$$

which implies $\mathcal{C}_{E_N} = (-\Delta_{0,N})^{-1}$. □

To approximate the density φ of μ with respect to ν as given in (3.3), we approximate the integral $\int_0^1 \Phi(x_s)\mathrm{d}s$ by its Riemann–sum: Set

$$\begin{aligned}
V_N(x) &:= \frac{1}{d_N} \sum_{i=0}^{d_N} \Phi(x_{s_i}), \\
\varphi_N(x) &:= \exp\left(-V_N(x)\right) \qquad \text{for } x \in W_N.
\end{aligned}$$

The following lemma presents some properties of this discretization.

Lemma 3.13. *For $x, y \in W_N$,*

$$\begin{aligned}
\langle \nabla_{S_N} V_N(x), y\rangle_{S_N} &= \frac{1}{d_N} \sum_{k=0}^{d_N} \nabla\Phi(x_{s_i}) \cdot y_{s_i}, \qquad \textit{and} \\
\nabla_{S_N} V_N(x) &= -\Delta_{0,N}^{-1} \nabla\Phi(x).
\end{aligned}$$

Proof. By definition of the gradient, it holds for $y \in W_N$,

$$\begin{aligned}
\langle \nabla_{S_N} V_N(x), y\rangle_{S_N} &= \frac{\partial}{\partial y} V_N(x) \\
&= \frac{1}{d_N} \sum_{k=0}^{d_N} \nabla\Phi(x_{s_i}) \cdot y_{s_i}. \qquad (3.19)
\end{aligned}$$

By (3.17), we also have

$$\langle \nabla_{S_N} V_N(x), y\rangle_{S_N} = \langle -\Delta_{0,N}(\nabla_{S_N} V_N(x)), y\rangle_{E_N}$$

for all $y \in W_N$ which combined with (3.19) implies

$$\nabla_{S_N} V_N(x) = -\Delta_{0,N}^{-1} \nabla \Phi(x).$$

□

Knowing the gradient of V_N, we can now construct the MALA–process as introduced in Chapter 3.1. The acceptance probability is given by (3.8), and G_h is identified in the next Lemma:

Lemma 3.14. *Let $x, y \in W_N$, then*

$$\begin{aligned} G_h(x,y) := \frac{1}{d_N}\sum_{i=0}^{d_N} \Phi(y_{s_i}) - \Phi(x_{s_i}) - \frac{1}{2d_N}\sum_{i=0}^{d_N} (\nabla\Phi(x_{s_i}) + \nabla\Phi(y_{s_i}))\cdot(y_{s_i} - x_{s_i}) \\ + \frac{h}{8-2h}\frac{1}{d_N}\sum_{i=0}^{d_N} (\nabla\Phi(y_{s_i}) - \nabla\Phi(x_{s_i}))\cdot(x_{s_i} + y_{s_i}) \\ + \frac{h}{8-2h}\frac{1}{d_N}\sum_{i=0}^{d_N} \nabla\Phi(y_{s_i})\cdot\left(-\Delta_{0,N}^{-1}\nabla\Phi(y)\right)_{s_i} - \nabla\Phi(x_{s_i})\cdot\left(-\Delta_{0,N}^{-1}\nabla\Phi(x)\right)_{s_i}. \end{aligned}$$

Proof. By Lemma 3.2, G_h satisfies

$$\begin{aligned} G_h(x,y) := V_N(y) - V_N(x) - \frac{1}{2}\langle \nabla_{S_N} V_N(x) + \nabla_{S_N} V_N(y), y - x\rangle_{S_N} \\ + \frac{h}{8-2h}\langle \nabla_{S_N} V_N(y) - \nabla_{S_N} V_N(x), x + y\rangle_{S_N} \\ + \frac{h}{8-2h}\left(\|\nabla_{S_N} V_N(y)\|_{S_N}^2 - \|\nabla_{S_N} V_N(x)\|_{S_N}^2\right). \end{aligned}$$

Lemma 3.13 gives for $y \in W_N$

$$\langle \nabla_{S_N} V_N(x), y\rangle_{S_N} = \frac{1}{d_N}\sum_{k=0}^{d_N} \nabla\Phi(x_{s_i})\cdot y_{s_i} \qquad \text{and}$$

$$\nabla_{S_N} V_N(x) = -\Delta_{0,N}^{-1}\nabla\Phi(x)$$

such that

$$\langle \nabla_{S_N} V_N(x) + \nabla_{S_N} V_N(y), y - x\rangle_{S_N} = \frac{1}{d_N}\sum_{i=0}^{d_N} (\nabla\Phi(x_{s_i}) + \nabla\Phi(y_{s_i}))\cdot(y_{s_i} - x_{s_i}),$$

$$\langle \nabla_{S_N} V_N(y) - \nabla_{S_N} V_N(x), x + y\rangle_{S_N} = \frac{1}{d_N}\sum_{i=0}^{d_N} (\nabla\Phi(y_{s_i}) - \nabla\Phi(x_{s_i}))\cdot(x_{s_i} + y_{s_i})$$

and

$$\begin{aligned}&\|\nabla_{S_N} V_N(y)\|_{S_N}^2 - \|\nabla_{S_N} V_N(x)\|_{S_N}^2 \\ &= \frac{1}{d_N}\sum_{i=0}^{d_N} \nabla\Phi(y_{s_i}) \cdot \left(-\Delta_{0,N}^{-1}\nabla\Phi(y)\right)_{s_i} - \nabla\Phi(x_{s_i}) \cdot \left(-\Delta_{0,N}^{-1}\nabla\Phi(x)\right)_{s_i}.\end{aligned}$$

□

3.3 Speed of convergence of the MALA–process

We now analyze the speed of convergence of the MALA–process. For this purpose, we state conditions such that the MALA–process X_n is contracting. The distance that measures the distance of the distributions of the coupled processes is the Wasserstein distance, see e.g. Villani [42] for an introduction in this topic.

This section strongly relies on the work of Eberle [16]. Eberle proves the convergence results using coupling techniques for $\mathbb{R}^d$–valued processes with invariant measures which is absolutely continuous to a Gaussian measure. We transfer his techniques to the space of infinite–dimensional Hilbert spaces. Since the result in [16] aim at situations where the finite–dimensional processes converge to an infinite–dimensional limit, this transfer is straightforward in most cases. For the sake of completeness, we nonetheless present the full proofs here.

We start with stating basic properties of couplings. Denote with $\mathcal{P}(W)$ the space of all probability measures on W, and by $\mathcal{B}(W)$ the set of all Borel sets of W.

Definition 3.1. *A probability measure $\pi : \mathcal{B}(W) \times \mathcal{B}(W) \to [0,1]$ is called coupling of the probability measures $\eta, \tilde{\eta} : \mathcal{B}(W) \to [0,1]$ if for all $B \in \mathcal{B}(W)$*

$$\begin{aligned}\int_{B\times W} \pi(\mathrm{d}(x,\tilde{x})) &= \eta(B), \\ \int_{W\times B} \pi(\mathrm{d}(x,\tilde{x})) &= \tilde{\eta}(B).\end{aligned}$$

Definition 3.2. *Given a metric $d : W \times W \to [0,\infty]$, the Wasserstein distance*

$$\mathcal{W}_d : \mathcal{P}(W) \times \mathcal{P}(W) \to [0,\infty)$$

is defined by

$$\mathcal{W}_d(\eta, \tilde{\eta}) := \inf_{\pi} \int_{W \times W} d(x, \tilde{x}) \mathrm{d}\pi(x, \tilde{x})$$

where the infimum is taken over all couplings π of η and $\tilde{\eta}$.

Definition 3.3. *Given stochastic kernels*

$$q : W \times \mathcal{B}(W) \to [0, 1] \text{ and}$$
$$c : (W \times W) \times \mathcal{B}(W \times W) \to [0, 1],$$

c is called pair coupling of q if for any $x, \tilde{x} \in W$, the distribution of the first and second component of $c((x, \tilde{x}), \mathrm{d}y\mathrm{d}\tilde{y})$ is $q(x, \mathrm{d}y)$ and $q(\tilde{x}, \mathrm{d}\tilde{y})$ respectively, that is for all $B \in \mathcal{B}(W)$,

$$\int_{W \times B} c((x, \tilde{x}), \mathrm{d}y\mathrm{d}\tilde{y}) = q(\tilde{x}, B),$$
$$\int_{B \times W} c((x, \tilde{x}), \mathrm{d}y\mathrm{d}\tilde{y}) = q(x, B).$$

We will use the following theorem to find an upper bound for the distance to equilibrium of the MALA–process:

Theorem 3.1. *Let $d : W \times W \to [0, R]$ be a metric with diameter $0 < R < \infty$, $c : (W \times W) \times \mathcal{B}(W \times W) \to [0, 1]$ a pair coupling of the stochastic kernel $q : W \times \mathcal{B}(W) \to [0, 1]$. Let $\gamma \in (0, 1)$, and assume $U \subset W$ is an open subset of W, such that for all $x, \tilde{x} \in U$, $d(x, \tilde{x}) < R$. If the contraction property*

$$\int_{W \times W} d(y, \tilde{y}) c((x, \tilde{x}), \mathrm{d}y\mathrm{d}\tilde{y}) \leq \gamma d(x, \tilde{x}) \tag{3.20}$$

holds, the Wasserstein distance of μq^n and νq can be bounded for any $n \in \mathbb{N}$ and every probability measure ν and μ:

$$\mathcal{W}_d(\mu q^n, \nu q^n) \leq \gamma^n \mathcal{W}_d(\mu, \nu) + \frac{R}{1 - \gamma}(C_n(U, \mu) + C_n(U, \nu)), \tag{3.21}$$

with

$$C_n(U, \eta) := \sup_{k \in \{0, \ldots, n\}} (\eta q)(S \setminus U). \tag{3.22}$$

for $\eta \in \{\nu, \mu\}$.

Proof. We start with $n = 1$. If ν and μ are probability measures on $\mathcal{B}(W)$, and $\eta : \mathcal{B}(W \times W)$ is a coupling of ν and μ, then the probability measure

$$(\eta c)(B) := \int_{W\times W} \eta(\mathrm{d}x\mathrm{d}\tilde{x})c((x,\tilde{x}),B), \qquad B \in \mathcal{B}(W \times W),$$

is a coupling of νq and μq. Indeed, given $A \in \mathcal{B}(W)$, we have

$$\begin{aligned}(\eta c)(A \times W) &= \int_{A\times W} \eta(\mathrm{d}x\mathrm{d}\tilde{x})q(x,A) &&= \int_W \nu(\mathrm{d}x)q(x,A) &&= (\nu q)(A),\\ (\eta c)(W \times A) &= \int_{W\times A} \eta(\mathrm{d}x\mathrm{d}\tilde{x})q(\tilde{x},A) &&= \int_W \mu(\mathrm{d}\tilde{x})q(\tilde{x},A) &&= (\mu q)(A)\end{aligned}$$

as c is a pair coupling of q.
Now assume (3.20) holds for all $x, \tilde{x} \in U$. Then (ηc) can be used to get an upper bound for the d_R–Wasserstein distance of νq and μq:

$$\begin{aligned}\mathcal{W}_d(\nu q, \mu q) &\le \int_{W\times W} d(y,\tilde{y})(\eta c)(\mathrm{d}y,\mathrm{d}\tilde{y})\\ &= \int_{W\times W}\int_{W\times W} d(y,\tilde{y})c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y})\eta(\mathrm{d}y\mathrm{d}\tilde{x})\\ &\le \int_{U\times U}\int_{W\times W} d(y,\tilde{y})c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y})\eta(\mathrm{d}y\mathrm{d}\tilde{x}) + R\,\eta((W\times W)\setminus(U\times U))\\ &\le \gamma\int_{U\times U} d(x,\tilde{x})\eta(\mathrm{d}x,\mathrm{d}\tilde{x}) + R\,(\nu(S\setminus U)+\mu(S\setminus U))\\ &\le \gamma\int_{W\times W} d(x,\tilde{x})\eta(\mathrm{d}x,\mathrm{d}\tilde{x}) + R\,(\nu(S\setminus U)+\mu(S\setminus U)).\end{aligned}$$

Taking the infimum over all couplings η of ν and μ, we get

$$\mathcal{W}_d(\mu q, \nu q) \le \gamma\mathcal{W}_d(\nu,\mu) + R\,(\nu(S\setminus U)+\mu(S\setminus U)).$$

With induction over n,

$$\mathcal{W}_d(\mu q^n, \nu q^n) \le \gamma^n\mathcal{W}_d(\nu,\mu) + \frac{R}{1-\gamma}(\nu(S\setminus U)+\mu(S\setminus U))$$

follows. □

A pair coupling of the kernel of a stochastic kernel of a Markov Process can be constructed by considering the distribution of a coupling of the Markov Process starting in different starting points.

Lemma 3.15. *Let $(X_n, \tilde{X}_n)_{n\in\mathbb{N}}$ be a coupling of time–homogeneous Markov Processes with kernel p, with $(X_0, \tilde{X}_0) = (x,\tilde{x})$ $\mathrm{P}_{x,\tilde{x}}$–almost surely. Let c be the kernel of $(X_n, \tilde{X}_n)_{n\in\mathbb{N}}$.*

Then c is a pair coupling of p.
In this case:

$$\int_{W\times W} d(y,\tilde{y})c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y}) = \mathbb{E}_{x,\tilde{x}}\left[d(X_1,\tilde{X}_1)\right].$$

Proof. As kernel of the process $(X_n,\tilde{X}_n)_{n\in\mathbb{N}}$, c is a function $c:(W\times W)\times\mathcal{B}(W\times W)\to[0,1]$. We need to show that for $B\in\mathcal{B}(W)$

$$\int_{W\times B} c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y}) = q(\tilde{x},B),$$
$$\int_{B\times W} c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y}) = q(x,B)$$

hold. This follows directly, because q is the kernel of $(X_n)_{n\in\mathbb{N}}$ as well as of $(\tilde{X}_n)_{n\in\mathbb{N}}$:

$$\begin{aligned}\int_{W\times B} c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y}) &= \mathrm{P}_{x,\tilde{x}}\left[X_1\in W,\tilde{X}_1\in B\right]\\ &= \mathrm{P}_{x,\tilde{x}}\left[\tilde{X}_1\in B\right]\\ &= q(\tilde{x},B)\end{aligned}$$

and

$$\begin{aligned}\int_{B\times W} c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y}) &= \mathrm{P}_{x,\tilde{x}}\left[X_1\in B\right]\\ &= q(x,B).\end{aligned}$$

□

3.3.1 Contraction property of the MALA–process

In this section, we present a pair coupling c of the kernel p of the MALA–process constructed in Section 3.1, and show that it is contracting. The pair coupling is constructed by a coupling $(X_n,\tilde{X}_n)_{n\in\mathbb{N}}$ of the MALA–process starting in x and $\tilde{x}$ respectively. Then the kernel of $(X_n,\tilde{X}_n)_{n\in\mathbb{N}}$ is a pair coupling of p by Lemma 3.15.

Coupling the proposals

We construct the coupling $(X_n,\tilde{X}_n)_{n\in\mathbb{N}}$ as follows: For a given i.i.d. sequence of ν–distributed random variable $(N_n)_{n\in\mathbb{N}}$ on W and for given $x\in W$ we define

$$Y_{h,n}(x) := \left(1-\frac{h}{2}\right)x - \frac{h}{2}\nabla_S V(x) + \sqrt{h}N_{n+1}. \tag{3.23}$$

Starting in $X_n, \tilde{X}_n \in W$, we propose to move to $Y_{h,n}(X_n)$ and $Y_{h,n}(\tilde{X}_n)$ respectively and accept each move with acceptance probability $a_h(X_n, Y_{h,n}(X_n))$ and $a_h(\tilde{X}_n, Y_{h,n}(\tilde{X}_n))$. For this, we take a sequence of uniformly distributed random variables $(U_n)_{n\in\mathbb{N}}$ on $[0,1]$ and set

$$X_{n+1} := \begin{cases} Y_{h,n}(X_n) & \text{if } U_{n+1} < a_h(X_n, Y_{h,n}(X_n)), \\ X_n & \text{otherwise.} \end{cases} \tag{3.24}$$

$$\tilde{X}_{n+1} := \begin{cases} Y_{h,n}(\tilde{X}_n) & \text{if } U_{n+1} < a_h(\tilde{X}_n, Y_{h,n}(\tilde{X}_n)), \\ \tilde{X}_n & \text{otherwise.} \end{cases} . \tag{3.25}$$

We use the same random variables $(N_n)_{n\in\mathbb{N}}$ and $(U_n)_{n\in\mathbb{N}}$ for X as well as for $\tilde{X}$. This leads to $Y_{h,n}(X_n) - Y_{h,n}(\tilde{X}_n)$ being independent of the noise term N_{n+1}. Furthermore, given the proposals $Y_{h,n}(X_n)$ and $Y_{h,n}(\tilde{X}_n)$, we minimize the probability that the proposal of one chain is accepted and the proposal of the other chain is rejected.
The goal of this section is to control the Wasserstein distance $\mathbb{E}_{x,\tilde{x}}[d(X_1, \tilde{X}_1)]$ of the coupling. The first step is the decomposition of the expectation value into four cases:

Proposition 3.16. *Let $d : W \times W \to [0, R]$ be a metric bounded by R, and $X_1, \tilde{X}_1$ the processes defined in (3.24),(3.25) respectively. Then*

$$\begin{aligned} \mathbb{E}_{x,\tilde{x}}[d(X_1, \tilde{X}_1)] \leq\ & \mathbb{E}_{x,\tilde{x}}[d(Y_{h,0}(x), Y_{h,0}(\tilde{x}))] \\ & + R\,\mathbb{E}_{x,\tilde{x}}[|G_h(x, Y_{h,0}(x)) - G_h(\tilde{x}, Y_{h,0}(\tilde{x}))|] \\ & + d(x,\tilde{x})\mathbb{E}_{x,\tilde{x}}\left[\min\{G_h(x, Y_{h,0}(x)), G_h(\tilde{x}, Y_{h,0}(\tilde{x}))\}\right]. \end{aligned} \tag{3.26}$$

Proof. We decompose the expectation value into four summands, distinguishing whether the processes $(X_n)_{n\in\mathbb{N}}$ and $(\tilde{X}_n)_{n\in\mathbb{N}}$ except their proposals at the first time step $n = 1$.

$$\begin{aligned} \mathbb{E}_{x,\tilde{x}}[d(X_1, \tilde{X}_1)] =\ & \mathbb{E}_{x,\tilde{x}}[d(Y_{h,0}(x), Y_{h,0}(\tilde{x})), U_1 < \min\{a_h(x, Y_{h,0}(x)), a_h(\tilde{x}, Y_{h,0}(\tilde{x}))\}] \\ & + \mathbb{E}_{x,\tilde{x}}[d(x, Y_{h,0}(\tilde{x})), a_h(x, Y_{h,0}(x)) < U_1 < a_h(\tilde{x}, Y_{h,0}(\tilde{x}))] \\ & + \mathbb{E}_{x,\tilde{x}}[d(Y_{h,0}(x), \tilde{x}), a_h(\tilde{x}, Y_{h,0}(\tilde{x})) < U_1 < a_h(x, Y_{h,0}(x))] \\ & + d(x,\tilde{x})\mathrm{P}_{x,\tilde{x}}[U_1 > \max\{a_h(x, Y_{h,0}(x)), a_h(\tilde{x}, Y_{h,0}(\tilde{x}))] \\ \leq\ & \mathbb{E}_{x,\tilde{x}}[d(Y_{h,0}(x), Y_{h,0}(\tilde{x}))] \\ & + R\mathbb{E}_{x,\tilde{x}}[|a_h(x, Y_{h,0}(x)) - a_h(\tilde{x}, Y_{h,0}(\tilde{x}))|] \\ & + d(x,\tilde{x})\mathbb{E}_{x,\tilde{x}}[\min\{1 - a_h(x, Y_{h,0}(x)), 1 - a_h(\tilde{x}, Y_{h,0}(\tilde{x}))\}]. \end{aligned}$$

With Remark 3.3 this proves the result. □

In the sequel, we will bound each of these terms to finally show the contraction property of the coupling. Our bounds are proven under the following two assumptions: We need to assume that the S–gradient of V satisfies a Lipschitz bound.

Assumption 3.3. *There exists $0 \leq L < 1$ such that*

$$\|\nabla_S V(x) - \nabla_S V(y)\|_W \leq L\|x - y\|_W \tag{3.27}$$

holds. Set $\delta := (1 - L)$.

Furthermore, we define constants $L_n(x, y)$ which control the derivatives of V in the interval $[x, y] := \{sx + (1 - s)y | s \in [0, 1]\}$.

$$L_n(x, y) := \sup_{z \in [x,y]} \|\mathrm{D}^n V(z)\|_{W^{\otimes n} \to \mathbb{R}}$$

where $\|\cdot\|_{W^{\otimes n} \to \mathbb{R}}$ is the norm on n–form given by

$$\|l\|_{W^{\otimes n} \to \mathbb{R}} := \sup \{l(\xi_1, \ldots, \xi_n) \mid \|\xi_1\|_W = \ldots = \|\xi_n\|_W = 1\}$$

We assume that the first four derivatives of V grow at most polynomial:

Assumption 3.4. *The potential V is four times differentiable in directions of W and there exists constants C_n, $p_n \in [0, \infty)$ such that the derivatives as operators from $W^{\otimes n}$ to $\mathbb{R}$ are bounded by a polynomial:*

$$|\mathrm{D}^n V(x)(\xi_1, \ldots, \xi_n)| \leq C_n \max\{1, \|x\|_W\}^{p_n}$$

for all $x \in W$, $\|\xi_1\|_W = \ldots = \|\xi_n\|_W = 1$, and $n \in \{2, 3, 4\}$.

Note that if Assumption 3.4 is satisfied, $L_n(x, y)$ bounded by

$$L_n(x, y) \leq C_n \max\{1, \|x\|_W, \|y\|_W\}^{p_n}, \qquad n \in \{2, 3, 4\}$$

for $x, y \in W$.
Furthermore, we can control $L_n(x, Y_{h,n}(x))$ in the following way:

Lemma 3.17. *Let Assumption 3.4 be satisfied. Then for $x \in W$, $h \in (0, 2)$, $n \in \{2, 3, 4\}$ and $p_n \in [1, \infty)$, the inequality*

$$L_n(x, Y_{h,n}(x)) \leq C_n 3^{p_n - 1} \left(\max\{1, \|x\|_W\}^{p_n} + \left(\frac{h}{2}\right)^{p_n} \|\nabla_S U(x)\|_W^{p_n} + h^{\frac{p_n}{2}} \|N_{n+1}\|_W^{p_n} \right)$$

holds.

Proof. Using the triangle inequality, we get

$$\|Y_{h,n}(x) - x\|_W \leq \frac{h}{2}\|\nabla_S U(x)\|_W + \sqrt{\tilde{h}}\|N_{n+1}\|_W.$$

Therefore we can bound

$$\begin{aligned} L_n(x, Y_{h,n}(x)) &\leq C_n \max\{1, \|x\|_W, \|Y_{h,n}(x)\|_W\}^{p_n} \\ &\leq C_n \left(\max\{1, \|x\|_W\} + \frac{h}{2}\|\nabla_S U(x)\|_W + \sqrt{\tilde{h}}\|N_{n+1}\|_W\right)^{p_n} \\ &\leq C_n 3^{p_n - 1} \left(\max\{1, \|x\|_W\}^{p_n} + \left(\frac{h}{2}\right)^{p_n} \|\nabla_S U(x)\|_W^{p_n} + h^{\frac{p_n}{2}} \|N_{n+1}\|_W^{p_n}\right) \end{aligned}$$

by Hölder's inequality. □

Contraction property of the proposal

In this section, we present how to control the terms arising in equation (3.26). We start with the distance of the proposals $Y_{h,n}(x)$ and $Y_{h,n}(\tilde{x})$. Under Assumption 3.3, we show that the coupling is contracting in the metric

$$d_\infty(x, \tilde{x}) := \|x - \tilde{x}\|_W \qquad \text{for } x, \tilde{x} \in W.$$

Lemma 3.18. *Let Assumption 3.3 be satisfied. Then for all $h \in (0, 2)$ and x, $\tilde{x} \in W$*

$$\mathbb{E}_{x,\tilde{x}}\left[d_\infty(Y_{h,n}(x), Y_{h,n}(\tilde{x}))\right] \leq \left(1 - \frac{1}{2}\delta h\right) d(x, \tilde{x}).$$

Proof. We firstly note that as $Y_{h,n}(x)$ and $Y_{h,n}(\tilde{x})$ are constructed with the same noise term N_{n+1}, the noise cancels in the difference so that $Y_{h,n}(x) - Y_{h,n}(\tilde{x})$ is deterministic:

$$\mathbb{E}_{x,\tilde{x}}\left[d_\infty(Y_{h,n}(x), Y_{n,h}(\tilde{x}))\right] = \|Y_{h,n}(x) - Y_{h,n}(\tilde{x})\|_W.$$

Inserting the definitions of $Y_{h,n}$ and δ as well as the assumptions of the lemma, we get for

$h \in (0,2)$ and x, $\tilde{x} \in W$:

$$\begin{aligned}
\|Y_{h,n}(x) - Y_{h,n}(\tilde{x})\|_W^2 &= \left\| \left(1 - \frac{h}{2}\right)(x - \tilde{x}) - \frac{h}{2}(\nabla_S V(x) - \nabla_S V(\tilde{x})) \right\|_W^2 \\
&= \left(1 - \frac{h}{2}\right)^2 \|x - \tilde{x}\|_W^2 - h\left(1 - \frac{h}{2}\right) \langle x - \tilde{x}, \nabla_S V(x) - \nabla_S V(\tilde{x}) \rangle_W \\
&\quad + \frac{h^2}{4} \|\nabla_S V(x) - \nabla_S V(\tilde{x})\|_W^2 \\
&\leq \|x - \tilde{x}\|_W^2 \left[\left(1 - h + \frac{h^2}{4}\right) + Lh\left(1 - \frac{h}{2}\right) + \frac{h^2}{4} L^2 \right] \\
&\leq \|x - \tilde{x}\|_W^2 \left(1 - h(1 - L) + \frac{h^2}{4}(L^2 - 2L + 1)\right) \\
&\leq \|x - \tilde{x}\|_W^2 \left(1 - \delta h\right).
\end{aligned}$$

As $\sqrt{1-z} \leq 1 - \frac{1}{2}z$, for $z \in (-\infty, 1]$, the result follows. □

3.3.2 Bound on the rejection probability

In this section we bound the average rejection probability of the MALA–process. The main result is the following Proposition which guarantees that the average rejection probability decreases with order $\frac{3}{2}$ of the step–size h. It is the Hilbert space version of [16, Proposition 1.7] which bounds the rejection probability for the MALA–process in a finite–dimensional setting. As the bound in [16] is already designed to scale well when the dimension converges to infinity the proof can be carried over to the Hilbert space setting almost unchanged.

Proposition 3.19. *If Assumption 3.4 is satisfied, then there exists a polynomial* $\mathcal{P} : \mathbb{R}^2 \to [0, \infty)$ *of degree* $\max\{p_3 + 3, 2p_2 + 2\}$ *such that*

$$\mathbb{E}\left[G_h(x, Y_{h,n}(x))^+\right] \leq \mathcal{P}(\|x\|_W, \|\nabla_S U(x)\|_W) \cdot h^{\frac{3}{2}}, \qquad \text{for all } x \in W, h \in (0,2).$$

To prove the Proposition, we use a couple of Lemmas. We apply the representation of G_h in Lemma 3.2 and reexpress the terms arising in this representation in terms of derivatives of V. Then the first line of the representation in Lemma 3.2 has the following structure:

Lemma 3.20. *Let Assumption 3.4 be satisfied. Then for all $x, y \in W$*

$$\begin{aligned} &V(y) - V(x) - \frac{1}{2}\langle y - x, \nabla_S V(y) + \nabla_S V(x)\rangle_S \\ &= -\frac{1}{2}\int_0^1 t(1-t)\mathrm{D}^3 V((1-t)x + ty)(y-x)^3 \mathrm{d}t. \end{aligned}$$

Proof. We expand the function $f(t) := V(x + t(y - x))$ with Taylor's formula. For this choice of f, we have

$$\begin{aligned} f'(t) &= \mathrm{D}V(x + t(y-x))(y-x) = \langle \nabla_S V(x), y - x\rangle_S, \\ f''(t) &= \mathrm{D}^2 V(x + t(y-x))(y-x)^2, \\ f'''(t) &= \mathrm{D}^3 V(x + t(y-x))(y-x)^3. \end{aligned}$$

With Taylor's fomula, $V(y) - V(x)$ can be expressed as

$$\begin{aligned} V(y) - V(x) &= \int_0^1 f'(t)\,\mathrm{d}t \\ &= \langle y - x, \nabla_S V(x)\rangle_S + \int_0^1 \int_0^t f''(s)\,\mathrm{d}s\mathrm{d}t \\ &= \langle y - x, \nabla_S V(x)\rangle_S + \int_0^1 (1-s) f''(s)\,\mathrm{d}s. \end{aligned}$$

Similarly, we get

$$\begin{aligned} V(x) - V(y) &= \int_1^0 f'(t)\,\mathrm{d}t \\ &= -\langle y - x, \nabla_S V(y)\rangle_S + \int_0^1 \int_t^1 f''(s)\,\mathrm{d}s\mathrm{d}t \\ &= -\langle y - x, \nabla_S V(y)\rangle_S + \int_0^1 s f''(s)\,\mathrm{d}s. \end{aligned}$$

Combining both equations, we obtain

$$V(x) - V(y) - \frac{1}{2}\langle y - x, \nabla_S V(x) + \nabla_S V(y)\rangle_S = \frac{1}{2}\int_0^1 (1-2s) f''(s)\,\mathrm{d}s.$$

As

$$\begin{aligned} \int_0^1 (1-2s) f''(s)\,\mathrm{d}s &= \int_0^1 (1-2s)\int_0^s f'''(t)\mathrm{d}t\mathrm{d}s \\ &= \int_0^1 \int_t^1 (1-2s)\mathrm{d}s f'''(t)\mathrm{d}t \\ &= -\int_0^1 t(1-t) f'''(t)\mathrm{d}t \end{aligned}$$

we conclude

$$V(x) - V(y) - \frac{1}{2}\langle y - x, \nabla_S V(x) + \nabla_S V(y)\rangle_S = -\frac{1}{2}\int_0^1 t(1-t)\mathrm{D}^3 V(x + t(y-x))(y-x)^3 \mathrm{d}t.$$

□

Lemma 3.21. *If V satisfies Assumption 3.4, we have for $x, y \in W$*

1) $$\left|V(y) - V(x) - \frac{1}{2}\langle y - x, \nabla_S V(y) + \nabla_S V(x)\rangle_S\right| \leq \frac{1}{12} L_3(x,y)\|y-x\|_W^3,$$

2) $$|\langle \nabla_S U(y) + \nabla_S U(x), \nabla_S V(y) - \nabla_S V(x)\rangle_S| \leq L_2(x,y)\|\nabla_S U(y) + \nabla_S U(x)\|_W \|y-x\|_W,$$

3) $$\|\nabla_S U(y) + \nabla_S U(x)\|_W \leq 2\|\nabla_S U(x)\|_W + (1 + L_2(x,y))\|y-x\|_W.$$

Proof. By the definition of the upper bounds $L_2(x,y)$ and $L_3(x,y)$ of the derivatives of V, we get for the inequalities 1) - 3):

1): With Lemma 3.20,

$$\begin{aligned}\left|V(y) - V(x) - \frac{1}{2}\langle y - x, \nabla_S V(y) + \nabla_S V(x)\rangle_S\right| &\leq \frac{1}{2}\sup_{z\in[x,y]} \mathrm{D}^3 V(z) \int_0^1 t(1-t)\mathrm{d}t \\ &\leq \frac{1}{12} L_3(x,y)\|x-y\|_W^3.\end{aligned}$$

2): The second equation follows from

$$\begin{aligned}&\langle \nabla_S U(y) + \nabla_S U(x), \nabla_S V(y) - \nabla_S V(x)\rangle_S \\ &= \mathrm{D}V(y)(\nabla_S U(y) + \nabla_S U(x)) - \mathrm{D}V(x)(\nabla_S U(y) + \nabla_S U(x)) \\ &= \int_0^1 \mathrm{D}^2 V((1-t)x + ty)(\nabla_S U(y) + \nabla_S U(x))\mathrm{d}t \\ &\leq L_2(x,y)\|\nabla_S U(y) + \nabla_S U(x)\|_W \|x-y\|_W.\end{aligned}$$

3): As

$$\begin{aligned}\|\nabla_S V(y) - \nabla_S V(x)\|_W &\leq \left\|\int_0^1 \mathrm{D}^2 V((1-t)x + ty)(y-x)\mathrm{d}t\right\|_W \\ &\leq L_2(x,y)\|y-x\|_W,\end{aligned}$$

we get

$$\begin{aligned}\|\nabla_S U(y) + \nabla_S U(x)\|_W &\leq 2\|\nabla_S U(x)\|_W + \|\nabla_S U(y) - \nabla_S U(x)\|_W \\ &\leq 2\|\nabla_S U(x)\|_W + \|y-x\|_W + \|\nabla_S V(y) - \nabla_S V(x)\|_W \\ &\leq 2\|\nabla_S U(x)\|_W + (1 + L_2(x,y))\|y-x\|_W.\end{aligned}$$

□

We are now have the tools to bound the average rejection probability as state in Proposition 3.19.

Proof. By Lemma 3.2, $G_h(x,y)$ has the form

$$\begin{aligned} G_h(x,y) := V(y) - V(x) - \frac{1}{2}\langle \nabla_S V(x) + \nabla_S V(y), y - x\rangle_S \\ + \frac{h}{8-2h}\langle \nabla_S V(y) - \nabla_S V(x), x + y\rangle_S \\ + \frac{h}{8-2h}\left(\|\nabla_S V(y)\|_S^2 - \|\nabla_S V(x)\|_S^2\right). \end{aligned}$$

Thus we can bound $G_h(x, Y_{h,n}(x))$ for $h \in (0,2)$ by

$$G_h(x, Y_{h,n}(x)) \leq I + \frac{h}{4} II,$$

where

$$\begin{aligned} I &:= \left| V(y) - V(x) - \frac{1}{2}\langle \nabla_S V(x) + \nabla_S V(y), y - x\rangle_S \right| \qquad \text{and} \\ II &:= |\langle \nabla_S V(y) - \nabla_S V(x), \nabla_S U(x) + \nabla_S U(y)\rangle_S|. \end{aligned}$$

With Lemma 3.21, we can bound the first term by

$$I \leq \frac{1}{12}\mathbb{E}\left[L_3(x, Y_{h,n}(x))\|Y_{h,n}(x) - x\|_W^3\right]$$

and the second term by

$$\begin{aligned} II \leq \mathbb{E}\left[2L_2(x, Y_{h,n}(x))\|\nabla_S U(x)\|_W \|Y_{h,n}(x) - x\|_W\right] \\ + \mathbb{E}\left[L_2(x, Y_{h,n}(x))(1 + L_2(x, Y_{h,n}(x)))\|Y_{h,n}(x) - x\|_W^2\right]. \end{aligned}$$

Applying the bounds stated in Lemmas 3.21 and 3.17, it follows

$$\begin{aligned} I &\leq \frac{1}{12}\mathbb{E}\left[L_3(x, Y_{h,n}(x))\|Y_{h,n}(x) - x\|_W^3\right] \\ &\leq \frac{1}{12}C_3 3^{p_3-1}\left((1 + \|x\|_W)^{p_3} + \left(\frac{h}{2}\right)^{p_n}\|\nabla_S U(x)\|_W^{p_n}\right)\mathbb{E}\left[\|Y_{h,n}(x) - x\|_W^3\right] \\ &\quad + \frac{1}{2}C_3 3^{p_3-1} h^{\frac{p_3}{2}}\mathbb{E}\left[\|N_{n+1}\|^{p_3}\|Y_{h,n}(x) - x\|_W^3\right] \\ &\leq h^{\frac{3}{2}} \cdot \left[\frac{1}{12}C_3 3^{p_3-1}\left(1 + \|x\|_W^{p_3} + \|\nabla_S U(x)\|_W^{p_3}\right)\mathbb{E}\left[\left(\|\nabla_S U(x)\|_W + \|N_{n+1}\|_W\right)^3\right]\right. \\ &\quad \left. + \frac{1}{2}C_3 3^{p_3-1} 2^{\frac{p_3}{2}}\mathbb{E}\left[\|N_{n+1}\|_W^{p_3}(\|\nabla_S U(x)\|_W + \|N_{n+1}\|)^3\right]\right] \\ &\leq h^{\frac{3}{2}} \cdot \mathcal{P}_1(\|x\|_W, \|\nabla_S U(x)\|_W) \end{aligned}$$

for a polynomial $\mathcal{P}_1(x,y)$ of degree p_3+3, which only depends on C_3, p_3 and the first p_3+3 moments of ν on W. Similarly, II can be bounded by

$$II \leq h^{\frac{1}{2}} \cdot \mathcal{P}_2(\|x\|_W, \|\nabla_S U(x)\|_W)$$

for a polynom $\mathcal{P}_2(x,y)$ of degree p_2+2, which only depends on C_2, p_2 and first p_2+3 moments of ν on W. Applying these bounds, we get

$$\begin{aligned}\mathbb{E}[G_h(x, Y_{h,n}(x))^+] &\leq I + \frac{h}{4} II \\ &\leq h^{\frac{3}{2}}(\mathcal{P}_1(x, Y_{h,n}(x)) + \frac{1}{4}\mathcal{P}_2(x, Y_{h,n}(x))) \\ &\leq h^{\frac{3}{2}}\mathcal{P}(x, Y_{h,n}(x)),\end{aligned}$$

where $\mathcal{P}(x,y) := \mathcal{P}_1(x, Y_{h,n}(x)) + \frac{1}{4}\mathcal{P}_2(x, Y_{h,n}(x))$ is a polynomial of degree $\max\{p_3+3, 2p_2+2\}$, which only depends on C_2, C_3, p_2, p_3 and the first four moments of ν on W. □

3.3.3 Bound on the derivative of the acceptance probability

In this section, we bound the derivative of average rejection probability of the MALA–process, to control the term

$$\mathbb{E}_{x,\tilde{x}}\left[a_h(x, Y_{h,n}(x)) - a_h(\tilde{x}, Y_{h,n}(\tilde{x}))\right]$$

in equation (3.26). The main result of this section is the following Proposition:

Proposition 3.22. *If Assumption 3.4 is satisfied, then there exists a polynomial $\mathcal{Q}: \mathbb{R}^2 \to \mathbb{R}$ of degree $\max\{p_4+3, p_3+p_2+2, 3p_2+1\}$ such that for all $x \in W$ the gradient of the acceptance probability is bounded by*

$$\mathbb{E}\left[\|\nabla_W G_h(x, Y_{h,n}(x))\|_W\right] \leq h^{\frac{3}{2}}\mathcal{Q}(\|x\|_W, \|\nabla_S U(x)\|_W) \qquad \text{for } h \in (0,2).$$

The coefficients of $\mathcal{Q}$ only depend on $C_2, C_3, C_4, p_2, p_3, p_4$ and the first $\max\{p_4+3, p_3+p_2+2, 3p_2+1\}$ moments of ν on W.

Again, this result is the Hilbert space version of [16, Proposition 1.9]. The proof carries over almost unchanged from the finite–dimensional version.

We start with some notation. Define for $x, w \in W$

$$F_h^w(x,w) := G_h\left(x, x - \frac{h}{2}\nabla_S U(x) + w\right).$$

Note that F_h is related to the acceptance probability given in (3.7) of the chain at position x with proposal $Y_{h,n}(x)$ by

$$a_h(x, Y_{h,n}(x)) = \exp\left(-F_h^w\left(x, \tilde{h}N_{n+1}\right)\right).$$

We define for fixed $w \in W$

$$y := y^w(x) := x - \frac{h}{2}\nabla_S U(x) + w.$$

Let for $x \in W$, $\nabla_W^2 V(x)$ be the linear operator on W defined by

$$\left\langle \xi, \nabla_W^2 V(x)\eta \right\rangle_W = \mathrm{D}^2 V(x)(\xi, \eta) \qquad \text{for all } \xi, \eta \in W$$

and $\nabla_S^2 V(x)$ the linear operator on S defined by

$$\left\langle \xi, \nabla_S^2 V(x)\eta \right\rangle_S = \mathrm{D}^2 V(x)(\xi, \eta) \qquad \text{for all } \xi, \eta \in S.$$

Furthermore, note that

$$\nabla_W^2 V(x) = \mathcal{C}^{-1}\nabla_S^2 V(x)$$

because of $\langle \xi, \eta \rangle_S = \left\langle \mathcal{C}^{-1}\xi, \eta \right\rangle_W$.

For the proof of Proposition 3.22, we first establish the following bounds on the operator norm of $\nabla_S^2 V(x)$ and $\nabla_W^2 V(x)$:

Lemma 3.23. *Let Assumptions 3.3 and 3.4 be satisfied. Then, for $x \in W$,*

$$\left\|\nabla_W^2 V(x)\right\|_{W\to W} \le L_2(x, x),$$
$$\left\|\nabla_S^2 V(x)\right\|_{W\to W} \le c_\pi L_2(x, x).$$

Proof. For $x, \xi, \eta \in W$, the inequalities are derived throught the following considerations: The first one is given by

$$\begin{aligned}\left\langle \xi, \nabla_W^2 V(x)\eta \right\rangle_W &= \mathrm{D}^2(x)(\xi, \eta) \\ &\le L_2(x, x)\|\xi\|_W\|\eta\|_W.\end{aligned}$$

The second one is due to

$$\begin{aligned}\left\langle \xi, \nabla_S^2 V(x)\eta \right\rangle_W &= \left\langle \mathcal{C}\xi, \nabla_S^2 V(x)\eta \right\rangle_S \\ &= \mathrm{D}^2(x)(\mathcal{C}\xi, \eta) \\ &\le L_2(x, x)\|\mathcal{C}\xi\|_W\|\eta\|_W \\ &\le c_\pi L_2(x, x)\|\xi\|_W\|\eta\|_W.\end{aligned}$$

□

With the notation introduced above, we can express the derivatives $\mathrm{D}\nabla_S V(x)$, $\mathrm{D}\nabla_S U(x)$ and $\mathrm{D}y(x)$ as operators from W to W as described in the next lemma.

Lemma 3.24. *The derivatives of $\nabla_S V(x)$, $\nabla_S U(x)$ and $y(x) : W \to W$ are given by:*

$$\begin{aligned}
\mathrm{D}\nabla_S V(x) &= \nabla_S^2 V(x) \\
\mathrm{D}\nabla_S U(x) &= \mathrm{I} + \nabla_S^2 V(x) \\
\mathrm{D}y^w(x) &= \left(1 - \frac{h}{2}\right)\mathrm{I} - \frac{h}{2}\nabla_S^2 V(x) \\
&=: \mathrm{I} - \frac{h}{2}\nabla_S^2 U(x).
\end{aligned}$$

Proof. For $\xi, \eta \in S$, we have

$$\begin{aligned}
\langle \eta, \mathrm{D}\nabla_S V(x)(\xi)\rangle_S &= \mathrm{D}\langle \eta, \nabla_S V(x)\rangle_S(\xi) \\
&= \mathrm{D}\langle \eta, \nabla_W V(x)\rangle_S(\xi) \\
&= \mathrm{D}(\mathrm{D}V(x)(\eta))(\xi) \\
&= \mathrm{D}^2 V(x)(\eta, \xi) \\
&= \left\langle \xi, \nabla_S^2 \eta \right\rangle_S.
\end{aligned}$$

Therefore, $\mathrm{D}\nabla_S V(x) = \nabla_S^2$. As $\nabla_S U(x) = x + \nabla_S V(x)$, and $y^w(x) = \left(1 - \frac{h}{2}\right)\mathrm{I} - \frac{h}{2}\nabla_S V(x) + w$, their derivatives are given by

$$\begin{aligned}
\mathrm{D}\nabla_S U(x) &= \mathrm{D}x + \mathrm{D}\nabla_S V(x) \\
&= \mathrm{I} + \nabla_S^2 V(x), \\
\mathrm{D}y^w(x) &= \left(1 - \frac{h}{2}\right)\mathrm{D}x - \frac{h}{2}\mathrm{D}\nabla_S V(x) \\
&= \left(1 - \frac{h}{2}\right)\mathrm{I} - \frac{h}{2}\nabla_S^2 V(x).
\end{aligned}$$

□

We now calculate the W–gradient of F_h^w.

Proposition 3.25. *Let Assumption 3.4 be satisfied. The W–gradient of $F_h^w(x)$ can be*

decomposed for $x, w \in W$:

$$\begin{aligned}\nabla_W F_h^w(x) = \nabla_W V(y) - \nabla_W V(x) - \frac{1}{2}(\nabla_W^2 V(y) - \nabla_W^2 V(x))(y-x) \\ - \frac{h}{4}\left(\nabla_W^2 V(x)\nabla_S^2 V(y) + \nabla_W^2 V(y)\right)(y-x) \\ + \frac{h}{8-2h}\left(\nabla_W^2 V(y) - \nabla_W^2 V(x)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right)) \\ - \frac{h^2}{16-4h}\left(\nabla_W^2 V(x)\nabla_S^2 V(y) + \nabla_W^2 V(y)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right).\end{aligned}$$

Proof. By Lemma 3.4, $F_h^w(x)$ is given by

$$F_h^w(x) = A_h^w(x) + \frac{h}{8-2h} B_h^w(x)$$

with

$$\begin{aligned}A_h^w(x) &= V(y(x)) - V(x) - \frac{1}{2}\langle \nabla_S V(y(x)) + \nabla_S V(x), y(x) - x\rangle_S, \\ B_h^w(x) &= \langle \nabla_S U(y(x)) - \nabla_S U(x), \nabla_S V(y(x)) - \nabla_S V(x)\rangle_S.\end{aligned}$$

First, we calculate the S–gradient $\nabla_S F_h^w(x)$ and derive the W–gradient from the identity

$$\nabla_W F_h^w(x) = \mathcal{C}^{-1} \nabla_S F_h^w(x).$$

The S–gradient of A_h^w is given by

$$\begin{aligned}\nabla_S A_h^w(x,w) = \mathrm{D}(y)^* \nabla_S V(y) - \nabla_S V(x) \\ - \frac{1}{2}(\mathrm{D}(y)^* \nabla_S^2 V(y) + \nabla_S^2 V(x))(y-x) \\ - \frac{1}{2}\mathrm{D}(y-x)^*(\nabla_S V(y) + \nabla_S V(x)),\end{aligned}$$

where $\mathrm{D}(y)^*$ denotes the adjoint operator of $\mathrm{D}(y)$ on S. Note that

$$\mathrm{D}(y) = \mathrm{I} - \frac{h}{2}\nabla_S^2 U(x)$$

and is self–adjoint on S as the sum of the identity and a second derivative operator. There-

fore, we can conclude

$$\begin{aligned}
\nabla_S A_h(x,w) &= \nabla_S V(y) - \nabla_S V(x) - \frac{1}{2}(\nabla_S^2 V(y) + \nabla_S^2 V(x))(y-x) \\
&\quad - \frac{h}{2}\nabla_S^2 U(x)\left(\nabla_S V(y) - \frac{1}{2}\nabla_S^2 V(y)(y-x)\right) \\
&\quad + \frac{h}{4}\nabla_S^2 U(x)(\nabla_S V(y) + \nabla_S V(x)) \\
&= \nabla_S V(y) - \nabla_S V(x) - \frac{1}{2}(\nabla_S^2 V(y) + \nabla_S^2 V(x))(y-x) \\
&\quad - \frac{h}{4}\nabla_S^2 U(x)\left(\nabla_S V(y) - \nabla_S V(x)\right) \\
&\quad + \frac{h}{4}\nabla_S^2 U(x)\nabla_S^2 V(y)(y-x).
\end{aligned}$$

The gradient of the second summand B_h^w is derived by similar calculations:

$$\begin{aligned}
\nabla_S B_h^w(x) &= \left(\mathrm{D}(y)^*\nabla_S^2 V(y) - \nabla_S^2 V(x)\right)\left(\nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad + \left(\mathrm{D}(y)^*\nabla_S^2 U(y) + \nabla_S^2 U(x)\right)\left(\nabla_S V(y) - \nabla_S V(x)\right) \\
&= \left(\nabla_S^2 V(y) - \nabla_S^2 V(x)\right)\left(\nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad + \left(\nabla_S^2 U(y) + \nabla_S^2 U(x)\right)\left(\nabla_S V(y) - \nabla_S V(x)\right) \\
&\quad - \frac{h}{2}\nabla_S^2 U(x)\nabla_S^2 U(y)(\nabla_S V(y) - \nabla_S V(x)) \\
&\quad - \frac{h}{2}\nabla_S^2 U(x)\nabla_S^2 V(y)(\nabla_S U(y) + \nabla_S U(x)) \\
&= \left(\nabla_S^2 V(y) - \nabla_S^2 V(x)\right)\left(\nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad + \left(\nabla_S^2 V(y) - \nabla_S^2 V(x)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad + 2\nabla_S^2 U(x)(\nabla_S V(y) - \nabla_S V(x)) \\
&\quad - \frac{h}{2}\nabla_S^2 U(x)(\nabla_S V(y) - \nabla_S V(x)) \\
&\quad - \frac{h}{2}\nabla_S^2 U(x)\nabla_S^2 V(y)(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)).
\end{aligned}$$

In $A_h^w(x)$ and $B_h^w(x)$, there are in total 3 terms of the type $\nabla_S^2 U(x)(\nabla_S V(y) - \nabla_S V(x))$. As

$$-\frac{h}{4} + \frac{h}{8-2h}\left(2 - \frac{h}{2}\right) = 0,$$

these terms all cancel and we can summarize the above results:

$$\begin{aligned}
\nabla_S F_h^w(x) &= \nabla_S A_h^w(x) + \frac{h}{8-2h}\nabla_S B_h^w(x) \\
&= \nabla_S V(y) - \nabla_S V(x) - \frac{1}{2}(\nabla_S^2 V(y) + \nabla_S^2 V(x))(y-x) \\
&\quad + \frac{h}{4}\nabla_S^2 U(x)\nabla_S^2 V(y)(y-x) \\
&\quad + \frac{h}{8-2h}\left(\nabla_S^2 V(y) - \nabla_S^2 V(x)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad - \frac{h^2}{16-4h}\nabla_S^2 U(x)\nabla_S^2 V(y)(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)) \\
&= \nabla_S V(y) - \nabla_S V(x) - \frac{1}{2}(\nabla_S^2 V(y) + \nabla_S^2 V(x))(y-x) \\
&\quad + \frac{h}{4}(\mathrm{I} + \nabla_S^2 V(x))\nabla_S^2 V(y)(y-x) \\
&\quad + \frac{h}{8-2h}\left(\nabla_S^2 V(y) - \nabla_S^2 V(x)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad - \frac{h^2}{16-4h}\left(\mathrm{I} + \nabla_S^2 V(x)\right)\nabla_S^2 V(y)(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)).
\end{aligned}$$

With $\nabla_W F_h^w(x) = \mathcal{C}^{-1}\nabla_S F_h^w(x)$, we now get the final depiction of the derivative:

$$\begin{aligned}
\nabla_W F_h^w(x) &= \nabla_W V(y) - \nabla_W V(x) - \frac{1}{2}(\nabla_W^2 V(y) + \nabla_W^2 V(x))(y-x) \\
&\quad + \frac{h}{4}\left(\nabla_W^2 V(y) + \nabla_W^2 V(x)\nabla_S^2 V(y)\right)(y-x) \\
&\quad + \frac{h}{8-2h}\left(\nabla_W^2 V(y) - \nabla_W^2 V(x)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right) \\
&\quad - \frac{h^2}{16-4h}\left(\nabla_W^2 V(y) + \nabla_W^2 V(x)\nabla_S^2 V(y)\right)\left(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\right).
\end{aligned}$$

□

Similarly to Lemma 3.26, we now bound the terms arising in the derivative of the acceptance probability in terms of the derivatives of the potential V.

Lemma 3.26. *Let Assumption 3.4 be satisfied. Then for $x, y \in W$, the following bounds*

hold:

1) $$\left\|\nabla_S V(y) - \nabla_S V(x) - \frac{1}{2}(\nabla_S^2 V(y) + \nabla_S^2 V(x))(y-x)\right\|_W \leq \frac{1}{12} L_4(x,y)\|y-x\|_W,$$

2) $$\left\|(\nabla_W^2 V(y) + \nabla_W^2 V(x))\nabla_S^2 V(y)(y-x)\right\|_W \leq L_2(y,y)(1 + c_\pi L_2(x,x))\|y-x\|_W,$$

3) $$\begin{aligned}&\left\|(\nabla_W^2 V(y) - \nabla_W^2 V(x))(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(x) + \nabla_S U(y))\right\|_W,\\ &\quad\leq L_3(x,y)(L_2(x,y)\|y-x\|_W + \|\nabla_S U(y) + \nabla_S U(x)\|_W)\|y-x\|_W\end{aligned}$$

4) $$\begin{aligned}&\left\|(\nabla_W^2 V(y) + \nabla_W^2 V(x)\nabla_S^2 V(y))(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x))\right\|_W,\\ &\quad\leq (1 + c_\pi L_2(x,x))L_2(y,y)(L_2(x,y)\|y-x\|_W + \|\nabla_S U(y) + \nabla_S U(x)\|_W).\end{aligned}$$

Proof. 1) Analogously to the proof of Lemma 3.20, we define for x, y and $\xi \in W$,

$$f_\xi(t) := \langle \nabla_W V(x + t(y-x)), \xi\rangle_W.$$

The derivative of f_ξ are given by

$$\begin{aligned}f_\xi(t) &= \mathrm{D}V(x + t(y-x))(\xi),\\ f_\xi'(t) &= \mathrm{D}^2 V(x + t(y-x))(\xi, y-x),\\ f_\xi''(t) &= \mathrm{D}^3 V(x + t(y-x))(\xi, y-x, y-x),\\ f_\xi'''(t) &= \mathrm{D}^4 V(x + t(y-x))(\xi, y-x, y-x, y-x).\end{aligned}$$

Like in Lemma 3.20, the above terms can be used to find the following bound

$$\begin{aligned}&\langle \nabla_W V(y), \xi\rangle_W - \langle \nabla_W V(x), \xi\rangle_W - \frac{1}{2}\big\langle (\nabla_W^2 V(y) + \nabla_W^2 V(x))(y-x), \xi\big\rangle_W\\ &= f_\xi(1) - f_\xi(0) - \frac{1}{2}(f_\xi'(0) + f_\xi'(1))\\ &= \frac{1}{2}\int_0^1 t(1-t) f_\xi'''(t)\mathrm{d}t\\ &\leq \frac{1}{12} L_4(x,y)\|\xi\|_W \|y-x\|_W^3\end{aligned}$$

such that

$$\begin{aligned}&\left\|\nabla_W V(y) - \nabla_W V(x) - \frac{1}{2}(\nabla_W^2 V(y) + \nabla_W^2 V(x))(y-x)\right\|_W\\ &\leq \frac{1}{12} L_4(x,y)\|y-x\|_W^3.\end{aligned}$$

2) The second statement follows from Lemma 3.23 and the two inequalities

$$\begin{aligned}\left\|(\nabla_W^2 V(y))(y-x)\right\|_W &\leq L_2(y,y)\|x-y\|_W,\\ \left\|\nabla_W^2 V(x)\nabla_S^2 V(y)(y-x)\right\|_W &\leq c_\pi L_2(x,x)L_2(y,y)\|x-y\|_W.\end{aligned}$$

3) For the third one, note that

$$\begin{aligned}
&\left\|(\nabla_W^2 V(y) - \nabla_W^2 V(x))(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(x) + \nabla_S U(y))\right\|_W \\
&= \sup_{\|\xi\|_W=1} \left\langle \xi, (\nabla_W^2 V(y) - \nabla_W^2 V(x))(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x))\right\rangle_W \\
&= \sup_{\|\xi\|_W=1} (\mathrm{D}^2 V(y) - \mathrm{D}^2 V(x))(\xi, \nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)) \\
&\leq \sup_{\|\xi\|_W=1} L_3(x,y)\|y-x\|_W\|\xi\|_W\|\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x)\|_W \\
&\leq L_3(y,x)\|y-x\|_W \left(L_2(x,y)\|y-x\|_W + \|\nabla_S U(y) + \nabla_S U(x)\|_W\right).
\end{aligned}$$

4) Finally, the two inequalities hold

$$\begin{aligned}
&\left\|\nabla_W^2 V(y)(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x))\right\|_W \\
&\leq L_2(y,y)(L_2(x,y)\|y-x\|_W + \|\nabla_S U(y) + \nabla_S U(x)\|_W) \qquad \text{and} \\
&\left\|\nabla_W^2 V(x)\nabla_S^2 V(y)(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x))\right\|_W \\
&\leq c_\pi L_2(x,x)L_2(y,y)(L_2(x,y)\|y-x\|_W + \|\nabla_S U(y) + \nabla_S U(x)\|_W).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\left\|(\nabla_W^2 V(x)\nabla_S^2 V(y) + \nabla_W^2 V(y))(\nabla_S V(y) - \nabla_S V(x) + \nabla_S U(y) + \nabla_S U(x))\right\|_W \\
&\leq (1 + c_\pi L_2(x,x))L_2(y,y)(L_2(x,y)\|y-x\|_W + \|\nabla_S U(y) + \nabla_S U(x)\|_W).
\end{aligned}$$

□

We are now ready to prove the main result of this section, Proposition 3.22.

Proof. (Proposition 3.19)
By the definition of G_h, and Proposition 3.25, the expectation of G_h can be bounded by

$$\begin{aligned}
\mathbb{E}\left[\|\nabla_W G_h(x, Y_{h,n}(x))\|_W\right] &\leq \mathbb{E}\left[\left\|\nabla_S F_h^{\tilde{h}N_n}(x)\right\|_W\right] \\
&\leq I + \frac{h}{4} II + \frac{h}{8-2h} III + \frac{h^2}{16-4h} IV,
\end{aligned}$$

where I–IV are given by

$$\begin{aligned}
I &:= \mathbb{E}\left[\left\|\nabla_W V(Y_{h,n}(x)) - \nabla_W V(x) - \frac{1}{2}(\nabla_W^2 V(Y_{h,n}(x)) - \nabla_W^2 V(x))(Y_{h,n}(x) - x)\right\|_W\right], \\
II &:= \mathbb{E}\left[\left\|(\nabla_W^2 V(Y_{h,n}(x)) + \nabla_W V(x)\nabla_S V(Y_{h,n}(x)))(Y_{h,n}(x) - x)\right\|_W\right], \\
III &:= \mathbb{E}\left[\left\|(\nabla_W^2 V(Y_{h,n}(x)) - \nabla_W^2 V(x))(\nabla_S V(Y_{h,n}(x)) - \nabla_S V(x) + \nabla_S U(Y_{h,n}(x)) + \nabla_S U(x))\right\|_W\right], \\
IV &:= \mathbb{E}\left[\|(\nabla_W V(Y_{h,n}(x)) \right. \\
&\qquad \left. + \nabla_W^2 V(x)\nabla_S^2 V(Y_{h,n}(x)))(\nabla_S V(Y_{h,n}(x)) - \nabla_S V(x) + \nabla_S U(Y_{h,n}(x)) + \nabla_S U(x))\right\|_W\right]
\end{aligned}$$

The results of Lemma 3.21 and 3.26 lead to

$$I \leq \frac{1}{12}\mathbb{E}[L_4(x, Y_{h,n}(x))\|Y_{h,n}(x) - x\|_W^3], \tag{3.28}$$

$$II \leq (1 + c_\pi L_2(x,x))\mathbb{E}[L_2(Y_{h,n}(x), Y_{h,n}(x))\|Y_{h,n}(x) - x\|_W,] \tag{3.29}$$

$$III \leq \mathbb{E}[L_3(x, Y_{h,n}(x))\|Y_{h,n}(x) - x\|_W((1 + L_2(x, Y_{h,n}(x)))\|Y_{h,n}(x) - x\|_W + 2\|\nabla_S U(x)\|_W)], \tag{3.30}$$

$$IV \leq (1 + L_2(x,x))\mathbb{E}[L_2(Y_{h,n}(x), Y_{h,n}(x))(1 + L_2(x, Y_{h,n}(x))\|Y_{h,n}(x) - x\|_W + 2\|\nabla_S U(x)\|_W)]. \tag{3.31}$$

Similarly to Proposition 3.19, we can bound (3.28) – (3.31) by

$$\begin{aligned} I &\leq h^{\frac{3}{2}}\, \mathcal{Q}_1(\|x\|_W, \|\nabla_S U(x)\|_W), \\ II &\leq h^{\frac{1}{2}}\, \mathcal{Q}_2(\|x\|_W, \|\nabla_S U(x)\|_W), \\ III &\leq h^{\frac{1}{2}}\, \mathcal{Q}_3(\|x\|_W, \|\nabla_S U(x)\|_W), \\ IV &\leq h^{\frac{1}{2}}\, \mathcal{Q}_4(\|x\|_W, \|\nabla_S U(x)\|_W) \end{aligned}$$

where $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3, \mathcal{Q}_4 : \mathbb{R}^2 \to \mathbb{R}$ are polynomials of degree $p_4 + 3$, $2p_2 + 1$, $p_3 + p_2 + 2$ and $3p_2 + 1$ respectively. Therefore, the expectation value of the gradient of the acceptance probability satisfies

$$\begin{aligned} \mathbb{E}\left[\|\nabla_S G_h(x, Y_{h,n}(x))\|_W\right] &\leq I + \frac{h}{4}II + \frac{h}{8-2h}III + \frac{h^2}{16-4h}IV \\ &\leq h^{\frac{3}{2}}\mathcal{Q}(\|x\|_W, \|\nabla_S U(x)\|_W) \end{aligned}$$

where $\mathcal{Q} : \mathbb{R}^2 \to \mathbb{R}$ is a polynomial of degree $d_{\mathcal{Q}} := \max\{p_4 + 3, p_3 + p_2 + 2, 3p_2 + 1\}$. □

As a direct consequence, we get a bound on the difference between the acceptance probabilities of the process starting in x and $\tilde{x}$. This lets us control the second summand in (3.26).

Corollary 3.27. *Let $a_h : W \times W \to [0,1]$ be the acceptance probability of the coupling of the MALA–process $(X_n, \tilde{X}_n)_{n\in\mathbb{N}}$ as constructed in Chapter 3.3.1. Then for all $h \in (0,2)$*

$$\begin{aligned} &\mathbb{E}\left[|a_h(x, Y_{h,n}(x)) - a_h(\tilde{x}, Y_{h,n}(\tilde{x}))|\right] \\ &\leq \mathbb{E}\left[|G_h(x, Y_{h,n}(x)) - G_h(\tilde{x}, Y_{h,n}(\tilde{x}))|\right] \\ &\leq h^{\frac{3}{2}}\|x - \tilde{x}\|_W\, \mathcal{Q}\left(\max\{\|x\|_W, \|\tilde{x}\|_W\}, \sup_{z\in[x,\tilde{x}]}\|\nabla_S U(z)\|_W\right) \end{aligned}$$

holds, where $\mathcal{Q} : \mathbb{R}^2 \to \mathbb{R}$ is the polynomial from Proposition 3.22.

Proof. By construction of the coupling and of G_h, we have

$$a(x, Y_{h,n}(x)) = \exp(-G_h(x, Y_{h,n}(x)) \wedge 0),$$
$$a(\tilde{x}, Y_{h,n}(\tilde{x})) = \exp(-G_h(\tilde{x}, Y_{h,n}(\tilde{x})) \wedge 0).$$

As $x \mapsto \exp(-x)$ is 1–Lipschitz for $x \in [0, \infty)$, we get

$$\begin{aligned}\mathbb{E}\left[|a(x, Y_{h,n}(x)) - a(\tilde{x}, Y_{h,n}(\tilde{x}))|\right] &\leq \mathbb{E}\left[|G_h(x, Y_{h,n}(x)) - G_h(\tilde{x}, Y_{h,n}(\tilde{x}))|\right] \\ &\leq \mathbb{E}\left[\|\nabla_W G_h(z, Y_{h,n}(z))\|_W\right] \|x - \tilde{x}\|_W.\end{aligned}$$

Proposition 3.22 bounds $\mathbb{E}\left[\|\nabla_W G_h(x, Y_{h,n}(z))\|_W\right]$ such that

$$\begin{aligned}&\mathbb{E}\left[|a(x, Y_{h,n}(x)) - a(\tilde{x}, Y_{h,n}(\tilde{x}))|\right] \\ &\leq h^{\frac{3}{2}} \|x - \tilde{x}\|_W \, \mathcal{Q}\left(\max\{\|x\|_W, \|\tilde{x}\|_W\}, \sup_{z \in [x,\tilde{x}]} \|\nabla_S U(z)\|_W\right).\end{aligned}$$

□

First bound on the Wasserstein distance of the MALA–process

Combining the results of the sections above where we controlled the acceptance probability of the MALA–process, we finally derive a bound on the Wasserstein distance of the coupling of the MALA–process. This bound depends on the Wasserstein distance of the metric $d_R(x, y) := \|y - x\|_W \wedge R$.

Proposition 3.28. *Let Assumptions 3.3 and 3.4 be satisfied, and $q : \mathcal{W} \times \mathcal{B}(W) \to [0, 1]$ be the transition kernel of the MALA–process. Define $d_R(x, y) := \|x - y\|_W \wedge R$ and $U_R := \{x \in W | d(x, 0) < \frac{R}{2}\}$. Then there exists $r \in \mathbb{N}$ and a pair coupling c of q such that for all $x, \tilde{x} \in U_R$*

$$\int_{W \times W} d_R(y, \tilde{y}) c((x, \tilde{x}), \mathrm{d}y\mathrm{d}\tilde{y}) \leq \left(1 - \frac{h}{2}\delta + h^{\frac{3}{2}}\gamma(1 + R^r)\right) d_R(x, \tilde{x}).$$

Proof. Let c be the pair coupling of $(X_n)_{n \in \mathbb{N}}$ and q as constructed in Section 3.3.1. $(X_n)_{n \in \mathbb{N}}$ is the MALA–process constructed in Section 3.1. By Proposition 3.16, we know that

$$\begin{aligned}\int_{W \times W} d(y, \tilde{y}) c((x, \tilde{x}), \mathrm{d}y\mathrm{d}\tilde{y}) &= \mathbb{E}_{x,\tilde{x}}[d_R(Y_{h,n}(x), Y_{h,n}(\tilde{x}))] \\ &\leq \mathbb{E}_{x,\tilde{x}}\left[d_R(Y_{h,n}(x), Y_{h,n}(\tilde{x}))\right] \\ &\quad + R\,\mathbb{E}_{x,\tilde{x}}\left[|G_h(x, Y_{h,n}(x)) - G_h(\tilde{x}, Y_{h,n}(\tilde{x}))|\right] \\ &\quad + d_R(x, \tilde{x})\,\mathbb{E}_{x,\tilde{x}}\left[\min\{G_h(x, Y_{h,n}(x))^+, G_h(\tilde{x}, Y_{h,n}(\tilde{x}))^+\}\right].\end{aligned}$$

These three summands can now be bounded by the results of the previous sections. As $x, \tilde{x} \in U_R$, we have $\|Y_{h,n}(x) - Y_{h,n}(\tilde{x})\|_W \leq R$ and

$$\begin{aligned} d_R(x, \tilde{x}) &= \|x - \tilde{x}\|_W, \\ d_R(Y_{h,n}(x), Y_{h,n}(\tilde{x})) &= \|Y_{h,n}(x) - Y_{h,n}(\tilde{x})\|_W \leq \left(1 - \frac{h}{2}\delta\right) \|x - \tilde{x}\|. \end{aligned}$$

Therefore, Corollary 3.27 leads for all $h \in (0, 2)$ to

$$\begin{aligned} &\mathbb{E}_{x,\tilde{x}} \left[|G_h(x, Y_{h,n}(x)) - G_h(\tilde{x}, Y_{h,n}(\tilde{x}))| \right] \\ &\leq h^{\frac{3}{2}} \|x - \tilde{x}\|_W \, \mathcal{Q}\left(\max\{\|x\|_W, \|\tilde{x}\|_W\}, \sup_{z \in [x,\tilde{x}]} \right). \end{aligned}$$

Applying Proposition 3.22 results in

$$\mathbb{E}\left[\|\nabla_W G_h(x, Y_{h,n}(x))\|_W \right] \leq h^{\frac{3}{2}} \mathcal{Q}(\|x\|_W, \|\nabla_S U(x)\|_W) \qquad \text{for } h \in (0, 2)$$

which implies

$$\begin{aligned} &R\, \mathbb{E}_{x,\tilde{x}} \left[|G_h(x, Y_{h,n}(x)) - G_h(\tilde{x}, Y_{h,n}(\tilde{x}))| \right] \\ &\leq R \sup_{z \in U_R} \mathcal{Q}(\|z\|_W, \|\nabla_S U(z)\|_W) \, \|x - \tilde{x}\|_W \\ &\leq R \sup_{z \in U_R} \mathcal{Q}(\|z\|_W, \|\nabla_S U(z)\|_W) \, d_R(x, \tilde{x}). \end{aligned}$$

Finally, the third summand is bounded by Proposition 3.19

$$\begin{aligned} &d_R(x, \tilde{x}) \mathbb{E}_{x,\tilde{x}} \left[\min\{G_h(x, Y_{h,n}(x))^+, G_h(\tilde{x}, Y_{h,n}(\tilde{x}))^+\} \right] \\ &\leq d_R(x, \tilde{x}) \sup_{z \in U_R} \mathcal{P}(\|z\|_W, \|\nabla_S U(z)\|_W) \cdot h^{\frac{3}{2}}. \end{aligned}$$

Moreover, Assumption 3.3 gives us $\|\nabla_S U(x)\|_W \leq (1 + L)\|x\|_W + \|U(0)\|_W$. As $\mathcal{P}$ and $\mathcal{Q}$ are polynomials, we can now choose $c > 0$, such that

$$c(1 + R^r) \geq \sup_{z \in U_R} R\, \mathcal{P}(\|z\|_W, \|\nabla_S U(z)\|_W) + \mathcal{Q}(\|z\|_W, \|\nabla_S U(z)\|_W) \quad \text{for all } R \geq 0 \tag{3.32}$$

with

$$r := \max\{\deg \mathcal{P} + 1, \deg \mathcal{Q}\}. \tag{3.33}$$

Therefore, as $\delta = 1 - L$,

$$\begin{aligned}
&\int_{W\times W} d_R(y,\tilde{y})c((x,\tilde{x}),\mathrm{d}y\mathrm{d}\tilde{y}) \\
&\leq \left(1 - \frac{1}{2}\delta h\right) d_R(x,\tilde{x}) \\
&\quad + R \sup_{z\in U_R} \mathcal{Q}(\|z\|_W, \|\nabla_S U(z)\|_W)\, d_R(x,\tilde{x}) \\
&\quad + d_R(x,\tilde{x}) \sup_{z\in U_R} \mathcal{P}(\|z\|_W, \|\nabla_S U(z)\|_W) \cdot h^{\frac{3}{2}} \\
&\leq \left(1 - \frac{h}{2}\delta + ch^{\frac{3}{2}}(1+R^r)\right) d_R(x,\tilde{x}).
\end{aligned}$$

□

Remark 3.29. *Note that r, the power of R in the remainder term, is bounded by*

$$\max\{\deg\mathcal{P}+1, \deg\mathcal{Q}\} = \max\{p_3+4, 2p_2+3, p_4+3, p_3+p_2+2, 3p_2+1\}.$$

Corollary 3.30. *Under the assumptions of Proposition 3.28, we have*

$$W_{d_R}(\mu q^n, \nu) \leq \left(1 - \frac{h}{2}\delta + h^{\frac{3}{2}}c(1+R^r)\right)^n + \frac{2R}{h\left(\delta + 2h^{\frac{1}{2}}c(1+R^r)\right)}(C_n(U_R,\mu) + C_n(U_R,\nu))$$

where C_n was defined in (3.22).

Proof. The result follows directly from Proposition 3.28 and Theorem 3.1. □

In order to use Corollary 3.30 to find a bound on the Wasserstein distance, it remains to get good bounds on the escape probabilities $C_n(U_R,\mu)$ of the MALA–process. This will be subject of the next section:

Controlling the escape probability

This section we show the existence of bounds for the escape probability of the MALA–process on a ball with radius R to control the terms $C_n(U_R,\mu)$ and $C_n(U_R,\nu)$ as arising in Corollary 3.30.

Theorem 3.2. *Let Assumption 3.3 be satisfied. Let $(X_i)_{i\in\mathbb{N}}$ be the MALA–process $(X_n)_{n\in\mathbb{N}}$, as constructed in Chapter 3.1. Then there exist constants $\theta_0 > 0$, $R_0 < \infty$ and $\kappa > 0$,*

independent of $n \in \mathbb{N}$*, and a polynomial function* $\mathcal{P}(R)$*, such that for all* $0 < \theta < \theta_0$*,* $R > R_0$*,* $\mathrm{x} \in W$ *with* $\|x\|_W < \frac{1}{2}R$*, and* $h \leq h_+(R)$*,*

$$\mathrm{P}_x \left[\|X_i\|_W < R \quad \forall i \in \{0, \dots, n\} \right] \geq 1 - n \left(\exp(\kappa) \exp\left(\theta \left(\|x\|_W^2 - \frac{R^2}{2} \right) \right) \right).$$

The upper bound for h *is given by*

$$h_+(R) := \frac{1}{4} L^2 \mathcal{P}(R)^{-2}. \tag{3.34}$$

The rough idea of the proof of Theorem 3.2 is to bound the exponential moment

$$\mathbb{E}_x \left[\exp\left(\theta(\|X_n\|_W^2) \right) \right] \leq \exp(\theta \|x\|_W^2 + \kappa) \tag{3.35}$$

for a constant κ independent of h, R and i using the contraction property established in the previous sections, and to apply Markov's inequality to bound the probability of the chain leaving the ball with radius R.

To this purpose, we need the following fact on Gaussian measures on Hilbert spaces:

Lemma 3.31. *Let* ν *be a Gaussian measure on* W *with covariance operator* Q*. Set* $c_w := 4 \operatorname{trace} Q$*, then for* $s \in \left[0, \frac{1}{c_w}\right)$

$$\int_W \exp(s\|z\|_W^2) \nu(\mathrm{d}z) \leq \exp\left(\frac{1}{2} c_w s \right)$$

holds. Furthermore, for all $s \geq 0$ *and all* $x \in W$*,*

$$\int_W \exp(s \langle x, y \rangle_W) \nu(\mathrm{d}y) \leq \exp\left(\frac{1}{8} c_w s^2 \|x\|_W^2 \right).$$

Proof. As $\|Q\|_{W \to W} \leq \operatorname{trace} Q$, we have for $0 \leq s < \frac{1}{c_w}$ by [12, Proposition 2.16],

$$\int_W \exp(s\|x\|_W^2) \nu(\mathrm{d}x) = \exp\left(\frac{1}{2} \sum_{k=1}^{\infty} \frac{(2s)^k}{k} \operatorname{trace} Q^k \right)$$

We now bound

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{(2s)^k}{k} \operatorname{trace} Q^k &\leq 2s \operatorname{trace} Q \left(1 + \sum_{k=1}^{\infty} \frac{(2s \operatorname{trace} Q)^k}{k} \right) \\ &\leq 2s \operatorname{trace} Q \sum_{k=0}^{\infty} \frac{2^k}{4^k} \\ &\leq 4s \operatorname{trace} Q, \end{aligned}$$

which proves the first part of the lemma. For the second one, note that $\langle x, Y\rangle_W$ is one–dimensional Gaussian with mean 0 and variance $\langle x, Qx\rangle_W$ if $x \in W$ and Y is a ν–distributed random variable. Therefore,

$$\begin{aligned}\int_W \exp(s\langle x, y\rangle_W)\nu(\mathrm{d}y) &= \frac{1}{\sqrt{2\pi\langle x, Qx\rangle_W}} \int_{-\infty}^{\infty} \exp(sz) \exp\left(-\frac{z^2}{2\langle x, Qx\rangle_W}\right) \mathrm{d}z \\ &= \exp\left(\frac{1}{2}s^2\langle x, Qx\rangle_W\right) \\ &\leq \exp\left(\frac{1}{2}s^2 \operatorname{trace} Q\|x\|_W^2\right) \\ &= \exp\left(\frac{1}{8}c_w s^2\|x\|_W^2\right).\end{aligned}$$

□

The following Lemma introduces a bound on the exponential moment of X_n conditioned on the event that the Markov Chain $(X_n)_{n\in\mathbb{N}}$ has not left the ball of radius R until time $n-1$. This is the key step for the proof of Theorem 3.2.
For $i \in \mathbb{N}$, define the events $\mathcal{B}_i$ by

$$\mathcal{B}_i := \{\|X_i\| < R \text{ for } i \in \{1, \dots, i\}\}. \tag{3.36}$$

Lemma 3.32. *Let $(X_i)_{i\in\mathbb{N}}$ be the MALA–process constructed in Section 3.1 with step size h, and h_+ given by (3.34). Let $n \in \mathbb{N}$. Then there exist constants $\theta_0 > 0$, $R_0 < \infty$ and $\kappa > 0$, independent of $i \in \mathbb{N}$, and a polynomial function $\mathcal{P}(R)$ such that for all $0 < \theta < \theta_0$, $R > R_0$, $x \in W$ with $\|x\|_W < \frac{1}{2}R$ and $h \leq h_+(R)$*

$$\mathbb{E}_x\left[\exp\left(\theta(\|X_n\|_W^2)\right) \mathrm{I}_{\mathcal{B}_{n-1}}\right] \leq \exp(\theta\|x\|_W^2 + \kappa).$$

Proof. To calculate the expectation value, we split it on the sets $\mathcal{A}_n(x) := \{U_n < a_h\left(x, Y_{h,n-1}(x)\right)\}$, where the proposal is accepted, and $\mathcal{A}_n^c(x)$ where it is rejected. Then for $\theta > 0$,

$$\begin{aligned}&\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}}\right] \\ &\leq \mathbb{E}_x\left[\exp\left(\theta\|Y_{h,n-1}(X_{n-1})\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}}, \mathcal{A}_n(X_{n-1})\right] + \mathbb{E}_x\left[\exp\left(\theta\|X_{n-1}\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}}, \mathcal{A}_n^c(X_{n-1})\right] \\ &\leq \mathbb{E}_x\left[\exp\left(\theta\|Y_{h,n-1}(X_{n-1})\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}}\right] + \mathbb{E}_x\left[\exp\left(\theta\|X_{n-1}\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}}\right] \sup_{\|z\|_W<R} \mathrm{P}_x[\mathcal{A}_n^c(z)],\end{aligned} \tag{3.37}$$

because $\|X_{n-1}\|_W < R$ almost surely on $\mathcal{B}_{n-1}$.
The acceptance part is the important one in our analysis, the rejection part is in the end treated as error. It can be controlled as by Assumption 3.3, $\|\nabla_S U(x)\|_W$ grows at most linearly in x. Therefore, we deduce from Proposition 3.19 the existence of a polynomial $\mathcal{P}(R)$ such that for all $x \in W$ with $\|x\|_W < \frac{1}{2}R$

$$\sup_{\|z\|_W < R} \mathrm{P}_x[\mathcal{A}_n^c(z)] \leq \mathcal{P}(R) h^{\frac{3}{2}}.$$

Now, we turn to the acceptance part. First, we define

$$Y_n^* := \left(1 - \frac{h}{2}\right) X_{n-1} + \frac{h}{2} \nabla_S V(X_{n-1}).$$

With Assumption 3.3, we can bound this term:

$$\begin{aligned}
\|Y_n^*\|_W^2 &\leq \left(\left(1 - \frac{h}{2}(1-L)\right) \|X_n\|_W + \frac{h}{2} a \right)^2 \\
&\leq \left(1 - \frac{h}{4}(1-L)\right) \|X_n\|_W^2 + \left(\frac{h}{2} a\right)^2 \frac{4}{(1-L)h} \\
&= (1 - \rho h) \|X_n\|_W^2 + h \frac{a^2}{4\rho}
\end{aligned}$$

where we set $\rho := \frac{1}{4}(1-L)$, $a := \|\nabla_S V(0)\|_W$, and used $(x+y)^2 \leq px^2 + qy^2$ for $\frac{1}{p} + \frac{1}{q} = 1$ with $p = \frac{1-\rho h}{1-2\rho h}$. We now calculate the conditional expectation $\mathbb{E}_x[\cdot|\mathcal{F}_{n-1}]$ with respect to the sigma–algebra $\mathcal{F}_{n-1} := \sigma(X_0, \ldots, X_{n-1})$. As N_n is independent of $\mathcal{F}_{n-1}$, we have

$$\begin{aligned}
&\mathbb{E}_x \left[\exp\left(\theta \|Y_{h,n}(X_{n-1})\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}} \middle| \mathcal{F}_{n-1} \right] \\
&\leq \exp\left(\theta \left\|Y_{n-1}^*\right\|_W^2\right) \mathrm{I}_{\mathcal{B}_{n-1}} \mathbb{E}_x \left[\exp\left(\theta \sqrt{\tilde{h}} \langle Y_{n-1}^*, N_n \rangle_W + \theta \tilde{h} \|N_n\|_W^2\right) \middle| \mathcal{F}_{n-1} \right] \\
&\leq \exp\left(\theta \left\|Y_{n-1}^*\right\|_W^2\right) \left(\mathbb{E}_x \left[\exp\left(2\theta \sqrt{\tilde{h}} \langle Y_{n-1}^*, N_n \rangle_W\right) \right] \mathbb{E}_x \left[\exp\left(2\theta \tilde{h} \|N_n\|_W^2\right) \right] \right)^{\frac{1}{2}} \\
&\leq \exp\left(\theta \left\|Y_{n-1}^*\right\|_W^2\right) \exp\left(2\theta^2 \frac{1}{8} c_w \tilde{h} \left\|Y_{n-1}^*\right\|_W^2\right) \exp\left(\frac{1}{2} \theta \tilde{h} c_w\right) \\
&\leq \exp\left(\theta(1-\rho h) \|X_{n-1}\|_W^2\right) \exp\left(\frac{1}{4} \theta^2 c_w h \|X_{n-1}\|_W^2\right) \exp\left(\theta h \left(\frac{a^2}{4\rho} + \frac{1}{16} \frac{a^2}{\rho} \theta c_w h + \frac{1}{2} c_w\right)\right)
\end{aligned}$$

for $\theta < \frac{1}{2c_w}$ applying Lemma 3.31. Setting

$$K := \frac{a^2}{4} \left(1 + \frac{1}{\rho}\right) + \frac{1}{2} c_w$$

and choosing $\theta \leq \min\left\{\frac{1}{2c_w}, \frac{2\rho}{c_w}\right\}$, we can sum up the derived bounds above:

$$\mathbb{E}_x\left[\exp\left(\theta\|Y_{h,n}(X_{n-1})\|_W^2\right)\middle|\,\mathcal{B}_{n-1}\right] \leq \exp\left(\theta\left(1-\frac{1}{2}\rho h\right)\|X_{n-1}\|_W^2 + \theta h K\right).$$

Using this estimate for (3.37) as well as the bounds already shown for $\sup_{\|z\|_W<R} P\left[\mathcal{A}_n^c(z)\right]$ leads to

$$\begin{aligned}
&\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right] \\
&\leq \exp\left(\theta\left(1-\frac{1}{2}\rho h\right)\|X_{n-1}\|_W^2 + \theta h K\right) + \mathcal{P}(R)h^{\frac{3}{2}}\exp\left(\theta\|X_{n-1}\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}.
\end{aligned}$$

By applying the expectation value and using $B_{i+1} \subset \mathcal{B}_i$ for $i \in \mathbb{N}$, we conclude:

$$\begin{aligned}
&\mathbb{E}_x\left[\exp\left(\theta\|X_{n+1}\|_W^2\right)\mathrm{I}_{\mathcal{B}_n}\right] \\
&= \mathbb{E}_x\left[\mathbb{E}_x\left[\exp\left(\theta\|X_{n+1}\|_W^2\right)\mathrm{I}_{\mathcal{B}_n}\middle|\,\mathcal{F}_n\right]\right] \\
&\leq \mathbb{E}_x\left[\exp\left(\theta\left(1-\frac{1}{2}\rho h\right)\|X_n\|_W^2 + \theta h K\right)\mathrm{I}_{\mathcal{B}_n} + \mathcal{P}(R)h^{\frac{3}{2}}\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_n}\right] \\
&\leq \exp(\theta h K)\mathbb{E}_x\left[\exp\left(\theta\left(1-\frac{1}{2}\rho h\right)\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right] + \mathcal{P}(R)h^{\frac{3}{2}}\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right] \\
&= \exp(\theta h K)\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right]^{\left(1-\frac{1}{2}\rho h\right)} \\
&\quad\cdot\left(1+\exp(-\theta h K)\mathcal{P}(R)h^{\frac{3}{2}}\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right]^{\frac{1}{2}\rho h}\right)
\end{aligned}$$

by Jensen's inequality. We now define for $n \in \mathbb{N}$

$$l_n := \log\left(\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right]\right).$$

Applying the logarithm to the results above gives us

$$\begin{aligned}
l_{n+1} &\leq \theta h K + \left(1-\frac{1}{2}\rho h\right) l_n \\
&\quad + \log\left(1+\exp(-\theta h K)\mathcal{P}(R)h^{\frac{3}{2}})\exp\left(\frac{1}{2}\rho h l_n\right)\right) \\
&\leq \theta h K + \left(1-\frac{1}{2}\rho h\right) l_n + \mathcal{P}(R)h^{\frac{3}{2}}\exp\left(\frac{1}{2}\rho h l_n\right).
\end{aligned}$$

Based on this inequality, we use an inductive argument to prove

$$l_n \leq \left(1-\frac{1}{2}\rho h\right)^n l_0 + \sum_{i=0}^{n-1}\left(1-\frac{1}{2}\rho h\right). \tag{3.38}$$

For $n = 0$, this is trivial, so assume

$$l_j \leq \left(1 - \frac{1}{2}\rho h\right)^j l_0 + \sum_{i=0}^{j-1} \left(1 - \frac{1}{2}\rho h\right)$$

to be true for $j = 0, \ldots, n$. Note that by this inductive assumption, $l_0 \leq \theta R^2$ implies $l_n \leq \theta R^2$ for sufficiently large R. Therefore, as $h \leq \frac{1}{16} L^2 \mathcal{P}(R)^{-2} \leq R^{-2}$ for sufficiently large R, we get

$$\exp\left(\frac{1}{2}\rho h l_n\right) \leq \exp\left(\frac{1}{2}\rho\theta\right) \leq e.$$

for $\theta \leq \frac{2}{\rho}$. This leads to

$$l_{n+1} \leq \theta h K + \left(1 - \frac{1}{2}\rho h\right) l_n + \mathcal{P}(R) h^{\frac{3}{2}} e.$$

As h is by assumption bounded from above by $h \leq \frac{1}{4} L^2 \mathcal{P}(R)^{-2}$

$$l_{n+1} \leq h\left(\theta K + \frac{1}{2} L e\right) + \left(1 - \frac{1}{2}\rho h\right) l_n$$

holds, which gives us

$$\begin{aligned} l_{n+1} &\leq \left(1 - \frac{1}{2}\rho h\right)^{n+1} l_0 + h\left(\theta K + \frac{1}{2} L e\right) \sum_{i=0}^{n} \left(1 - \frac{1}{2}\rho h\right) \\ &\leq l_0 + \frac{2K\theta + eL}{\rho} e \end{aligned}$$

by applying the inductive assumption. We now set

$$\kappa := \frac{2K\theta_0 + L}{\rho} e$$

and get

$$\mathbb{E}_x\left[\exp\left(\theta \|X_n\|_W^2\right)\right] \leq \exp\left(\theta \|x\|_W^2 + \kappa\right)$$

for $h \leq \frac{1}{4} L^2 \mathcal{P}(R)^{-2}$. □

Theorem 3.2 is now a consequence of Markov's inequality.

Proof. (Theorem 3.2)
We prove this by induction. For $n = 0$, the statement is trivial as we assumed $x < \frac{R}{2}$. So

assume the statement is true for $i \in \{0, \dots, n-1\}$. Then

$$\begin{aligned}
&\mathrm{P}_x\left[\|X_i\|_W < R \quad \forall i \in \{0,\dots,n\}\right] \\
&= \mathrm{P}_x\left[\|X_i\|_W < R \quad \forall i \in \{0,\dots,n-1\}\right] - \mathrm{P}_x\left[\|X_n\|_W > R, \|X_i\|_W < R \quad \forall i \in \{0,\dots,n-1\}\right] \\
&\geq 1 - (n-1)\left(\exp(\kappa)\exp\left(\theta\left(\|x\|_W^2 - \frac{R^2}{2}\right)\right)\right) - \mathrm{P}_x\left[\|X_n\|_W > R, \mathcal{B}_{n-1}\right],
\end{aligned}$$

where $\mathcal{B}_n$ was defined in (3.36). By the assumptions of Theorem, the conditions of Lemma 3.32 are fulfilled. Therefore, there exist constants κ, R_0 and $\theta_0 > 0$ such that for all $0 < \theta < \theta_0$, $R > R_0$, all $x \in W$ with $\|x\|_W < \frac{1}{2}R$ and $h \leq h_+(R)$

$$\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\right)\mathrm{I}_{\mathcal{B}_{n-1}}\right] \leq \exp\left(\theta\|x\|_W^2 + \kappa\right). \tag{3.39}$$

Markov's inequality bounds the probability that for $\|x\|_W < \frac{R}{2}$ the process X started in x does not leave the ball at step n by

$$\begin{aligned}
\mathrm{P}_x\left[\|X_n\|_W > R, \mathcal{B}_{n-1}\right] &\leq \exp(-\theta R^2)\mathbb{E}_x\left[\exp\left(\theta\|X_n\|_W^2\mathrm{I}_{\mathcal{B}_{n-1}}\right)\right] \\
&\leq \exp\left(-\theta R^2\right)\exp\left(\theta\|x\|_W^2 + \kappa_\theta\right).
\end{aligned}$$

This implies

$$\mathrm{P}_x\left[\|X_i\|_W < R \quad \forall i \in \{0,\dots,n\}\right] \geq 1 - n\left(\exp(\kappa)\exp\left(\theta\left(\|x\|_W^2 - \frac{R^2}{2}\right)\right)\right).$$

□

Final result on the Wasserstein distance of the MALA–process

The bounds on the escape probabilities are the final piece we need to prove the main theorem.

Theorem 3.3. *Let q_h be the kernel of the MALA–process with step-size $h \in (0,2)$. Let Assumptions 3.3 and 3.4 be satisfied. Then there exist $C > 0$ and $n_0 > 0$ such that for given $n \geq n_0$ there exists $h(n) > 0$ with*

$$\mathcal{W}_{d_1}(\nu q_{h(n)}^n, \mu) \leq \exp\left(-cn^{\frac{2}{1+r}}\right)(\mathcal{W}_{d_\infty}(\mu,\nu) + C).$$

Proof. As Assumptions 3.3 and 3.4 hold, we can apply Corollary 3.30:

$$\begin{aligned}
\mathcal{W}_{d_R}(\nu q_{h(n)}^n, \mu) \leq &\left(1 - \frac{h}{2}L + h^{\frac{3}{2}}\mathcal{P}(R)\right)^n d_R(\mu,\nu) \\
&+ \frac{2R}{h\left(L + 2h^{\frac{1}{2}}\mathcal{P}(R)\right)}(C_n(U_R,\mu) + C_n(U_R,\nu)).
\end{aligned}$$

Theorem 3.2 provides bounds for the escape probabilities C_n:
For $R > R_0$ and $\theta < \min\left\{\frac{1}{2c_w}, \frac{2\rho}{c_w}, 2\rho\right\}$

$$\begin{aligned} C_n(U_R, \nu) &= \sup_{i\in\{1,\dots,n\}} \mathrm{P}_\nu\left[\|X_i\|_W > R\right] \\ &\leq \mathrm{P}_\nu\left[\exists i \in \{1,\dots,n\} : \|X_i\|_W > R\right] \\ &\leq n\int_W \exp(\kappa)\exp\left(\theta\left(\|x\|_W^2 - R^2\right)\right)\nu(\mathrm{d}x) + \nu\left(\left\{\|x\|_W > \frac{R}{2}\right\}\right) \end{aligned}$$

These two terms can each be bounded based on the results from Lemma 3.31:

$$\begin{aligned} &\int_W \exp(\kappa)\exp\left(\theta\left(\|x\|_W^2 - R^2\right)\right)\nu(\mathrm{d}x) \\ &\leq \exp(\kappa)\exp\left(-\theta R^2\right)\int_W \exp\left(\theta\|x\|_W^2\right)\nu(\mathrm{d}x) \\ &\leq \exp(\kappa)\exp\left(-\theta R^2\right)\exp\left(\frac{1}{2}c_w\theta\right) \\ &\leq \exp(\kappa)\exp\left(-\theta R^2\right)\exp\left(\rho\right), \end{aligned}$$

and

$$\begin{aligned} &\nu\left(\left\{\|x\|_W > \frac{R}{2}\right\}\right) \\ &\leq \exp\left(-\frac{1}{2c_w}\frac{R^2}{4}\right)\int_W \exp\left(\frac{1}{2c_w}\|x\|_W^2\right)\nu(\mathrm{d}x) \\ &\leq \exp\left(-\frac{1}{8c_w}R^2 + \frac{1}{4}\right). \end{aligned}$$

This leads to the final bound of the escape probability C_n:

$$\begin{aligned} C_n(U_R, \nu) &\leq n\exp(\kappa)\exp\left(-\theta R^2\right)\exp\left(\rho\right) + \exp\left(-\frac{1}{8c_w}R^2 + \frac{1}{4}\right) \\ &\leq 2n\exp(\tilde{\kappa})\exp\left(-\theta R^2\right) \end{aligned}$$

for $\theta < \min\left\{\frac{1}{8c_w}, \frac{2\rho}{c_w}, 2\rho\right\}$ and $\tilde{\kappa} := \max\left\{\frac{1}{4}, \kappa\right\}$.
The escape probability starting in μ can be bounded by

$$C_n(U_R, \mu) \leq \int_W n\exp(\kappa)\exp\left(\theta\left(\|x\|_W^2 - R^2\right)\right)\mu(\mathrm{d}x) + \mu\left(\left\{\|x\|_W > \frac{R}{2}\right\}\right)$$

Again, we consider both summands separately. For the first one, it holds

$$\begin{aligned}
&\int_W \exp(\kappa)\exp\left(\theta\left(\|x\|_W^2 - R^2\right)\right)\mu(\mathrm{d}x)\\
&\leq \frac{1}{Z}\exp(\kappa)\exp\left(-\theta R^2\right)\int_W \exp\left(\theta\|x\|_W^2 - V(x)\right)\nu(\mathrm{d}x)\\
&\leq \frac{1}{Z}\exp(\kappa)\exp\left(-\theta R^2\right)\left(\int_W \exp\left(2\theta\|x\|_W^2\right)\nu(\mathrm{d}x)\int_W \exp\left(-2V(x)\right)\nu(\mathrm{d}x)\right)^{\frac{1}{2}}\\
&\leq \Psi\exp(\kappa)\exp\left(-\theta R^2\right)\exp\left(\rho\right)
\end{aligned}$$

where Ψ is given by

$$\Psi := \frac{\left(\int_W \exp(-2V(x))\nu(\mathrm{d}x)\right)^{\frac{1}{2}}}{\int_W \exp(-V(x))\nu(\mathrm{d}x)}.$$

For the second summand, we have

$$\mu\left(\left\{\|x\|_W > \frac{R}{2}\right\}\right) \leq \Psi\nu\left(\left\{\|x\|_W > \frac{R}{2}\right\}\right) \leq \Psi\exp\left(-\frac{1}{8c_w}R^2 + \frac{1}{4}\right)$$

by Jensen's inequality. Adding both terms results in

$$C_n(U_R,\mu) \leq 2n\Psi\exp(\tilde{\kappa})\exp\left(-\theta R^2\right),$$

analogously to $C_n(U_R,\nu)$.

In particular, C_n is bounded independently of h. Set

$$\begin{aligned}
C_R &:= C_n(U_R,\mu) + C_n(U_R,\nu) \qquad \text{and}\\
c_1 &:= (1+\Psi)\exp(\kappa)\exp(\rho)\\
&= C_R\exp(\theta R^2).
\end{aligned}$$

We now specify $R := R(n) := n^{\frac{1}{2(1+r)}}$ and $h := \frac{1}{16}L^2(1+R^r)^{-2}$ and choose n_0 such that $R(n_0) > R_0$. Thus, we achieve for $n \geq n_0$

$$\begin{aligned}
\frac{2R(n)}{h\left(L + 2h^{\frac{1}{2}}c(1+R(n)^r)\right)} &= \frac{4R(n)}{hL}\\
&= \frac{64}{L^3}c^2R(1+R(n)^r)^2\\
&= \frac{64}{L^3}c^2 n^{\frac{1}{2(1+r)}}\left(1 + n^{\frac{r}{2(1+r)}}\right)^2
\end{aligned}$$

and

$$\begin{aligned}
C_R &\leq c_1 n\exp\left(-\theta R(n)^2\right)\\
&\leq c_1 n\exp\left(-\theta n^{\frac{1}{1+r}}\right).
\end{aligned}$$

Furthermore, it follows that

$$\begin{aligned}\left(1-\frac{h}{2}L+h^{\frac{3}{2}}c(1+R(n)^r)\right)^n &\leq \left(1-\frac{1}{4}Lh\right)^n \\ &\leq \exp\left(-\frac{1}{4}Lhn\right) \\ &\leq \exp\left(-\frac{1}{64}L^3c^{-2}R(n)^{-2r}n\right) \\ &\leq \exp\left(-bn^{1-\frac{2r}{2(1+r)}}\right) \\ &= \exp\left(-bn^{\frac{1}{1+r}}\right)\end{aligned}$$

where $b := \frac{1}{64}L^3c^{-2}$ such that we get

$$\begin{aligned}\mathcal{W}_{d_R}(\nu q^n_{h(n)},\mu) \leq \exp\left(-bn^{\frac{1}{1+r}}\right)\mathcal{W}_{d_R}(\nu,\mu) \\ + \frac{64}{L^3}c^2n^{\frac{1}{2(1+r)}}\left(1+n^{\frac{r}{2(1+r)}}\right)^2 nc_1\exp\left(-\theta n^{\frac{1}{1+r}}\right).\end{aligned}$$

Setting $a := \frac{1}{2}\min\{b,\theta\}$,

$$C := \frac{64}{L^3}c^2c_1\sup_{n\in\mathbb{N}}\left\{n^{1+\frac{1}{2(1+r)}}\left(1+n^{\frac{r}{2(1+r)}}\right)^2\exp\left(-\frac{1}{4}\theta n^{\frac{1}{1+r}}\right)\right\}$$

and using $\mathcal{W}_{d_1}(\nu,\mu) \leq \mathcal{W}_{d_R}(\nu,\mu) \leq W_{d_\infty}(\nu,\mu)$ results in the depiction of the bound for the Wasserstein distance as stated in the theorem:

$$\mathcal{W}_{d_1}(\nu q^n_{h(n)},\mu) \leq \exp\left(-bn^{\frac{1}{1+r}}\right)(\mathcal{W}_{d_\infty}(\nu,\mu)+C).$$

□

3.4 Speed of convergence in Transition Path Sampling

We now use the results from Chapter 3.3 to bound the speed of convergence of the MALA–process in the Transition Path Sampling setting as constructed in Section 3.2. We start with the infinite–dimensional setting and show that the conditions can be satisfied there. Later, we discuss finite–dimensional approximations that could be used for simulations on a computer. In this setting, we can derive uniform estimates that do not depend on the discretization level.

We remind on the choices of the spaces W and S from Section 3.2. In principle, other choices on these metrics are possible. We will discuss this briefly in Section 3.5.

We now discuss conditions under which we can use the results of the previous chapters in the Transition Path Sampling setting. First, we apply Theorem 3.3 to the infinite–dimensional process to bound its speed of convergence. After that, we analyze the sequence of finite–dimensional approximations constructed in Section 3.2.2 and use Theorem 3.3 to obtain a uniform bound on its speed of convergence independent of the dimension of the approximation. In both cases, Assumptions 3.2 and 3.1 are sufficient.

3.4.1 Application to the infinte–dimensional case

When we choose the weak metric $\|\cdot\|_W$ to be the W_α–norm, the estimation of the derivatives on path space is possible by imposing bounds on the derivatives of the finite–dimensional potential.
We show that Assumptions 3.2 and 3.1 imply Assumption 3.3 and 3.4. This allows us to apply Theorem 3.3.

Lemma 3.33. *Let Assumption 3.2 be satisfied. Then for $\frac{1}{3} \leq \alpha < \frac{1}{2}$, Assumption 3.3 is satisfied, i.e. there exists $L < 1$ such that*

$$\|\nabla_S V(x) - \nabla_S V(\tilde{x})\|_{W_\alpha} \leq L\|x - \tilde{x}\|_{W_\alpha}.$$

Proof. For bounded $\mathrm{D}^2\Phi$ and $\eta^1, \eta^2 \in L^2([0,1], \mathbb{R}^d)$, the second derivative of $V(x) = \int_0^1 \Phi(x_s)\mathrm{d}s$ is given by

$$\mathrm{D}^2 V(x)(\eta^1, \eta^2) = \int_0^1 \mathrm{D}^2\Phi(x_s)(\eta_1(s), \eta_2(s))\mathrm{d}s.$$

The difference of the gradients is then bounded by

$$\begin{aligned}
\|\nabla_S V(x) - \nabla_S V(\tilde{x})\|_{W_\alpha} &\leq \|\nabla_S V(x) - \nabla_S V(\tilde{x})\|_S \\
&\leq \sup_{\|\xi\|_S = 1} \sup_{z \in [x, \tilde{x}]} \mathrm{D}^2 V(z)(\xi, x - \tilde{x}) \\
&\leq L_\Phi \sup_{\|\xi\|_S = 1} \|\xi\|_{L^2} \|x - \tilde{x}\|_{L^2} \\
&\leq \frac{C_{\alpha,2}}{\pi} L_\Phi \|x - \tilde{x}\|_{W_\alpha} \\
&\leq L\|x - \tilde{x}\|_{W_\alpha}
\end{aligned}$$

where the constants L_Φ and $C_{\alpha,2}$ were introduced in Assumption 1.5 and equation (3.13), and

$$L := \frac{\sqrt{2}}{\pi} L_\Phi < 1.$$

To show this result, we need Lemma 3.6:

$$\begin{aligned} \|x\|_{L^2} &\leq C_{\alpha,2} \|x\|_{W_\alpha} && \text{for } x \in W_\alpha \text{ and} \\ \|x\|_{L^2} &\leq \frac{1}{\pi} \|x\|_{H_0^1([0,1],\mathbb{R}^d)} && \text{for } x \in S. \end{aligned}$$

Furthermore, for $\alpha \geq \frac{1}{3}$, $C_{\alpha,2}$ can be bounded by

$$\begin{aligned} C_{\alpha,2} &\leq \left(\sum_{i=1}^{\infty} 2^{i\left(1-\frac{2}{3}-1\right)} \right)^{\frac{1}{2}} \\ &= \left(\sum_{i=1}^{\infty} 2^{-\frac{2}{3}i} \right)^{\frac{1}{2}} \\ &= \left(\frac{1}{4^{\frac{1}{3}} - 1} \right)^{\frac{1}{2}} \\ &< \sqrt{2}. \end{aligned} \tag{3.40}$$

□

Lemma 3.34. *Let Assumption 3.1 be satisfied. Then there exists $\frac{1}{3} \leq \alpha \leq \frac{1}{2}$ such that Assumption 3.4 is satisfied, i.e.*

$$\left| \mathrm{D}^n V(x)(\xi^1, \ldots, \xi^n) \right| \leq C_n \left(\max\{1, \|x\|_{W_\alpha}\} \right)^{p_n}$$

for all $\xi^1, \ldots, \xi^n$ with $\left\|\xi^i\right\|_{W_\alpha} = 1$, $i \in \{1, \ldots, n\}$.

Proof. The n-th derivative of V is given by

$$\mathrm{D}^n V(x)(\xi^1, \ldots, \xi^n) = \int_0^1 \mathrm{D}^n \Phi(x_s)(\xi_s^1, \ldots, \xi_s^n) \mathrm{d}s.$$

Assumption 3.1 implies

$$\mathrm{D}^n \Phi(x_s)(\xi_s^1, \ldots, \xi_s^n) \leq C_n \max\left\{1, \|x_s\|_{\mathbb{R}^d}\right\}^{p_n},$$

so we can bound

$$\begin{aligned} \mathrm{D}^n V(x)(\xi^1, \ldots, \xi^n) &\leq C_n \int_0^1 \max\left\{1, \|x_s\|_{\mathbb{R}^d}\right\}^{p_n} \left\|\xi_s^1\right\|_{\mathbb{R}^d} \cdot \ldots \cdot \|\xi_s^n\|_{\mathbb{R}^d} \, \mathrm{d}s \\ &\leq C_n \left(\left(1 + \int_0^1 \|x_s\|_{\mathbb{R}^d}^{(1+n)p_n} \, \mathrm{d}s \right) \int_0^1 \left\|\xi_s^1\right\|_{\mathbb{R}^d}^{1+n} \mathrm{d}s \cdot \ldots \cdot \int_0^1 \|\xi_s^n\|_{\mathbb{R}^d}^{1+n} \right)^{\frac{1}{1+n}} \\ &\leq C_n \left(1 + \|x\|_{L^{p_n(1+n)}}^{p_n} \right) \cdot \left\|\xi^1\right\|_{L^{1+n}} \cdot \ldots \cdot \|\xi^n\|_{L^{1+n}}. \end{aligned}$$

Since $n \in \{1,\dots,4\}$, we can choose $q := \max_{n\in\{0,\dots,4\}} \{p_n(1+n), 6\}$ and by applying Lemma 3.6 conclude for some constant $c_1 < \infty$

$$\begin{aligned}
\mathrm{D}^n V(x)(\xi^1,\dots,\xi^n) &\le c_1 \left(1 + \|x\|_{L^q}^{p_n}\right) \cdot \left\|\xi^1\right\|_{L^q} \cdot \ldots \cdot \|\xi^n\|_{L^q} \\
&\le c_2 \left(1 + \|x\|_{W_\alpha}^{p_n}\right) \cdot \left\|\xi^1\right\|_{W_\alpha} \cdot \ldots \cdot \|\xi^n\|_{W_\alpha} \\
&\le 2c_2 \max\left\{1, \|x\|_{W_\alpha}^{p_n}\right\} \cdot \left\|\xi^1\right\|_{W_\alpha} \cdot \ldots \cdot \|\xi^n\|_{W_\alpha}
\end{aligned}$$

for $\alpha > \frac{1}{2} - \frac{1}{q} > \frac{1}{3}$. □

After showing the validity of Assumptions 3.2 and 3.3, we now explicitly calculate the constant c_w from the bounds on the exponential moments of the Gaussian measure ν in Lemma 3.31:

Lemma 3.35. *c_w satisfies*

$$c_w = \frac{d}{1 - 2^{2\alpha-1}}.$$

Proof. By definition, $c_w = 4 \operatorname{trace} \mathcal{C}$, where $\mathcal{C}$ is the covariance operator of the Gaussian measure ν. Furthermore, let $\tilde{e}_{i,k,j}$ be the orthonormal basis constructed in (3.12). Then, as $\alpha < \frac{1}{2}$,

$$\begin{aligned}
\operatorname{trace} \mathcal{C} &= \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} \langle \tilde{e}_{i,k,j}, \mathcal{C}\tilde{e}_{i,k,j} \rangle_W \\
&= \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{k=1}^{2^{i-1}} 2^{-2(1-\alpha)i} \\
&= d \sum_{i=1}^{\infty} 2^i \, 2^{-2(1-\alpha)i} \\
&= d \sum_{i=1}^{\infty} 2^{(2\alpha-1)i} \\
&= \frac{d}{1 - 2^{2\alpha-1}}.
\end{aligned}$$

□

We are now in a position to prove that the distance of the MALA–process to its equilibrium measure in Transition Path Sampling decreases with rate $\exp\left(-cn^{\frac{1}{1+r}}\right)$, which allows to specify the number of needed steps of this proces to achieve a given error.

Theorem 3.4. *Let q_h be the kernel of the process $(X_n)_{n\in\mathbb{N}}$ with step–size h as constructed in Section 3.1. Let d_R be the Wasserstein distance with respect to the distance $(x,y) \mapsto \|x-y\|_{W_\alpha} \wedge R$. Let Assumptions 3.2 and 3.1 be satisfied. Then there exists $C > 0$ such that for given $n \in \mathbb{N}$, there exist $h(N) > 0$ such that*

$$\mathcal{W}_{d_1}(\nu q^n_{h(n)}, \mu) \leq \exp\left(-c_1 n^{\frac{1}{1+r}}\right)(\mathcal{W}_{d_\infty}(\mu,\nu) + C),$$

where $c_1 := \frac{1}{64}L^3c^{-2}$, r is given by (3.33) and c is constructed in equation (3.32).

Proof. Lemmas 3.33 and 3.34 show that Assumptions 3.1 and 3.2 imply Assumptions 3.3 and 3.4, so that we can apply Theorem 3.3. As in the proof of Theorem 3.3, we choose $R := n^{\frac{1}{2(1+r)}}$ and $h(n) := \frac{1}{16}L^2c^{-2}(1+R^r)^{-2}$. Then, the result follows immediately from Theorem 3.3. □

3.4.2 Application to the finite–dimensional approximations

We now analyze the implications of Assumptions 3.1 and 3.2 for a finite–dimensional approximation in the Transition Path Sampling setting. We show for the approximations constructed in Section 3.2.2 that these conditions imply uniform bounds in Assumptions 3.3 and 3.4 and thus a uniform bound for the speed of convergence of the MALA–process in Theorem 3.3.

We recall the notation

$$d_N := 2^{N-1} \qquad \text{and}$$
$$s_i := \frac{i}{d_N}$$

from Section 3.2.2.

Lemma 3.36. *Let Assumption 3.2 be satisfied. Assume furthermore that $L_\Phi < \frac{\pi}{3}$. Then for $N \in \mathbb{N}$, V_N satisfies Assumption 3.3, i.e. there exists $L < 1$ such that*

$$\|\nabla_S V_N(x) - \nabla_S V_N(\tilde{x})\|_{W_N} \leq L\|x-\tilde{x}\|_{W_N}.$$

α and L are independent of N.

Proof. For bounded $\mathrm{D}^2\Phi$ and $\eta^1, \eta^2 \in E$, the second derivative of $V_N(x) := \frac{1}{d_N}\sum_{i=1}^{d_N}\Phi(x_{s_i})$ is given by

$$\mathrm{D}^2 V_N(x)(\eta^1,\eta^2) = \frac{1}{d_N}\sum_{i=1}^{d_N}\mathrm{D}^2\Phi\left(x_{s_i}\right)\left(\eta^1_{s_i},\eta^2_{s_i}\right).$$

The difference of the gradients is now bounded by

$$\begin{aligned}
\|\nabla_S V_N(x) - \nabla_S V_N(\tilde{x})\|_{W_N} &\leq \|\nabla_S V_N(x) - \nabla_S V_N(\tilde{x})\|_{S_N} \\
&\leq \sup_{\|\xi\|_{S_N}=1}\sup_{z\in[x,\tilde{x}]}\mathrm{D}^2 V_N(z)(\xi, x-\tilde{x}) \\
&\leq L_\Phi \sup_{\|\xi\|_{S_N}=1}\frac{1}{d_N}\sum_{i=1}^{d_N}\xi_{s_i}\cdot(x_{s_i}-\tilde{x}_{s_i}) \\
&\leq L_\Phi\|\xi\|_{E_N}\|x-\tilde{x}\|_{E_N} \\
&\leq L\|x-\tilde{x}\|_{W_N},
\end{aligned}$$

where

$$L := \frac{3}{\pi}L_\Phi < 1.$$

To show this result, we used Lemmas 3.6 and 3.11 as well as inequality (3.40): For $x \in E_N$,

$$\begin{aligned}
\|x\|_{E_N} &\leq \sqrt{\frac{3}{2}}\|x\|_E \leq \frac{\sqrt{3}}{\sqrt{2}\pi}\|x\|_{S_N} \\
\|x\|_{E_N} &\leq \sqrt{\frac{3}{2}}\|x\|_E \leq \sqrt{\frac{3}{2}}C_{\alpha,2}\|x\|_{W_N} \leq \sqrt{6}\|x\|_{W_N}.
\end{aligned}$$

□

Lemma 3.37. *Let Assumption 3.1 be satisfied. Then there exists $\alpha < \frac{1}{2}$ such that Assumption 3.4 is satisfied, i.e.*

$$\left|\mathrm{D}^n V_N(x)(\xi^1,\ldots,\xi^n)\right| \leq C_n\left(\max\{1,\|x\|_{W_N}\}\right)^{p_n}$$

for all $\xi^1,\ldots,\xi^n$ with $\left\|\xi^i\right\|_{W_\alpha} = 1$, $i \in \{1,\ldots,n\}$.

Proof. The n-th derivative of V_N is given by

$$\mathrm{D}^n V_N(x)(\xi^1,\ldots,\xi^n) = \frac{1}{d_N}\sum_{i=1}^{d_N}\mathrm{D}^n\Phi(x_{s_i})(\xi^1_{s_i},\ldots,\xi^n_{s_i}).$$

Assumption 3.1 implies

$$\mathrm{D}^n\Phi(x_{s_i})(\xi^1_{s_i},\dots,\xi^n_{s_i}) \le C_n \max\left\{1, \|x_{s_i}\|_{\mathbb{R}^d}\right\}^{p_n} \left\|\xi^1_{s_i}\right\|_{\mathbb{R}^d} \cdot \ldots \cdot \left\|\xi^n_{s_i}\right\|_{\mathbb{R}^d},$$

so by applying Hölder's inequality, this leads to

$$\begin{aligned}
&\mathrm{D}^n V(x)(\xi^1,\dots,\xi^n)\\
&\le C_n \frac{1}{d_N}\sum_{i=1}^{d_N} \max\left\{1, \|x_{s_i}\|_{\mathbb{R}^d}\right\}^{p_n} \left\|\xi^1_{s_i}\right\|_{\mathbb{R}^d} \cdot \ldots \cdot \left\|\xi^n_{s_i}\right\|_{\mathbb{R}^d}\\
&\le C_n \left(\left(1+\frac{1}{d_N}\sum_{i=1}^{d_N}\|x_{s_i}\|_{\mathbb{R}^d}^{(1+n)p_n}\right)\left(\frac{1}{d_N}\sum_{i=1}^{d_N}\left\|\xi^1_{s_i}\right\|_{\mathbb{R}^d}^{1+n}\right)\cdot\ldots\cdot\left(\frac{1}{d_N}\sum_{i=1}^{d_N}\left\|\xi^n_{s_i}\right\|_{\mathbb{R}^d}^{1+n}\right)\right)^{\frac{1}{1+n}}.
\end{aligned}$$

As for $m \in \mathbb{N}$ and $\xi \in E_N$,

$$\|\xi\|^m_{L^m([0,1],\mathbb{R}^d)} \ge \frac{2}{m+1}\frac{1}{d_N}\sum_{i=1}^{d_N}\|\xi_{s_i}\|^m_{\mathbb{R}^d},$$

we get

$$\begin{aligned}
\frac{1}{d_N}\sum_{i=1}^{d_N}\|\xi_{s_i}\|^m_{\mathbb{R}^d} &\le \frac{m+1}{2}\|\xi\|^m_{L^m([0,1],\mathbb{R}^d)}\\
&\le \frac{m+1}{2}C^m_{\alpha,m}\|\xi\|^m_{W_N},
\end{aligned}$$

where $C_{\alpha,m}$ was definied in (3.13). Set

$$D_{\alpha,m} := \left(\frac{m+1}{2}\right)^{\frac{1}{m}} C_{\alpha,m},$$

then

$$\begin{aligned}
&\mathrm{D}^n V(x)(\xi^1,\dots,\xi^n)\\
&\le C_n \left(\left(1+\frac{1}{d_N}\sum_{i=1}^{d_N}\|x_{s_i}\|_{\mathbb{R}^d}^{(1+n)p_n}\right)\left(\frac{1}{d_N}\sum_{i=1}^{d_N}\left\|\xi^1_{s_i}\right\|_{\mathbb{R}^d}^{1+n}\right)\cdot\ldots\cdot\left(\frac{1}{d_N}\sum_{i=1}^{d_N}\left\|\xi^n_{s_i}\right\|_{\mathbb{R}^d}^{1+n}\right)\right)^{\frac{1}{1+n}}\\
&\le C_n D_{\alpha,(1+n)p_n} D^n_{\alpha,(1+n)}\left(1+\|x\|^{p_n}_{W_N}\right)\left\|\xi^1\right\|_{W_N}\cdot\ldots\cdot\|\xi^n\|_{W_N}\\
&\le 2C_n D_{\alpha,(1+n)p_n} D^n_{\alpha,(1+n)} \max\left\{1,\|x\|^{p_n}_{W_N}\right\}\left\|\xi^1\right\|_{W_N}\cdot\ldots\cdot\|\xi^n\|_{W_N}.
\end{aligned}$$

□

We can now prove a result on the distance of the MALA–process in Transition Path Sampling to its equilibrium measure:

Theorem 3.5. *Let $q_{N,h}$ be the kernel of the process $(X_n^N)_{n\in\mathbb{N}}$ with step–size h as constructed in Section 3.1. Let d_R be the Wasserstein distance with respect to the distance $(x,y) \mapsto \|x-y\|_{W_\alpha} \wedge R$. Let Assumptions 3.2 and 3.1 be satisfied. Additionally assume $L_\Phi \leq \frac{\pi}{3}$. Then for given $n \in \mathbb{N}$, there exists $h(n) > 0$ such that*

$$\mathcal{W}_{d_1}(\nu q_{N,h(n)}^n, \mu) \leq \exp\left(-c_1 n^{\frac{1}{1+r}}\right)\left(\mathcal{W}_{d_\infty}(\mu,\nu)+1\right)$$

where $c_1 := \frac{1}{64}L^3c^{-2}$, r is given by (3.33) and c by (3.32).

Proof. Lemma 3.36 and 3.37 show that Assumptions 3.1 and 3.2 with $L_\Phi \leq \frac{\pi}{3}$ imply Assumptions 3.3 and 3.4, so that we can apply Theorem 3.3. As in the proof of Theorem 3.3, we choose $R := n^{\frac{1}{2(1+r)}}$ and $h(n) := \frac{1}{16}L^2c^{-2}(1+R^r)^{-2}$. Then, and the result follows immediately from Theorem 3.3. □

3.5 Further choices for Markov Chain Monte Carlo processes

Given a probability measure

$$\mu(\mathrm{d}x) = \frac{1}{Z}\exp(-V(x))\nu(\mathrm{d}x),$$

different processes that are reversible with respect to μ can be constructed. For each positive self–adjoint linear operator Q, the solution of the stochastic differential equation

$$\mathrm{d}x_t = -Qx_t\mathrm{d}t - Q\nabla_S V(x_t)\mathrm{d}t + \sqrt{2Q^{-1}}\mathrm{d}w_t \tag{3.41}$$

driven by a S-Wiener–process w is reversible with respect to μ. The case $Q = \mathrm{I}$ corresponds to the H^1–case considered above for Transition Path Sampling, $Q = (-\Delta_0)$ corresponds to the "non–preconditioned case" of [24] and [8]. The H^1–case is called "pre–conditioned case" in these works. We additionally consider $Q = (-\Delta_0)^\alpha$ for $\alpha \in [0,1]$ to analyze the effect of different noises on the contraction property. It turns out that the case $Q = \mathrm{I}$ is the only one where the proposal of the MALA–process is contracting, and thus the only one where our analysis can be applied.

In the next section, we consider discrete–time processes with reversible measure μ on the space $E := L^2([0,1],\mathbb{R})$, driven by varying noise, and analyze the contraction

property of these processes. Although it is nevertheless necessary to work with discretizations for numerical simulations, the contraction property on the infinite–dimensional space strongly indicates that is possible to find a sequence of discretizations of the process which possess a uniform contraction constant.

Contraction properties of discrete–time processes

First, we consider the Gaussian case $V = 0$. Formally, discrete–time schemes for the s.d.e. (3.41) with $V = 0$ are given for $\theta \in [0, 1]$ by

$$X_{n+1} = X_n - \theta h Q X_n - (1-\theta) h Q X_{n+1} + \sqrt{2hQ^{-1}(-\Delta_0)^{-1}} N_n$$

where $(N_n)_{n\in\mathbb{N}}$ are i.i.d. Gaussian random variables with covariance induced by $\|\cdot\|_W$. A rigorous implementation of the scheme is given by

$$X_{n+1} = (Q + (1-\theta)h\mathrm{I})^{-1}(Q - \theta h\mathrm{I})X_n + \sqrt{2h}(Q + (1-\theta)h\mathrm{I})^{-1} Q^{\frac{1}{2}}(-\Delta_0)^{-\frac{1}{2}} N_n. \quad (3.42)$$

A semi–implicit discretization with $\theta = \frac{1}{2}$ is the only one which is reversible with respect to ν, cf. the analysis in [8] for the case $\alpha \in \{0, 1\}$. It corresponds to the process studied above. We now show this statement for general α.

Proposition 3.38. *Let q^θ be the kernel induced by (3.42). Then q^θ is reversible with respect to ν if and only if $\theta = \frac{1}{2}$.*

Proof. We consider the characteristic function of the measure $\nu\, q^\theta$:

$$\begin{aligned}
&\int_{E\times E} \exp\left(-i\langle (l_1, l_2), (x, y)\rangle_S\right) \nu(\mathrm{d}x) q^\theta(x, \mathrm{d}y) \\
&= \int_{E\times E} \exp\left(-i\langle (l_1, l_2), (x, Ax + y)\rangle_S\right) \nu(\mathrm{d}x) q^\theta(0, \mathrm{d}y) \\
&= \exp\left(-\frac{1}{2}\|l_1 + Al_2\|_S - h\left\|(Q + (1-\theta)h\mathrm{I})^{-1} Q^{\frac{1}{2}} l_2\right\|_S\right)
\end{aligned}$$

where

$$A := (Q + (1-\theta)h\mathrm{I})^{-1}(Q - \theta h\mathrm{I}).$$

The characteristic function of νq^θ is symmetric in l_1, l_2 if and only if q^θ is reversible with

respect to ν. The exponent can be written as

$$\begin{aligned}&\|l_1 + Al_2\|_S + 2h\left\|(Q + (1-\theta)h\mathrm{I})^{-1}Q^{\frac{1}{2}}l_2\right\|_S \\ &= \|l_1\|_S + 2\langle l_1, Al_2\rangle_S + \|Al_2\|_S + 2\left\|(Q + (1-\theta)h\mathrm{I})^{-1}Q^{\frac{1}{2}}l_2\right\|_S\end{aligned}$$

As A is self–adjoint on S, $\langle l_1, Al_2\rangle_S$ is symmetric. So the characteristic function is symmetric if and only if

$$\|Al_2\|_S + 2h\left\|(Q + (1-\theta)h\mathrm{I})^{-1}Q^{\frac{1}{2}}l_2\right\|_S = \|l_2\|_S.$$

As powers of the Laplacian, all operators commute, and we get

$$\begin{aligned}&\|Al_2\|_S + 2h\left\|(Q + (1-\theta)h\mathrm{I})^{-1}Q^{\frac{1}{2}}l_2\right\|_S \\ &= \left\langle l_2, (Q + (1-\theta)h\mathrm{I})^{-2}\left((Q - \theta h\mathrm{I})^2 + 2hQ\right)l_2\right\rangle_S.\end{aligned}$$

Moreover, we can rewrite

$$\begin{aligned}&(Q + (1-\theta)h\mathrm{I})^{-2}(Q - \theta h\mathrm{I})^2 + 2hQ \\ &= (Q + (1-\theta)h\mathrm{I})^{-2}\left((Q + (1-\theta)h\mathrm{I})^2 - 2hQ - (1-\theta)^2h^2\mathrm{I} + \theta^2h^2\mathrm{I} + 2hQ\right) \\ &= \mathrm{I} + (Q + (1-\theta)h\mathrm{I})^{-2}\left(h^2\left(-(1-\theta)^2 + \theta^2\right)\mathrm{I}\right).\end{aligned}$$

We have symmetry if and only if the second summand vanishes, this is the case for $\theta = \frac{1}{2}$. □

Proposition 3.38 states that ν is the reversible measure of the process $(X_n)_{n\in\mathbb{N}}$ defined by

$$X_{n+1} = \left(Q + \frac{1}{2}h\mathrm{I}\right)^{-1}\left(Q - \frac{1}{2}h\mathrm{I}\right)X_n + \sqrt{2h}\left(Q + \frac{1}{2}h\mathrm{I}\right)^{-1}Q^{\frac{1}{2}}(-\Delta_0)^{-\frac{1}{2}}\xi_n. \tag{3.43}$$

It even follows that the distribution of the proposals $\nu(\mathrm{d}x)q^\theta(x,\mathrm{d}y)$ is not absolutely continuous to $\nu(\mathrm{d}y)q^\theta(y,\mathrm{d}x)$ for $\theta \neq \frac{1}{2}$, as two Gaussian measures with the same mean and covariance operator Q_1, Q_2 respectively are absolutely continuous only if the operator $(Q_1^{-\frac{1}{2}}Q_2^{-\frac{1}{2}})(Q_1^{-\frac{1}{2}}Q_2^{-\frac{1}{2}})^* - \mathrm{I}$, is a Hilbert–Schmidt operator, see e.g. [12, Theroem 2.23]. Thus $\theta = \frac{1}{2}$ is the only possible choice four MALA–process, because its acceptance probability is defined as the relative density of $\nu(\mathrm{d}x)q^\theta(x,\mathrm{d}y)$ and $\nu(\mathrm{d}y)q^\theta(y,\mathrm{d}x)$.

We now analyze the contraction properties of a coupling of two processes $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ of (3.43) starting in different positions $X_0 = x_0$ and $Y_0 = y_0$. Analogously to

the construction in Section 3.3.1, the processes are driven by the same W–Gaussian noise $(N_n)_{n\in\mathbb{N}}$. They are given by

$$X_{n+1} := \left(Q+\frac{1}{2}h\mathrm{I}\right)^{-1}\left(Q-\frac{1}{2}h\mathrm{I}\right)X_n + \sqrt{2h}\left(Q+\frac{1}{2}h\mathrm{I}\right)^{-1}Q^{\frac{1}{2}}(-\Delta_0)^{-\frac{1}{2}}N_n, \tag{3.44}$$
$$Y_{n+1} := \left(Q+\frac{1}{2}h\mathrm{I}\right)^{-1}\left(Q-\frac{1}{2}h\mathrm{I}\right)Y_n + \sqrt{2h}\left(Q+\frac{1}{2}h\mathrm{I}\right)^{-1}Q^{\frac{1}{2}}(-\Delta_0)^{-\frac{1}{2}}N_n.$$

Remark 3.39. *Note that in the case $\alpha = 0$ and $V \equiv 0$, the coupling $(X_n, Y_n)_{n\in\mathbb{N}}$ coincides with the one analyzed in chapter (3.3.1). If we set $h = \frac{2\varepsilon}{8-\varepsilon}$, then (3.44) reads*

$$X_{n+1} := \left(1-\frac{\varepsilon}{2}\right)X_n + \sqrt{\varepsilon-\frac{\varepsilon^2}{4}}(-\Delta_0)^{-\frac{1}{2}}N_n,$$
$$Y_{n+1} := \left(1-\frac{\varepsilon}{2}\right)Y_n + \sqrt{\varepsilon-\frac{\varepsilon^2}{4}}(-\Delta_0)^{-\frac{1}{2}}N_n,$$

and $(-\Delta_0)^{-\frac{1}{2}}N_n$ is a ν–distributed random variable.

The next proposition states the contraction properties of the coupling for different values of α.

Proposition 3.40. *For $\alpha = 0$, the coupling $(X_n, Y_n)_{n\in\mathbb{N}}$ given by (3.44) is contracting in every norm $\|\cdot\|_{H_0^\beta([0,1],\mathbb{R})}$ for $\beta \in \left[0, \frac{1}{2}\right)$:*

$$\|X_1 - Y_1\|_\beta \le \frac{2-h}{2+h}\|x_0 - y_0\|_{H_0^\beta([0,1],\mathbb{R})} \quad \text{for all } x_0, y_0 \in H_0^\beta([0,1],\mathbb{R}).$$

For $\alpha > 0$, the coupling $(X_n, Y_n)_{n\in\mathbb{N}}$ is not contracting in $\|\cdot\|_{H_0^\beta([0,1],\mathbb{R})}$ for each $\beta \in \left[0, \frac{1}{2}\right)$: There exists $\mathrm{x}^\varepsilon, y^\varepsilon \in H_0^\beta([0,1],\mathbb{R})$, such that

$$\|X_1^\varepsilon - Y_1^\varepsilon\|_{H_0^\beta([0,1],\mathbb{R})} \ge (1-\varepsilon)\|x_0^\varepsilon - y_0^\varepsilon\|_{H_0^\beta([0,1],\mathbb{R})}.$$

Proof. Define

$$A = \left(Q+\frac{1}{2}h\mathrm{I}\right)^{-1}\left(Q-\frac{1}{2}h\mathrm{I}\right).$$

Then for each β,

$$\|X_1 - Y_1\|_{H_0^\beta([0,1],\mathbb{R})} = \|A(x_0 - y_0)\|_{H_0^\beta([0,1],\mathbb{R})}.$$

For $\alpha = 0$,

$$A = \frac{2-h}{2+h}\mathrm{I},$$

which clearly satisfies

$$\|A\phi\|_{H_0^\beta([0,1],\mathbb{R})} \leq \frac{2-h}{2+h}\|\phi\|_{H_0^\beta([0,1],\mathbb{R})}$$

for all $\phi \in H_0^\beta([0,1],\mathbb{R})$.

For $\alpha > 0$, let $\phi_i \in H_0^\beta([0,1],\mathbb{R})$ be the ith eigenfunction of $(-\Delta_0)^{-1}$ with respect to the Fourier basis

$$\phi_i(t) = \sin\left(i\pi\frac{t}{T}\right) \in H_0^1([0,1],\mathbb{R}^d) \subset W_\beta \qquad \text{for } \beta \in \left[0, \frac{1}{2}\right).$$

The corresponding eigenvalues are given by

$$Q\phi_i = \frac{1}{i^{2\alpha}}\phi_i,$$

so we see that

$$\begin{aligned} A\phi_i &= \left(Q + \frac{1}{2}h\mathrm{I}\right)^{-1}\left(Q - \frac{1}{2}h\mathrm{I}\right)\phi_i \\ &= \frac{\frac{2}{i^{2\alpha}} - h}{\frac{2}{i^{2\alpha}} + h}\phi_i. \end{aligned}$$

For $i \to \infty$ and $\alpha > 0$

$$\frac{\frac{2}{i^{2\alpha}} - h}{\frac{2}{i^{2\alpha}} + h} \to -1.$$

So for given $\varepsilon > 0$, we can find a ϕ_ε such that $A\phi_\varepsilon = -(1-\varepsilon)\phi_\varepsilon$ which results in

$$\|A\phi_\varepsilon\|_{H_0^\beta([0,1],\mathbb{R})} = (1-\varepsilon)\|\phi_\varepsilon\|_{H_0^\beta([0,1],\mathbb{R})}.$$

Setting $x_0^\varepsilon = 0$, $y_0^\varepsilon = \phi_\varepsilon$ leads to the stated property. □

Propositions 3.38 and 3.40 show that the choice of the proposal of the MALA–process in (3.6) was a natural choice. While there is the possibility of choosing different processes which are still reversible with respect to ν, Proposition 3.40 shows that one can not expect them to have the contraction properties used in the proof of Theorem 3.3.

Bibliography

[1] Cécile Ané, Sébastien Blachère, Djalil Chafai, Pierre Fougères, Ivan Gentil, Florent Malrieu, Cyril Roberto, and Grégory Scheffer. *On logarithmic Sobolev inequalities. With a preface of Dominique Bakry and Michel Ledoux. (Sur les inégalités de Sobolev logarithmiques.)*. Panoramas et Synthèses. 10. Paris: Société Mathématique de France. xiii, 217 p. , 2000.

[2] Mylène Bédard and Jeffrey S. Rosenthal. Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *Canadian Journal of Statistics*, 36(4):483–503, 2008.

[3] A. Beskos, G. Roberts, and A. Stuart. Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions. *The Annals of Applied Probability*, 19(3):863–898, 2009.

[4] A. Beskos and A. Stuart. MCMC methods for sampling function space. *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, European Mathematical Society*, pages 337–364, 2009.

[5] A. Beskos and A.M. Stuart. MCMC methods for sampling function space. In R. Jeltsch and G. Wanner, editors, *ICIAM Invited Lecture 2007*. European Mathematical Society, 2008.

[6] Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O. Roberts. A Factorisation of Diffusion Measure and Finite Sample Path Constructions. *Methodology and Computing in Applied Probability*, 10(1):85–104, 2008.

[7] Alexandros Beskos and Gareth O. Roberts. Exact Simulation of Diffusions. *The Annals of Applied Probability*, 15(4):pp. 2422–2444, 2005.

[8] Alexandros Beskos, Gareth O. Roberts, Andrew M. Stuart, and Jochen Voss. MCMC Methods for Diffusion Bridges. *Stochastics and Dynamics*, 8(3):319–350, 2008.

[9] Nawaf Bou-Rabee, Martin Hairer, and Eric Vanden-Eijnden. Non-asymptotic mixing of the MALA algorithm. Technical Report arXiv:1008.3514, Aug 2010. preprint.

[10] Jakob Creutzig, Steffen Dereich, Thomas Müller-Gronbach, and Klaus Ritter. Infinite-dimensional quadrature and approximation of distributions. *Foundations of Computational Mathematics*, 9(4):391–429, 2009.

[11] Giuseppe Da Prato. *An introduction to infinite-dimensional analysis*. Universitext. Springer-Verlag, Berlin, 2006. Revised and extended from the 2001 original by Da Prato.

[12] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge University Press, 1992.

[13] Giuseppe Da Prato and Jerzy Zabczyk. *Ergodicity for infinite dimensional systems*, volume 229. Cambridge University Press, 1996.

[14] C. Dellago, P.G. Bolhuis, F.S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *The Journal of chemical Physics*, 108:1964–1977, 1998.

[15] Christoph Dellago and Peter G. Bolhuis. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. In Christian Holm and Kurt Kremer, editors, *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, volume 221 of *Advances in Polymer Science*, pages 167–233. Springer Berlin Heidelberg, 2009.

[16] Andreas Eberle. Metropolis-Hastings algorithms for perturbations of Gaussian measures in high dimensions: Contraction properties and error bounds in the logconcave case. *Annals of Applied Probability*. to appear.

[17] Charles J. Geyer. Markov Chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

[18] Michael B. Giles. Multilevel Monte–Carlo Path Simulation. *Operations Research*, 56(3):607–617, 2008.

[19] Mike Giles. Improved Multilevel Monte Carlo Convergence using the Milstein scheme. In Alexander Keller, Stefan Heinrich, and Harald Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2006*, pages 343–358. Springer Berlin Heidelberg, 2008.

[20] Jonathan Goodman and Alan D. Sokal. Multigrid Monte Carlo method. conceptual foundations. *Phys. Rev. D*, 40(6):2035–2071, Sep 1989.

[21] M. Hairer. Singular perturbations to semilinear stochastic heat equations. *Probability Theory and Related Fields*, pages 1–33, 2011.

[22] M. Hairer, A.M. Stuart, and J. Voss. Sampling conditioned hypoelliptic diffusions. *The Annals of Applied Probability*, 21(2):669–698, 2011.

[23] M. Hairer, A.M. Stuart, J. Voss, and P. Wiberg. Analysis of SPDEs arising in path sampling. part I: The Gaussian case. *Communications in Mathematical Sciences*, 3(4):587–603, 2005.

[24] Martin Hairer, Andrew M. Stuart, and Jochen Voss. Analysis of SPDEs arising in Path Sampling, Part II: The Nonlinear case. *Annals of Applied Probability*, 17(5):1657–1706, 2007.

[25] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1969.

[26] S. Heinrich. Monte Carlo Complexity of Global Solution of Integral Equations. *Journal of Complexity*, 14(2):151 – 175, 1998.

[27] Stefan Heinrich. Multilevel monte carlo methods. In Svetozar Margenov, Jerzy Waśniewski, and Plamen Yalamov, editors, *Large-Scale Scientific Computing*, volume

2179 of *Lecture Notes in Computer Science*, pages 58–67. Springer Berlin Heidelberg, 2001.

[28] Stefan Heinrich and Eugène Sindambiwe. Monte Carlo Complexity of Parametric Integration. *Journal of Complexity*, 15(3):317 – 341, 1999.

[29] V.H. Hoang, C. Schwab, and A.M. Stuart. Sparse MCMC gpc finite element methods for Bayesian inverse problems. Submitted.

[30] M. Hutzenthaler, A. Jentzen, and P.E. Kloeden. Divergence of the multilevel Monte Carlo method. *Arxiv preprint arXiv:1105.0226*, 2011.

[31] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus (Graduate Texts in Mathematics)*. Springer, 1 edition, 11 1987.

[32] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[33] Enzo Marinari and Giorgio Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19(6):451–458, 1992.

[34] J.C. Mattingly, N.S. Pillai, and A.M. Stuart. Diffusion limits of the random walk Metropolis algorithm in high dimensions. *The Annals of Applied Probability*, 22(3):881–930, 2012.

[35] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[36] T. Müller-Gronbach and K. Ritter. Variable subspace sampling and multi-level algorithms. In P. L'Ecuyer and A. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 131–156. Springer-Verlag, 2009.

[37] Gareth O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120, 1997.

[38] Gareth O. Roberts and Jeff S. Rosenthal. General state space Markov Chains and MCMC algorithms. *Probability Surveys*, 1, 2004.

[39] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc. B*, 60:255–268, 1997.

[40] G.O. Roberts and R.L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[41] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 1.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Foundations, Reprint of the second (1994) edition.

[42] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. AMS Bookstore, 2003.