Non-asymptotic Error Bounds for Sequential MCMC Methods

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Nikolaus Schweizer
aus
Köln

Bonn 2011

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

1. Gutachter: Prof. Dr. Andreas Eberle
2. Gutachter: Prof. Dr. Patrik Ferrari

Tag der Promotion: 15.02.2012
Erscheinungsjahr: 2012

To my parents

Acknowledgments

When I became the student of Andreas Eberle all I knew about probability theory was that I wanted to learn about Monte Carlo methods. I am deeply indebted to him for giving me that chance and more, and I thank him so much for his generosity and patience and for the trust and effort he put into me. I thank Patrik Ferrari for not even thinking about agreeing to read all this (or thinking very quickly). Another really big thanks goes to Mei-Ling Wang for always having an eye on the big picture of the probability group (and for including me in that picture); and to Daniel Gruhlke for being the office-mate all the kids dream of. There are many more people I have to thank for hospitality, proof-reading, and various other forms of support, mathematical, moral and material, all of which are, of course, indispensable; more than one of those in most cases. A non-exhaustive list includes Sergio Albeverio, Stefan Ankirchner, Giada Basile, Anton Bovier, Flurin Casura, Sergio Conti, René Frings, Markus Hirsch, Nicola Kistler, Carlo Marinelli, Daniel Nagel, Erich Novak, Evangelia Petrou, Daniel Rudolf, Justus Schweizer, Karoline Schweizer, Margarete Schweizer, Martin Slowik, Adela Svejda, Nora Szech, Mario Ullrich and Shi-Dong Wang. On my way here I listened to thousands of hours of music, I have to thank Terry Callier, Sean Levitt, Gene Clark and Emmylou Harris for recording the albums that guided me through the most productive phases.

Abstract

Sequential MCMC methods are a class of stochastic numerical integration methods for target measures μ which cannot feasibly be attacked directly with standard MCMC methods due to the presence of multiple well-separated modes. The basic idea is to approximate the target distribution μ with a sequence of distributions $\mu_0, \ldots, \mu_n$ such that $\mu_n = \mu$ is the actual target distribution and such that μ_0 is easy to sample from. The algorithm constructs a system of N particles which sequentially approximates the measures μ_0 to μ_n. The algorithm is initialized with N independent samples from μ_0 and then alternates two types of steps, Importance Sampling Resampling and MCMC: In the Importance Sampling Resampling steps, a cloud of particles approximating μ_k is transformed into a cloud of particles approximating μ_{k+1} by randomly duplicating and eliminating particles in a suitable way depending on the relative density between μ_{k+1} and μ_k. In the MCMC steps, particles move independently according to an MCMC dynamics for the current target distribution in order to adjust better to the changed environment.

Our main question is the following: How well does the mean $\eta_n^N(f)$ of an integrand f with respect to the empirical measure η_n^N of the particle system approximate the integral of interest $\mu_n(f)$? We address this question by proving non-asymptotic error bounds of the type

$$\mathbb{E}[(\mu_n(f) - \eta_n^N(f))^2] \leq \frac{C_n(f)}{N},$$

where $\mathbb{E}$ is the expectation with respect to the randomness in the particle system and $C_n(f)$ is a moderately-sized constant depending on the model parameters and on the function f in an explicit way. More specifically, our results center around two main questions: 1) Under which conditions can the smoothing effect of the MCMC steps balance the additional variance introduced into the system through the resampling step? 2) Under which conditions does the particle dynamics work well in multimodal settings where conventional MCMC methods are trapped in local modes? We address both questions by proving suitable non-asymptotic error bounds which depend on a) an upper bound on relative densities, b) constants associated with global or local mixing properties of the MCMC dynamics, and c) the amount of probability mass shifted between effectively disconnected components of the state space as we move from μ_0 to μ_n.

Contents

1 Introduction

1.1 Overview

Since the 1950s, Markov Chain Monte Carlo (MCMC) methods have become an increasingly popular tool for challenging numerical integration problems in a wide variety of fields ranging from chemical physics to financial econometrics. The basic idea is to approximate the integral of a function f with respect to a measure μ by simulating a Markov chain with ergodic measure μ and to calculate the ergodic average of f evaluated at the positions visited by the Markov chain. By construction, MCMC methods only work well if the simulated Markov chain reaches equilibrium sufficiently quickly. Roughly speaking, this is the case when μ is essentially unimodal and it is not the case when μ is severely multimodal in the sense of being characterized by several well-separated modes. In the latter case, MCMC methods tend to get stuck in local modes for very long times and therefore approach their equilibrium μ only on time-scales well beyond those that can feasibly be simulated. This metastability phenomenon is a serious drawback of standard MCMC methods in many applications.

Sequential MCMC methods are a class of algorithms which try to overcome this problem. The basic idea is to approximate the target distribution μ with a sequence of distributions $\mu_0, \ldots, \mu_n$ such that $\mu_n = \mu$ is the actual target distribution and such that μ_0 is easy to sample from. The algorithm constructs a system of N particles which sequentially approximates the measures μ_0 to μ_n. The algorithm is initialized with N independent samples from μ_0 and then alternates two types of steps, Importance Sampling Resampling and MCMC: In the Importance Sampling Resampling steps, a cloud of particles approximating μ_k is transformed into a cloud of particles approximating μ_{k+1} by randomly duplicating and eliminating particles in a suitable way depending on the relative density between μ_{k+1} and μ_k. This step is similar to the selection step in models of population genetics where particles form the population and where the relative density takes the role of a fitness function guiding the number of off-spring a particle has. In the MCMC steps, particles move independently according to an MCMC dynamics for the current target distribution in order to adjust better to the changed environment. This step resembles the mutation step in models of population genetics.

The main question studied in this thesis is the following: How well does the mean $\eta_n^N(f)$ of f with respect to the empirical measure η_n^N of the particle system approximate the integral of interest $\mu_n(f)$? We address this question by proving non-asymptotic error bounds of the type

$$\mathbb{E}[(\mu_n(f) - \eta_n^N(f))^2] \leq \frac{C_n(f)}{N},$$

where $\mathbb{E}$ is the expectation with respect to the randomness in the particle system and $C_n(f)$ is a moderately-sized constant depending on the model parameters and on the function f in an explicit way. More specifically, our results center around two main questions: 1) Under which conditions can the smoothing effect of the MCMC steps balance the additional variance introduced into the system through the resampling step? 2) Under which conditions does the particle dynamics work well in multimodal settings where conventional MCMC methods are trapped in local modes? We address both questions by proving suitable non-asymptotic error bounds which depend on a) an upper bound on relative densities, b) constants associated with global or local mixing properties of the MCMC dynamics, and c) the amount of probability mass shifted between effectively disconnected components of the state space as we move from μ_0 to μ_n.

Outline

The remainder of the introduction is structured as follows: To fix ideas and notation, Section 1.2 introduces the Sequential MCMC algorithm analyzed subsequently. Section 1.3 sets the algorithm into perspective by discussing its relation to other Multilevel MCMC algorithms such as Tempering algorithms. Section 1.4 gives an overview of our main results, relates them to known results and discusses the underlying assumptions. Notably, while the later chapters of the text contain more complete statements of our results as well as their proofs, most of the discussion is found already in Section 1.4. The only major exception are two extended examples in Sections 3.5 and 4.4. Sections 1.3 and 1.4 can be read independently.

The relation between Section 1.4 and the later chapters is as follows: Section 1.4.1 introduces our basic error bounds proved in Chapter 2. Section 1.4.2 presents our results on stability of the algorithm under global mixing conditions found in Chapter 3. Section 1.4.3 discusses our results on the algorithm's performance on multimodal state spaces which are found in Chapters 4 and 5. Chapters 2 to 5 can be read independently with basically two exceptions: Some elementary notation is introduced only at the beginning of Chapter 2; and suitable corollaries of the error bounds of Chapter 2 are restated in the later chapters but not proved again.

1.2 Sequential MCMC

We now briefly introduce the Sequential MCMC algorithm studied subsequently. The more technical details of the framework are postponed to Section 2.1.

Let μ_n be a probability distribution on a state space E, e.g. $E = \mathbb{R}^d$, and let f :

$E \to \mathbb{R}$ be a bounded, measurable function. Our aim is to numerically approximate

$$\mu_n(f) = \int_E f(x)\mu_n(dx).$$

For this purpose we construct a system of N particles $(\xi_n^i)_{i=1}^N$, $\xi_n^i \in E$, which are each approximately distributed according to μ_n and estimate $\mu_n(f)$ by $\eta_n^N(f)$ which is defined as the empirical mean

$$\eta_n^N(f) = \frac{1}{N}\sum_{i=1}^N f(\xi_n^i).$$

We are interested in settings where it is not tractable to generate samples from μ_n directly or through a standard MCMC algorithm. For instance, this is typical of situations where μ_n possesses multiple well-separated modes. Instead we assume there is a distribution μ_0 on E which can easily be sampled and a sequence of probability distributions $(\mu_k)_{k=1}^{n-1}$ which interpolate between μ_0 and μ_n in the following sense: For $k = 0, \ldots, n-1$, μ_k and μ_{k+1} are mutually absolutely-continuous and $\overline{g}_{k,k+1}$ is the relative density of μ_{k+1} with respect to μ_k, i.e.

$$\mu_{k+1}(f) = \mu_k(\overline{g}_{k,k+1} f)$$

for any bounded, measurable function $f : E \to \mathbb{R}$. To capture the idea of interpolation, we assume that there exists a constant $\gamma > 1$ such that $\overline{g}_{k,k+1}(x) < \gamma$ for all $x \in E$. Thus, the weight assigned to a point in E by μ_{k+1} can be bounded by γ times the weight assigned by μ_k. To formulate our algorithm we also need a sequence $K_k(x, dy)$ of transition kernels on E where K_k has stationary distribution μ_k.

To fix ideas, K_k can be thought of as many steps of a local Metropolis dynamics with respect to μ_k. Typical values of the parameters could be $N = 1000$, $\gamma = 2$ and $n = 10$ so that μ_0 and μ_n can differ locally by a factor of 2^{10}.

The algorithm proceeds by constructing a sequence of particle approximations to the measures μ_k moving from the tractable μ_0 to our target μ_n. The algorithm alternates between two steps: 1) an Importance Sampling Resampling step which moves from μ_{k-1} to μ_k and 2) MCMC steps with respect to μ_k.

We start with N particles $(\xi_0^i)_{i=1}^N$ drawn independently from μ_0. Then, for $k = 1, \ldots, n$ we generate particles $(\hat{\xi}_k^i)_{i=1}^N$ approximately distributed according to μ_k through an Importance Sampling step with Multinomial Resampling: The particles $\hat{\xi}_k^i$ are drawn conditionally independently from the empirical distribution of the particles $(\xi_{k-1}^i)_{i=1}^N$ weighted with the relative density $\overline{g}_{k-1,k}$,

$$\mathbb{P}[\hat{\xi}_k^i = \xi_{k-1}^j | \xi_{k-1}^1, \ldots, \xi_{k-1}^N] = \frac{\overline{g}_{k-1,k}(\xi_{k-1}^j)}{\sum_{l=1}^N \overline{g}_{k-1,k}(\xi_{k-1}^l)}.$$

Next, the particles $\hat{\xi}_k^i$ are each moved conditionally independently with the MCMC kernel K_k to generate new particle positions ξ_k^i, i.e.,

$$\mathbb{P}[\xi_k^i \in dx | \hat{\xi}_k^1, \ldots, \hat{\xi}_k^N] = K_k(\hat{\xi}_k^i, dx).$$

This procedure is iterated until we obtain the particles $(\xi_n^i)_{i=1}^N$. Note that in order to run this algorithm it is sufficient to know the densities up to a normalizing factor. Therefore, we denote in the following by $g_{k-1,k}$ an unnormalized version of $\overline{g}_{k-1,k}$ and state the algorithms in terms of $g_{k-1,k}$.

To close this section, a few words on the name of the algorithm are in order. We refer to it as the "Sequential MCMC algorithm" since it addresses the same problem as MCMC with similar methods and since – unlike some related algorithms such as Parallel Tempering – it moves sequentially from one distribution to the next. We refer to it as "our" algorithm to distinguish it from other algorithms. This should not obscure the fact that the algorithm is not our invention but rather an algorithm which has been invented and generalized under several names in multiple applications. For instance, the algorithm is a) an adaption of the Bootstrap Filter of Gordon, Salmond and Smith (1993) to the problem of numerical integration, b) a simple special case of the Sequential Importance Sampling with Resampling (SISR) algorithm discussed in Cappé, Moulines and Rydén (2005), and c) a simple special case of the the Sequential Monte Carlo Samplers of Del Moral, Doucet and Jasra (2006).

1.3 Multilevel MCMC Methods

Section 1.3.1 introduces MCMC algorithms and their limitations concerning integration with respect to multimodal target distributions. Sections 1.3.2 to 1.3.5 are concerned with multilevel MCMC methods which aim at overcoming these limitations, namely Tempering algorithms (Section 1.3.3) and Sequential MCMC methods (Section 1.3.5). These algorithms are conceptually and historically linked to algorithms which address different problems, namely, the Simulated Annealing algorithm for optimization and particle MCMC methods for filtering. These algorithms are presented, respectively, in Sections 1.3.2 and 1.3.4. While there is no separate discussion of Importance Sampling which is another important ingredient of our Sequential MCMC algorithm, the idea is introduced in the context of Umbrella Sampling at the end of Section 1.3.3.

1.3.1 MCMC and Multimodality

Consider the problem of approximating numerically the integral

$$\mu(f) = \int_E f(x)\mu(dx)$$

of the function $f : E \to \mathbb{R}$ over a state space E with respect to some probability measure μ. In many applications of interest, no feasible deterministic methods are available for this problem. Notably, this is the case when E is in some sense large, being, e.g., a complicated graph or a subset of $\mathbb{R}^d$ where d is large. For instance, when $E \subset \mathbb{R}^d$, the computational cost of approximating $\mu(f)$ to a fixed precision using numerical integration methods based on regular grids increases exponentially in the dimension d.

Monte Carlo methods are a class of widely-used solutions to this dilemma. The simplest Monte Carlo algorithm consists in drawing N independent samples $\xi_1, \dots, \xi_N$ from the distribution μ and to estimate $\mu(f)$ by

$$\eta^N(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi_i).$$

Then we have $\mathbb{E}[\eta^N(f)] = \mu_n(f)$ and, under weak additional conditions, by the weak law of large numbers $\eta^N(f)$ will converge stochastically to $\mu(f)$. Moreover, by Chebyshev's inequality the approximation error is of order $O(N^{-\frac{1}{2}})$ regardless of the dimension of E.

The applicability of this simple Monte Carlo method is severely limited by the fact that drawing samples from μ is not feasible in many applications. Notably, this is the case when the distribution μ is only known up to a normalizing factor: Calculating the normalizing constant involves an integration over E and is thus typically not easier than the original integration problem. Markov Chain Monte Carlo (MCMC) Methods, first developed by Metropolis et al. (1953), offer a possibility to approximately sample a distribution which is known only up to a normalizing constant. The basic idea is to run a Markov chain with ergodic distribution μ and to use its values after sufficiently many steps as approximate samples from μ. Roughly speaking, this is an easier problem than calculating a normalizing constant, since deciding whether the chain should jump from its present location to a new one is a much more local problem than integration over the whole state space. Basically, there are two main ideas for constructing MCMC chains in practice: Updating the current state of the chain only on a suitable lower dimensional sub-space of E (Gibbs-Sampling, Geman and Geman (1984)) or slowing down an a priori unrelated Markov chain in an appropriate way (Metropolis-Sampling, Metropolis et al. (1953), Hastings (1970)) See Diaconis (2009) for a brief recent introduction to MCMC.

For future reference and for the sake of concreteness, we take a closer look at how to construct a Metropolis chain with stationary distribution μ. Let $r(x, dy)$ be a Markov transition kernel on E. Then the transition kernel of the associated Metropolis chain is defined by

$$p(x, dy) = r(x, dy) \min\left(1, \frac{\mu(dy) r(y, dx)}{\mu(dx) r(x, dy)}\right).$$

This means that the proposals of the kernel r are sometimes rejected with probabilities chosen in a way such that p fulfills the detailed balance condition

$$\mu(dx)p(x,dy) = \min(\mu(dy)r(y,dx), \mu(dx)r(x,dy)) = \mu(dy)p(y,dx)$$

with respect to μ. Note that in order to construct p it is sufficient to know μ up to a constant factor. Under suitable regularity conditions, see, e.g, Chapter 6 of Robert and Casella (2004) the states of a chain with transition kernel p will eventually be distributed approximately according to μ.

While it is fairly easy to guarantee that a Markov chain with transition kernel p will eventually converge, it is typically much harder to assess the speed of convergence and thus the required running time of the associated MCMC algorithm. Even worse, in many examples of interest, this convergence happens on a time-scale which is much larger than any reasonable running time of the algorithm. Roughly, this occurs when the target distribution μ is strongly multimodal in the sense that it assigns large weights to regions of the state space which are separated by areas of little weight. Then a Metropolis chain which proposes small changes to the current state in each step will typically get trapped near one mode for a long time and will only move to another mode after making many transitions which have a low acceptance probability.

This leads to the question why the proposal kernels r are usually chosen as kernels which propose comparatively small changes to the current state, e.g., as local random walk kernels. The reason for this is the following: In high dimensions, distributions tend to be concentrated in relatively small areas. When the chain is within one mode, we typically only have a substantial probability of proposing again a reasonably likely state if the proposal is not too far from the current state. If proposals are not local enough, the chain will thus be stuck in the same state for long periods of time. Obviously, this is only one side of a trade-off: If proposals are too local then the chain will accept most moves but it will move too slowly to explore the state space well enough. There has been quite a lot of work on how to optimally scale the proposal in the literature. For instance, the rule of thumb that an average acceptance probability of 0.234 is optimal for random walk Metropolis chains has been derived in several ways by now, see, e.g., Roberts and Rosenthal (2001) and Sherlock and Roberts (2009).

To see how multimodality enters in a natural Bayesian estimation problem, consider the problem of estimating a Gaussian mixture distribution[1]

$$\rho = 0.5\mathcal{N}(\mu_1, 0.3) + 0.5\mathcal{N}(\mu_2, 0.6)$$

on $\mathbb{R}$ where $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. Assume that there are observed data X in the form of k draws from ρ and we would like to estimate μ_1 and μ_2. Assume that the true values of μ_1 and μ_2 are

[1] A similar example is discussed in much more detail in Marin, Mengersen and Robert (2005). See also Frühwirth-Schnatter (2007) which gives an extensive introduction to Bayesian mixture estimation.

given by $(\mu_1, \mu_2) = (-1, 1)$. A standard Bayesian approach to this problem is to assume a prior on μ_1 and μ_2, e.g., $\mu_1 \sim \mathcal{N}(0, 1)$ and $\mu_2 \sim \mathcal{N}(0, 1)$. Then, a posterior distribution $\pi(\mu_1, \mu_2|X)$ of (μ_1, μ_2) is calculated from the data by updating the prior accordingly. To explore this distribution, an MCMC chain on $\mathbb{R}^2$ with target distribution $\pi(\mu_1, \mu_2|X)$ is simulated. If there are sufficiently many data points, the largest mode of $\pi(\mu_1, \mu_2|X)$ is near the true values $(-1, 1)$. However, there will be a second mode near $(1, -1)$ since combining the "right" means with the wrong variances still yields a better fit to the data than most of the other configurations: If a local MCMC dynamics is trapped in $(1, -1)$ it will take very long for it to get near the true values $(-1, 1)$.

While this two-dimensional example may look rather harmless, the problem gets much difficult quickly if we increase the number of mixture components: The dimension grows linearly in the number of mixture components and the number of local modes grows exponentially. This behavior is typical for MCMC problems in Gaussian mixture estimation. Multimodal problems are prevalent in many other fields as well including spin-systems below the critical temperature in statistical physics, or molecular simulations in chemical physics, see Liu (2001) for short introductions to various applications.

Note also that "convergence diagnostics" such as auto-correlation times derived from the observed dynamics will typically fail to detect slow mixing in examples such as this one since the part of the state space which is visited at all by the dynamics is explored rather well. Moreover, there is by now a sizeable mathematical literature on proving fast or slow mixing of MCMC chains, see, e.g., the overviews in Diaconis and Saloff-Coste (1998) and Levin, Peres and Wilmer (2009). Yet such results can, of course, only detect slow mixing but they cannot prevent it.

Some authors have suggested to combine a local Metropolis dynamics with occasional long range proposals (see, e.g., Guan and Krone (2007) and Bassetti and Leisen (2007)). This leads to a dynamics which explores the state space well locally despite the presence of long range proposals. However, unless the problem is good-natured (e.g. low-dimensional, or symmetric in a way that can be exploited), there seems to be little hope that these random long range proposals will discover distant modes in a high-dimensional state space with non-negligible probability. The following sections present a number of algorithms which try to bridge distinct modes of the target distribution in a more systematic fashion.

1.3.2 Simulated Annealing

Simulated Annealing (Kirkpatrick, Gelatt, and Vecchi (1983)) is possibly the most widely used MCMC algorithm. Unlike the MCMC algorithms discussed in the previous section, it is intended as an optimization algorithm and not as a method for numerical integration. We discuss it here since it can be seen as an important precursor of both, the Tempering algorithms introduced in Section 1.3.3 and the Sequential MCMC algorithm of Section 1.2.

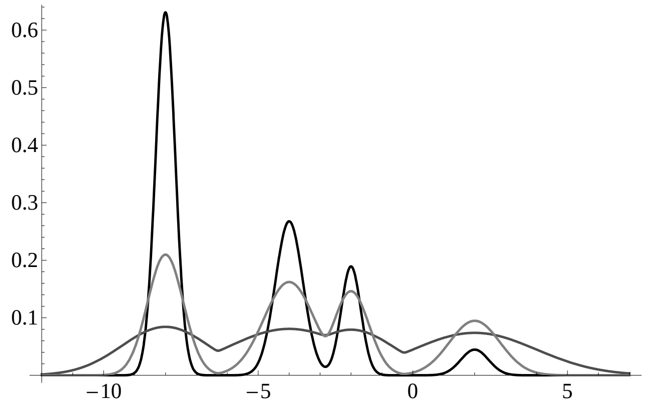

Figure 1.1: The effect of varying β

Assume we would like to find the global minima of a function $H : E \to \mathbb{R}$ which is bounded from below. Consider the family of probability distributions $(\mu_\beta)_{\beta \geq 0}$ given by

$$\mu_\beta(dx) = \frac{1}{Z_\beta}\, e^{-\beta H(x)} \pi(dx)$$

where π is a reference measure such that the support of H is contained in the support of π and Z_β is a normalizing constant. In analogy with statistical physics, the parameter β is called the inverse temperature. Assume that π is such that μ_β is easy to sample from either directly or through MCMC for small values of β. The larger β is, the more do small changes in H influence the distribution μ_β. In the limit $\beta \to \infty$, μ_β converges to the uniform distribution on the global minima of H.

Since an MCMC dynamics with target μ_β can be expected to be stuck in local modes for large β, the idea behind Simulated Annealing is to start with small values of β where the chain can move freely and to increase β only gradually. Specifically, the Simulated Annealing algorithm consists in running an MCMC chain for the target μ_β and while gradually "cooling down" the system by increasing the value of β.

For an illustration of how a higher temperature can bridge components of the state space which are effectively disconnected at lower temperatures consider Figure 1.1. Depicted are distributions μ_β for $\beta \in \{0.1, 0.3, 1\}$ where the reference measure π is the Lesbesgue measure on $\mathbb{R}$ and where H is chosen such that μ_1 is the following Gaussian mixture distribution,

$$\mu_1 = 0.05\mathcal{N}(2, 0.2) + 0.15\mathcal{N}(-2, 0.1) + 0.3\mathcal{N}(-4, 0.2) + 0.5\mathcal{N}(-8, 0.1).$$

To see why Simulated Annealing is generally not suitable for approximating integrals with respect to μ_β, i.e., to see why the distribution of the chain at inverse temperature β is generally not close to μ_β, consider a situation where H has two local minima. Assume that there is a β^* such that for $\beta < \beta^*$ a local Metropolis dynamics with respect to μ_β mixes well while for $\beta > \beta^*$ the state space is divided into two areas E_1 and E_2 around the respective modes which are essentially unconnected by the MCMC dynamics with respect to μ_β. Such behavior is typical e.g. for spin systems exhibiting a first order phase transition such as the mean field Ising model, see, e.g. Madras and Piccioni (1999). In this case, for $\beta > \beta^*$ the probability that the chain is in the region E_1 will approximately equal $\mu_{\beta^*}(E_1)$, the probability of the chain entering the region before it is separated, instead of the equilibrium value $\mu_\beta(E_1)$. For similar reasons, convergence of the algorithm to a global minimum can generally only be guaranteed when the process of cooling down is much slower than any schedule that it feasible to implement (see Holley, Kusuoka and Stroock (1989)). Nevertheless, Simulated Annealing is a widely used heuristic method for solving challenging optimization problems.

Note that if we run multiple parallel copies of Simulated Annealing we almost arrive at the Sequential MCMC method of Section 1.2 (with a less general but common choice of the interpolating distributions μ_k). The crucial difference between the algorithms lies in the Importance Sampling Resampling step employed in Sequential MCMC. The purpose of this step can be seen in balancing the numbers of particles between effectively disconnected components of the state space to avoid the problem just described. See Section 1.3.5 for more discussion along these lines.

1.3.3 Simulated Tempering and Parallel Tempering

As discussed at the end of the preceding section, Sequential MCMC can be seen as a modification of Simulated Annealing which adds a reweighting step to make it an algorithm suitable for numerical integration. Another class of algorithms which modify Simulated Annealing in order to change it in this direction are Tempering algorithms, namely, Simulated Tempering and Parallel Tempering. The idea behind these algorithms is to substitute the deterministic movements between temperature levels from Simulated Annealing by an MCMC dynamics. As will become clear below, both algorithms can be interpreted as Metropolis chains on an augmented state space.

Simulated Tempering (Marinari and Parisi (1992), Geyer and Thompson (1995)) and Parallel Tempering (Geyer (1991), Hukushima and Nemoto (1996)) were both developed independently in the statistics and statistical physics literatures. Interestingly however, Simulated Tempering was introduced as an improvement over Parallel Tempering in the statistics literature, while the opposite was the case in statistical physics. As discussed below, there are sound arguments in support of both views.

Like in Section 1.2 we consider a probability distribution μ_n we would like to sample from, a probability distribution μ_0 on E which can easily be sampled and a sequence of probability distributions $(\mu_k)_{k=1}^{n-1}$ which interpolate between μ_0 and μ_n in the following sense: For $k = 0, \dots, n-1$, assume that μ_k and μ_{k+1} are mutually absolutely-continuous. Denote by $\overline{g}_{k,k+1}$ the relative density of μ_{k+1} with respect to μ_k. Most of the literature on Tempering algorithms has considered the case where $\mu_k(dx) \sim \exp(-\beta_k H(x))\mu_0(dx)$ for a suitable increasing sequence $(\beta_k)_k$ of inverse temperatures – thus the name Tempering – but as already pointed out in Geyer (1991) this is not a necessity.

Simulated Tempering

The Simulated Tempering algorithm consists in running an MCMC dynamics on the augmented state space $E \times I$ where $I = \{0, \dots, n\}$. At each point in time the state (x, k) of the chain consists of a position x in E and a label k which denotes the current "temperature level". From position (x, k), the Simulated Tempering chain makes two types of moves: Level moves which vary x and keep k fixed and temperature moves which change k to $k+1$ or $k-1$ and keep x fixed. There are many possibilities for choosing between these two types of moves, for instance, one can flip a coin in each step to decide whether to make a level move or a temperature move. Level moves are steps of a standard MCMC dynamics with target distribution μ_k as introduced in Section 1.3.1. Temperature moves are essentially Metropolis moves on I with a random walk proposal: When making a level move from position (x, k), $0 < k < n$, the chain moves to $(x, k + 1)$ with probability

$$\frac{1}{2} \min(1, \overline{g}_{k,k+1}(x))$$

and to $(x, k - 1)$ with probability

$$\frac{1}{2} \min(1, \overline{g}_{k,k-1}(x))$$

where $\overline{g}_{k,k-1} = 1/\overline{g}_{k-1,k}$ is the relative density of μ_{k-1} with respect to μ_k. With the remaining probability, the chain stays in (x, k). The transitions from $(x, 0)$ to $(x, 1)$ and from (x, n) to $(x, n - 1)$ are defined accordingly.

It is straightforward to check that the reversible distribution π of this dynamics is given by

$$\pi(dx, k) = \frac{1}{n + 1} \pi_k(dx),$$

so that in equilibrium the chain spends equal amounts of time at each temperature level. Moreover, conditional on being currently at level k, the position x of the chain is approximately a μ_k distributed random variable. The main idea behind the algorithm is that the chain can move between well-separated modes of μ_n by moving from level n to level 0 where the chain mixes quickly and then back to level n.

There are many non-trivial design choices in implementing this algorithm, for example: 1) How to choose the proportions of level moves and temperature moves is an intricate question governed by the following trade-off: With too many temperature moves the chain spends too little time at each level to equilibrate locally. With too few temperature moves, the chain moves too slowly between temperature levels and thus it takes too long to move from n to 0 and then back to n. Accordingly, one should aim at an intermediate choice. See 2) There are virtually endless possibilities in choosing μ_0 and the interpolating distributions $(\mu_k)_{k=1}^{n-1}$, see the discussion in Section 1.4.3.4. For more discussion of the practical design of Tempering algorithms, see, e.g., Predescu, Predescu and Ciobanu (2004), Nadler and Hansmann (2007) and Atchadé, Rosenthal and Roberts (2011).

A major disadvantage of Simulated Tempering lies in the fact that in order to run the algorithm the relative densities $\overline{g}_{k,k+1}$ need to be known explicitly – and not only up to a normalizing constant. The reason is that the normalized relative densities appear in the transition probabilities for the temperature moves stated above. It should however be noted that there are efficient methods for estimating these normalizing constants while the algorithm is running, see e.g. Liang (2005) and Park and Pande (2007). Thus, Simulated Tempering is indeed utilized in applications. In consequence, a practical application of Simulated Tempering will typically involve acceptance probabilities which change over time so that it is a special case of an Adaptive MCMC algorithm.[2] Since the Simulated Tempering chain with adaptive estimation of normalizing constants is not a time-homogeneous Markov chain – it is not even a Markov chain – its convergence behavior is not easy to analyze rigorously. In fact, this seems to be an open problem.

Parallel Tempering

The Parallel Tempering algorithm, also known, e.g., as Swapping, as Exchange Monte Carlo and as Replica Exchange, overcomes the problem of requiring normalized densities in the transition probabilities. The Parallel Tempering chain is a Metropolis chain on E^{n+1} with target distribution

$$\pi(dx) = \prod_{k=0}^{n} \mu_k(dx_k).$$

Therefore, in equilibrium the component x_k of the state $x = (x_0, \dots, x_n) \in E^{n+1}$ of the chain is approximately distributed according to μ_k for all k. Accordingly, the Parallel Tempering chain can be used as an MCMC chain for calculating integrals with respect to all μ_k and, in particular, with respect to our distribution of interest μ_n. Like the Simulated Tempering chain, the Parallel Tempering chain alternates level moves and temperature moves. In a level move, each component x_k of the current state $(x_0, \dots, x_n)$ is updated conditionally independently with an MCMC dynamics with target μ_k. In a temperature move, the chain uniformly picks a pair of adjacent levels $(k, k+1)$, $k \in \{0, \dots, n-1\}$ and proposes to move from the current state $x = (x_0, \dots, x_k, x_{k+1}, \dots, x_n)$ to $\widetilde{x} = (x_0, \dots, x_{k+1}, x_k, \dots, x_n)$. This proposal is accepted with probability

$$\min\left(1, \frac{\pi(d\widetilde{x})}{\pi(dx)}\right) = \min(1, g_{k,k+1}(x_k)g_{k+1,k}(x_{k+1})) = \min\left(1, \frac{g_{k,k+1}(x_k)}{g_{k,k+1}(x_{k+1})}\right).$$

An intuitive, slightly different way of describing the algorithm is to say that we run an MCMC chain at each level π_k and swap the states of adjacent chains with suitably chosen probabilities.

In order to calculate the acceptance probabilities of Parallel Tempering it is sufficient to know the relative densities $g_{k,k+1}$ up to a normalizing constant. This comparative

[2] For an introduction to Adaptive MCMC algorithms see, e.g., Andrieu and Thoms (2008) or Atchadé, Fort, Moulines, Priouret (2011).

advantage over Simulated Tempering comes at the following cost: In a temperature move, Parallel Tempering always proposes two changes at once, i.e. moving x_k from level k to level $k+1$ and moving x_{k+1} in the opposite direction. If we assume that x_k lies in a location which is typical for the distribution μ_k (and not for μ_{k+1}), and that x_{k+1} lies in a location which is typical for μ_{k+1} (and not for μ_k), it is intuitive to expect that Simulated Tempering moves more easily between temperature levels than Parallel Tempering since only one temperature move has to be accepted in each step. See Liang (2005) for an extensive discussion of this intuition and a number of numerical examples in its support.

Since the Parallel Tempering chain is a time-homogeneous Markov chain, it can be analyzed with the same techniques used for studying mixing properties of single-level MCMC chains. This is done in a small literature starting with Madras and Zheng (2002), Zheng (2003) and Bhatnagar and Randall (2004). The question of whether a certain Tempering algorithm yields an improvement over single-level MCMC is not trivial. For instance, Madras and Zheng (2002) showed that Tempering algorithms mix rapidly (i.e., in polynomial time with respect to the system size) for the Mean-Field Ising model at any target temperature. Single-level MCMC chains mix torpidly (i.e., in exponential time) in this case. In contrast, Bhatnagar and Randall (2004) showed for the Mean-Field Potts model that Tempering algorithms mix torpidly under a natural choice of interpolating distributions. See Section 1.4.3.3 for more discussion.

Comparison with Sequential MCMC

Zuckerman and Lyman (2006) point out that Parallel Tempering can only yield an improvement over single-level MCMC if the mixing time of the dynamics at "high temperature" levels is at least a factor $n+1$ shorter than the mixing time of the dynamics at the "low temperature" level n. It seems safe to assert that the same intuition also holds for Simulated Tempering and that it also holds with a factor of $(n+1)^2$. The main argument behind this is that a nearest-neighbor random walk on $\{0,\dots,n\}$ mixes in $O(n^2)$ steps. There seems to be no reason to expect the two Tempering chains to exhibit a better dependence on n since both incorporate a stochastic nearest neighbor dynamics between levels which assigns the same weight to all levels, see also Liu (2001, p. 211). This is unfortunate since we are actually not interested in movements of the dynamics from level n to level 0: All we want is a chain which mixes well at level 0 and then carries this good mixing down to level n. While the movements in the opposite direction only serve the purpose of "balancing" the algorithm, they can be expected to lead to a considerable slowing down of the chain. The Sequential MCMC method of Section 1.2 can be seen as overcoming this problem since it proceeds deterministically from level 0 to level n.

The fact that unlike Sequential MCMC the Tempering algorithms move between levels without taking into account our preferred direction of moving from level 0 to level n also has an advantage: It is fairly easy to extend these algorithms to more complicated systems of probability distributions than sequences $(\mu_0,\dots,\mu_n)$. This is done in the so-called Hyper-Parallel Tempering algorithms applied in chemical physics, see Section 2.4 in the survey of Earl and Deem (2005) for an introduction and references. In these algorithms, swapping moves are proposed not only with respect to temperature but also with respect to other model parameters. Accordingly, there is a system $(\mu_\alpha)_\alpha$ of probability distributions where $\alpha \in \{0, \dots, n\}^k$ is a k-dimensional multi-index for some $k \in \mathbb{N}$. For each value of α, an MCMC dynamics with target μ_α is simulated and additionally there are swapping moves which exchange the chain positions between "levels" α and α' where α and α' differ by 1 in one component and are identical otherwise. Thus, k denotes the number of directions in which swapping moves are proposed and the choice $k = 1$ corresponds to the simple Parallel Tempering algorithm. Simulated Tempering can be extended in a similar way. This is not the case for Sequential MCMC.

Another potential advantage of Tempering algorithms lies in the fact that they have smaller requirements on computer memory than Sequential MCMC. This may be an issue when the state space is, e.g., very high-dimensional so that the cost of storing an element of the state space is nonnegligible: For Simulated Tempering, only one state at a time has to be stored while for Parallel Tempering we need one state at each "temperature" level. In contrast, in Sequential MCMC we need one state for each particle in the particle system. Since typical applications work with several thousands of particles while the number of levels will usually be somewhere between 4 and 100, this may be a good argument in favor of Tempering algorithms in some cases.

Further Discussion

One advantage which Tempering algorithms and Sequential MCMC have in common is that they do not only sample μ_n but all the distributions $(\mu_0, \dots, \mu_n)$. This may be of interest, e.g., in physical applications where a model is to be studied at different temperatures. Multilevel MCMC methods offer a solution to this problem which may have the added advantage of overcoming slow mixing at low temperatures.

A similar idea was already discovered by Torrie and Valleau (1977) who proposed the *Umbrella Sampling* algorithm. In Umbrella Sampling, one runs a single MCMC chain whose target ν is, e.g., a symmetric mixture of the distributions of interest $(\mu_0, \dots, \mu_n)$. ν is called the Umbrella distribution since all the μ_i are absolutely continuous with respect to ν with a density which is bounded by $n + 1$. The approximate samples from ν obtained through MCMC are then used as proposals in an Importance Sampling step that samples from the μ_k: Assume we would like to approximate $\mu_k(f)$ for some $f : E \to \mathbb{R}$. Denote by $\overline{g}_{\nu,k}$ the relative density of μ_k with respect to ν and by $\xi_1, \dots, \xi_N$ our approximate samples from ν. Then the Umbrella Sampling estimator of $\mu_k(f)$ is given by

$$\mu_k(f) = \nu(f\,\overline{g}_{\nu,k}) \approx \frac{1}{N} \sum_{i=1}^{N} f(\xi_i) \overline{g}_{\nu,k}(\xi_i).$$

As in Simulated Tempering, some modifications of this method will typically be necessary to take into account that the relative densities $\overline{g}_{\nu,k}$ are not known explicitly. Usually, if this is the case the normalizing constant is approximated by

$$\nu(g_{\nu,k}) \approx \frac{1}{N} \sum_{i=1}^{N} g_{\nu,k}(\xi_i)$$

where $g_{\nu,k}$ denotes an unnormalized version of $\overline{g}_{\nu,k}$, and $\mu_k(f)$ is approximated by the quotient

$$\mu_k(f) \approx \frac{\sum_{i=1}^{N} f(\xi_i) g_{\nu,k}(\xi_i)}{\sum_{i=1}^{N} g_{\nu,k}(\xi_i)} \tag{1.1}$$

Umbrella Sampling can be seen as a special case of *Importance Sampling* with a particular recipe for choosing the proposal distribution ν. See Madras and Piccioni (1999) for a more detailed discussion of Umbrella Sampling. A similar Importance Sampling idea can also be applied within a Tempering or Sequential MCMC algorithm for calculating expectations with respect to a distribution $\widetilde{\mu}$ which lies between the steps μ_k and μ_{k+1} of our sequence of distributions, using, e.g, approximate samples from μ_k. It has also been suggested to use particle positions from levels other than the target level as Importance Sampling proposals to improve the number of available samples, see Gramacy, Samworth, and King (2010) and the references therein.

The Sequential MCMC algorithm of Section 1.2 can be seen as a modification of a Tempering algorithm where the stochastic dynamics on "temperature levels" is substituted by a sequence of Importance Sampling steps. Finally, note that the inevitable "self-normalization" with the sum of weights as in (1.1) makes algorithms considerably harder to analyze, since there is no longer a sum of independent random variables on the right hand side.

1.3.4 Particle Filters

In this section we briefly discuss particle methods for the filtering problem. As will become clear below, Sequential MCMC methods and Particle Filters are in fact virtually identical methods with the main distinction being which parameters are choice parameters in the algorithm and which parameters are part of the problem. For instance, our Sequential MCMC algorithm essentially corresponds to the Bootstrap Filter of Gordon, Salmond and Smith (1993) which was one of the first filtering algorithms to combine MCMC dynamics with an Importance Sampling Resampling step. For introductions to particle methods and filtering, see, e.g., Doucet, De Freitas, Gordon (2001) and Bain and Crisan (2009). In the following we will mainly discuss what it means to interpret the setting of Section 1.2 as a filtering setting. A discussion of algorithm design (such as the choice of the resampling method) is postponed to Section 1.3.5 where we return to the problem of approximating integrals with respect to a fixed target distribution.

Filtering is the problem of extracting information about the current state of some unobservable variable, the so-called signal, from noisy or partially revealing observations. Practical examples of this problem include determining the current position of an airplane from radar data, or estimating the volatility in a stochastic volatility model from stock market data. In this interpretation, the state space E is the set of possible values of the signal and the distribution μ_k is the distribution of the signal at time k conditional on all information available at time k. A common feature of many filtering problems is that the evolution of the measures μ_k is driven not only by the incoming information but also by assumptions on the evolution of the signal. For instance, even in the absence of new information coming in between times $k-1$ and k we would not expect an airplane to remain in the same position during the time interval but rather to follow a trajectory extrapolated from its previous movements.

This informal exposition is sufficient to point out a number of important differences between filtering and our Sequential MCMC setting: In Sequential MCMC only the distribution μ_n is given while the sequence $(\mu_k)_{k=0}^{n-1}$ is a choice parameter which can in principle be chosen in such a way that the algorithm works best. In contrast, in a filtering problem the entire sequence $(\mu_k)_{k=0}^{n}$ is given. Notably, while the distributions μ_k can be thought of as smoothed versions of μ_n in the Sequential MCMC context, the same is not true in most filtering settings. Furthermore, in filtering the sequence $(\mu_k)_{k=0}^{n}$ only becomes available over time. In typical applications integrals with respect to μ_k need to be approximated at time k before μ_{k+1} is known.

For the last reason, it is a desirable feature of a filtering algorithm that it exploits the fact that good approximations of μ_k are already available when it derives an approximation of μ_{k+1}. Indeed, the algorithm of Section 1.2 has this property: If we have already run the algorithm for the sequence $(\mu_0, \ldots, \mu_k)$, it is sufficient to just add one more step once the distribution μ_{k+1} becomes available. In contrast, a "Tempering" algorithm stochastically moves forwards and backwards in "time" in this interpretation: In order to apply, e.g., Parallel Tempering in a filtering problem, we would need to run the algorithm for the sequence of levels $(\mu_0, \ldots, \mu_k)$ at time k and would need to run it again with one additional level at time $k+1$.

Thus, in filtering settings there is a second major disadvantage of the fact that Simulated Tempering and Parallel Tempering move forwards and backwards between levels beyond the one already discussed in Section 1.3.3. In this light, it is not surprising that Sequential MCMC methods were initially developed for the filtering problem and were applied to the problem of integration with respect to a fixed distribution only later on.

1.3.5 Sequential MCMC

We are now prepared to give a proper motivation and discussion of the Sequential MCMC algorithm of Section 1.2. Some motivations for the design of the algorithm already follow from the discussion of the previous sections: We saw in Section 1.3.1 that multimodality of the target distribution is a serious problem of simple MCMC algorithms. In Section 1.3.2 we saw that approximating the target with a sequence of smoother distributions is a promising recipe for keeping MCMC dynamics from being trapped in local modes. The idea was to transport the good mixing properties of some initial distribution μ_0 step-wise over to the distribution of interest μ_n. In Section 1.3.3 we saw the Tempering algorithms which, figuratively speaking, solve this transportation task by relying on the services of a random walker. Now obviously, a random walker is not the ideal person to rely on when facing a well-defined transportation task. This consideration suggested looking for an alternative method which moves from μ_0 in μ_n in a quicker and more predictable fashion. Finally, in Section 1.3.4 we saw that, for a variety of reasons, developing such methods was a highly natural problem in the filtering literature. This problem was solved to a remarkable extent by the bootstrap filter of Gordon, Salmond and Smith (1993) and the subsequent literature.

In this light, there are arguably two main issues about the algorithm which need further discussion: the resampling step and the choice of the approximating sequence of distributions. At this point, the majority of the related work has analyzed the algorithm in the filtering context and not as an MCMC algorithm. Notable exceptions include Del Moral, Doucet and Jasra (2006) and a number of precursors and followers, see the references therein and the discussion below. For this reason, the question of resampling has been discussed much more extensively than the question of choosing the distributions μ_k which basically does not arise in filtering. Accordingly, we will focus on resampling in the following. The second issue will be discussed in the light of our results and of related results for Tempering algorithms in Section 1.4.3.

Resampling

To gain more intuition for the resampling steps, note first that these are essentially equivalent to selection steps found in models from mathematical biology: All particles (individuals) are weighted with the relative density (the fitness function) and the number of new particles (offspring) replacing a given particle depends on how large the relative density is at that particle (how fit the individual is). Similarly, the MCMC steps can be interpreted as mutation steps. Therefore, it is not surprising that the first remotely similar algorithms came up in the literature on genetic algorithms (Fogel, Owens and Walsh (1996), Rechenberg (1973), Holland (1975)) which, roughly speaking, develops algorithms based on biological ideas, see Man, Tang and Kwong (1999) for an introduction.

Before we start discussing the benefits of the resampling step in the algorithm, we give the following example which shows that it has to be applied with some caution. Consider the following simple Monte Carlo setting: We have samples $\xi_0^1, \ldots, \xi_0^N$ from a distribution μ. Thus, we can approximate integrals with respect to μ by integrals with respect to the empirical distribution of the ξ_0^i. Now assume we resample, i.e., we generate a new sample $\xi_1^1, \ldots, \xi_1^N$ by drawing N times independently and uniformly from $\{\xi_0^1, \ldots, \xi_0^N\}$. The empirical distribution of the new sample still approximates μ but the quality of the sample is worse since some values from the original sample are lost while others are duplicated. This effect gets stronger when the procedure is iterated until at some point all ξ_k^i have the same value. Therefore, taken by itself resampling leads to a degeneration of the sample. These observations show that the MCMC steps serve at least two purposes in our algorithm: Besides helping to explore the target distributions better, they also decrease the dependence between the particles by moving apart particles which duplicate the same predecessor. Studying how the degeneration introduced through resampling is balanced through the MCMC steps will be one of the main topics of subsequent chapters.

Resampling or Weighting?

To understand the benefits of the resampling step it is worthwhile to consider the most common alternative: Instead of duplicating particles with a high “fitness” we can consider a suitably weighted particle approximation. Algorithms of this type are called (among others) *Sequential Importance Sampling* methods and are – like their counterparts with resampling – found under a variety of names in many applications (see, e.g., the introduction in Cappé, Moulines, and Rydén (2005)). For discussions of these methods as MCMC algorithms, see Jarzynski (1997a, 1997b), Neal (2001) and the more general framework of Del Moral, Doucet and Jasra (2006).

Consider the following basic Sequential Importance Sampling algorithm which corresponds to the *Annealed Importance Sampling* of Neal (2001). We start with generating N independent runs of Simulated Annealing: First, we generate for $i = 1, \dots, N$ a particle $\widetilde{\xi}_0^i$ distributed according to μ_0. These particles are the starting points of our runs of Simulated Annealing. Then we generate for all i and for $k = 0, \dots, n-1$ a particle $\widetilde{\xi}_{k+1}^i$ from $\widetilde{\xi}_k^i$ using the transitional kernel K_{k+1}. As discussed in Section 1.3.2, the particles $\widetilde{\xi}_n^i$ do not approximate our target distribution μ_n. This discrepancy is taken into account by assigning each particle $\widetilde{\xi}_n^i$ a weight $\overline{w}_n^i$ as follows: Define the unnormalized weight of particle i by the product of unnormalized relative densities along the trajectory $(\widetilde{\xi}_0^i, \dots, \widetilde{\xi}_n^i)$

$$w_n^i = \prod_{k=0}^{n-1} g_{k,k+1}\left(\widetilde{\xi}_k^i\right),$$

and normalize the weights by their sum

$$\overline{w}_n^i = \frac{w_n^i}{\sum_{j=0}^N w_n^j}.$$

Then, we can approximate $\mu_n(f)$ by

$$\mu_n(f) \approx \sum_{i=1}^n \overline{w}_n^i f\left(\widetilde{\xi}_n^i\right),$$

see, e.g., Neal (2001) for a heuristic justification.

If we wanted to add one more level $n+1$ to the algorithm, it would be sufficient to multiply the weights w_n^i by $g_{n,n+1}(\widetilde{\xi}_n^i)$ and to normalize again. For this reason, the algorithm is often written in a way where weights are calculated recursively. However this distracts from the following crucial observation: The movements of the particles are not influenced by the weights. Thus, Annealed Importance Sampling is a standard Importance Sampling algorithm where the proposal distribution is constructed using Simulated Annealing.

To see the advantage of the resampling step over weighting, it is instructive to consider the bimodal example from Neal (2001) where Annealed Importance Sampling is tested on an asymmetric mixture of two Gaussian distributions in $\mathbb{R}^6$. While the algorithm gives a fairly good approximation to the integral of interest, one of the two modes which carries two thirds of the probability mass contains only 27 out of the 1000 particles. This shows clearly that a particle approximation based on unweighted Simulated Annealing would have lead to a disastrous approximation. However it also unveils a fundamental problem of Annealed Importance Sampling: Since one of the two modes is explored by only 27 particles, the quality of the Monte Carlo approximation is much worse than the total number of 1000 particles would suggest: Exploring the other mode with 977 particles is essentially a waste of computational effort since this cannot make up for the error made by the 27 particles in the other mode. Put differently, while Annealed Importance Sampling allocates probability mass largely correctly over the two modes, it makes no effort to adjust the amounts of MCMC computations in the two modes accordingly. In this light, the resampling step in our Sequential MCMC algorithm can be seen as a method of allocating the MCMC computations proportionally to probability mass in every step of the algorithm.

The behavior observed in this example is a serious drawback of Sequential Importance Sampling methods for at least two reasons: First, this type of weight degeneration, i.e., the concentration of most probability mass in few particles, is a well-documented property of the algorithm, see, e.g. Cappé, Moulines and Rydén (2005, p. 231). Basically, it arises from the fact that in many natural models a particle which gains a lot of mass in one step has a high probability of gaining a lot of mass again in future steps. To obtain some intuition for this, note that if we choose the distributions μ_k as $\mu_k(dx) \sim \exp(-\beta_k H(x))\mu_0(dx)$, then unnormalized relative densities $g_{k,k+1}$ are given by

$$g_{k,k+1} = \exp(-(\beta_{k+1} - \beta_k)H(x)).$$

For all k, these functions $g_{k,k+1}$ have their local maxima in the same points, namely, in the local minima of H. Thus, if we think of our particles as moving only locally, we see that there is a substantial probability that most weight is concentrated in only few particles in promising locations after a modest number of weighting steps. Second, a main reason for the fairly good performance of the Annealed Importance Sampling algorithm in Neal's example lies in the fact that the particles responsible for discovering each of the two modes face a relatively simple MCMC problem which is essentially equivalent to MCMC for a single Gaussian distribution. With a more complicated target distribution, Annealed Importance Sampling may easily miss important parts of the state space. For instance, if a mode with only 27 particles was split into two well-separated modes again as the algorithm proceeds, there is a substantial probability that the particle approximation cannot keep track of this since one of the modes is missed. The latter argument may be the main reason for including a resampling step in the algorithm. A more rigorous study of this intuition will be one of the main subjects of subsequent chapters, see Section 1.4.3.

How to Resample

We have thus seen that the resampling step is an important ingredient in the algorithm, but we have also seen that it may lead to a degeneration of the sample. For this reason, a number of alternative resampling procedures have been proposed which introduce less variance into the system than the so-called Multinomial Resampling step found in the algorithm of Section 1.2. For instance, in Residual Resampling, each particle $\hat{\xi}^i_{k-1}$ is replaced with at least M^i_k particles where

$$M^i_k = \left\lfloor N \frac{\overline{g}_{k-1,k}(\hat{\xi}^i_{k-1})}{\sum_{l=1}^N \overline{g}_{k-1,k}(\hat{\xi}^l_{k-1})} \right\rfloor ,$$

is the largest integer number smaller than the expected number of successors in Multinomial Resampling. The remaining $N - \sum_i M^i_k$ particles are assigned using a Multinomial Resampling procedure with appropriately chosen weights. This choice of resampling procedure eliminates some unnecessary resamplings. In the above example, where the relative density $\overline{g}_{k-1,k}$ is constant, each particle is deterministically replaced by exactly one successor. The situation is however less clear in general: If $M^i_k > 1$ for all particles but one, $N-1$ particles are assigned deterministically. In contrast, if $M^i_k < 1$ for all but one particle and $M^i_k < 2$ for all particles, Residual Resampling hardly differs from Multinomial Resampling. For such reasons, it is difficult to rigorously quantify the advantage of Residual Resampling over Multinomial Resampling in more general settings. Cappé, Douc and Moulines (2005) and Hol, Schön and Gustafsson (2006) have compared several resampling schemes heuristically and numerically. From these studies, it seems reasonable to assert that switching from Multinomial Resampling to, e.g., Residual Resampling is strongly advised but does not lead to a dramatic improvement. Other resampling schemes such as Systematic Resampling can lead to small further improvements. Nevertheless we will consider Multinomial Resampling in the following since it is by far the easiest to analyze and since we are mostly interested in upper bounds on the error.

Further Discussion

There are various other generalizations of Sequential MCMC. For instance, in practice it is rather common to consider an adaptive variant of the algorithm which calculates weights like in Annealed Importance Sampling and employs the resampling step only when the effective sample size (for a definition, see Cappé, Moulines and Rydén (2005) p. 235) falls below some threshold. Yet this adaptivity may be more important, e.g., in the context of filtering – where information comes in small portions barely influencing the distribution – than in Sequential MCMC where the sequence of distributions can be chosen as needed. Adaptive MCMC algorithms are naturally harder to analyze, see Del Moral, Doucet and Jasra (2011) and Atchadé, Fort, Moulines, Priouret (2011). Another adaptive variant of Sequential MCMC is the Equi-Energy-Sampler of Kou, Zhou and Wong (2006) which constructs proposal distributions for the MCMC steps by keeping track of previously visited states. Del Moral, Doucet and Jasra (2006) introduce our Sequential MCMC algorithm within a wide class of algorithms, the so-called Sequential Monte Carlo Samplers, which replace the MCMC steps with respect to the target by more general transition kernels. Finally, the number of particles can be varied at each resampling step, and the resampling step can be replaced by a branching step which leads to a random number of particles. For instance, in minimal variance branching (see, e.g., Crisan, Gaines and Lyons (1998)), each particle is replaced by either M_k^i or M_k^i+1 successors with probabilities chosen such that the expected number of successors is the same as in Multinomial Resampling.

1.3.6 Notes

In the previous sections we introduced the Sequential MCMC method which will be studied in the following, introduced the problem of integrating with respect to multimodal target distributions and introduced and discussed a number of alternative algorithms addressing this problem. Since this discussion was far from exhaustive, we close with a number of references to more detailed introductions: Liu (2001) gives an introduction to MCMC with many applications and also covers the Tempering algorithms of Section 1.3.3. Cappé, Moulines and Rydén (2005) give a detailed introduction to Sequential Monte Carlo methods, covering both theory and many applications. An introduction to MCMC on a general state space and an extensive treatment of MCMC in Bayesian statistics are found in Robert and Casella (2004).

1.4 Main Results

We now give an overview of our main results. Our aim is to prove explicit, non-asymptotic error bounds for the Sequential MCMC algorithm of Section 1.2. When analyzing the algorithm we assume suitable mixing conditions for the MCMC dynamics we employ. For this reason, our results do not depend explicitly on the underlying MCMC dynamics and the state space. For instance, they apply both to a discrete and to a continuous state space. Nevertheless, our motivation lies in proving bounds that can handle settings with the following three properties: multimodality, high dimensions and non-compact state space. Multimodality is important in practice since its presence is the main motivation for relying on a multilevel MCMC algorithm like ours instead of, e.g., a standard MCMC algorithm. Similarly, high dimensions are important since in low-dimensional integration problems deterministic algorithms can be expected to yield better results than Monte Carlo methods.

Finally, it is desirable to have results which can handle non-compact state spaces such as $\mathbb{R}^d$ since otherwise we cannot address many important examples such as Gaussian distributions and Gaussian mixture distributions.

Our results fall into three basic categories: We first prove a relatively general error bound and then apply it to the study of two questions, stability and multimodality. Stability refers to the algorithm's ability to reduce the variance introduced in the resampling steps through the smoothing MCMC steps. Multimodality refers to the algorithm's ability to handle multimodal target distributions.

Obviously, an exhaustive answer to the question of multimodality must implicitly contain an answer to the question of stability. Thus, some clarification is in order. Stability has been studied comparatively much in the previous literature, see Section 1.4.2.2 below. Research has mostly focused on studying the error under assumptions of global mixing of the MCMC dynamics at all levels k, thus abstracting from the problem of multimodality. There are basically two justifications for this: First, multimodality is not quite as important in other applications of Sequential Monte Carlo, such as filtering, as it is in MCMC integration. Second, a good understanding of the algorithm's performance in the more interesting multimodal case cannot be achieved without a good understanding of the case with good global mixing. Our stability results follow this line of research and provide error bounds under global mixing conditions.

In contrast, multimodality has been studied curiously little in the literature and our results can be seen as first steps in this direction. While a number of convincing heuristics about the algorithm's ability to cope with multimodal target distributions can be derived from our results, the assumptions are still too restrictive to actually apply to most models of interest. In short: Our stability results are non-asymptotic error bounds under global mixing conditions. Our results on multimodality are non-asymptotic error bounds under local mixing conditions and under somewhat restrictive technical assumptions otherwise.

Sections 1.4.1.1, 1.4.2.1 and 1.4.3.2 present our results on, respectively, general error bounds, stability and multimodality. Each section is followed by a discussion of the results and of their relation to the literature. For convenience, the results of Section 1.4.1.1 are presented in the framework introduced in Section 1.2 although the actual results in Chapter 2 are proved in a more general setting allowing for, e.g., a sequence E_k of state spaces instead of a fixed state space E.

1.4.1 Basic Error Bounds

1.4.1.1 Results

In order to present the basic error bounds for the algorithm, which are the subject of Chapter 2, we need some more notation in addition to that of Section 1.2. Denote by $B(E)$ the bounded, measurable functions from E to $\mathbb{R}$ and define for $f \in B(E)$

$$K_k(f)(x) = \int_E f(z) K_k(x, dz),$$

i.e., $K_k(f)$ is the transition kernel K_k applied to f. Moreover, define the mapping $q_{k-1,k} : B(E) \to B(E)$ by

$$q_{k-1,k}(f) = \frac{g_{k-1,k} K_k(f)}{\mu_{k-1}(g_{k-1,k})}$$

and define for $0 \leq j < k \leq n$ the mapping $q_{j,k} : B(E) \to B(E)$ by

$$q_{j,k}(f) = q_{j,j+1}(q_{j+1,j+2}(\dots q_{k-1,k}(f)))$$

and $q_{k,k}(f) = f$. By these definitions we have for all $f \in B(E)$ the relation

$$\mu_j(q_{j,k}(f)) = \mu_k(f) \quad \text{for } 0 \leq j \leq k \leq n.$$

So to say, the operator $q_{j,k}$ shifts the flow of probability mass from μ_j to μ_k, including the MCMC steps, from the measures to the integrand f. Basically, the error bounds stated below relate the integration error

$$\mathbb{E}[(\eta_n^N(f) - \mu_n(f))^2]$$

to stability properties of the semigroup $q_{j,k}$. As before, $\eta_k^N(f)$ denotes the empirical distribution of the N particles $(\xi_k^i)_{i=1}^N$.

Weighting the particle system

Denote by $\mathbb{E}$ the expectation with respect to the particle dynamics of the algorithm and denote by $\mathcal{F}_k$ the sigma algebra generated by the particle systems' dynamics up to step k. Then we have

$$\mathbb{E}[\eta_k^N(f)|\mathcal{F}_{k-1}] = \frac{\eta_{k-1}^N(q_{k-1,k}(f))}{\eta_{k-1}^N(g_{k-1,k})}.$$

Since η_{k-1}^N appears both in the numerator and in the denominator on the right hand side, we see that the non-linearity in the resampling step would lead to fairly complicated expressions when iterating these expectations over a number of steps. To avoid this difficulty, our analysis focuses on the approximation error of the weighted empirical measures

$$\nu_k^N(f) = \varphi_k \, \eta_k^N(f), \quad \text{where } \varphi_k = \prod_{j=0}^{k-1} \eta_j^N(g_{j,j+1}).$$

The idea behind the measures ν_k^N is to introduce in every step a factor correcting the distortion caused by normalizing with the sum of current particle weights.[3] These correction factors have expectation 1, $\mathbb{E}[\varphi_k] = 1$, and it holds that

$$\mathbb{E}[\nu_k^N(f)|\mathcal{F}_{k-1}] = \nu_{k-1}^N(q_{k-1,k}(f)).$$

This demonstrates that the measures ν_k^N are considerably easier to handle than the unweighted particle measures η_k^N. Moreover, we prove the inequality

$$\mathbb{E}[(\eta_n^N(f) - \mu_n(f))^2] \leq 2\operatorname{Var}(\nu_n^N(f)) + 2\,\|f - \mu_n(f)\|_{\sup}^2 \operatorname{Var}(\nu_n^N(1)) \tag{1.2}$$

where $\|\cdot\|_{\sup}$ denotes the supremum norm on $B(E)$. This inequality shows that for bounded functions f it is sufficient to control the approximation errors

$$\operatorname{Var}(\nu_n^N(f)) = \mathbb{E}[(\nu_n^N(f) - \mu_n(f))^2]$$

of ν_n^N in order to control the errors with respect to the algorithm's output η_n^N.

The error bound

Our first main step consists in using martingale techniques to obtain an explicit expression for this error. We show that

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] = \frac{1}{N}\operatorname{Var}_{\mu_n}(f) + \frac{1}{N}\mathbb{E}\left[\sum_{j=0}^{n-1} V_{j,n}^N(f)\right] \tag{1.3}$$

where

$$V_{j,n}^N(f) = \nu_j^N(1)\nu_j^N(q_{j,n}(f)^2) - \nu_j^N(q_{j,n}(f))^2 + \nu_j^N(q_{j,j+1}(1) - 1)\nu_j^N(q_{j,n}(f^2)). \tag{1.4}$$

The basic idea behind our error bounds now lies in the following observation: On the left hand side of (1.3) we have $\nu_n^N(f)^2$ as the most problematic term, since $\mathbb{E}[\nu_n^N(f)] = \mu_n(f)$. On the right hand side we have terms of the form $\frac{1}{N}\nu_j^N(g)\nu_j^N(h)$ where the functions g and h depend on f and on the operators $q_{j,k}$. Roughly, our strategy is to bound $\frac{1}{N}\nu_j^N(g)\nu_j^N(h)$ by $\frac{1}{N}\nu_n^N(f)^2$ times a constant depending on the semigroup. Then we apply a fixed-point type argument, using the $\frac{1}{N}$ on the right hand side to show that the error must be small. Going through the details of this idea leads to the following error bound:[4]

Theorem 1.1. *For $0 \leq j \leq n$, let $\|\cdot\|_j$ be a norm on the function space $B(E)$ with $\|f\|_j < \infty$ for all $f \in B(E)$. For $0 \leq j < k \leq n$, let $c_{j,k}$ be a constant such that for all $f \in B(E)$, the following inequality is satisfied*

$$\max(\|1\|_j\|q_{j,k}(f)^2\|_j, \|q_{j,k}(f)\|_j^2, \|q_{j,k}(f^2)\|_j) \leq c_{j,k}\|f\|_k^2. \tag{1.5}$$

[3] Notably, "weighted" refers to the pre-factor φ_k and not to individual weights for each particle as in, e.g., the Annealed Importance Sampling algorithm presented in Section 1.3.5.

[4] This error bound is a special case of the one in Theorem 2.1

Define

$$\overline{c}_k = \max_{l \le k} \sum_{j=0}^{l-1} c_{j,l} \Big(2 + \| q_{j,j+1}(1) - 1 \|_j \Big)$$

and

$$\overline{v}_k = \max_{l \le k} \sup \left\{ \sum_{j=0}^{l} V_{j,l}(f) \,\middle|\, \|f\|_l \le 1 \right\}.$$

where

$$V_{j,l}(f) = \mathrm{Var}_{\mu_j}(q_{j,l}(f)).$$

Then for $N \ge 2\overline{c}_n$ and for all $f \in B(E)$ we have

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \le \frac{\sum_{j=0}^n V_{j,n}(f)}{N} + \frac{2\overline{v}_n \overline{c}_n \|f\|_n^2}{N^2}. \tag{1.6}$$

The first order term on the right hand side of (1.6) is exact in the sense that the coefficient $\sum_{j=0}^n V_{j,n}(f)$ corresponds to the asymptotic variance in the central limit theorem for ν_n^N found in Del Moral and Miclo (2000, p. 45). The theorem thus reduces the problem of bounding the approximation error to the problem of choosing suitable norms and then proving the inequality (1.5).[5]

1.4.1.2 Related Work

The results and techniques up to the formula (1.4) for the expected approximation error of the weighted empirical measure ν_n^N are known in the literature, see, e.g., Del Moral and Miclo (2000, Section 2). They are stated here (and proved below) for the sake of completeness and to present exactly what we need in a coherent way and in the form in which we need it. The basic strategy of proof for the non-asymptotic error bound is adapted from Eberle and Marinelli (2011). Eberle and Marinelli consider a continuous time dynamics on a finite state space where – instead of our resampling step – particles replace particles with a lower “fitness” at suitably chosen rates. Notably, in our algorithm there is there is a freely chosen number of MCMC steps between two resampling steps. This is not the case in the setting of Eberle and Marinelli where a short time interval before the endpoint is controlled with different techniques that do not rely on mixing properties of the MCMC dynamics. The case of a discrete time and a general state space treated here is closer to the algorithmic applications of interest.

Most of the non-asymptotic error bounds in the literature, see Whiteley (2011) for a recent example and Section 1.4.3.3 for more references, rely on additive corrections in place of the multiplicative correcting factors φ_k. Cérou, Del Moral and Guyader (2011) consider tree-based expansions of ν_n^N: They explicitly study the particle system's genealogy, considering, e.g., how many particles were resampled from the same “ancestor” how many steps back. While this approach has great potential

[5] In principle, one also needs bounds on $\overline{v}_k$ and $\overline{c}_k$. In the settings we consider later on such bounds can also be derived from (1.5).

for yielding a good understanding of the particle system, it does not seem to give the opportunity to separate the properties of the semigroup $q_{j,k}$ from the particle dynamics as cleanly as in our approach.

1.4.1.3 Discussion

The advantages of our error bound should become clear from its applications presented in the following sections. Thus we focus on a brief discussion of its disadvantages here: Our overall approach relies on studying the weighted empirical measure ν_n^N and then transferring the results to the algorithm's output η_n^N using the inequality (1.2). Since (1.2) depends on the supremum of the integrand f, we are limited to considering bounded integrands. Note however that our results on the weighted particle system ν_n^N do not rely on this boundedness. There is some intrinsic interest in such results as well, see the discussion in Cérou, Del Moral and Guyader (2011).

Another disadvantage of the approach is introduced through the fixed-point-type argument outlined in the discussion below (1.4): We rely on the fact that we can express the quadratic error of $\nu_n^N(f)$ in terms of the measures $(\nu_k^N)_k$ again. This works out nicely for the algorithm we have here but seems to get lost easily when considering resampling schemes[6] other than Multinomial Resampling. For instance, in the case of Residual Resampling some particles are replaced directly while others are replaced with Multinomial Resampling using residual weights. Thus the expected error is constituted by two rather different terms which do not seem straightforward to reassemble in terms of ν_k^N as in the case of Multinomial Resampling. However, our present results can also be seen as upper bounds for the error of more sophisticated resampling schemes.

1.4.2 Stability

1.4.2.1 Results

Our stability results for the algorithm, found in Chapter 3, are derived from Theorem 1.1 by choosing $\|\cdot\|_j = \|\cdot\|_{L_p(\mu_j)}$ for some $p > 2$ where the L_p-norm $\|\cdot\|_{L_p(\mu_j)}$ is defined by

$$\|f\|_{L_p(\mu_j)} = \mu_j(|f|^p)^{\frac{1}{p}}.$$

With this choice of norms, the crucial inequality (1.5) in Theorem 1.1 is fulfilled for all $j < k$ with constant

$$c(p) = \max\left(\widetilde{c}\left(p, \frac{p}{2}\right), \widetilde{c}\left(2p, p\right)^2\right),$$

where we define $\widetilde{c}(p,q)$ as a constant in an L_p-L_q inequality for the semigroup $q_{j,k}$, i.e., a constant such that

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \widetilde{c}(p,q)\|f\|_{L_q(\mu_k)}$$

[6] See Section 1.3.5 for more discussion of different resampling schemes.

is satisfied for all $f \in B(E)$ and all $j < k$. Given such constants $\widetilde{c}(p,q)$ we immediately obtain an error bound from Theorem 1.1. Therefore, the main work lies in deriving these constants.

It is fairly easy to derive L_p-bounds based on the observation that

$$|q_{j,k}(f)(x)| \leq \gamma^{k-j}|f(x)|$$

for all $x \in E$ where γ is our uniform upper bound on $\overline{g}_{l,l+1}$. L_p-bounds for the semigroup $q_{j,k}$ relying entirely on this bound would degenerate exponentially fast in the length of the time interval $k-j$. We aim at bounds with constants which are stable over time. To achieve this, we need to assume and exploit suitable mixing conditions for the kernels K_k.

To this end, we assume the following: There are constants $\alpha > 0$ and $\beta \in [0,1]$ such that for all $f \in B(E)$ and all k we have the following L_2-bound,

$$\|K_k(\overline{g}_{k,k+1}f)\|^2_{L_2(\mu_k)} \leq \alpha\|f\|^2_{L_2(\mu_{k+1})} + \beta\mu_{k+1}(f)^2. \tag{1.7}$$

If K_k is reversible, the inequality (1.7) with constants $\alpha = (1-\rho)\gamma$ and $\beta = \rho$ follows, e.g., from a Poincaré-like inequality

$$\mathrm{Var}_{\mu_k}(K_k(f)) \leq (1-\rho)\mathrm{Var}_{\mu_k}(f)$$

with constant $\rho \in (0,1)$. Recall that the kernels K_k each represent many steps of MCMC and that thus the constant α can be controlled by varying the number of MCMC steps.

Our main result states that if (1.7) holds with a sufficiently small α, i.e., if the dynamics K_k all mix sufficiently well, we obtain an L_p-L_p-bound with time-independent constant:[7]

Proposition 1.1. *For $r \in \mathbb{N}$, consider $p \in [2^r, 2^{r+1}]$ and assume that (1.7) holds with an α for which $\alpha\gamma^{2^r-2} < 1$. Then we have for $0 \leq j < k \leq n$ and $f \in B(E)$ the inequality*

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \widetilde{c}(p,p)\|f\|_{L_p(\mu_k)}$$

with

$$\widetilde{c}(p,p) = \frac{\gamma^{r+1}}{1-\alpha\gamma^{2^r-2}}$$

This shows that L_p-norms remain stable under sufficiently good mixing. However in order to verify (1.5), we also need to bound, respectively, the L_{2p} and L_p norms of $q_{j,k}(f)$ against the L_p and $L_{p/2}$ norms of f. To achieve this, we need an additional assumption of hyperboundedness:[8]

[7] The following result corresponds to Proposition 3.1 below.

[8] The following result is found as Corollary 3.4 below.

Theorem 1.2. *For $r \in \mathbb{N}$, consider $p \in [2^r, 2^{r+1}]$ and assume that (1.7) holds with an α for which $\alpha\gamma^{2^r-2} < 1$. Assume furthermore that for a fixed $q \leq p$ there is a $\theta(p,q) > 0$ such that for all $f \in B(E)$ and all $1 \leq k \leq n$,*

$$\|K_k(f)\|_{L_p(\mu_k)} < \theta(p,q)\|f\|_{L_q(\mu_k)}. \tag{1.8}$$

Then we have for $0 \leq j < k \leq n$ and $f \in B(E)$ the inequality

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \widetilde{c}(p,q)\|f\|_{L_q(\mu_k)}$$

with

$$\widetilde{c}(p,q) = \theta(p,q)\frac{\gamma^{r+2}}{1-\alpha\gamma^{2^r-2}}$$

Note that in (1.8) we do not require $\theta(p,q) < 1$ so that we only require hyperboundedness but not hypercontractivity. Since we can also bound the other constants in Theorem 1.1 by

$$\overline{c}_k \leq k\, c(p)\, (3 \vee (1+\gamma)) \quad \text{and} \quad \overline{v}_k \leq k\, \widetilde{c}(2,2)^2,$$

we thus obtain a non-asymptotic error bound which is explicit and polynomial in n, in γ and in α. Finally, note that the latter bound on the asymptotic variance $\overline{v}_n$ follows already from Proposition 1.1 and thus does not rely on hyperboundedness: It can be derived using, e.g., a Poincaré inequality and an upper bound on relative densities.

1.4.2.2 Related Work and Discussion

As seen in the previous sections, our stability results depend basically on three conditions: two mixing conditions, namely, an L_2-mixing condition (1.7) and hyperboundedness (1.8), and a uniform upper bound on relative densities. Both mixing conditions are implied by Logarithmic Sobolev inequalities for the MCMC dynamics, since the latter implies hypercontractivity, i.e., hyperboundedness with a constant smaller than 1, and a Poincaré inequality. See Ané et al. (2000) for background and the example in Section 3.5 for concreteness. Our approach to proving stability of the Feynman-Kac semigroups $q_{j,k}$ is an adaption to the discrete-time case of the results derived by Eberle and Marinelli (2010) for the case of continuous time.

Most of the previous literature on stability of Sequential MCMC (see, e.g., Del Moral and Miclo (2000), Theorem 7.4.4 of Del Moral (2005), Cérou, Del Moral and Guyader (2011)) has instead relied on conditions which, in our setting, correspond to the mixing condition

$$K_k(x,\cdot) \leq \lambda\, K_k(y,\cdot) \tag{1.9}$$

for all k, for all x and y in E and for some $\lambda > 1$ and the boundedness condition

$$\overline{g}_{k,k+1}(x) < \kappa\, \overline{g}_{k,k+1}(y) \tag{1.10}$$

for all k, for all x and y in E and for some $\kappa > 1$. As will be pointed out in the following, neither of these conditions is well-suited for the study of MCMC on high-dimensional non-compact state spaces. We begin with discussions of conditions (1.9) and (1.10) and their counterparts in our analysis. We close with a remark on the dimension-dependence of our error bounds.

Before we start, it should be pointed out that, beginning with the central limit theorems in Del Moral (1996), Chopin (2004) and Künsch (2005), there is also by now a rich literature on asymptotic error bounds for the limit $N \to \infty$. See Del Moral (2005) for an overview and many results, and Douc and Moulines (2008) for a recent contribution.

Mixing Conditions

The mixing condition (1.9) is fairly restrictive with regards to the applications we have in mind. For instance, it is never satisfied for dynamics which remain in their initial position with a positive probability and which are continuously distributed otherwise. This is the case, e.g., for Metropolis dynamics on $\mathbb{R}^d$. Moreover, (1.9) is typically not fulfilled for local dynamics on an infinite state space over a finite time horizon: Consider for instance the case where K_k corresponds to t steps of an MCMC dynamics on $\mathbb{R}$ which moves to a new state within a ball of radius r around the current state x in each step. Then $K_k(x, \cdot)$ and $K_k(y, \cdot)$ only have an overlap in their supports if $|x - y| < \frac{r\,t}{2}$. This shows that such a kernel K_k will never satisfy (1.9). Qualitatively, this type of problem persists if we substitute these bounded jumps by other local dynamics: Typically, (1.9) will either be violated or fulfilled with a huge constant.

Unfortunately, so far the literature on discrete-time Markov chains on $\mathbb{R}^d$ does not provide readily available techniques for proving our mixing conditions (1.7) and, especially, (1.8). Exceptions come from the literature on estimating the volume of a convex body, see, e.g., Lovász, Kannan and Simonovits (1997) and – for an exposition of related results which aim at integration instead of volume computation – Rudolf (2009). This literature has largely focused on deriving Poincaré inequalities using conductance techniques and has mainly worked with continuous but compact state spaces. Our mixing conditions can however be verified for continuous-time processes such as Ornstein-Uhlenbeck processes and Langevin diffusions, both of which generally do not satisfy (1.9). Results on these processes can be seen as indicators of the performance we can expect from actual MCMC dynamics, see the example of Section 3.5. Moreover, there is some hope that future research may close this gap.

Recently, Whiteley (2011) proved non-asymptotic error bounds which – unlike those based on (1.9) and (1.10) – can be expected to be applicable to non-trivial models with non-compact state spaces. In Whiteley's setting, the mixing condition (1.9) is replaced by conditions of minorization on a small set and drift conditions outside the small set, applying results of Douc, Moulines and Rosenthal (2004). Minorization and drift conditions are a popular technique for deriving mixing conditions for Markov chains on a general state space, see Roberts and Rosenthal (2004) for an introduction, and Jasra and Doucet (2008) for an earlier application in a Sequential Monte Carlo setting. Additionally, the relative densities $\overline{g}_{k,k+1}$ are assumed to be uniformly bounded from above as in our setting. Finally, there is an assumption which, roughly speaking, bounds the values of $q_{k,n}(1)$ from below outside the tails. Due to the latter condition and due to the fact that Whiteley's proofs switch within an abstract family of minorization and drift conditions, the error bounds are not explicit enough to quantitatively assess the dependence on the model's parameters such as dimension, see the discussion below.

Whiteley's error analysis does however incorporate two problems left open by ours: It includes the case of unbounded integrands f and it treats the case of an initial error, i.e., the case where an error is made when sampling from μ_0, showing that the contribution from the initial error decreases exponentially.

Conditions on relative densities

We now turn to the boundedness condition (1.10). This condition is violated already in simple cases such as the following: Assume that μ_k and μ_{k+1} are Gaussian distributions on $\mathbb{R}$ with means 0 and variances $\sigma_k^2 = 1$ and $\sigma_{k+1}^2 = \delta < 1$. Then the relative density is given by

$$\overline{g}_{k,k+1}(x) = \frac{1}{\sqrt{\delta}} \exp\left(-\left(\frac{1}{\delta} - 1\right)\frac{x^2}{2}\right) \tag{1.11}$$

so that $\overline{g}_{k,k+1}(0)/\overline{g}_{k,k+1}(x)$ is unbounded as $|x|$ gets large. Thus (1.10) is violated. In contrast, our condition of an upper bound γ on $\overline{g}_{k,k+1}$ is fulfilled with $\gamma = 1/\sqrt{\delta}$.

Despite the fact that our assumption of bounded relative densities is weaker than what is usually found in the literature, namely, (1.10), it is still fairly restrictive in a general Sequential Monte Carlo setting: Assume that we wish to track the position of a flying object on $\mathbb{R}$. The initial position is distributed according to $\mu_k = \mathcal{N}(0,1)$ and the position at the next step is distributed according to $\mu_{k+1} = \mathcal{N}(\varepsilon, 1)$ where $\varepsilon \neq 0$ is extracted from some incoming data. In this case, the relative density is given by

$$\overline{g}_{k,k+1}(x) \sim \exp(\varepsilon x) \tag{1.12}$$

which is unbounded in one of the tails. This example is only the most elementary manifestation of a rather general problem when dealing with "moving" probability distributions on $\mathbb{R}^d$.

In Section 3.5 we apply our bounds to a sequence μ_k of d-dimensional Gaussian distributions restricted to a compact box in $\mathbb{R}^d$ where this restriction is necessary to keep relative densities bounded. One crucial observation there is that the one case where the restriction is not necessary is the case where the variance of the distribution is decreased. This can be seen already from comparing examples (1.11) and (1.12). The good news is now that when dealing with Sequential MCMC applications where we freely choose the sequence μ_k to gradually move to a more concentrated distribution μ_n, the case exemplified in (1.11) is indeed the most relevant one: If we choose $\mu_k(dx) \sim \exp(-\beta_k H(x))\mu_0(dx)$ for some increasing sequence β_k, the relative densities $\overline{g}_{k,k+1}$ are given by

$$\overline{g}_{k,k+1} \sim \exp(-(\beta_{k+1} - \beta_k)H(x)).$$

These expressions are bounded from above whenever H is bounded from below. Moreover, the upper bound can be controlled by the choice of $\beta_{k+1} - \beta_k$.

Finally, note that while typically only the unnormalized relative densities $g_{k,k+1}$ are available a priori, the normalized relative densities can be estimated by

$$\overline{g}_{k,k+1}(x) \approx \frac{g_{k,k+1}(x)}{\eta_k^N(g_{k,k+1})}$$

before the resampling step with respect to $g_{k,k+1}$ takes place in the algorithm.

While our overall approach is similar to that of Eberle and Marinelli (2010, 2011) where a continuous-time dynamics on a finite state space is considered, their approach does rely on an assumption similar to (1.10) and thus on both upper and lower bounds on relative densities. Unlike previous results, e.g. those in Del Moral and Miclo (2000), the error bounds in these works are however logarithmic and not polynomial in the constant κ in (1.10). This difference between our results and those in Eberle and Marinelli (2010, 2011) comes from the different resampling schemes: In our algorithm, there is a fixed amount of resampling at fixed times separated by MCMC steps of the dynamics. In Eberle and Marinelli's continuous time particle system, resampling takes the form of particles copying each others' locations at rates which depend on the (log) ratio of relative densities. An upper bound on the ratio of relative densities is used to control the particle system in a short final time interval where only a small number of MCMC steps occurs. This is not necessary in our discrete-time framework.

Dimension Dependence

A particular advantage of our error bounds is that they allow for deriving the algorithm's dimension dependence fairly explicitly for the case where the measures μ_k are product measures on $\mathbb{R}^d$. This can be seen as a first step to understanding the algorithm's overall dimension dependence.

We demonstrate that if we have a one-dimensional setting where our bounds apply, then we can obtain bounds of the same order for the d-dimensional product of the one-dimensional target distribution by increasing the computational effort by a factor of order $O(d^3)$ and thus by a factor which is polynomial in d. Consider a sequence of distributions $\mu_0,\ldots, \mu_n$ on $\mathbb{R}$ such that relative densities are bounded by γ and such that we can obtain sufficiently good mixing properties for the dynamics K_k from Logarithmic Sobolev inequalities.[9] It is well-known that the constants in

[9] The Logarithmic Sobolev constant does not only yield hypercontractivity. It can also be used to bound the Poincaré constant in the right direction, see Ané et al. (2000).

Logarithmic Sobolev inequalities are not dimension-dependent for product measures, see Ané et al. (2000) and the example in Section 3.5. Thus, we obtain the same constants in the mixing conditions for the d-dimensional product dynamics $K_k^{(d)}$ with target $\mu_k^{\otimes d}$ as for the one-dimensional dynamics K_k. Since the d-dimensional relative densities $\overline{g}_{k,k+1}^{(d)}$ are d-fold products of the one-dimensional densities $\overline{g}_{k,k+1}$, we need to increase the number of interpolating distributions by a factor d when switching from $\mathbb{R}$ to $\mathbb{R}^d$. This can be done by inserting $d-1$ additional distributions $\mu_{k,1}^{\otimes d}$ to $\mu_{k,d-1}^{\otimes d}$ between $\mu_k^{\otimes d}$ and $\mu_{k+1}^{\otimes d}$ where, for $1 \leq j < d$, $\mu_{k,j}^{\otimes d}$ is given by

$$\mu_{k,j}^{\otimes d}(dx_1, \dots, dx_d) = \left(\prod_{l=1}^{d} \overline{g}_{k,k+1}(x_l) \right)^{j\, d^{-1}} \mu_k^{\otimes d}(dx_1, \dots, dx_d).$$

Note that $\mu_{k,0}^{\otimes d} = \mu_k^{\otimes d}$ and $\mu_{k,d}^{\otimes d} = \mu_{k+1}^{\otimes d}$. Moreover, the relative densities between $\mu_{k,j}^{\otimes d}$ and $\mu_{k,j+1}^{\otimes d}$ are bounded by γ. Assume furthermore, that the Logarithmic Sobolev constant for the $d-1$ interpolating measures "inserted" between μ_k and μ_{k+1} lies in between the Logarithmic Sobolev constants for $\mu_k^{(d)}$ and $\mu_{k+1}^{(d)}$ so that MCMC with respect to the inserted distributions is not more difficult than MCMC with respect to the original distributions. Now re-index our sequence of $n^{(d)} = n \cdot d + 1$ product measures on $\mathbb{R}^d$ to $\mu_0^{(d)}, \dots \mu_{n_d}^{(d)}$ and denote the associated transition kernels and relative densities by $K_k^{(d)}$ and $\overline{g}_{k,k+1}^{(d)}$. Then, we obtain dimension-independent constants $c_{j,k}(p)$ in the inequalities (1.5) for the semigroup $q_{j,k}^{(d)}$ derived from $K_k^{(d)}$ and $\overline{g}_{k,k+1}^{(d)}$.[10] Notably, these constants coincide with those derived for the original one-dimensional sequence $\mu_0, \dots, \mu_n$ and we have $\mu_n^{\otimes d} = \mu_{n_d}^{(d)}$.

Overall, we then find that the algorithm's error depends on the dimension like $O(d^3)$ in this example: We need one factor d since each step of the product dynamics is more costly to simulate, one factor d due to the fact that we increase the number of levels by a factor d, and one factor d because we have to adjust the number of particles N by the same factor as the number of levels n as can be seen from Theorem 1.1. It is an open question whether similar results can be derived from minorization and drift conditions as in Whiteley (2011).

1.4.3 Multimodality

1.4.3.1 A Motivating Example

We now present our results about the algorithm's ability to cope with multimodality. Since we have already discussed the problem and the algorithm extensively in Section 1.3, this brief motivating section is mainly dedicated to motivating our approach to the problem. Our results and a discussion are provided afterwards.

Figure 1.2 depicts the particle positions before the resampling step in a run of the algorithm where the target μ_n is the Gaussian mixture distribution in Figure

[10] Recall that our bounds on the constants $c_{j,k}(p)$ were independent of j and k.

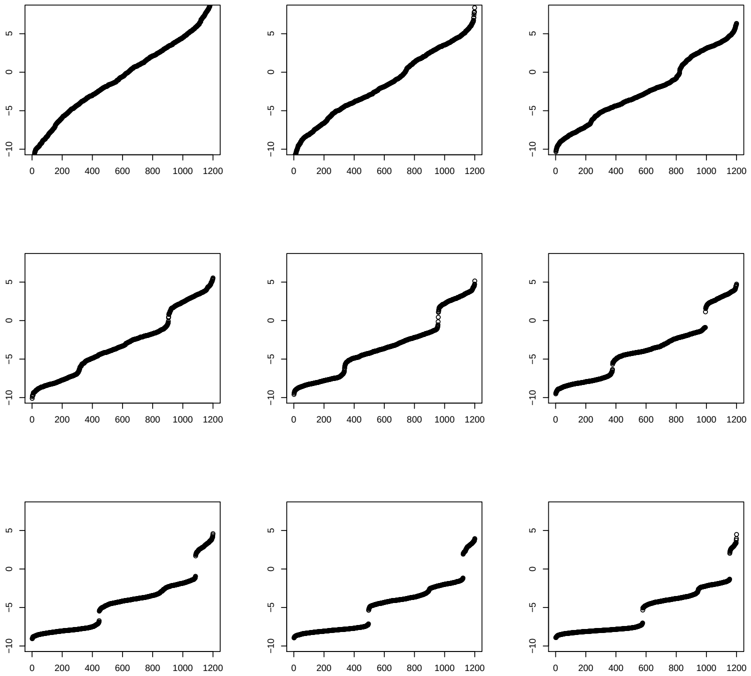

Figure 1.2: Particle Movements

1.1 on page 8 above.[11] Each picture corresponds to a level μ_k, $k = 0, \ldots, 8$. On the horizontal axis are the particles ordered by the position. On the vertical axis are the particle positions. Several features of the algorithm can be seen from the picture: All four modes are covered by the algorithm and in the final picture the numbers of particles in the four modes roughly correspond to the target proportions $(0.5, 0.3, 0.15, 0.05)$ in the modes at $(-8, -4, -2, 2)$. We can see that the target distribution's four modes become visible successively: Until μ_3, all particles seem to move in one big cloud. At μ_5, the particles in the mode at 2 are separated by a gap from the rest. At $\mu_/$, the same happens for the particles at the mode in -8. While the particles rarely move between modes through the MCMC steps, there is still a substantial exchange of probability mass between modes. For instance, there are about 400 particles in the mode in -8 when the mode is disconnected from the rest. At the end, there is approximately the right number of 600 particles in that mode.

Comparing with different specifications of the algorithm not reported here, one thing becomes quite apparent: The particles within one mode usually give a fairly accurate picture of that mode, even with much fewer particles and MCMC steps. The critical question is whether each mode has the correct weight, i.e. whether

[11] We chose $n = 9$. The nine values of β were $(0.02, 0.05, 0.1, 0.18, 0.3, 0.4, 0.64, 0.8, 1)$. There were 1200 particles. At each level, each particle took 400 steps of Metropolis with proposals from a Gaussian distribution with variance 0.2.

it contains approximately the right number of particles. This seems to be the first thing to go wrong, and it is a serious problem because, generally, there is little use to calculating the correct integral within each mode but then to assemble these "local" integrals to a "global" integral with the wrong weights.

1.4.3.2 Results

Motivated by observations such as the example just presented, our first error bound for the multimodal case focuses on a simplified model where we project each component of the state space in which the MCMC dynamics moves freely to a single point and only consider the algorithm on this projection. These are the results of Chapter 4. A second error bound which takes into account local mixing is presented in the second part of this section. Since the results here may need more context than the previous ones, part of the discussion is provided directly in this section while the rest is in Sections 1.4.3.3 and 1.4.3.4 below.

Taking into account that the state space will typically separate into more and more effectively disconnected components as we move from μ_0 to μ_n, we consider our Sequential MCMC algorithm on trees: At each level k the (projected) state space we consider now consists of a number of points. Each point at level k has at least one successor at level $k+1$. Each successor stands for one disjoint component of the original state space which can only be reached from its predecessor component at level k.[12] The role of the transition kernels K_{k+1} is limited to allocating particles from one point at level k to its successors at the next level $k+1$. We assume that particles cannot move between points at the same level. The latter assumption is in accordance with the fact that transitions between effectively disconnected components of the original unprojected state are only very rarely observed in practical examples. We do not use any mixing properties of the MCMC dynamics since such properties can only be expected to have an effect *within* each disconnected component – they will not help to correct errors made in allocating particles to modes.

To make this concrete, denote by $(\mu_0, \ldots, \mu_n)$ a sequence of probability measures on a sequence of finite state spaces $(I_0, \ldots, I_n)$. Consider a successor relation $s_k : \bigcup_{l<k} I_l :\to \mathcal{P}(I_k)$ which maps $x \in I_l$, $l < k$, to its set of successors $s_k(x) \subseteq I_k$ at level k. s_k has the following properties: For $l < k$ and $x \neq y \in I_l$ we have $s_k(x) \cap s_k(y) = \emptyset$, i.e., each point in I_k is a successor of at most one point in I_l. Moreover for each $u \in I_k$ there exists a $z \in I_l$ with $u \in s_k(z)$, i.e., each point in I_k has a predecessor in I_l. Additionally, we make the transitivity assumption that, for $j < k < l$ and $x \in I_j$, $y \in s_k(x)$ implies $s_l(y) \subseteq s_l(x)$. Thus we have indeed a tree structure (or, more accurately, a forest structure since we do not assume $|I_0| = 1$). Finally, we assume that no branch of the tree dies out, $s_n(x) \neq \emptyset$ for all $x \in I_j$ and all $0 \leq j < n$.

[12] This should make clear that the trees we consider here are only very loosely related to the genealogical trees of the particle system studied in Cérou, Del Moral and Guyader (2011).

For instance, to capture what happens in the example of Figure 1.2 we could choose $I_0 = \ldots = I_3 = \{0\}$, $I_4 = \{1, 2\}$ and $I_5 = \ldots = I_8 = \{1, 3, 4\}$ where each state stands for one (effectively) disconnected region. The measures μ_k on these reduced state spaces associate with each state the mass of the respective region at level k under the original target distribution on $\mathbb{R}$. Moreover, the kernels K_k take into account that in the transition from μ_4 to μ_5, when the number of states increases from two to three, two of the states in I_5 can be reached only from one of the states in I_4 and are thus its successors. 1 and 2 are successors of 0, and 3 and 4 are successors of 2 (and of 0).

Denote by $B(I_k)$ the bounded functions from I_k to $\mathbb{R}$. Let K_{k+1} be a transition kernel from I_k to I_{k+1} with the property that for $x \in I_k$, $K_{k+1}(x, \cdot)$ is a probability distribution on $s_{k+1} \subseteq I_{k+1}$, i.e., transitions only go to successors of a state. Moreover, let the unnormalized relative density $g_{k,k+1} \in B(I_k)$ be such that

$$\mu_{k+1}(f) = \frac{\mu_k(g_{k,k+1} K_{k+1}(f))}{\mu_k(g_{k,k+1})} \quad \text{for all } \ f \in B(I_{k+1}).$$

Define $q_{k,k+1} : B(I_{k+1}) \to B(I_k)$ by

$$q_{k,k+1}(f) = \frac{g_{k,k+1} K_{k+1}(f)}{\mu_k(g_{k,k+1})}.$$

Moreover for $j < k$ let $q_{j,k} : B(I_k) \to B(I_j)$ be given by

$$q_{j,k}(f) = q_{j,j+1}(q_{j+1,j+2}(\ldots q_{k-1,k}(f))) \quad \text{and} \quad q_{k,k}(f) = f.$$

Define the weighted particle measures ν_j^N as in Section 1.4.1.1. We study again the approximation error of ν_j^N and use it to bound the approximation error of the particle measure η_j^N from the algorithm using (1.2).[13]

Our main result shows that in this reduced setting the algorithm's approximation error can be controlled in terms of a constant which captures how strongly the components gain probability mass over time. Roughly speaking, we show that the algorithm works well if for all $j < k$ no disconnected component under μ_j carries much less weight than its successors under μ_k. The intuition for this is straightforward: If a component $x \in I_j$ is much less important than its successors in I_k, there is a substantial probability that there are no particles in x. If $\mu_j(x)$ is small, we may then still have a reasonable particle approximation of μ_j. But if we miss x we also miss its successors in I_k and if these are important we obtain a bad approximation of μ_k: Transition states with small weight create a bottleneck for the particle dynamics.

The error bound is based on a variation of Theorem 1.1 where we choose as the sequence of norms $\|\cdot\|_j$ the supremum norms on the spaces $B(I_j)$. Let M be the

[13] We cannot directly apply these results, since in the present setting the μ_k do not live on the same state space and since K_k is not stationary with respect to μ_k. The more general results proved in Chapter 2 are however completely analogous and cover this case.

largest gain in weight of a state in comparison with its successors:

$$M = \max_{0 \le j < k \le n} \max_{x \in I_j} \frac{\mu_k(s_k(x))}{\mu_j(x)}.$$

M has the property that for all $f \in B(I_k)$

$$\|q_{j,k}(f)^2\|_j = \|q_{j,k}(f)\|_j^2 \le M \, \|f\|_k^2.$$

Thus M can be used as a constant in condition (1.5) of Theorem 1.1 and we obtain the following result:[14]

Theorem 1.3. *Let $N \ge 2\, M\, (n+1)$. Then for $f \in B(I_n)$ we have*

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \le \frac{\sum_{j=0}^n \mathrm{Var}_{\mu_j}(q_{j,n}(f))}{N} + \frac{2\, M^3 (n+1)^2 \|f\|_n^2}{N^2}.$$

Moreover, we have the bound

$$\sum_{j=0}^n \mathrm{Var}_{\mu_j}(q_{j,n}(f)) \le M\, (n+1)\, \|f\|_n^2.$$

Thus we can indeed control the error in terms of M, n and the maximum of f. While this theorem is not designed for proving good performance of Sequential MCMC in concrete examples, it allows to draw a number of heuristic conclusions: Sequential MCMC can only work well if no component of the state space becomes important after it has been essentially separated from the rest of the state space. Moreover, in Section 4.4 we provide a detailed analysis of a tree model in which Sequential MCMC with Resampling works well while the error of Sequential Importance Sampling increases exponentially over time. This shows that resampling with a finite number of particles can indeed overcome difficulties associated with multimodality in settings where Sequential Importance Sampling fails.

Our second error bound for the case without global mixing aims at combining the ideas behind Theorem 1.3 with the L_p-stability analysis outlined in Section 1.4.2.1. These are the results of Chapter 5. We return to a sequence of mutually absolutely continuous measures $(\mu_k)_k$ on a common state space E with unnormalized relative densities $g_{k,k+1} \in B(E)$ and transition kernels $K_k : B(E) \to B(E)$ which are stationary with respect to μ_k. The main difference to the analysis of Section 1.4.2.1 is that we rely on a different set of norms under which stability of the semigroup $q_{j,k}$ can be derived from *local* mixing properties. Instead of the tree structure in Theorem 1.3 we now consider a sequence of increasingly finer partitions of E.

To this end, denote by $(F_i)_{i \in I}$ a collection of subsets of E where I is a finite index set and $\mu_0(F_i) > 0$ for all $i \in I$. Consider a sequence $(I_k)_{k=0}^n$ of index sets with

[14] The theorem corresponds to Corollary 4.1 below. It relies on a variation of Theorem 1.1, namely, Theorem 2.2.

$I_k \subseteq I$ and with the property that $(F_i)_{i \in I_k}$ is a disjoint partition of E for all k. Assume that partitions become increasingly finer in the sense that for $j < k$ and for all $i \in I_j$ there exists a subset $s_k(i) \subset I_k$ such that $(F_l)_{l \in s_k(i)}$ is a disjoint partition of F_i. This is basically the tree structure of Theorem 1.3 again. Yet this time it appears only as a sequence of index sets and not as the algorithm's state space.

We want to apply Theorem 1.1 to a setting where the MCMC kernels K_k mix well locally on each set F_i, $i \in I_k$ but not globally. To formulate this we need some additional notation. For $i \in I$ and $0 \leq k \leq n$ denote by $\mu_{k,i}$ the restriction of μ_k to F_i, i.e., the measure on E given by

$$\mu_{k,i}(f) = \frac{\mu_k(f 1_{F_i})}{\mu_k(F_i)}$$

for all $f \in B(E)$. We choose the following sequence of norms on E:

$$\|f\|_{k,p} = \max_{i \in I_k} \|f\|_{L_p(\mu_{k,i})} \quad \text{for } f \in B(E), p \geq 1, 0 \leq k \leq n.$$

The norms $\|\cdot\|_{k,p}$ thus combine local L_p norms with a maximum norm over components.

With this choice of norms, proving the crucial inequality (1.5) in Theorem 1.1 and thus a non-asymptotic error bound is reduced to finding constants

$$c_{j,k}(p) = \max\left(\widetilde{c}_{j,k}\left(p, \frac{p}{2}\right), \widetilde{c}_{j,k}\left(2p, p\right)^2\right),$$

where for $p \geq q \geq 1$ we define $\widetilde{c}(p,q)$ as a constant in an L_p-L_q inequality for the semigroup $q_{j,k}$, i.e., a constant such that

$$\|q_{j,k}(f)\|_{j,p} \leq \widetilde{c}_{j,k}(p,q) \|f\|_{k,q}$$

is satisfied for all $f \in B(E)$ and all $j < k$.

We make the following assumptions: Denote by $\overline{g}_{k,k+1,i} \in B(E)$ the normalization of $\overline{g}_{k,k+1}$ which makes it a relative density between the restricted measures $\mu_{k,i}$ and $\mu_{k+1,i}$, i.e.,

$$\overline{g}_{k,k+1,i}(x) = \frac{\mu_k(F_i)}{\mu_{k+1}(F_i)} \overline{g}_{k,k+1}(x) 1_{F_i}(x) \quad \text{for all} \quad x \in E.$$

We assume the following uniform upper bound on (local) relative densities: There exists a γ such that for all $0 \leq k < n$ and for all $i \in I_k$

$$\overline{g}_{k,k+1,i}(x) \leq \gamma \quad \text{for all} \quad x \in F_i.$$

We assume that the MCMC dynamics K_k does not move between the components

$(F_i)_{i\in I_k}$ of the state space at level k: For all $0 \le k \le n$ and for all $j \in I_k$

$$K_k(1_{F_j})(x) = 0 \quad \text{for all } \; x \in E \setminus F_j.$$

This assumption considerably simplifies the analysis, since it implies that if $f \in B(E)$ has support only in F_j, $j \in I_k$, then $K_k(f)$ also has support only in F_j. Moreover, while it will not be satisfied in most applications, it is very much in line with the idea that the sets F_j are separated by barriers which preclude good global mixing of the K_k: We make the technical assumption that transitions between separated local modes – which are possible but very rare – never occur. This leads to the following stability result for the semigroup $q_{j,k}$ which can be used to determine $c_{j,k}(p)$ and thus to deduce a non-asymptotic error bound from Theorem 1.1:[15]

Theorem 1.4. *Assume there are constants $\alpha > 0$ and $\beta \in [0,1]$ such that for all $f \in B(E)$ for all $0 \le k \le n$ and all $i \in I_k$ we have the following L_2-bound,*

$$\|K_k(\overline{g}_{k,k+1,i}f)\|^2_{L_2(\mu_{k,i})} \le \alpha\|f\|^2_{L_2(\mu_{k+1,i})} + \beta\mu_{k+1,i}(f)^2. \tag{1.13}$$

For $r \in \mathbb{N}$, consider $p \in [2^r, 2^{r+1}]$ and assume that (1.13) holds with an α for which $\alpha\gamma^{2^r-2} < 1$. Assume furthermore that for $q > p$ there is a $\theta(q,p) > 0$ such that for all $f \in B(E)$ and for all $i \in I_j$

$$\|K_j(f)\|_{L_q(\mu_{j,i})} < \theta(q,p)\|f\|_{L_p(\mu_{j,i})}. \tag{1.14}$$

Then we have for $0 \le j < k \le n$ and $f \in B(E)$ the inequality

$$\|q_{j,k}(f)\|_{j,q} \le \widetilde{c}_{j,k}(q,p)\|f\|_{k,p}$$

with

$$\widetilde{c}_{j,k}(q,p) = A_{j,k}\,\theta(q,p)\,\frac{\gamma^{r+2}}{1-\alpha\gamma^{2^r-2}}$$

where

$$A_{j,k} = \max_{i\in I_k}\prod_{l=j}^{k-1}\frac{\mu_{l+1}(F_{p_l(i)})}{\mu_l(F_{p_l(i)})}. \tag{1.15}$$

Here, for $l < k$, $p_l(i) \in I_l$ denotes the predecessor of $i \in I_k$ in I_l, i.e., the unique element $p_l(i) \in I_l$ for which $i \in s_k(p_l(i))$.

Comparing with the result under global mixing in Theorem 1.2, the present result differs in the following ways: We have replaced the global mixing conditions (1.7) and (1.8) by analogous local conditions (1.13) and (1.14) which, so to say, require good mixing only around each mode. Specifically, we require L_2-mixing and hyperboundedness of the K_k with respect to the restricted measures $\mu_{k,i}$ for all $i \in I_k$. Both conditions can be derived from local Logarithmic Sobolev inequalities. Similarly to the constant M in Theorem 1.3 we get an additional factor $A_{j,k}$ in our

[15] The result is an immediate consequence of Corollary 3.4 below.

bound which takes into account how probability mass is shifted between the separated components of the state space. $A_{j,k}$ is the maximal product of relative mass changes one has to go through when moving from some component F_i, $i \in I_j$ to one of its successors F_l, $l \in s_k(j)$. $\max_{j<k} A_{j,k}$ is generally larger than M since, roughly speaking, we compare F_i to its worst successor and not to its set of successors. This is the price we pay for taking into account local mixing and local variations in relative densities from which we abstracted in Theorem 1.3. Note also that in the case with only one component at each level, $|I_k| = 1$ for all k, we have $A_{j,k} = 1$ and the norms reduce to L_p-norms on E. Theorem 1.2 is thus a special case of Theorem 1.4.

1.4.3.3 Related Work

The idea of reducing complicated multimodal distributions to trees, also known as disconnectivity graphs, has been studied extensively in the chemical physics literature, see Chapter 5 of Wales (2003) for an introduction. In the Sequential Monte Carlo literature, the only precursors of our results appear to be in Eberle and Marinelli (2010, 2011) who consider the continuous time case and restrict attention to the case of forests where each disconnected component under μ_0 has exactly one successor component at each level. This corresponds to the case of $I_0 = \ldots = I_n$ in the setting of our Theorem 1.4. The constant $A_{j,k}$ in Theorem 1.4 reduces to the constant they find in this case, cf. Theorem 2.10 in Eberle and Marinelli (2011).

A number of related results for the Parallel Tempering algorithm[16] have been proved, in increasing generality, by Madras and Zheng (2002), Bhatnagar and Randall (2004) and Woodard, Schmidler and Huber (2009a, 2009b). Technically, these results rely on decomposition results for bounding spectral gaps of Markov chains, namely on an unpublished result of Caracciollo, Pelissetto and Sokal (1992) which was first published and extended in Madras and Randall (2002), see also Jerrum, Son, Tetali and Vigoda (2004). These decomposition results have the advantage that they do not rely on the assumption that the MCMC dynamics does not move between effectively disconnected components which we made. Therefore, these results can be applied directly to some simple of interest such as the mean field Ising model.

All these results on Tempering are restricted to simple trees with one node at the origin and a number of branches which do not branch further at later levels. This corresponds to the case of $I_0 = 1$ and $I_1 = \ldots = I_n$ in the setting of Theorem 1.4. Figure 1.2 above demonstrates that our more general trees arise naturally in applications, see also Wales (2003) for many examples from chemical physics.

1.4.3.4 Discussion

As pointed out previously, a drawback of Theorems 1.3 and 1.4 is that we assume that the MCMC dynamics never moves between components of the state space

[16] See Section 1.3.3 for a discussion of Tempering algorithms and their relation to Sequential MCMC.

which are separated by regions of low probability. This technical assumptions has the consequence that we cannot hope to actually apply the results even to the usual toy models. Thus, some words of defense seem in order.

First, in the majority of practical applications and even for simple MCMC algorithms there is no way to just check a number of mixing assumptions and then deduce a computationally feasible required running time of the algorithms. Two reasons for this are that most available error bounds are not sharp enough and that in many applications the target distribution is essentially a "black box" for which quantities such as a spectral gap are difficult to obtain. One might argue that most existing research on error bounds for MCMC aims at deriving abstract characterizations of settings in which the algorithms work well or not so well. This is also the spirit behind Theorem 1.3 and Theorem 1.4.

Second, there seems to be little reason to expect that we improve the error by, essentially, setting some (small) transition probabilities of the MCMC dynamics to zero: While the assumption of no transitions between disconnected components makes the error easier to bound, it should rather make the error itself larger than smaller. Basically, we can derive a model fulfilling our assumptions from another model by setting, e.g., all probabilities below a certain threshold to zero. With very high probability, this change would not even affect our simulations. Rigorous results along these lines would be an important next step.

Since Theorem 1.3 abstracts from most of the local structure – and thus from some possible problems – it should be seen as a rough but intuitive criterion for identifying settings where the algorithm works or does not work. Consider for instance the mean field Potts model for which slow mixing of the Tempering algorithms was proved by Bhatnagar and Randall (2004), see their paper also for more details about the model. Basically, in this model there is a distribution μ_0 which is unimodal and a distribution μ_n which has four modes of roughly equal weight. Along the transition from μ_0 to μ_n, three additional modes arise which are immediately well-separated from the initial one and have a tiny initial mass, say ε. Thus, we obtain a huge constant M of order $O(\varepsilon^{-1})$ in our error bound.[17]

For the mean field Ising model, Madras and Zheng (2002) proved rapid mixing of Tempering algorithms. Theorem 1.3 suggests that the same should be true for Sequential MCMC: In the mean field Ising model there is basically one mode which is split into two modes of equal weight at some point as we move from μ_0 to μ_n. In this case, each mode has exactly the same weight as its successors. Thus we can expect a good performance of the algorithm.

Another example, where a similar behavior can be expected, is the problem of estimating the parameters of mixture distributions described in Celeux, Hurn and Robert (2000). There, the target distribution μ_n is a distribution on the parameter

[17] Recall that the leading coefficient in the error bound corresponds to the asymptotic variance so that we have indeed more than just an upper bound. For more discussion, see the end of Section 4.2.2 beginning with Proposition 4.1.

space $\mathbb{R}^{l\times k}$ with the symmetry property that all permutations of the rows of a given $\theta \in \mathbb{R}^{l\times k}$ have the same probability under μ_n. Without going into further detail here, the source of the multimodality problem is that, e.g., the Gaussian mixture distributions

$$0.25\mathcal{N}(5,1) + 0.75\mathcal{N}(0,1) \text{ and } 0.75\mathcal{N}(0,1) + 0.25\mathcal{N}(5,1)$$

are identical, i.e., that some permutations of the parameters correspond to the same mixture distribution. The target distribution μ_n of MCMC is a posterior distribution on the parameter space and thus assigns the same weight to $\theta_1 = (0.25, 5, 1; 0.75, 0, 1) \in \mathbb{R}^{2\times 3}$ and $\theta_2 = (0.75, 0, 1; 0.25, 5, 1) \in \mathbb{R}^{2\times 3}$. See also the related example in Section 1.3.1. Theorem 1.3 suggests that if this type of permutation symmetry is the sole source of multimodality, Sequential MCMC should work well, since the symmetry is retained when tempering the target distribution. Thus, the areas around each local mode have the same weight at all "temperatures". This intuition is confirmed by simulations of Celeux, Hurn and Robert (2000) who study an example along these lines and demonstrate that Simulated Tempering can move between local modes while simple MCMC cannot. Permutation symmetries are also one source of multimodality in models from chemical physics, see Wales (2003). Another message of Theorem 1.3 is however that multimodality caused by permutation symmetries is one of the easiest to deal with cases of multimodality. Thus, examples of this type are rather limited toy examples for testing a multilevel MCMC algorithm's ability to move between disconnected modes.

Theorem 1.4 conveys basically the same intuition as Theorem 1.3 but it explicitly takes into account the aspects left out in Theorem 1.3, notably, local mixing and sufficient similarity of μ_k and μ_{k+1} *within* disconnected components. Both aspects are important to keep in mind: Assume we choose $n = 1$, let μ_0 be a distribution for which we have excellent global mixing and let μ_1 be an arbitrary other distribution which is strongly multimodal. This setting can be projected to a tree where a number of leafs branch from a single root. Then Theorem 1.3 seems to suggest, that Sequential MCMC with only these two distributions should work very well, since the root of the tree has mass 1 under μ_0 and its successors have mass 1 under μ_1. This – obviously false – conclusion can be drawn from Theorem 1.3 since the theorem does not take into account local variations in relative densities which may lead to huge errors in the resampling step. Basically, Theorem 1.3 makes the – implicit – assumption that relative densities are constant within each component. Similar "wrong intuitions" can be derived from Theorem 1.3 by disregarding the fact that local mixing has to be guaranteed within each component.

On a related note, Theorem 1.4 also shows that problems of the algorithm which stem from disconnected components gaining mass can generally not be alleviated by increasing the number of interpolating distributions: Adding additional steps in the sequence $\mu_0, \ldots, \mu_n$ can only increase the constant M. This separates this type of problem from problems associated with large local variations in relative densities in the presence of good global mixing, see the discussion at the end of Section 1.4.2.2.

The only way to control the constants M and $A_{j,k}$ seems to be to choose an entirely different sequence $\mu_0, \dots, \mu_{n-1}$.

To conclude, an important message of Theorems 1.3 and 1.4 is that, generally, a bad performance of Sequential MCMC is not a property of the target distribution μ_n but a property of the approximating sequence $\mu_0, \dots, \mu_{n-1}$ which is a parameter in the algorithm, not in the problem of interest. So far, the mathematical literature on multilevel MCMC algorithms has largely focused on flattening a target distribution by tempering. In the applied literature, there are many more, sometimes model-specific, proposals for choosing a sequence of distributions such as cutting off the Hamiltonian at chosen minimum levels in addition to tempering, varying the system size or spatial coarse graining, see, e.g. Kou, Zhou and Wong (2006), Liu and Sabatti (1998) or Lyman, Ytreberg and Zuckerman (2006). A more systematic study of methods for approximating target distributions seems to be an important and highly challenging task for future research.

2 The Quadratic Error of Sequential MCMC

In this chapter we introduce the Sequential MCMC setting we are interested in and derive general versions of our non-asymptotic error bounds in terms of stability properties of the Feynman-Kac semigroup $q_{j,k}$ associated with the algorithm's particle dynamics. Basically, the analysis of the later chapters is dedicated to verifying these stability properties in more concrete settings and to then apply the error bounds proved below. Section 2.1 introduces the notation, the model and the interacting particle system simulated in the algorithm. Section 2.2 introduces a suitably weighted version of the particles' occupation measure. We show how to control the approximation error with respect to the original particle measures in terms of the error of these weighted particle measures and give an explicit formula for the quadratic error of the weighted measures. Section 2.3 proves our main error bound. Section 2.4 gives an alternative error bound which yields slightly better constants in the setting of Chapter 4 and provides some discussion.

The setting we introduce here is slightly more general than the one discussed in the introduction: We consider a sequence of distributions $(\mu_k)_k$ which live on a sequence of state spaces instead of a common state space. Moreover, we do not assume that the transition kernels K_k are stationary with respect to the measures μ_k. Instead, we only assume that the combination of applying the weight function $\overline{g}_{k-1,k}$ and then the kernel K_k leads from μ_{k-1} to μ_k, see Section 2.1.2 for details. This more general framework has the advantage that it covers both the Sequential MCMC algorithm of Section 1.2, which is studied in Chapters 3 and 5, and the stylized version of the algorithm studied in Chapter 4. In addition, the more general framework here may be of interest in some further problems besides the study of our algorithm, see, e.g., Section 1.4 of Del Moral (2005) for a short overview of other applications.

Throughout, we assume the integrands f to be bounded. This assumption becomes necessary at a crucial place in this chapter, namely when transferring results from the weighted particle measure to the original one in Lemma 2.2. For this reason, we restrict attention to bounded integrands, avoiding to introduce systems of integrability conditions (as is done, e.g., in Chapter 9.2 of Cappé, Moulines and Rydén). However, the results on the weighted particle measures ν_n^N do not rely on this boundedness and could be generalized along such lines.

2.1 Preliminaries

2.1.1 Notation

Let (E, r) be a Polish space and let $\mathcal{B}(E)$ be the σ-algebra of Borel subsets of E. Denote by $M(E)$ the space of finite signed Borel measures on E. Let $M_1(E) \subset M(E)$ be the subset of all probability measures. Let $B(E)$ be the space of bounded, measurable, real-valued functions on E.

For $\mu \in M(E)$ and $f \in B(E)$ define $\mu(f)$ by

$$\mu(f) = \int_E f(x)\mu(dx)$$

and $\mathrm{Var}_\mu(f)$ by

$$\mathrm{Var}_\mu(f) = \mu(f^2) - \mu(f)^2.$$

Let $(\widetilde{E}, \widetilde{r})$ be another Polish space. Consider an integral operator $K(x, A)$ with $K(x, \cdot) \in M(\widetilde{E})$ for $x \in E$ and $K(\cdot, A) \in B(E)$ for $A \in \mathcal{B}(\widetilde{E})$. We define for $\mu \in M(E)$ the measure $\mu K \in M(\widetilde{E})$ by

$$\mu K(A) = \int_E K(x, A)\mu(dx) \quad \forall A \in \mathcal{B}(\widetilde{E}).$$

For $f \in B(\widetilde{E})$ we denote by $K(f) \in B(E)$ the function given by

$$K(f)(x) = K(x, f) = \int_E f(z)K(x, dz) \quad \forall x \in E.$$

2.1.2 The Measure-Valued Model

Consider a sequence of Polish spaces (E_k, r_k) and a sequence of probability measures $(\mu_k)_{k=0}^n$, $\mu_k \in M_1(E_k)$. This is the sequence of measures we wish to approximate with the algorithm introduced in Section 2.1.3. The measures μ_k are related through

$$\mu_k(f) = \frac{\mu_{k-1}(g_{k-1,k}K_k(f))}{\mu_{k-1}(g_{k-1,k})} \quad \forall f \in B(E_k)$$

for positive functions $g_{k-1,k} \in B(E_{k-1})$ and transition kernels K_k with $K_k(x, \cdot) \in M_1(E_k)$ for $x \in E_{k-1}$ and $K_k(\cdot, A) \in B(E_{k-1})$ for $A \in \mathcal{B}(E_k)$. We define the probability distribution $\hat{\mu}_k \in M_1(E_{k-1})$ by

$$\hat{\mu}_k(f) = \frac{\mu_{k-1}(g_{k-1,k}f)}{\mu_{k-1}(g_{k-1,k})} \quad \forall f \in B(E_{k-1}).$$

This implies $\hat{\mu}_k(K_k(f)) = \mu_k(f)$ for $f \in B(E_k)$.

Next we introduce the Feynman-Kac semigroup $q_{j,k}$ which will be the central object of our error analysis. Define the mapping $q_{k-1,k} : B(E_k) \to B(E_{k-1})$ by

$$q_{k-1,k}(f) = \frac{g_{k-1,k}K_k(f)}{\mu_{k-1}(g_{k-1,k})}$$

Observe that this implies

$$\mu_k(f) = \mu_{k-1}(q_{k-1,k}(f))$$

Furthermore define for $0 \leq j < k \leq n$ the mapping $q_{j,k} : B(E_k) \to B(E_j)$ by

$$q_{j,k}(f) = q_{j,j+1}(q_{j+1,j+2}(\dots q_{k-1,k}(f)))$$

and $q_{k,k}(f) = f$. Note that for $f \in B(E_k)$ we have the relation

$$\mu_j(q_{j,k}(f)) = \mu_k(f) \quad \text{for } 0 \leq j \leq k \leq n$$

and the semigroup property

$$q_{j,l}(q_{l,k}(f)) = q_{j,k}(f) \quad \text{for } 0 \leq j < l < k \leq n.$$

2.1.3 The Interacting Particle System

In the Sequential MCMC algorithm, we approximate the measures $(\mu_k)_k$ by simulating the interacting particle system introduced in the following. We start with N independent samples $\xi_0 = (\xi_0^1, \dots, \xi_0^N)$ from μ_0. The particle dynamics alternates two steps: Importance Sampling Resampling and Mutation: A vector of particles ξ_{k-1} approximating μ_{k-1} is transformed into a vector $\hat{\xi}_k$ approximating $\hat{\mu}_k$ by drawing N conditionally independent samples from the empirical distribution of ξ_{k-1} weighted with the functions $g_{k-1,k}$. Afterwards, $\hat{\xi}_k$ is transformed into a vector ξ_k approximating μ_k by moving the particles $\hat{\xi}_k^i$ independently with the transition kernel K_k.

We thus have two arrays of random variables $(\xi_k^j)_{0\leq k\leq n, 1\leq j\leq N}$ and $(\hat{\xi}_k^j)_{1\leq k\leq n, 1\leq j\leq N}$ where ξ_k^j and $\hat{\xi}_{k+1}^j$ take values in E_k. Denote respectively by $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ probabilities and expectations taken with respect to the randomness in the particle system, i.e., with respect to the random variables $(\xi_k^j)_{k,j}$ and $(\hat{\xi}_k^j)_{k,j}$. The random variables $\xi_0^1, \dots, \xi_0^N$ are independent and distributed according to μ_0. The distributions of the remaining $\hat{\xi}_k^j$ and ξ_k^j are pinned down by the transition probabilities

$$\mathbb{P}[\hat{\xi}_k \in dx | \xi_{k-1} = z] = \prod_{j=1}^{N} \sum_{i=1}^{N} \frac{g_{k-1,k}(z^i)}{\sum_{l=1}^{N} g_{k-1,k}(z^l)} \delta_{z^i}(dx^j)$$

and

$$\mathbb{P}[\xi_k \in dx | \hat{\xi}_k = z] = \prod_{j=1}^{N} K_k(z^j, dx^j).$$

Denote by $\mathcal{F}_k$ the σ-algebra generated by $\xi_0, \dots \xi_k$ and $\hat{\xi}_1, \dots \hat{\xi}_k$ and denote by η_k^N

the empirical measure of ξ_k, i.e.

$$\eta_k^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_k^i}.$$

In the following we will study, how well η_k^N approximates μ_n. We end the preliminary observations with the following lemma:

Lemma 2.1. *We have for $f \in B(E_k)$ and $1 \leq k \leq n$*

$$\mathbb{E}[\eta_k^N(f)|\mathcal{F}_{k-1}] = \frac{\eta_{k-1}^N(q_{k-1,k}(f))}{\eta_{k-1}^N(q_{k-1,k}(1))}$$

and for $1 \leq j \leq N$

$$\mathbb{E}[f(\xi_k^j)|\mathcal{F}_{k-1}] = \frac{\eta_{k-1}^N(q_{k-1,k}(f))}{\eta_{k-1}^N(q_{k-1,k}(1))}$$

Proof. Denote by $\hat{\mathcal{F}}_k$ the σ-algebra generated by $\xi_0, \dots \xi_{k-1}$ and $\hat{\xi}_1, \dots \hat{\xi}_k$. Note that $\mathcal{F}_{k-1} \subseteq \hat{\mathcal{F}}_k \subseteq \mathcal{F}_k$. We can thus write

$$\begin{aligned}
\mathbb{E}[f(\xi_k^j)|\mathcal{F}_{k-1}] &= \mathbb{E}[\mathbb{E}[f(\xi_k^j)|\hat{\mathcal{F}}_k]|\mathcal{F}_{k-1}] \\
&= \mathbb{E}[K_k(\hat{\xi}_k^j, f)|\mathcal{F}_{k-1}] \\
&= \frac{\sum_{i=1}^N g_{k-1,k}(\xi_{k-1}^i)K_k(\xi_{k-1}^i, f)}{\sum_{l=1}^N g_{k-1,k}(\xi_{k-1}^l)} \\
&= \frac{\eta_{k-1}^N(g_{k-1,k}K_k(f))}{\eta_{k-1}^N(g_{k-1,k})} \\
&= \frac{\eta_{k-1}^N(q_{k-1,k}(f))}{\eta_{k-1}^N(q_{k-1,k}(1))}.
\end{aligned}$$

This proves the claim for $\mathbb{E}[f(\xi_k^j)|\mathcal{F}_{k-1}]$ and immediately implies the claim for $\mathbb{E}[\eta_k^N(f)|\mathcal{F}_{k-1}]$. □

2.2 Variances of Weighted Empirical Averages

We are interested in finding efficient upper bounds for the quantities

$$\mathbb{E}[|\eta_n^N(f) - \mu_n(f)|^2]$$

and

$$\mathbb{E}[|\eta_n^N(f) - \mu_n(f)|].$$

As is shown in Lemma 2.2 below, these quantities can be controlled in terms of the approximation error of a weighted empirical measure $\nu_n^N(f)$ which is easier to handle. In this section, we introduce the measure $\nu_n^N(f)$ and establish an explicit formula for its quadratic error

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2].$$

Define for $0 \leq k \leq n$

$$\nu_k^N(f) = \varphi_k \, \eta_k^N(f)$$

where φ_k is given by

$$\varphi_k = \prod_{j=0}^{k-1} \eta_j^N(q_{j,j+1}(1))$$

for $1 \leq k \leq n$ and $\varphi_0 = 1$. Note that typically ν_k^N is not a probability distribution and that φ_k is $\mathcal{F}_{k-1}$-measurable. The factor φ_k is chosen in a way that from Lemma 2.1 we have

$$\mathbb{E}[\nu_k^N(f)|\mathcal{F}_{k-1}] = \nu_{k-1}^N(q_{k-1,k}(f)). \tag{2.1}$$

By the unbiasedness proved in Proposition 2.1 below it also holds that

$$\mathbb{E}[\varphi_k] = \mathbb{E}[\nu_k^N(1)] = \mathbb{E}[\mu_k(1)] = 1.$$

Furthermore, the following relation will prove useful:

$$\nu_{k+1}^N(1) = \varphi_{k+1} = \varphi_k \eta_k^N(q_{k,k+1}(1)) = \nu_k^N(q_{k,k+1}(1)). \tag{2.2}$$

The connection between the approximation errors of $\eta_n^N(f)$ and $\nu_n^N(f)$ is established in the following lemma:

Lemma 2.2. *For $f \in B(E_n)$ we have the bounds*

$$\mathbb{E}[(\eta_n^N(f) - \mu_n(f))^2] \leq 2\,\mathrm{Var}(\nu_n^N(f)) + 2\,\|f - \mu_n(f)\|_{\mathrm{sup},n}^2 \mathrm{Var}(\nu_n^N(1)) \tag{2.3}$$

and

$$\begin{aligned} & \mathbb{E}[|\eta_n^N(f) - \mu_n(f)|] & (2.4) \\ \leq \quad & \mathrm{Var}(\nu_n^N(f))^{\frac{1}{2}} + \sqrt{2}\|f - \mu_n(f)\|_{\mathrm{sup},n}\mathrm{Var}(\nu_n^N(1)) + \sqrt{2}\,Var(\nu_n^N(f))^{\frac{1}{2}}\mathrm{Var}(\nu_n^N(1))^{\frac{1}{2}}, \end{aligned}$$

where $\|\cdot\|_{\mathrm{sup},n}$ denotes the supremum norm on $B(E_n)$.

Proof of Lemma 2.2. Define $f_n = f - \mu_n(f)$ and observe that for $a, b \in \mathbb{R}$ the fact that $(a - 2b)^2 \geq 0$ implies

$$a^2 \leq 2(a-b)^2 + 2b^2.$$

We can thus prove (2.3) as follows:

$$\begin{aligned} \mathbb{E}[\eta_n^N(f_n)^2] & \leq 2\mathbb{E}[(\eta_n^N(f_n) - \nu_n^N(f_n))^2] + 2\mathbb{E}[\nu_n^N(f_n)^2] \\ & \leq 2\|f_n\|_{\mathrm{sup},n}^2 \mathrm{Var}(\nu_n^N(1)) + 2\mathrm{Var}(\nu_n^N(f)) \end{aligned}$$

where the last step uses the unbiasedness of ν_n^N proved in Proposition 2.1 below. To show (2.4), observe that by the triangle inequality, by the definition of ν_n^N and by the Cauchy-Schwarz inequality we have

$$\begin{aligned}\mathbb{E}[|\eta_n^N(f_n)|] &\leq \mathbb{E}[|\eta_n^N(f_n) - \nu_n^N(1)\eta_n^N(f_n)|] + \mathbb{E}[|\nu_n^N(f_n)|] \\ &\leq \mathbb{E}[\eta_n^N(f_n)^2]^{\frac{1}{2}}\mathbb{E}[(\nu_n^N(1)-1)^2]^{\frac{1}{2}} + \mathbb{E}[\nu_n^N(f_n)^2]^{\frac{1}{2}}.\end{aligned}$$

Inserting (2.3) completes the proof. □

Thus we can indeed control the approximation error of η_n^N in terms of the approximation error of ν_n^N.

The main result of this section shows that $\nu_k^N(f)$ is an unbiased estimator for $\mu_k(f)$ and gives an explicit expression for its variance which is well-suited for deriving our later error bounds:

Proposition 2.1. *For all $f \in B(E_n)$,*

$$\mathbb{E}[\nu_n^N(f)] = \mu_n(f)$$

and

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] = \frac{1}{N}\mathrm{Var}_{\mu_n}(f) + \frac{1}{N}\mathbb{E}\left[\sum_{j=0}^{n-1} V_{j,n}^N(f)\right]$$

where

$$V_{j,n}^N(f) = \nu_j^N(1)\nu_j^N(q_{j,n}(f)^2) - \nu_j^N(q_{j,n}(f))^2 + \nu_j^N(q_{j,j+1}(1)-1)\nu_j^N(q_{j,n}(f^2)). \quad (2.5)$$

The proof of the proposition is based on martingale methods and proceeds in a number of lemmas which make up the remainder of this section. The actual proof of the proposition follows at the end. Note first that for any fixed $f \in B(E_n)$ the process $(A_j)_{j=0}^n$ defined by

$$A_j = \nu_j^N(q_{j,n}(f))$$

is an $(\mathcal{F}_n)$-martingale by (2.1) and by the semigroup property of the mappings $q_{j,n}$. Recall that by the Doob-Meyer decomposition the process H_j given by

$$H_j = A_j^2 - A_0^2 - \sum_{k=0}^{j-1}\mathbb{E}[A_{k+1}^2 - A_k^2|\mathcal{F}_k] \quad (2.6)$$

is a martingale. We next derive a more explicit expression for H_j.

Lemma 2.3. *We have*

$$H_j = A_j^2 - A_0^2 - \frac{1}{N}\sum_{k=0}^{j-1}\nu_k^N(q_{k,k+1}(1))\nu_k^N(q_{k,k+1}(q_{k+1,n}(f)^2)) - \nu_k^N(q_{k,n}(f))^2.$$

Proof. By the definitions of A_k, ν_k^N and η_k^N and since the random variables $\xi_{k+1}^1, \ldots, \xi_{k+1}^N$ are conditionally (on $\mathcal{F}_k$) independent we can write

$$\begin{aligned}
\mathbb{E}[A_{k+1}^2|\mathcal{F}_k] &= \frac{\varphi_{k+1}^2}{N^2}\mathbb{E}\left[\left(\sum_{j=1}^N q_{k+1,n}(f)(\xi_{k+1}^j)\right)^2 \middle| \mathcal{F}_k\right] \\
&= \frac{\varphi_{k+1}^2}{N^2}\left[\sum_{j=1}^N \mathbb{E}[q_{k+1,n}(f)(\xi_{k+1}^j)^2|\mathcal{F}_k] - \sum_{j=1}^N \mathbb{E}[q_{k+1,n}(f)(\xi_{k+1}^j)|\mathcal{F}_k]^2\right. \\
&+ \left.\left(\sum_{j=1}^N \mathbb{E}[q_{k+1,n}(f)(\xi_{k+1}^j)|\mathcal{F}_k]\right)^2\right] \\
&= \frac{\varphi_{k+1}^2}{N^2}\left(N\mathbb{E}[q_{k+1,n}(f)(\xi_{k+1}^1)^2|\mathcal{F}_k] - N\mathbb{E}[q_{k+1,n}(f)(\xi_{k+1}^1)|\mathcal{F}_k]^2\right. \\
&+ \left. N^2\,\mathbb{E}[q_{k+1,n}(f)(\xi_{k+1}^1)|\mathcal{F}_k]^2\right)
\end{aligned}$$

Thus by Lemma 2.1 and the semigroup property of the $q_{j,k}$ we have

$$\begin{aligned}
\mathbb{E}[A_{k+1}^2|\mathcal{F}_k] &= \frac{\varphi_{k+1}^2}{N^2}\left[N\frac{\eta_k^N(q_{k,k+1}(q_{k+1,n}(f)^2))}{\eta_k^N(q_{k,k+1}(1))} - N\frac{\eta_k^N(q_{k,n}(f))^2}{\eta_k^N(q_{k,k+1}(1))^2} + N^2\frac{\eta_k^N(q_{k,n}(f))^2}{\eta_k^N(q_{k,k+1}(1))^2}\right] \\
&= \frac{1}{N}[\nu_k^N(q_{k,k+1}(1))\nu_k^N(q_{k,k+1}(q_{k+1,n}(f)^2)) - \nu_k^N(q_{k,n}(f))^2] + A_k^2
\end{aligned}$$

where in the last step we used (2.2). Inserting the resulting expression for $\mathbb{E}[A_{k+1}^2 - A_k^2|\mathcal{F}_k]$ into (2.6) concludes the proof. □

We can use Lemma 2.3 to derive an explicit expression for $\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2]$. In order to make this expression more tractable, concretely, in order to remove the terms $q_{k,k+1}(q_{k+1,n}(f)^2)$, we use the following lemma:

Lemma 2.4. *For all $f \in B(E_n)$ the processes*

$$\begin{aligned}
L_k &= \nu_k^N(1)\nu_k^N(q_{k,n}(f)^2) - \nu_0^N(1)\nu_0^N(q_{0,n}(f)^2) \\
&- \sum_{j=0}^{k-1}\nu_j^N(q_{j,j+1}(1))\nu_j^N(q_{j,j+1}(q_{j+1,n}(f)^2)) + \sum_{j=0}^{k-1}\nu_j^N(1)\nu_j^N(q_{j,n}(f)^2) \qquad (2.7)
\end{aligned}$$

and

$$M_k = \nu_k^N(1)\nu_k^N(q_{k,n}(f^2)) - \nu_0^N(1)\nu_0^N(q_{0,n}(f^2)) - \sum_{j=0}^{k-1}\nu_j^N(q_{j,j+1}(1) - 1)\nu_j^N(q_{j,n}(f^2)) \qquad (2.8)$$

are $(\mathcal{F}_k)$-martingales.

Proof. By the Doob-Meyer decomposition, for $B_k = \nu_k^N(1)\nu_k^N(q_{k,n}(f)^2)$ the process

$$L_k = B_k - B_0 - \sum_{j=0}^{k-1} \mathbb{E}[B_{j+1}|\mathcal{F}_j] + \sum_{j=0}^{k-1} B_j$$

is a martingale. To obtain the expression in (2.7) it is sufficient to note that by Lemma 2.1 and by (2.2)

$$\begin{aligned} \mathbb{E}[B_{j+1}|\mathcal{F}_j] &= \varphi_{j+1}^2 \mathbb{E}[\eta_{j+1}^N(q_{j+1,n}(f)^2)|\mathcal{F}_j] = \varphi_{j+1}^2 \frac{\eta_j^N(q_{j,j+1}(q_{j+1,n}(f)^2))}{\eta_j^N(q_{j,j+1}(1))} \\ &= \nu_j^N(q_{j,j+1}(1))\nu_j^N(q_{j,j+1}(q_{j+1,n}(f)^2)). \end{aligned}$$

Likewise for $\widetilde{B}_k = \nu_k^N(1)\nu_k^N(q_{k,n}(f^2))$ the process

$$M_k = \widetilde{B}_k - \widetilde{B}_0 - \sum_{j=0}^{k-1} \mathbb{E}[\widetilde{B}_{j+1} - \widetilde{B}_j|\mathcal{F}_j]$$

is a martingale. To obtain the expression in (2.8) note that by (2.1) and by (2.2)

$$\begin{aligned} \mathbb{E}[\widetilde{B}_{j+1} - \widetilde{B}_j|\mathcal{F}_j] &= \nu_{j+1}^N(1)\mathbb{E}[\nu_{j+1}^N(q_{j+1,n}(f^2))|\mathcal{F}_j] - \nu_j^N(1)\nu_j^N(q_{j,n}(f^2)) \\ &= (\nu_{j+1}^N(1) - \nu_j^N(1))\nu_j^N(q_{j,n}(f^2)) = \nu_j^N(q_{j,j+1}(1) - 1)\nu_j^N(q_{j,n}(f^2)). \end{aligned}$$

□

With these lemmas, we have established the tools needed to prove Proposition 2.1:

Proof of Proposition 2.1. For the unbiasedness, note that

$$\begin{aligned} \nu_n^N(f) - \mu_n(f) &= \nu_n^N(q_{n,n}(f)) - \nu_0^N(q_{0,n}(f)) + \nu_0^N(q_{0,n}(f)) - \mu_0(q_{0,n}(f)) \\ &= A_n - A_0 + \nu_0^N(q_{0,n}(f)) - \mu_0(q_{0,n}(f)). \end{aligned}$$

This implies

$$\mathbb{E}[\nu_n^N(f)] = \mathbb{E}[\mu_n(f)]$$

since the martingale property of A_n implies $\mathbb{E}[A_n - A_0] = 0$, and since we have

$$\mathbb{E}[\nu_0^N(q_{0,n}(f))] = \mathbb{E}[\mu_0(q_{0,n}(f))]$$

because of $\nu_0^N = \eta_0^N$ and because the particles ξ_0^j are independent samples from μ_0.

Now note that by conditional independence

$$\begin{aligned} \mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] &= \mathbb{E}[|(A_n - A_0) + (\nu_0^N(q_{0,n}(f)) - \mu_0(q_{0,n}(f)))|^2] \\ &= \mathbb{E}[A_n^2 - A_0^2] + \mathbb{E}[(\nu_0^N(q_{0,n}(f)) - \mu_0(q_{0,n}(f)))^2]. \quad (2.9) \end{aligned}$$

Note that the second summand equals $\frac{1}{N}\mathrm{Var}_{\mu_0}(q_{0,n}(f))$. Using Lemma 2.3 and, in the second step, (2.7) we can thus write

$$\begin{aligned}
\mathbb{E}[|\nu_n^N(f)-\mu_n(f)|^2] &= \frac{1}{N}\Big(\mathrm{Var}_{\mu_0}(q_{0,n}(f)) \\
&+ E\Big[\sum_{k=0}^{n-1}\nu_k^N(q_{k,k+1}(1))\nu_k^N(q_{k,k+1}(q_{k+1,n}(f)^2)) - \nu_k^N(q_{k,n}(f))^2\Big]\Big) \\
&= \frac{1}{N}\Big(\mathrm{Var}_{\mu_0}(q_{0,n}(f)) + E\Big[\nu_n^N(1)\nu_n^N(q_{n,n}(f)^2) - \nu_0^N(1)\nu_0^N(q_{0,n}(f)^2) \\
&+ \sum_{k=0}^{n-1}\nu_k^N(1)\nu_k^N(q_{k,n}(f)^2) - \nu_k^N(q_{k,n}(f))^2\Big]\Big). \qquad (2.10)
\end{aligned}$$

Now observe that

$$\begin{aligned}
\mathrm{Var}_{\mu_0}(q_{0,n}(f)) - \mathbb{E}[\nu_0^N(1)\nu_0^N(q_{0,n}(f)^2)] &= \mathrm{Var}_{\mu_0}(q_{0,n}(f)) - \mu_0(q_{0,n}(f)^2) \\
&= -\mu_0(q_{0,n}(f))^2 = -\mu_n(f)^2. \qquad (2.11)
\end{aligned}$$

Moreover, by (2.8) we have

$$\begin{aligned}
\mathbb{E}[\nu_n^N(1)\nu_n^N(f^2)] - \mu_n(f)^2 &= \mathbb{E}[\nu_0^N(1)\nu_0^N(q_{0,n}(f^2))] - \mu_n(f)^2 \\
&+ \sum_{j=0}^{n-1}\mathbb{E}[\nu_j^N(q_{j,j+1}(1)-1)\nu_j^N(q_{j,n}(f^2))] \\
&= \mathrm{Var}_{\mu_n}(f) + \sum_{j=0}^{n-1}\mathbb{E}[\nu_j^N(q_{j,j+1}(1)-1)\nu_j^N(q_{j,n}(f^2))]
\end{aligned} \qquad (2.12)$$

Inserting (2.11) and then (2.12) into (2.10) yields

$$\mathbb{E}[|\nu_n^N(f)-\mu_n(f)|^2] = \frac{1}{N}\left(\mathrm{Var}_{\mu_n}(f) + \mathbb{E}\left[\sum_{j=0}^{n-1}V_{j,n}^N(f)\right]\right)$$

with

$$V_{j,n}^N(f) = \nu_j^N(1)\nu_j^N(q_{j,n}(f)^2) - \nu_j^N(q_{j,n}(f))^2 + \nu_j^N(q_{j,j+1}(1)-1)\nu_j^N(q_{j,n}(f^2))$$

so we are done. □

2.3 Non-asymptotic Error Bounds

In this section we derive non-asymptotic error bounds from our expression for the variance of $\nu_n^N(f)$ derived in the previous section. For $0 \le j \le n$, let $\|\cdot\|_j$ be a norm on the function space $B(E_j)$ such that $\|f\|_j < \infty$ for all $f \in B(E_j)$. For $0 \le j < k \le n$, let $c_{j,k}$ be a constant such that for all $f \in B(E_k)$, the following stability inequality for the semigroup $q_{j,k}$ is satisfied:

$$\max(\|1\|_j \|q_{j,k}(f)^2\|_j, \|q_{j,k}(f)\|_j^2, \|q_{j,k}(f^2)\|_j) \leq c_{j,k}\|f\|_k^2. \tag{2.13}$$

In the following we show that the quadratic approximation error of ν_n^N can essentially be controlled through the constants $c_{j,k}$.

Recall that the approximation error was given by the expected sum over j of the expressions $V_{j,n}^N$ defined in the previous sections. We first show how $V_{j,n}^N$ can be bounded through $V_{j,n}$ defined by

$$V_{j,n}(f) = \mathrm{Var}_{\mu_j}(q_{j,n}(f)).$$

and an error term. Note that $V_{j,n}$ is what we obtain when substituting ν_j^N by μ_j in our expression for $V_{j,n}^N$. Define

$$\varepsilon_j^N = \sup\left\{\mathbb{E}[|\nu_j^N(f) - \mu_j(f)|^2]\middle| \|f\|_j \leq 1\right\}.$$

Then we have the following result:

Proposition 2.2. *For $0 \leq j < n$ we have*

$$\mathbb{E}[V_{j,n}^N(f)] \leq V_{j,n}(f) + c_{j,n}\|f\|_n^2\Big(2 + \|q_{j,j+1}(1) - 1\|_j\Big)\varepsilon_j^N.$$

Proof. Note first that by the Cauchy-Schwarz inequality and since $\nu_j^N(\cdot)$ is an unbiased estimator for $\mu_j(\cdot)$, we have for any $g, h \in B(E_j)$

$$\begin{aligned} & |\mathbb{E}[\nu_j^N(g)\nu_j^N(h) - \mu_j(g)\mu_j(h)]| \\ \leq\ & |\mu_j(g)\mathbb{E}[\nu_j^N(h) - \mu_j(h)] + \mu_j(h)\mathbb{E}[\nu_j^N(g) - \mu_j(g)]| + \mathbb{E}[|\nu_j^N(g) - \mu_j(g)||\nu_j^N(h) - \mu_j(h)|] \\ \leq\ & \|g\|_j\|h\|_j\ \varepsilon_j^N. \end{aligned} \tag{2.14}$$

Adding $\pm V_{j,n}(f)$ to the definition (2.5) of $V_{j,n}^N(f)$ and applying (2.14) three times yields

$$\mathbb{E}[V_{j,n}^N(f)] \leq V_{j,n}(f) + R_{j,n}(f)\ \varepsilon_j^N$$

with

$$R_{j,n}(f) \ = \ \|1\|_j\|q_{j,n}(f)^2\|_j + \|q_{j,n}(f)\|_j^2 + \|q_{j,n}(f^2)\|_j\|q_{j,j+1}(1) - 1\|_j.$$

Applying (2.13) yields

$$R_{j,n}(f) \leq c_{j,n}\|f\|_n^2(2 + \|q_{j,j+1}(1) - 1\|_j)$$

and thus the desired inequality. □

In order to state the main result of this section, we need a few more definitions.

Define $\widehat{c}_k$ and $\widehat{v}_k$ by

$$\widehat{c}_k = \sum_{j=0}^{k-1} c_{j,k}\Big(2 + \|q_{j,j+1}(1) - 1\|_j\Big)$$

and

$$\widehat{v}_k = \sup\left\{ \sum_{j=0}^{k} \mathrm{Var}_{\mu_j}(q_{j,k}(f)) \,\middle|\, \|f\|_k \leq 1 \right\}.$$

Furthermore define

$$\overline{c}_k = \max_{j\leq k} \widehat{c}_j, \qquad \overline{v}_k = \max_{j\leq k} \widehat{v}_j \quad \text{and} \quad \overline{\varepsilon}_k^N = \max_{j\leq k} \varepsilon_j^N.$$

Then we have the following bound on the approximation error:

Theorem 2.1. *Let $N \geq 2\overline{c}_n$. Then for $f \in B(E_n)$ we have*

$$N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + \|f\|_n^2 \widehat{c}_n \, \overline{\varepsilon}_n^N \tag{2.15}$$

and

$$\overline{\varepsilon}_n^N \leq 2\frac{\overline{v}_n}{N} \tag{2.16}$$

Proof of Theorem 2.1. Note that by Propositions 2.1 and 2.2 and by the definition of $V_{j,n}(f)$ we get

$$\begin{aligned} & N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \\ \leq \; & \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + \|f\|_n^2 \sum_{j=0}^{n-1} c_{j,n}(2 + \|q_{j,j+1}(1) - 1\|_j)\, \varepsilon_j^N. \end{aligned}$$

Bounding ε_j^N by $\overline{\varepsilon}_n^N$ and inserting the definition of $\widehat{c}_n$ shows (2.15). Optimizing (2.15) over f with $\|f\|_n \leq 1$ and over n yields

$$N\overline{\varepsilon}_n^N \leq \overline{v}_n + \overline{c}_n \, \overline{\varepsilon}_n^N.$$

Choosing $N \geq 2\,\overline{c}_n$ and thus $N - \overline{c}_n \geq \frac{N}{2}$ gives (2.16). □

2.4 Another Non-asymptotic Error Bound

In this section, we prove an alternative to Theorem 2.1. Basically, we obtain this result by stopping the proof of Proposition 2.1 at formula (2.11), thus getting a different expression for the variance of $\nu_n^N(f)$, and then continuing the remaining steps as in the proof of Theorem 2.1. This leads to an error bound where the crucial inequality (2.13) is replaced by

$$\max(\|1\|_j \|q_{j,k}(f)^2\|_j, \|q_{j,k}(f)\|_j^2) \leq d_{j,k}\|f\|_k^2 \tag{2.17}$$

for $0 \leq j \leq k \leq n$. The main difference between this condition and (2.13) is that (2.17) includes the case $j = k$. This implies that we need a constant which allows to bound $\|f^2\|_k$ against $\|f\|_k^2$. This is generally difficult since unlike in the cases $j < k$ there are no transition kernels K_l on the left hand side whose smoothing properties may be exploited. An important exception is the case where $\|\cdot\|_k$ is a supremum norm. In that case we have $\|f^2\|_k = \|f\|_k^2$. The analysis of this section thus serves two purposes: For one thing, it makes clearer why it was necessary to rewrite the variance further in Proposition 2.1. For another, it yields an alternative error bound which has better constants in the setting of Sequential MCMC on trees analyzed in Chapter 4 where we rely on supremum norms.

The variance of $\nu_n^N(f)$ can be written as follows:

Lemma 2.5. *For all* $f \in B(E_n)$

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] = \frac{1}{N}\mathbb{E}[\nu_n^N(1)\nu_n^N(f^2) - \mu_n(f)^2] + \frac{1}{N}\mathbb{E}\left[\sum_{j=0}^{n-1} U_{j,n}^N(f)\right]$$

where

$$U_{j,n}^N(f) = \nu_j^N(1)\nu_j^N(q_{j,n}(f)^2) - \nu_j^N(q_{j,n}(f))^2 \tag{2.18}$$

Proof of Lemma 2.5. The result follows immediately by inserting (2.11) into (2.10) in the Proof of Proposition 2.1. □

Define ε_j^N, $\overline{\varepsilon}_j^N$ and $\overline{v}_j$ as in the previous section. Analogously to Proposition 2.2 we can then prove the following:

Lemma 2.6. *For* $0 \leq j < n$ *we have*

$$\mathbb{E}[U_{j,n}^N(f)] \leq V_{j,n}(f) + 2\,d_{j,n}\|f\|_n^2\varepsilon_j^N .$$

Moreover

$$\mathbb{E}[\nu_n^N(1)\nu_n^N(f^2) - \mu_n(f)^2] \leq V_{n,n}(f) + d_{n,n}\|f\|_n^2\varepsilon_j^N .$$

Proof of Lemma 2.6. Arguing as in the proof of Proposition 2.2, we obtain for $j < n$

$$\mathbb{E}[U_{j,n}^N(f)] \leq V_{j,n}(f) + S_{j,n}(f)\varepsilon_j^N$$

where

$$S_{j,n}(f) = \|1\|_j\|q_{j,k}(f)^2\|_j + \|q_{j,k}(f)\|_j^2.$$

The same argument also yields

$$\mathbb{E}[\nu_n^N(1)\nu_n^N(f^2) - \mu_n(f)^2] \leq V_{n,n}(f) + \widetilde{S}_{n,n}(f)\varepsilon_j^N .$$

where $\widetilde{S}_{n,n}(f) = \|1\|_n\|f^2\|_n$. Applying (2.17) completes the proof. □

Now define constants

$$\widehat{d}_k = 2 \sum_{j=0}^{k} d_{j,k}$$

and $\overline{d}_k = \max_{j \leq k} \widehat{d}_k$. Then we obtain the following alternative bound on the approximation error:

Theorem 2.2. *Let $N \geq 2\overline{d}_n$. Then for $f \in B(E_n)$ we have*

$$N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + \|f\|_n^2 \widehat{d}_n \, \overline{\varepsilon}_n^N \tag{2.19}$$

and

$$\overline{\varepsilon}_n^N \leq 2\frac{\overline{v}_n}{N} \tag{2.20}$$

Proof of Theorem 2.2. The proof is parallel to the one of Theorem 2.1: By Lemmas 2.5 and 2.6 and by the definition of $V_{j,n}(f)$ we obtain

$$\begin{aligned} & N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \\ \leq \; & \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + 2\,\|f\|_n^2 \sum_{j=0}^{n} d_{j,n} \, \varepsilon_j^N . \end{aligned}$$

Bounding ε_j^N by $\overline{\varepsilon}_n^N$ and inserting the definition of $\widehat{d}_n$ shows (2.19). Optimizing (2.19) over f with $\|f\|_n \leq 1$ and over n yields

$$N\overline{\varepsilon}_n^N \leq \overline{v}_n + \overline{d}_n \, \overline{\varepsilon}_n^N .$$

Choosing $N \geq 2\,\overline{d}_n$ and thus $N - \overline{d}_n \geq \frac{N}{2}$ gives (2.20). □

Comparing Theorems 2.1 and 2.2, we thus see that the additional steps in rewriting the variance in the proof of Proposition 2.1 have the following two effects: We lose by getting the additional summand $\|q_{j,j+1}(1) - 1\|_j$ in our error bound and we gain by obtaining condition (2.13) instead of (2.17). The latter is generally an advantage since, unlike for (2.17), smoothing properties of the kernels K_k can be used to prove (2.17). Finally, note that in both theorems the coefficient of the leading term in the error bound corresponds to the asymptotic variance in the central limit theorem for ν_n^N found in Del Moral and Miclo (2000, p. 45).

3 L_p-Bounds under Global Mixing

In this chapter, we show how L_p-stability of the Feynman-Kac semigroups $q_{j,k}$ can be derived from global mixing properties of the MCMC kernels K_k. We apply these results to derive non-asymptotic error bounds for Sequential MCMC using the results of the previous chapter. We now consider a more restricted setting which covers the algorithm introduced in Section 1.2. Concretely, we assume that the measures μ_k live on a common state space and that the transition kernels K_k are stationary with respect to the measures μ_k. From Section 3.4 on, we add the assumption of reversibility of the kernels K_k.

Sections 3.1 and 3.2 introduce the special case of the model of Chapter 2 studied subsequently and reformulate the main results of that chapter into the present setting. The central part of this chapter is Section 3.3 which studies how L_p-bounds with time-independent constants can be derived from mixing properties of the MCMC dynamics. Section 3.4 shows how the stability results of Section 3.3 can be derived from global Poincaré and Logarithmic Sobolev inequalities and gives explicit expressions for the resulting error bounds. Finally, Section 3.5 studies an example where the measures μ_k are Gaussian measures restricted to a ball in $\mathbb{R}^d$ and where the transition kernels K_k are those associated with reflected Langevin diffusions with the appropriate target distributions.

3.1 The Model

Let (E, r) be a Polish space and let $\mathcal{B}(E)$ be the σ-algebra of Borel subsets of E. Denote by $M(E)$ the space of finite signed Borel measures on E. Let $M_1(E) \subset M(E)$ be the subset of all probability measures. Let $B(E)$ be the space of bounded, measurable, real-valued functions on E.

Consider the sequence of probability measures $(\mu_k)_{k=0}^n$, $\mu_k \in M_1(E)$. The μ_k are related through

$$\mu_k(f) = \frac{\mu_{k-1}(g_{k-1,k}f)}{\mu_{k-1}(g_{k-1,k})}$$

for strictly positive (unnormalized) relative densities $g_{k-1,k} \in B(E)$. In the notation of Section 2.1.2 this implies $\hat{\mu}_k = \mu_k$ for all k.

For $1 \leq k \leq n$, let $K_k(x, A)$ be an integral operator with $K_k(\cdot, f) \in B(E)$ for all $f \in B(E)$, with $K_k(x, \cdot) \in M_1(E)$ for all $x \in E$ and with stationary distribution μ_k, i.e.,

$$\mu_k K_k(A) = \mu_k(A) \qquad \text{for all} \quad A \in \mathcal{B}(E).$$

K_k can be thought of, e.g., as many steps of a Metropolis chain with respect target μ_k. Define the mapping $q_{k-1,k} : B(E) \to B(E)$ by

$$q_{k-1,k}(f) = \frac{g_{k-1,k}K_k(f)}{\mu_{k-1}(g_{k-1,k})}$$

Observe that this choice implies

$$\mu_k(f) = \mu_{k-1}(q_{k-1,k}(f))$$

Furthermore define for $0 \leq j < k \leq n$ the mapping $q_{j,k} : B(E) \to B(E)$ by

$$q_{j,k}(f) = q_{j,j+1}(q_{j+1,j+2}(\dots q_{k-1,k}(f)))$$

and $q_{k,k}(f) = f$. We have the relation

$$\mu_j(q_{j,k}(f)) = \mu_k(f) \quad \text{for } 0 \leq j \leq k \leq n$$

and the semigroup property

$$q_{j,l}(q_{l,k}(f)) = q_{j,k}(f) \quad \text{for } 0 \leq j < l < k \leq n.$$

3.2 Sequential MCMC

We now introduce the interacting particle system simulated in the Sequential MCMC algorithm for this model and state our non-asymptotic bounds on the approximation error in this case.

3.2.1 The Interacting Particle System

We want to approximate the sequence of measures μ_k by an interacting particle system. We start with N independent samples $\xi_0 = (\xi_0^1, \dots, \xi_0^N)$ from μ_0. The particle dynamics alternates two steps: Importance Sampling Resampling and Mutation: A vector of particles ξ_{k-1} approximating μ_{k-1} is transformed into a vector $\hat{\xi}_k$ approximating μ_k by drawing N conditionally independent samples from the empirical distribution of ξ_{k-1} weighted with the functions $g_{k-1,k}$. Afterwards, in order to reduce the variance introduced through resampling, $\hat{\xi}_k$ is transformed into a vector ξ_k (still approximating μ_k) by moving the particles $\hat{\xi}_k^i$ independently with the transition kernel K_k.

Accordingly, we have two arrays of E-valued random variables $(\xi_k^j)_{0 \leq k \leq n, 1 \leq j \leq N}$ and $(\hat{\xi}_k^j)_{1 \leq k \leq n, 1 \leq j \leq N}$. The random variables $\xi_0^1, \dots, \xi_0^N$ are independent and distributed according to μ_0. The distributions of the remaining $\hat{\xi}_k^j$ and ξ_k^j are pinned down by the transition probabilities

$$\mathbb{P}[\hat{\xi}_k \in dx | \xi_{k-1} = z] = \prod_{j=1}^{N} \sum_{i=1}^{N} \frac{g_{k-1,k}(z^i)}{\sum_{l=1}^{N} g_{k-1,k}(z^l)} \delta_{z^i}(dx^j)$$

and

$$\mathbb{P}[\xi_k \in dx | \hat{\xi}_k = z] = \prod_{j=1}^{N} K_k(z^j, dx^j).$$

3.2.2 Error Bounds in L_p

Denote by $\mathcal{F}_k$ the σ-algebra generated by $\xi_0, \ldots \xi_k$ and $\hat{\xi}_1, \ldots \hat{\xi}_k$ and denote the empirical measure of ξ_k by η_k^N, i.e.

$$\eta_k^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_k^i}.$$

We are interested in the question how well η_k^N approximates μ_k.

Recall that by Lemma 2.1 we have for $f \in B(E)$ and $1 \leq k \leq n$ that

$$\mathbb{E}[\eta_k^N(f) | \mathcal{F}_{k-1}] = \frac{\eta_{k-1}^N(q_{k-1,k}(f))}{\eta_{k-1}^N(q_{k-1,k}(1))},$$

which implies that $\eta_k^N(f)$ is a biased estimator for $\eta_{k-1}^N(q_{k-1,k}(f))$. It thus proves to be useful to remove this bias following the analysis of Section 2.2: Define for $0 \leq k \leq n$ the sequence of (unnormalized) measures

$$\nu_k^N(f) = \varphi_k \, \eta_k^N(f)$$

on E where φ_k is given by

$$\varphi_k = \prod_{j=0}^{k-1} \eta_j^N(q_{j,j+1}(1)).$$

Then we have for $f \in B(E)$

$$\mathbb{E}[\nu_k^N(f) | \mathcal{F}_{k-1}] = \nu_{k-1}^N(q_{k-1,k}(f)).$$

By Proposition 2.1, $\nu_k^N(f)$ is an unbiased estimator for $\mu_k(f)$ with quadratic error given in that proposition. Moreover, we can control the approximation error of η_k^N through the approximation error of ν_k^N by Lemma 2.2.

We now want to apply the error bound of Theorem 2.1. We thus need to define a series of norms $\| \cdot \|_j$ on $B(E)$ and find constants $c_{j,k}$ such that the inequality

$$\max\left(\|q_{j,k}(f)\|_j^2, \|q_{j,k}(f)^2\|_j, \|q_{j,k}(f^2)\|_j\right) \leq c_{j,k} \, \|f\|_k^2. \tag{3.1}$$

is satisfied for all $f \in B(E)$.

We fix $p > 2$ and choose $\|\cdot\|_j = \|\cdot\|_{L_p(\mu_j)}$, i.e.,

$$\|f\|_j = \|f\|_{L_p(\mu_j)} = \mu_j(|f|^p)^{\frac{1}{p}}.$$

Note that all bounded measurable functions are in $L_p(\mu_j)$ and that for $f \in B(E)$ we have $q_{j,k}(f) \in B(E)$ since we assumed the relative densities $g_{k-1,k}$ to be bounded. Now define $\widetilde{c}_{j,k}(p,q)$ to be the constant in an L_p-L_q bound for the semigroup $q_{j,k}$: For $p > q > 1$ and $0 \le j < k \le n$ we have

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \le \widetilde{c}_{j,k}(p,q)\|f\|_{L_q(\mu_k)} \quad \text{for all } f \in B(E)$$

Such constants will be studied in Section 3.3 below. Furthermore define

$$c_{j,k}(p) = \max\left(\widetilde{c}_{j,k}\left(p, \frac{p}{2}\right),\ \widetilde{c}_{j,k}(2p,p)^2\right).$$

This choice of $c_{j,k}(p)$ satisfies (3.1):

Lemma 3.1. *For $p > 2$, $0 \le j < k \le n$ and $f \in B(E)$ we have*

$$\max\left(\|1\|_{L_p(\mu_j)}\|q_{j,k}(f)^2\|_{L_p(\mu_j)}, \|q_{j,k}(f)\|^2_{L_p(\mu_j)}, \|q_{j,k}(f^2)\|_{L_p(\mu_j)}\right) \le c_{j,k}(p)\ \|f\|^2_{L_p(\mu_k)}.$$

Proof. Observe first that we have

$$\|q_{j,k}(f^2)\|_{L_p(\mu_j)} \le \widetilde{c}_{j,k}\left(p, \frac{p}{2}\right)\ \|f^2\|_{L_{\frac{p}{2}}(\mu_k)} = \widetilde{c}_{j,k}\left(p, \frac{p}{2}\right)\ \|f\|^2_{L_p(\mu_k)}.$$

Furthermore we have $\|1\|_{L_p(\mu_j)} = 1$ and

$$\|q_{j,k}(f)^2\|_{L_p(\mu_j)} = \|q_{j,k}(f)\|^2_{L_{2p}(\mu_j)} \le \widetilde{c}_{j,k}(2p,p)^2\|f\|^2_{L_p(\mu_k)}.$$

Finally, observing that

$$\|q_{j,k}(f)\|^2_{L_p(\mu_j)} \le \|q_{j,k}(f)\|^2_{L_{2p}(\mu_j)}$$

concludes the proof. □

In order to state our error bound we define another series of constants following the definitions of Section 2.3: Define

$$\widehat{c}_k(p) = \sum_{j=0}^{k-1} c_{j,k}(p)\left(2 + \|q_{j,j+1}(1) - 1\|_{L_p(\mu_j)}\right)$$

and

$$\widehat{v}_k(p) = \sup\left\{\sum_{j=0}^{k} \operatorname{Var}_{\mu_j}(q_{j,k}(f)) \,\middle|\, f \in B(E), \|f\|_{L_p(\mu_k)} \le 1\right\}$$

and

$$\varepsilon_k^N(p) = \sup\left\{\mathbb{E}[|\nu_k^N(f) - \mu_k(f)|^2]\middle| f \in B(E), \|f\|_{L_p(\mu_k)} \leq 1\right\}.$$

Moreover define

$$\overline{c}_k(p) = \max_{j\leq k} \widehat{c}_j(p), \quad \overline{v}_k(p) = \max_{j\leq k} \widehat{v}_j(p) \quad \text{and} \quad \overline{\varepsilon}_k^N(p) = \max_{j\leq k} \varepsilon_j^N(p).$$

Then the following error bound is an immediate consequence of Theorem 2.1.

Corollary 3.1. *Let $p > 2$ and $N \geq 2\overline{c}_n(p)$. Then for $f \in B(E)$ we have*

$$N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + \|f\|^2_{L_p(\mu_n)} \widehat{c}_n(p)\, \overline{\varepsilon}_n^N(p)$$

and

$$\overline{\varepsilon}_n^N(p) \leq 2\frac{\overline{v}_n(p)}{N}$$

3.3 Stability of Feynman-Kac Semigroups under Global Mixing

In this section we show how to derive inequalities of the type

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \widetilde{c}_{j,k}(p,q)\|f\|_{L_q(\mu_k)} \quad p \leq q, j < k \tag{3.2}$$

from suitable mixing conditions on the kernels $K_j, \ldots, K_k$. These are exactly the inequalities we need in order to make the error bounds of the previous section explicit. The constants $\widetilde{c}_{j,k}$ we derive are independent of the length of the time interval $k - j$. The central intermediate step is Proposition 3.1 which derives (3.2) for the case $p = q = 2^r$, $r \in \mathbb{N}$, and for a modified semigroup $\hat{q}_{j,k}$ from an L_2-mixing condition for the kernels $K_j, \ldots, K_k$. The proof of Proposition 3.1 proceeds in a number of lemmas starting with only one step, $k - j = 1$, and $p = 2$ and then gradually generalizing the result by showing how to proceed from L_p-stability to L_{2p}-stability and to more than one step, $k - j > 1$. Proposition 3.1 is followed by a number of corollaries, showing how to transfer the result to the original semigroup $q_{j,k}$, and, using an additional assumption of hyperboundedness, to the case $p > q$. Corollary 3.4 collects these observations and states the resulting version of inequality (3.2) which is applied later on. We close the chapter with an additional result, Proposition 3.2, which shows that for functions f with $\mu_k(f) = 0$ we can obtain a version of inequality (3.2) where the constants $\widetilde{c}_{j,k}(p,q)$ decay exponentially in the length of the time-interval $k - j$.

In principle, all results of this section except for Proposition 3.2 are corollaries of the results for the case of local mixing proved in Section 5.3 below. The proofs in the present setting are however considerably easier.

A central quantity in our analysis is a uniform upper bound γ on $\overline{g}_{k-1,k}$: Assume $\gamma > 1$ is such that

$$\overline{g}_{k-1,k}(x) \leq \gamma \tag{3.3}$$

for all $x \in E$ and all k with $1 \leq k \leq n$. γ is a rough measure of how strongly the measures μ_k differ from each other.

For the analysis of this section it proves to be convenient not to work with $q_{j,k}$ directly but to work with the semigroup $\hat{q}_{j,k}$ defined as follows: For $1 \leq k \leq n$ define $\hat{q}_{k-1,k} : B(E) \to B(E)$ by

$$\hat{q}_{k-1,k}(f) = K_{k-1}\left(\overline{g}_{k-1,k} f\right)$$

where $\overline{g}_{k-1,k}$ is the normalized density given by

$$\overline{g}_{k-1,k} = \frac{g_{k-1,k}}{\mu_{k-1}(g_{k-1,k})}.$$

Furthermore, define for $1 \leq j < k \leq n$ the mapping $\hat{q}_{j,k} : B(E) \to B(E)$ by

$$\hat{q}_{j,k}(f) = \hat{q}_{j,j+1}(\hat{q}_{j+1,j+2}(\ldots \hat{q}_{k-1,k}(f))) \quad \text{and} \quad \hat{q}_{k,k}(f) = f.$$

By definition $\hat{q}_{j,k}$ is a semigroup. $q_{j,k}$ and $\hat{q}_{j+1,k}$ are related through

$$q_{j,k}(f) = \overline{g}_{j,j+1}\hat{q}_{j+1,k}(K_k(f)).$$

Results for the semigroup $\hat{q}_{j,k}$ can be transfered to $q_{j,k}$ using Lemma 3.5 which is proved later in this section.

L_1-stability of $\hat{q}_{j,k}$ simply follows from

$$\|\hat{q}_{j,k}(f)\|_{L_1(\mu_j)} = \mu_j(|\hat{q}_{j,k}(f)|) \leq \mu_j(\hat{q}_{j,k}(|f|)) = \mu_k(|f|) = \|f\|_{L_1(\mu_k)}. \tag{3.4}$$

From (3.3) and from the fact that K_k is stationary with respect to μ_k it is easy to conclude bounds such as

$$\|\hat{q}_{j,k}(f)\|^2_{L_2(\mu_j)} \leq \gamma^{k-j}\|f\|^2_{L_2(\mu_k)}.$$

This bound has the strong disadvantage that it degenerates exponentially in $k-j$ since $\gamma > 1$. In the following, we assume and exploit mixing properties of the kernels K_k in order to obtain L_p-bounds for $p > 1$ which do not degenerate in $k-j$.

We assume the following mixing condition: We have constants $\alpha > 0$ and $\beta \in [0,1]$ such that for all $f \in B(E)$ we have the following L_2-bound for $\hat{q}_{k-1,k}$:

$$\|\hat{q}_{k-1,k}(f)\|^2_{L_2(\mu_{k-1})} \leq \alpha\|f\|^2_{L_2(\mu_k)} + \beta\mu_k(f)^2. \tag{3.5}$$

Additionally, our results will impose conditions that α is sufficiently small. In Section 3.4.1 below it is shown that one way to ensure that (3.5) holds with a sufficiently small α is to assume that the kernels K_k satisfy Poincaré inequalities associated with a sufficiently large spectral gap. Note that (3.5) is a global mixing condition which can only be hoped to hold with reasonable constants for a standard MCMC dynamics K_k if the distributions μ_k are essentially unimodal. In Chapter 5 below we extend the present analysis to cases where only local mixing conditions are fulfilled.

Our first step is to iterate the bound (3.5) in order to obtain L_2-bounds for $\hat{q}_{j,k}$:

Lemma 3.2. *Assume $\alpha < 1$. Then for $1 \leq j < k \leq n$ and $f \in B(E)$ we have the bounds*

$$\|\hat{q}_{j,k}(f)\|^2_{L_2(\mu_j)} \leq \alpha^{k-j}\, \|f\|^2_{L_2(\mu_k)} + \frac{\beta}{1-\alpha}\, \mu_k(f)^2 \tag{3.6}$$

and

$$\|\hat{q}_{j,k}(f)\|_{L_2(\mu_j)} \leq \frac{1}{(1-\alpha)^{\frac{1}{2}}}\, \|f\|_{L_2(\mu_k)} \tag{3.7}$$

Proof. Iterating the bound (3.5) and utilizing that $\mu_j(\hat{q}_{j,k}(f)) = \mu_k(f)$ we get

$$\|\hat{q}_{j,k}(f)\|^2_{L_2(\mu_j)} \leq \alpha^{k-j}\|f\|^2_{L_2(\mu_k)} + \sum_{i=0}^{k-j-1} \beta\alpha^i \mu_k(f)^2.$$

Applying to this the geometric series inequality immediately implies (3.6). Furthermore, since we assumed $\beta \leq 1$ and $\mu_k(f)^2 \leq \|f\|^2_{L_2(\mu_k)}$ we have

$$\|\hat{q}_{j,k}(f)\|^2_{L_2(\mu_j)} \leq \sum_{i=0}^{k-j} \alpha^i \|f\|^2_{L_2(\mu_k)}.$$

Applying again the geometric series inequality yields (3.7). □

We now turn to L_p-bounds for the case of $p = 2^r$ with $r \in \mathbb{N}$. We proceed inductively, deducing the bound for $p = 2^r$ from the bound for $p = 2^{r-1}$. We begin by deriving an L_{2p}-bound for $\hat{q}_{k-1,k}$ from (3.5).

Lemma 3.3. *For $1 \leq k \leq n$, $f \in B(E)$ and $p \geq 1$ we have*

$$\|\hat{q}_{k-1,k}(f)\|^{2p}_{L_{2p}(\mu_{k-1})} \leq \alpha\, \gamma^{2p-2}\|f\|^{2p}_{L_{2p}(\mu_k)} + \beta\, \gamma^{2p-2}\|f\|^{2p}_{L_p(\mu_k)}.$$

Proof. Note that we have

$$\begin{aligned} \|\hat{q}_{k-1,k}(f)\|^{2p}_{L_{2p}(\mu_{k-1})} &= \mu_{k-1}(|K_{k-1}(\overline{g}_{k-1,k}f)|^{2p}) \leq \mu_{k-1}(K_{k-1}(\overline{g}^p_{k-1,k}|f|^p)^2) \\ &= \|\hat{q}_{k-1,k}(\overline{g}^{p-1}_{k-1,k}|f|^p)\|^2_{L_2(\mu_{k-1})} \end{aligned}$$

Applying now the bound (3.5) yields

$$\begin{aligned} \|\hat{q}_{k-1,k}(f)\|^{2p}_{L_{2p}(\mu_{k-1})} &\leq \alpha\|\overline{g}^{p-1}_{k-1,k}|f|^p\|^2_{L_2(\mu_k)} + \beta\|\overline{g}^{p-1}_{k-1,k}|f|^p\|^2_{L_1(\mu_k)} \\ &\leq \alpha\gamma^{2p-2}\|f\|^{2p}_{L_{2p}(\mu_k)} + \beta\gamma^{2p-2}\|f\|^{2p}_{L_p(\mu_k)}. \end{aligned} \tag{3.8}$$

□

Our next step is to show how by applying Lemma 3.3 we get an L_{2p}-bound for $\hat{q}_{j,k}$ from an L_p-bound.

Lemma 3.4. *Assume that $\alpha\gamma^{2p-2} < 1$ and that for $\delta(p) \geq 1$ we have for $1 \leq j < k \leq n$ and $f \in B(E)$ the inequality*

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \leq \delta(p)\|f\|_{L_p(\mu_k)}.$$

Then we have

$$\|\hat{q}_{j,k}(f)\|_{L_{2p}(\mu_j)} \leq \delta(2p)\|f\|_{L_{2p}(\mu_k)}$$

with

$$\delta(2p) = \delta(p)\frac{\gamma^{1-\frac{1}{p}}}{(1-\alpha\gamma^{2p-2})^{\frac{1}{2p}}}.$$

Proof. Define $\theta = \alpha\gamma^{2p-2}$. Iterating the inequality of Lemma 3.3 and utilizing that $\beta \leq 1$, we get

$$\|\hat{q}_{j,k}(f)\|^{2p}_{L_{2p}(\mu_j)} \leq \theta^{k-j}\|f\|^{2p}_{L_{2p}(\mu_k)} + \gamma^{2p-2}\sum_{i=j+1}^{k}\theta^{i-1-j}\|\hat{q}_{i,k}(f)\|^{2p}_{L_p(\mu_i)} \tag{3.9}$$

Using our assumption on $\|\hat{q}_{j,k}\|_{L_p(\mu_j)}$ and the facts that $\gamma \geq 1$, $\delta(p) \geq 1$ and $\|f\|_{L_p(\mu_k)} \leq \|f\|_{L_{2p}(\mu_k)}$, we get that

$$\|\hat{q}_{j,k}(f)\|^{2p}_{L_{2p}(\mu_j)} \leq \|f\|^{2p}_{L_{2p}(\mu_k)}\gamma^{2p-2}\delta(p)^{2p}\sum_{i=0}^{k-j}\theta^i.$$

Thus, since we assumed $\theta < 1$, by the geometric series inequality we have

$$\|\hat{q}_{j,k}(f)\|_{L_{2p}(\mu_j)} \leq \delta(2p)\|f\|_{L_{2p}(\mu_k)}$$

with

$$\delta(2p) = \delta(p)\frac{\gamma^{1-\frac{1}{p}}}{(1-\alpha\gamma^{2p-2})^{\frac{1}{2p}}}.$$

□

Combining Lemmas 3.2 and 3.4 we can state the key result of this section as follows:

Proposition 3.1. *For $r \in \mathbb{N}$, consider $p = 2^r$ and assume that $\alpha\gamma^{p-2} < 1$. Then we have for $1 \leq j < k \leq n$ and $f \in B(E)$ the inequality*

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \leq \delta(p)\|f\|_{L_p(\mu_k)}$$

with

$$\delta(p) = \prod_{i=1}^{r} \frac{\gamma^{1-2^{-(i-1)}}}{(1-\alpha\gamma^{2^i-2})^{2^{-i}}} < \frac{\gamma^{r-2+2^{-(r-1)}}}{1-\alpha\gamma^{2^r-2}}$$

Proof. The case $r = 0$ follows from (3.4). In the case $r = 1$, the inequality coincides with (3.7). The inequalities for $r > 1$ follow because Lemma 3.4 implies that we can choose

$$\delta(2^r) = \delta(2) \prod_{i=2}^{r} \frac{\gamma^{1-2^{-(i-1)}}}{(1-\alpha\gamma^{2^i-2})^{2^{-i}}}.$$

We can apply Lemma 3.4 iteratively, since $\alpha\gamma^{p-2} < 1$ implies $\alpha\gamma^{q-2} < 1$ for all $q \le p$. For the upper bound on $\delta(p)$, we apply the geometric series equality in the nominator, bound the term in brackets under the exponent in the denominator by $1 - \alpha\gamma^{p-2}$ and apply the geometric series inequality to the product. □

Since the constants $\delta(2^r)$ are monotonically increasing in r, we can immediately extend the bounds of Proposition 3.1 to general $p \ge 1$ using the Riesz-Thorin interpolation theorem (see Davies (1990), §1.1.5):

Corollary 3.2. *Consider $p \in [2^r, 2^{r+1}]$ for $r \in \mathbb{N}$ and assume $\alpha\gamma^{2^{r+1}-2} < 1$. Then for $1 \le j < k \le n$ and $f \in B(E)$ we have*

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \le \delta(p)\|f\|_{L_p(\mu_k)}$$

with $\delta(p)$ given by

$$\delta(p) = \delta(2^{r+1})$$

where $\delta(2^{r+1})$ is defined as in Proposition 3.1.

Since we need L_{2p}-L_p-bounds in the error bounds of Section 3.2.2 we now show that given that we have an L_p-L_q-bound for K_k, we can immediately conclude from Corollary 3.2 an L_p-L_q-bound for $\hat{q}_{j,k}$:

Corollary 3.3. *Consider $p \ge 1$ and $q \ge 1$. Let $q \in [2^r, 2^{r+1}]$ for $r \in \mathbb{N}$ and assume $\alpha\gamma^{2^{r+1}-2} < 1$. Assume that for $1 \le j < n$ we have a constant $\theta_j(p,q) \ge 0$ such that*

$$\|K_j(f)\|_{L_p(\mu_j)} \le \theta_j(p,q)\|f\|_{L_q(\mu_j)} \tag{3.10}$$

Then for $j < k \le n$ and $f \in B(E)$ we have

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \le \theta_j(p,q)\gamma^{\frac{q-1}{q}}\delta(q)\|f\|_{L_q(\mu_k)}$$

with $\delta(q)$ as defined in Corollary 3.2.

Proof. By (3.10) we have

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \le \theta_j(p,q)\|\overline{g}_{j,j+1}\hat{q}_{j+1,k}(f)\|_{L_q(\mu_j)}$$

and thus by Corollary 3.2

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \leq \theta_j(p,q)\gamma^{\frac{q-1}{q}}\delta(q)\|f\|_{L_q(\mu_k)}.$$

□

The following lemma shows, that L_p-L_q-bounds for $\hat{q}_{j,k}$ can be used to obtain L_p-L_q-bounds for the original semigroup $q_{j,k}$.

Lemma 3.5. *Assume that for some $p \geq 1$ and $q \geq 1$ we have a $\delta \geq 0$ such that for all $f \in B(E)$ and for all $1 \leq j < k \leq n$*

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \leq \delta\, \|f\|_{L_q(\mu_k)}.$$

Then we have

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \delta\, \gamma^{\frac{p-1}{p}}\, \|f\|_{L_q(\mu_k)}.$$

Proof. Note that we have

$$\begin{aligned}
\|q_{j,k}(f)\|_{L_p(\mu_j)} &= \mu_j\left(|\overline{g}_{j,j+1}\hat{q}_{j+1,k}(K_k(f))|^p\right)^{\frac{1}{p}} \\
&\leq \gamma^{\frac{p-1}{p}}\mu_{j+1}\left(|\hat{q}_{j+1,k}(K_k(f))|^p\right)^{\frac{1}{p}} \\
&\leq \gamma^{\frac{p-1}{p}}\delta\, \|K_k(f)\|_{L_q(\mu_k)} \\
&\leq \gamma^{\frac{p-1}{p}}\delta\, \|f\|_{L_q(\mu_k)}
\end{aligned}$$

where in the last step we used that by Jensen's inequality $|K_k(f)|^q \leq K_k(|f|^q)$ and that K_k is stationary with respect to μ_k. □

Combining Proposition 3.1 with Corollary 3.3 and Lemma 3.5 we immediately obtain the type of bound needed in the error bounds of Section 3.2.2.

Corollary 3.4. *Consider $p \geq 1$ and $q \geq 1$. Let $q \in [2^r, 2^{r+1}]$ for $r \in \mathbb{N}$ and assume $\alpha\gamma^{2^{r+1}-2} < 1$. Assume that for all $1 \leq j \leq n$ we have a constant $\theta(p,q) \geq 0$ such that*

$$\|K_j(f)\|_{L_p(\mu_j)} \leq \theta(p,q)\|f\|_{L_q(\mu_j)}$$

Then for all $1 \leq j < k \leq n$ and $f \in B(E)$ we have

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \widetilde{c}_{j,k}(p,q)\|f\|_{L_q(\mu_k)}$$

with

$$\widetilde{c}_{j,k}(p,q) = \theta(p,q)\gamma^{\frac{p-1}{p}}\gamma^{\frac{q-1}{q}}\delta(q)$$

where $\delta(q)$ as defined in Corollary 3.2.

To round out the analysis of this section, we show that for functions f with $\mu_k(f) = 0$ we can moreover show the following result of exponential decay of $\|q_{j,k}(f)\|_{L_p(\mu_j)}$:

Proposition 3.2. *Let $p \geq 2$ with $p = 2^r$ for $r \in \mathbb{N}$. Assume that $\theta_p = \alpha\gamma^{2p-2} < 1$. Then for $1 \leq j < k \leq n$ and $f \in B(E)$ with $\mu_k(f) = 0$ we have*

$$\|\hat{q}_{j,k}(f)\|^p_{L_p(\mu_j)} \leq \lambda_p \, \theta_p^{k-j} \|f\|^p_{L_p(\mu_k)}$$

where the constants λ_p can be calculated recursively from $\lambda_2 = 1$ and

$$\lambda_{2p} = 1 + \lambda_p^2 \left(\alpha\left(1 - \frac{\alpha}{\gamma^2}\right)\right)^{-1}.$$

Moreover,

$$\lambda_p \leq \left[2\left(\alpha\left(1 - \frac{\alpha}{\gamma^2}\right)\right)^{-1}\right]^{\frac{p}{2}-1}$$

Proof. From (3.6) and $\mu_k(f) = 0$ we immediately get the result for $p = 2$. Now we proceed inductively concluding from a bound for p a bound for $2p$. Assume thus $\theta_{2p} < 1$ and that we have

$$\|\hat{q}_{j,k}(f)\|^p_{L_p(\mu_j)} \leq \lambda_p \theta_p^{k-j} \|f\|^p_{L_p(\mu_k)}. \tag{3.11}$$

for θ_p as defined above and for some $\lambda_p \geq 1$. Observe that $\theta_{2p} < 1$ implies immediately $\theta_p < 1$. From (3.9) and (3.11) and since we assumed $\lambda_p \geq 1$ we have the inequality

$$\|\hat{q}_{j,k}(f)\|^{2p}_{L_{2p}(\mu_j)} \leq \theta_{2p}^{k-j} \|f\|^{2p}_{L_{2p}(\mu_k)} + \gamma^{2p-2} \sum_{i=j+1}^{k} \theta_{2p}^{i-1-j} \lambda_p^2 \theta_p^{2(k-i)} \|f\|^{2p}_{L_p(\mu_k)}.$$

Thus we have

$$\|\hat{q}_{j,k}(f)\|^{2p}_{L_{2p}(\mu_j)} \leq \lambda_{2p} \, \theta_{2p}^{k-j} \|f\|^{2p}_{L_{2p}(\mu_k)}$$

with

$$\widetilde{\lambda}_{2p} = 1 + \lambda_p^2 \gamma^{2p-2} \theta_{2p}^{-1} \sum_{i=j+1}^{k} \left(\frac{\theta_p^2}{\theta_{2p}}\right)^{k-i} = 1 + \lambda_p^2 \gamma^{2p-2} \theta_{2p}^{-1} \sum_{i=0}^{k-j-1} \left(\frac{\theta_p^2}{\theta_{2p}}\right)^{i}$$

Observing that

$$\gamma^{2p-2} \, \theta_{2p}^{-1} = \frac{1}{\alpha}$$

and

$$\frac{\theta_p^2}{\theta_{2p}} = \frac{\alpha}{\gamma^2} < 1,$$

and applying the geometric series inequality thus yields

$$\widetilde{\lambda}_{2p} \leq 1 + \lambda_p^2 \left(\alpha\left(1 - \frac{\alpha}{\gamma^2}\right)\right)^{-1}.$$

Choosing

$$\lambda_{2p} = 1 + \lambda_p^2 \left(\alpha \left(1 - \frac{\alpha}{\gamma^2} \right) \right)^{-1}$$

we have thus shown the desired decay inequality. Moreover, observe that, since $\gamma > 1$ and $\alpha < 1$, $\lambda_p \geq 1$ implies that $\lambda_{2p} \geq 1$ and thus we have $\lambda_p \geq 1$ for all $p = 2^r$. To show the upper bound on the coefficients λ_p, define

$$\kappa = 2 \left(\alpha \left(1 - \frac{\alpha}{\gamma^2} \right) \right)^{-1}.$$

Since $\kappa > 2$ and $\lambda_p \geq 1$, we have $\lambda_{2p} \leq \kappa \lambda_p^2$. Since $\lambda_2 = 1$, this implies

$$\lambda_{2p} \leq \kappa^{p-1}.$$

□

By the Riesz-Thorin interpolation theorem, Proposition 3.2 immediately generalizes to the case $p \neq 2^r$. Corollary 3.3 and Lemma 3.5 can be used to extend Proposition 3.2 to L_p-L_q-bounds and to the semigroup $q_{j,k}$.

3.4 Error Bounds from Poincaré and Logarithmic Sobolev Inequalities

In Section 3.4.1 we show that assuming (global) Poincaré inequalities associated with a sufficiently large spectral gaps for the kernels K_k is sufficient for guaranteeing that the results on L_p-Stability of the Feynman-Kac semigroup $q_{j,k}$ from Section 3.3 can be applied. In Section 3.4.2 we then give a more explicit version of our error bound for Sequential MCMC in terms of the constants in Poincaré and Logarithmic Sobolev Inequalities.

We add one additional assumption for the remainder of Chapter 3: Assume that K_k is reversible with respect to μ_k, i.e., for all $f, g \in B(E)$

$$\mu_k(gK_k(f)) = \mu_k(fK_k(g)).$$

Reversibility is, for instance, fulfilled by construction for Metropolis chains.

3.4.1 Poincaré Inequalities and Stability of Feynman-Kac Semigroups

Our stability results of Section 3.3 relied on the assumption (3.5), namely,

$$\|\hat{q}_{k-1,k}(f)\|^2_{L_2(\mu_{k-1})} \leq \alpha \|f\|^2_{L_2(\mu_k)} + \beta \mu_k(f)^2.$$

for coefficients $\alpha > 0$ and $\beta \in [0, 1]$ where

$$\hat{q}_{k-1,k}(f) = K_{k-1}\left(\overline{g}_{k-1,k}f\right) \quad \forall f \in B(E)$$

and on the assumption (3.3) of an upper bound γ on the normalized relative densities $\overline{g}_{k-1,k}$. Additionally, we needed conditions assuming that α is sufficiently small.

In the following we relate condition (3.5) to Poincaré inequalities for the transition kernels K_k. We first show that (3.5) holds, provided that the following L_2-inequalities for the kernels K_k are satisfied: For $1 \leq k \leq n$, and some $\rho \in (0, 1)$ assume that for all $f \in B(E)$,

$$\mu_k(K_k(f - \mu_k(f))^2) \leq (1 - \rho)\mathrm{Var}_{\mu_k}(f). \tag{3.12}$$

Then we observe the following:

Lemma 3.6. *Assume that* (3.12) *is satisfied for some* $\rho \in (0, 1)$. *Then* (3.5) *holds with*

$$\alpha = (1 - \rho)\gamma \text{ and } \beta = \rho$$

Proof. Note that we can write

$$\begin{aligned} \|\hat{q}_{k-1,k}(f)\|^2_{L_2(\mu_{k-1})} &= \mu_{k-1}(K_{k-1}(\overline{g}_{k-1,k}f)^2) \\ &= \mu_{k-1}(K_{k-1}(\overline{g}_{k-1,k}f - \mu_{k-1}(\overline{g}_{k-1,k}f))^2) + \mu_{k-1}(\overline{g}_{k-1,k}f)^2. \end{aligned}$$

Thus by (3.12) we have

$$\begin{aligned} \|\hat{q}_{k-1,k}(f)\|^2_{L_2(\mu_{k-1})} &\leq (1 - \rho)\left(\mu_{k-1}((\overline{g}_{k-1,k}f)^2) - \mu_{k-1}(\overline{g}_{k-1,k}f)^2\right) + \mu_{k-1}(\overline{g}_{k-1,k}f)^2 \\ &\leq (1 - \rho)\gamma\mu_k(f^2) + \rho\mu_k(f)^2, \end{aligned}$$

which proves the claim. □

We next show how the constant ρ from (3.12) can be controlled in terms of lower bounds on the spectral gaps of the kernels K_k.

Lemma 3.7. *Assume that for all* $1 \leq k \leq n$ *we have a* $\lambda_k \in (0, 1)$ *such that* K_k *fulfills a Poincaré inequality with constant* λ_k*:*

$$\lambda_k\, \mu_k(f^2) \leq \mu_k(f\,(I - K_k)(f)) \tag{3.13}$$

for all $f \in B(E)$ *with* $\mu(f) = 0$ *where* I *denotes the identity mapping on* E. *Then we have*

$$\mu_k(K_k(f - \mu_k(f))^2) \leq (1 - \lambda_k)^2\, Var_{\mu_k}(f)$$

for all $f \in B(E)$. *In particular, (3.12) holds with*

$$\rho = \min_k\, (1 - \lambda_k)^2$$

Proof. By (3.13) we have for $f \in B(E)$ with $\mu(f) = 0$ and $f \not\equiv 0$,

$$\frac{\mu_k(f\, K_k(f))}{\mu_k(f^2)} \le 1 - \lambda_k,$$

and thus the second largest eigenvalue of K_k is bounded from above by $1-\lambda_k$. Thus the second largest eigenvalue of K_k^2 is bounded from above by $(1-\lambda_k)^2$, i.e.,

$$\frac{\mu_k(f\, K_k^2(f))}{\mu_k(f^2)} \le (1 - \lambda_k)^2.$$

By the reversibility of K_k, this is equivalent to

$$\mu_k(K_k(f)^2) \le (1-\lambda_k)^2 \mu_k(f^2).$$

To conclude the proof observe that thus for $f \in B(E)$

$$\mu_k(K_k(f - \mu(f))^2) \le (1-\lambda_k)^2 \mu_k((f - \mu_k(f))^2).$$

□

Thus, assuming (3.12) is essentially equivalent to assuming a Poincaré inequality. In the algorithm, ρ can be controlled by varying the number of MCMC steps: A sufficiently large number of MCMC steps makes ρ large and accordingly it makes α small. For future reference, we also give a version of Proposition 3.1 under the stronger assumption that (3.12) holds for some sufficiently large $\rho \in (0,1)$. This follows immediately from the Proposition by inserting the values of α and β from Lemma 3.6.

Corollary 3.5. *Assume that* (3.12) *holds for some* ρ *with* $(1-\rho)\gamma^{p-1} < 1$. *Consider* $p = 2^r$ *for* $r \in \mathbb{N}$. *Then we have for* $1 \le j < k \le n$ *and* $f \in B(E)$ *the inequality*

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \le \delta(p) \|f\|_{L_p(\mu_k)}$$

with

$$\delta(p) = \prod_{j=1}^{r} \frac{\gamma^{1-2^{-(j-1)}}}{(1-(1-\rho)\gamma^{2^j-1})^{2^{-j}}} < \frac{\gamma^{r-2+2^{-(r-1)}}}{1-(1-\rho)\gamma^{2^r-1}}$$

3.4.2 Explicit Error Bounds

In this section we introduce a further parameter t_k for our transition operators K_k which is thought to be the running time or number of MCMC steps contained in K_k. We write $K_k^{t_k}$ in the following to make this dependence clear. We assume the following two inequalities: the hypercontractivity inequality

$$\|K_k^{t_k}(f)\|_{L_{q(p,t_k)}(\mu_k)} \le \|f\|_{L_p(\mu_k)} \tag{3.14}$$

for $f \in B(E)$ and $q(p, t_k) = 1 + (p-1)\exp(2a_k^* t_k)$ and the L_2-L_2 inequality

$$\|K_k^{t_k}(f) - \mu_k(f)\|^2_{L_2(\mu_k)} \leq \exp(-2b_k^* t_k)\ \|f - \mu_k(f)\|^2_{L_2(\mu_k)}. \tag{3.15}$$

for $f \in B(E)$ and for positive constants a_k^* and b_k^*. These two inequalities follow, respectively, from a Logarithmic Sobolev inequality and a Poincaré inequality for the underlying MCMC dynamics, see e.g. Deuschel and Stroock (1990) or Ané et al (2000) and the example of Section 3.5.2.

Furthermore we assume as before that the μ_k are chosen in a way that $\gamma > 1$ is a uniform upper bound on the relative densities, i.e., for $0 \leq k < n$ and for all $x \in E$ we assume

$$\overline{g}_{k,k+1}(x) = \frac{g_{k,k+1}(x)}{\mu_k(g_{k,k+1})} \leq \gamma. \tag{3.16}$$

We then have the following bounds for $q_{j,k}$ provided that the running times $t_j, \dots, t_k$ are chosen sufficiently large.

Proposition 3.3. *Fix $0 \leq j < k \leq n$, $\gamma > 1$ and $\tau \in (0,1)$, $1 \leq s \in \mathbb{N}$ and $p = 2^s$. Assume that (3.16) holds for all $x \in E$ and for all $j \leq l \leq k-1$ and that the contraction inequalities (3.14) and (3.15) are satisfied for $j \leq l \leq k$. Assume furthermore that for $j \leq l \leq k$,*

$$t_l \geq \frac{1}{2b_l^*}\left[(p-1)\log(\gamma) - \log(1-\tau)\right]. \tag{3.17}$$

Then we have for $f \in L_p(\mu_k)$

$$\|q_{j,k}(f)\|_{L_p(\mu_j)} \leq \widetilde{c}_{j,k}(p,p)\|f\|_{L_p(\mu_k)}$$

with

$$\widetilde{c}_{j,k}(p,p) = \frac{\gamma^{s-1+1/p}}{\tau}.$$

If in addition we have for $p' > p$ and $j \leq l \leq k$,

$$t_j \geq \frac{1}{2a_j^*}\left[\log(p'-1) - \log(p-1)\right], \tag{3.18}$$

then we have for $f \in L_p(\mu_k)$

$$\|q_{j,k}(f)\|_{L_{p'}(\mu_j)} \leq \widetilde{c}_{j,k}(p',p)\|f\|_{L_p(\mu_k)}$$

with

$$\widetilde{c}_{j,k}(p',p) = \frac{\gamma^{s-1+1/p}}{\tau}\gamma^{\frac{p'-1}{p'}}.$$

Here and in the following, it is straightforward to relax the requirement of $p = 2^s$, see Corollary 3.2. The constant τ controls the contractivity in L^2 of the MCMC steps in the following sense: In order to apply our error bounds of Section 3.4.1, namely, Corollary 3.5, we need to ensure that

$$0 < 1 - \gamma^{p-1}(1-\rho_l) \overset{(3.15)}{=} 1 - \gamma^{p-1} e^{-a_l^* t_l}.$$

τ is a uniform measure of by how much this inequality is satisfied, i.e., τ is assumed to be a constant with

$$\tau < 1 - \gamma^{p-1}(1 - e^{-a_l^* t_l}) \quad \text{for all} \quad j \le l \le k$$

Our next step is to utilize the constants $\widetilde{c}_{j,k}(p',p)$ in order to bound the constants that arise in the error bound of Corollary 3.1.

Corollary 3.6. *Fix* $0 \le j < k \le n$, $\gamma > 1$ *and* $\tau \in (0,1)$, $2 \le s \in \mathbb{N}$ *and* $p = 2^s$. *Assume that* (3.16) *holds for all* $x \in E$ *and for all* $0 \le l \le n-1$ *and that the contraction inequalities* (3.14) *and* (3.15) *are satisfied for* $1 \le l \le n$. *Assume furthermore that for* $1 \le l \le n$,

$$t_l \ge \frac{1}{2b_l^*}\left[(2p-1)\log(\gamma) - \log(1-\tau)\right], \tag{3.19}$$

and

$$t_l \ge \frac{1}{2a_j^*}\left[\log(p-1) - \log\left(\frac{p}{2}-1\right)\right]. \tag{3.20}$$

Define

$$h(p) = \frac{\gamma^{2s+\frac{1}{p}}}{\tau^2}.$$

Then we have

$$c_{j,k}(p) \le h(p).$$

Furthermore

$$\widehat{c}_k(p) \le \overline{c}_k(p) \le ((1+\gamma) \vee 3)\, k\, h(p), \qquad \textit{and} \qquad \widehat{v}_k(p) \le \overline{v}_k(p) \le (k+1)\frac{\gamma}{\tau^2}.$$

Adding the final observations that for $f \in L_p(\mu_n)$ and $0 \le j < n$,

$$\mathrm{Var}_{\mu_j}(q_{j,n}(f)) \le b_{j,n}(2,2)^2 \|f\|^2_{L_p(\mu_n)},$$

and

$$\mathrm{Var}_{\mu_n}(q_{n,n}(f)) \le \|f\|^2_{L_p(\mu_n)},$$

we are now in the position to bound all the terms in the error bound of Corollary 3.1 through γ, τ, p, N and n. Thus we arrive at the following version of the corollary:

Corollary 3.7. *Fix* $0 < n \in \mathbb{N}$, $\gamma > 1$, $\tau \in (0,1)$, $2 \le s \in \mathbb{N}$ *and* $p = 2^s$. *Assume that* (3.16) *holds for all* $x \in E$ *and for all* $0 \le l \le n-1$ *and that the contraction inequalities* (3.14) *and* (3.15) *are satisfied for* $1 \le l \le n$. *Assume furthermore that* *for* $1 \leq l \leq n$,

$$t_l \geq \frac{1}{2b_l^*} \left[(2p-1)\log(\gamma) - \log(1-\tau)\right]$$

and

$$t_l \geq \frac{1}{2a_l^*} \left[\log(p-1) - \log\left(\frac{p}{2} - 1\right)\right].$$

Finally assume that

$$N \geq 2((1+\gamma) \vee 3) n \gamma^{2s+p^{-1}} \tau^{-2}. \tag{3.21}$$

Then for $f \in L_p(\mu_n)$ *we have*

$$N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2]$$

$$\leq \|f\|^2_{L_p(\mu_n)} \left[1 + n\gamma\tau^{-2} + n((1+\gamma) \vee 3)\gamma^{2s+p^{-1}} \tau^{-2} \overline{\varepsilon}_n^{N,p}\right]$$

and

$$\overline{\varepsilon}_n^{N,p} \leq 2 \frac{1 + n\gamma\tau^{-2}}{N}$$

Finally, for the sake of illustration we also state these bounds for a concrete choice of parameters, namely $\gamma = 2$, $\tau = 0.8$, $p = 4$ and thus $s = 2$. After rounding the coefficients to improve readability (in a way that makes the inequality slightly worse) and inserting the bound on $\overline{\varepsilon}_n^{N,p}$, this yields the bound

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \ \leq \ \|f\|^2_{L_4(\mu_n)} \left[\frac{1+4n}{N} + \frac{180n + 560n^2}{N^2}\right]$$

The required lower bound (3.21) on N is given by

$$N \geq 180$$

in this case.

Proof of Proposition 3.3. Comparing inequalities (3.12) and (3.15) shows that the terms $\exp(-2b_l^* t)$ play the role of $(1-\rho)$ in the setting of Section 3.4.1. Thus assuming for all $j \leq l \leq k$

$$1 - e^{-2b_l^* t_l} \gamma^{p-1} > \tau$$

or, equivalently, (3.17), ensures that we can apply Corollary 3.5 to obtain the bound

$$\|\hat{q}_{j,k}(f)\|_{L_p(\mu_j)} \leq \delta(p) \|f\|_{L_p(\mu_k)} \tag{3.22}$$

with

$$\delta(p) = \frac{\gamma^{s-2+2/p}}{\tau}.$$

Now applying Lemma 3.5 allows to conclude from this bound for $\hat{q}_{j,k}$ the desired

L_p-L_p-bound for $q_{j,k}$ with

$$\widetilde{c}(p,p) = \frac{\gamma^{s-2+2/p}}{\tau}\gamma^{\frac{p-1}{p}} = \frac{\gamma^{s-1+1/p}}{\tau}.$$

We next turn to the $L_{p'}$-L_p-bound. By (3.14), ensuring

$$p' \leq 1 + (p-1)e^{2a_j^* t}$$

or, equivalently, (3.18) is a sufficient condition for

$$\|K_j^{t_j}(f)\|_{L_{p'}(\mu_j)} \leq \|f\|_{L_p(\mu_j)}.$$

Thus we conclude from applying first Corollary 3.3 and then Lemma 3.5 to (3.22) the desired $L_{p'}$-L_p-bound for $q_{j,k}$ with

$$\widetilde{c}_{j,k}(p',p) = \frac{\gamma^{s-1+1/p}}{\tau}\gamma^{\frac{p-1}{p}}\gamma^{\frac{p'-1}{p'}} = \frac{\gamma^{s-1+1/p}}{\tau}\gamma^{\frac{p'-1}{p'}}.$$

□

Proof of Corollary 3.6. Choose $\widetilde{c}_{j,k}(p',p)$ as in Proposition 3.3. Since

$$c_{j,k}(p) = \max\left(1, \widetilde{c}_{j,k}\left(p, \frac{p}{2}\right), \widetilde{c}_{j,k}(2,p)^2\right),$$

we need to apply Proposition 3.3 for $(p, p/2)$ and $(2p, p)$. Note that if inequality (3.17) holds for $2p$ it also holds for p. Conversely, if (3.18) holds for $(p, p/2)$ it also holds for $(2p, p)$ since

$$\frac{2x-1}{x-1} = 2 + \frac{1}{x-1}$$

is decreasing in x. This motivates our assumption of (3.19) and (3.20). Now observe that

$$1 \leq \widetilde{c}_{j,k}\left(p, \frac{p}{2}\right) \leq \widetilde{c}_{j,k}(2p,p)^2 = \frac{\gamma^{2s+\frac{1}{p}}}{\tau^2} = h(p)$$

and thus we have

$$c_{j,k}(p) \leq h(p).$$

Now observe that for all $x \in E$ we have

$$-1 \leq q_{j,j+1}(1)(x) - 1 = \overline{g}_{j,j+1}(x) - 1 \leq \gamma - 1$$

and thus we have

$$\|q_{j,j+1}(1) - 1\|_{L_p(\mu_j)} \leq ((1+\gamma) \vee 3)$$

and can bound $\widehat{c}_k(p)$ as follows:

$$\widehat{c}_k(p) = \sum_{j=0}^{k-1} c_{j,k}(p)(2 + \|q_{j,j+1}(1)(x) - 1\|_{L_p(\mu_j)}) \leq ((1+\gamma) \vee 3)\, h(p)\, k.$$

Since this upper bound is monotonically increasing in k, it also applies to $\overline{c}_k(p)$. We now turn to $\widehat{v}_k(p)$. Observe that we have

$$\begin{aligned}\widehat{v}_k(p) &= \sup\left\{ \sum_{j=0}^{k} \mathrm{Var}_{\mu_j}(q_{j,k}(f)) \,\middle|\, f \in B(E), \|f\|_{L_p(\mu_k)} \leq 1 \right\} \\ &= \sup\left\{ \sum_{j=0}^{k} \mu_j(q_{j,k}(f)^2) \,\middle|\, f \in B(E), \|f\|_{L_2(\mu_k)} \leq 1 \right\} \\ &\leq \sum_{j=0}^{k} \widetilde{c}_{j,k}(2,2)^2.\end{aligned}$$

As we have

$$\widetilde{c}_{j,k}(2,2) = \frac{\sqrt{\gamma}}{\tau},$$

we get the desired upper bound on $\widehat{v}_k(p)$. Since this upper bound is increasing in k, it also applies to $\overline{v}_k(p)$. □

3.5 Example: Moving Gaussians

In this section we apply our quantitative convergence bounds to an example where the distributions μ_k are Gaussian distributions moving in $\mathbb{R}^d$. For technical reasons, namely to guarantee bounded relative densities, we consider only Gaussians restricted to a bounded set. In Section 3.5.1, we prove uniform bounds on the relative density of μ_{k+1} with respect to μ_k in the case where μ_k and μ_{k+1} do not differ too much. In Section 3.5.2, we specify concrete operators K_k, namely, the transition kernels associated with Langevin diffusions, and recall their contraction properties. Together, these results allow to ensure that the bound on relative densities (3.16) and the contraction inequalities (3.14) and (3.15) are satisfied in this class of examples so that the error bound of Corollary 3.7 can be applied.

3.5.1 Bounds on Relative Densities

We begin with a number of definitions. Fix a dimension d. Denote by $\mathcal{D}$ the set of diagonal matrices in $\mathbb{R}^{d\times d}$ with strictly positive diagonal entries. Denote by $\mathcal{R}$ the rotation matrices in $\mathbb{R}^{d\times d}$, i.e., the orthogonal matrices with determinant $+1$.

For $x, m \in \mathbb{R}^d$, $A \in \mathcal{D}$ and $Q \in \mathcal{R}$, denote by $h(x,m,A,Q)$ the density of the Gaussian distribution with mean m, and inverse covariance matrix $Q^T A Q$, i.e.,

$$h(x,m,A,Q) = \frac{\sqrt{\det A}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}(x-m)^T Q^T A Q (x-m)\right).$$

Note that any Gaussian density in $\mathbb{R}^d$ can be written in this form.

Since our previous results require absolute bounds on the relative densities between the μ_k and since the quotient between, e.g., $h(x, m, A, Q)$ and $h(x, m', A, Q)$ is unbounded for $m \neq m'$, we resort in the following to Gaussian distributions restricted to a ball around zero. In order to control the normalizing constants arising from restricting the distribution it proves to be helpful to assume a lower bound on the diagonal entries of A.

Denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^d$. Denote by $B_r(x)$ the d-dimensional Euclidean ball around x with radius $r > 0$. For $\underline{a} > 0$ define by $\mathcal{D}_{\underline{a}} \subset \mathcal{D}$ the diagonal matrices with diagonal entries weakly greater than $\underline{a}$, i.e.,

$$\mathcal{D}_{\underline{a}} = \{(a_{ij})_{1 \leq i,j \leq d} \in \mathbb{R}^{d \times d} | a_{ii} \geq \underline{a}, a_{ij} = 0 \text{ for } i \neq j\}.$$

Define $Z(r, m, A, Q)$ as the mass put by $h(\cdot, m, A, Q)$ in $B_{2r}(0)$, i.e.,

$$Z(r, m, A, Q) = \int\limits_{B_{2r}(0)} h(x, m, A, Q) dx.$$

Recall that $h(x, m, A, Q)$ is a normalized density on $\mathbb{R}^d$ and thus $Z(r, m, A, Q)$ only takes into account the change in mass due to restricting the state space to $B_{2r}(0)$. We will in the following consider movements within the following class $G_{\underline{a},r}$ of probability measures on $E = B_{2r}(0) \subset \mathbb{R}^d$,

$$G_{\underline{a},r} = \Big\{ \mu \in M_1(E) \;\Big|\; \mu(dx) = \frac{1}{Z(r, m, A, Q)} h(x, m, A, Q)\, 1_{\{x \in E\}}\, dx$$
$$\text{where } A \in \mathcal{D}_{\underline{a}}, Q \in \mathcal{R}, m \in B_r(0) \Big\}$$

where $\underline{a} > 0$ and $r > 0$. This the class on Gaussian distributions restricted to $B_{2r}(0)$ with center $m \in B_r(0)$ and with, roughly speaking, the variance in each (rotated) coordinate direction bounded from above by $\underline{a}^{-1}$. Observe however that due to the restriction on $B_{2r}(0)$, m and $Q^T A Q$ are not identical to the mean and the inverse covariance matrix of μ.

Since the elements of the matrices A are bounded from below and since we assumed m to be bounded away from the boundary, we can control the variation of Z by choosing r sufficiently large: For sufficiently large r, most mass is contained in $B_r(m) \subset E$ and thus restricting does not change the normalizing constant much. This is made precise in the following lemma:

Lemma 3.8. *Fix real-valued constants $\theta_1 > 1$, $r > 0$ and $\underline{a} > 0$. Assume the following inequality*

$$\nu\left(\left[-\frac{r\underline{a}}{\sqrt{d}}, \frac{r\underline{a}}{\sqrt{d}}\right]\right)^d > \frac{1}{\theta_1} \tag{3.23}$$

where ν denotes the standard Gaussian distribution on $\mathbb{R}$ with mean 0 and variance

1. *Then for all $m \in B_r(0)$, $A \in \mathcal{D}_{\underline{a}}$ and $Q \in \mathcal{R}$*

$$\frac{1}{\theta_1} \leq Z(r, m, A, Q) \leq 1.$$

All proofs are at the end of the section. Note that for fixed d and $\underline{a}$ inequality (3.23) is always fulfilled for sufficiently large r.

We now fix a finite sequence $(\mu_k)_{k=0}^n$ in $G_{\underline{a},r}$ and define A_k, Q_k and m_k as the diagonal matrix, rotation matrix and offset vector associated with μ_k. By f_k we denote the density of μ_k with respect to the Lebesgue measure on $\mathbb{R}^d$, i.e.,

$$f_k(x) = \frac{1}{Z(r, m_k, A_k, Q_k)} h(x, m_k, A_k, Q_k) 1_{\{x \in E\}}$$

We denote the diagonal elements of A_k by a_k^1,..., a_k^d.

We are interested in uniform upper bounds on $\overline{g}_{k,k+1} = f_{k+1}/f_k$ for the case where μ_k and μ_{k+1} do not differ too much in a sense that is made precise now. We restrict attention to three types of movements: *(i)* Shifts of m_k by a vector, *(ii)* changing one diagonal entry of A_k, and *(iii)* applying a rotation in the (i, j)-plane to Q_k. Note that we can interpolate between any two elements of $G_{\underline{a},r}$ using a sequence of these three types of movements.

In order to state our result we need one more definition: For $1 \leq i, j \leq d$ and $\varphi \in \mathbb{R}$, denote by $R_{ij}(\varphi) \in \mathcal{R}$ the rotation by the angle φ in the (i, j)-plane: $R_{ij}^{ii} = R_{ij}^{jj} = \cos(\varphi)$, $R_{ij}^{ij} = -\sin(\varphi)$, $R_{ij}^{ji} = \sin(\varphi)$, $R_{ij}^{kk} = 1$ for $k \notin \{i, j\}$ and $R_{ij}^{kl} = 0$ for $k \neq l$ with $\{k, l\} \neq \{i, j\}$ where R_{ij}^{kl} denotes the $(k, l)^{\text{th}}$ entry of R_{ij}.

Proposition 3.4. *Fix real-valued constants $\theta_1 > 1$, $\theta_2 > 1$, $r > 0$ and $\underline{a} > 0$. Assume that θ_1, r and $\underline{a}$ fulfill (3.23) and that $\mu_k \in G_{\underline{a},r}$ and $\mu_{k+1} \in G_{\underline{a},r}$ stand in one of the following relationships:*

(i) $A_{k+1} = A_k$, $Q_{k+1} = Q_k$ *and* $m_{k+1} = m_k + v$ *with*

$$\|v\| \leq \frac{\log \theta_2}{3r \max_i a_k^i}, \tag{3.24}$$

(iia) $Q_{k+1} = Q_k$, $m_{k+1} = m_k$, $a_{k+1}^j = a_k^j$ *for* $j \neq i$ *and* $a_{k+1}^i = \alpha a_k^i$ *with*

$$1 < \alpha < (\theta_1 \theta_2)^2, \tag{3.25}$$

(iib) $Q_{k+1} = Q_k$, $m_{k+1} = m_k$, $a_{k+1}^j = a_k^j$ *for* $j \neq i$ *and* $a_{k+1}^i = \alpha a_k^i$ *with*

$$\max\left(1 - \frac{2 \log \theta_2}{9r^2 a_k^i}, 0\right) < \alpha < 1, \text{ or} \tag{3.26}$$

(iii) $A_{k+1} = A_k$*,* $m_{k+1} = m_k$ *and* $Q_{k+1} = R_{ij}(\varphi)Q_k$ *with*

$$|\sin(\varphi)| \leq \frac{2\log(\theta_2)}{9r^2|a_k^j - a_k^i|} \tag{3.27}$$

for some for $1 \leq i \neq j \leq d$*,*

Then for all $x \in E$*,*

$$\frac{f_{k+1}(x)}{f_k(x)} \leq \theta_1\theta_2. \tag{3.28}$$

Recall that we have further restricted the admissible movements by requiring $a_k^i \geq \underline{a}$ and $m_k \in B_r(0)$ for all k.

The larger r gets, the smaller are the changes we can make to the distribution while retaining our bound on the relative density. An exception are movements of type *(iia)* which decrease the variance by increasing a diagonal element of A_k. In that case, the unnormalized density decreases everywhere and only the (global) change in normalizing constants has to be bounded. Notably, for considering this type of movement the restriction to a bounded domain is not necessary. As argued in Section 1.4.2.2, this case is luckily the most relevant one for the MCMC applications of the algorithm we have in mind.

For case *(iii)* of a rotation by an angle φ in the (i, j)-plane, note that we have to choose φ smaller when a_k^i and a_k^j differ more strongly. When a_k^i and a_k^j are sufficiently similar, the upper bound is large enough to make no restriction.

Together, Lemma 3.8 and Proposition 3.4 give a guideline on how to choose the distributions μ_k in order to ensure that

$$\overline{g}_{k,k+1} \leq \theta_1\theta_2$$

for $0 \leq k \leq n-1$ and thus to ensure that inequality (3.16) in the prerequisites of Corollary 3.7 is fulfilled.

Proof of Lemma 3.8. Since $h(\cdot, m, A, Q)$ is a probability density on $\mathbb{R}^d$,

$$Z(r, m, A, Q) \leq 1$$

follows immediately. Denote by $C_s(m)$ the d-dimensional cube with side-length s and center m. Observe that for any rotation matrix $Q \in \mathcal{R}$ and $m \in B_r(0)$ we have

$$Q\ C_{\frac{2r}{\sqrt{d}}}(m) \subseteq B_r(m) \subseteq B_{2r}(0) = E.$$

Therefore we have

$$Z(r, m, A, Q) \geq \int\limits_{C_{\frac{2r}{\sqrt{d}}}(0)} h(x, 0, A, I)\, dx$$

where I denotes the d-dimensional identity matrix. We can thus conclude the lower bound

$$Z(r,m,A,Q) > \prod_{i=1}^{d} \nu\left(\left[-\frac{r\sqrt{a^i}}{\sqrt{d}}, \frac{r\sqrt{a^i}}{\sqrt{d}}\right]\right) \geq \nu\left(\left[-\frac{r\sqrt{a}}{\sqrt{d}}, \frac{r\sqrt{a}}{\sqrt{d}}\right]\right)^d$$

where a^i denotes the i^{th} diagonal entry of A. □

Proof of Proposition 3.4. We first consider case *(i)*. In that case we have for $x \in E$

$$\frac{f_{k+1}(x)}{f_k(x)} = \frac{Z(r,m_k,A_k,Q_k)}{Z(r,m_k+v,A_k,Q_k)} e^{\frac{1}{2}(Q_k(x-m_k))^T A_k (Q_k(x-m_k)) - \frac{1}{2}(Q_k(x-m_k-v))^T A_k (Q_k(x-m_k-v))}. \tag{3.29}$$

By Lemma 3.8 we have

$$\frac{Z(r,m_k,A_k,Q_k)}{Z(r,m_k+v,A_k,Q_k)} \leq \theta_1.$$

Now define $z = Q_k(x-m_k)$ and $w = -Q_k v$. Note that by the triangle inequality and since Q_k is a rotation matrix,

$$\|z\| = \|x - m_k\| \leq \|x\| + \|m_k\| \leq 3r.$$

Moreover $\|v\| = \|w\|$. We can now rewrite and bound the term in the exponent in (3.29) as follows:

$$\begin{aligned} \frac{1}{2}z^T A_k z - \frac{1}{2}(z+w)^T A_k (z+w) &= -w^T A_k z - \frac{1}{2}w^T A_k w \leq -w^T A_k z \\ &\leq \sqrt{w^T A_k w}\sqrt{z^T A_k z} \leq \max_i a_k^i \|z\|\|w\| \\ &\leq \max_i a_k^i 3r\|v\|, \end{aligned}$$

where the first inequality is Cauchy-Schwarz. Thus assuming

$$\max_i a_k^i \, 3r\|v\| \leq \log(\theta_2)$$

which is equivalent to (3.24) implies that from (3.29) we can conclude (3.28).

We now turn to cases *(iia)* and *(iib)*. Defining again $z = Q_k(x-m_k)$ and observing that

$$\frac{\sqrt{\det A_{k+1}}}{\sqrt{\det A_k}} = \sqrt{\alpha}$$

we get

$$\begin{aligned}\frac{f_{k+1}(x)}{f_k(x)} &= \frac{Z(r,m_k,A_k,Q_k)}{Z(r,m_k,A_{k+1},Q_k)}\sqrt{\alpha}\; e^{\frac{1}{2}z^TA_kz-\frac{1}{2}z^TA_{k+1}z} \\ &= \frac{Z(r,m_k,A_k,Q_k)}{Z(r,m_k,A_{k+1},Q_k)}\sqrt{\alpha}\; e^{\frac{1}{2}(1-\alpha)a_k^iz_i^2}. \qquad (3.30)\end{aligned}$$

Consider first the case $\alpha > 1$, i.e., decreasing the variance in one direction. Then, the exponential term in (3.30) can be bounded by 1. Furthermore, in that case

$$Z(r,m_k,A_k,Q_k) < Z(r,m_k,A_{k+1},Q_k)$$

since μ_{k+1} is more concentrated than μ_k and thus closer to the unrestricted Gaussian distribution. Thus we have

$$\frac{f_{k+1}(x)}{f_k(x)} \leq \sqrt{\alpha}$$

so that assuming $\alpha \leq \theta_1^2\theta_2^2$ ensures (3.28) to hold as desired. In the case $\alpha < 1$, we bound the quotient of normalizing constants by θ_1 using Lemma 3.8 and bound $\sqrt{\alpha}$ by 1. Since $z \in B_{3r}(0)$ we have $z_i^2 \leq 9r^2$. Thus assuming

$$\frac{9}{2}(1-\alpha)\,a_i\,r^2 \leq \log(\theta_2)$$

or, equivalently, (3.26) ensures that (3.30) implies (3.28).

We now turn to case *(iii)*. Again we introduce some short-hand notation: $z = Q_k(x-m_k)$, $R = R_{ij}(\varphi)$, $c = \cos(\varphi)$ and $s = \sin(\varphi)$. Applying Lemma 3.8 to bound the normalizing constants yields

$$\frac{f_{k+1}(x)}{f_k(x)} \leq \theta_1 e^{\frac{1}{2}z^TA_kz-\frac{1}{2}(Rz)^TA_kRz}.$$

An elementary calculation yields that, using the relation $c^2+s^2=1$, the exponent can be rewritten as follows:

$$\begin{aligned}&\frac{1}{2}z^TA_kz-\frac{1}{2}(Rz)^TA_kRz \\ =\;& \frac{1}{2}a_k^iz_i^2+\frac{1}{2}a_k^jz_j^2-\frac{1}{2}(a_k^i(cz_i+sz_j)^2+a_k^j(cz_j+sz_i)^2) \\ =\;& \frac{1}{2}(a_k^j-a_k^i)s(2cz_iz_j+sz_j^2-sz_i^2) \leq \frac{1}{2}|a_k^j-a_k^i|\,|s|\,|\hat{z}^T\hat{Q}\hat{z}|,\end{aligned}$$

where $\hat{z} \in \mathbb{R}^2$ is defined as $\hat{z} = (z_i, z_j)$ and $\hat{Q} \in \mathbb{R}^{2\times 2}$ is given by $\hat{Q}_{11} = -s$,$\hat{Q}_{22} = s$, $\hat{Q}_{12} = c$ and $\hat{Q}_{21} = c$. Since $\hat{Q}$ is an orthogonal matrix and since $\hat{z}^T\hat{z} \leq 9r^2$ we get from applying the Cauchy-Schwarz inequality that

$$|\hat{z}^T \hat{Q} \hat{z}| \leq \sqrt{\hat{z}^T \hat{z}} \sqrt{(\hat{Q}\hat{z})^T \hat{Q} \hat{z}} \leq 9r^2.$$

We thus have

$$\frac{1}{2} z^T A_k z - \frac{1}{2} (Rz)^T A_k z \leq \frac{1}{2} |a_k^j - a_k^i| |\sin(\varphi)| 9r^2.$$

Therefore (3.27) is a sufficient condition for (3.28) in case *(iii)*. □

3.5.2 Reflected Langevin Diffusions

In order to apply our explicit error bounds, we need L_{2p}-L_p and L_2-L_2 bounds on the semigroup $q_{j,k}$. In Section 3.4.1 we showed how to derive such bounds from a uniform upper bound γ on relative densities and from L_{2p}-L_p and L_2-L_2 bounds on the MCMC kernels K_k. In this subsection we thus introduce concrete kernels K_k for our example of Gaussian distributions restricted to a ball and recall their contraction properties.

Instead of a discrete-time MCMC dynamics, we choose to move the particles with Langevin dynamics reflected at the boundary of $E = B_{2r}(0)$ since for this type of dynamics the required mixing properties can be verified in a straightforward way. We thus choose

$$(K_k\, f)(x) = \mathbb{E}[f(X_{t_k}^{x,k})]$$

where the d-dimensional diffusion process $X_t^{x,k}$ is the Langevin diffusion with start in $x \in E$, target measure μ_k and with reflection at the boundary of E. $t_k > 0$ is the running time of the diffusion process. t_k thus corresponds to the number of MCMC steps we make. For the corresponding non-reflected diffusion process, i.e., $r = \infty$, it is well-known (see e.g. Ané et al. (2000, Chapter 5) or Deuschel and Stroock (1990)) that by the Bakry-Émery criterion the semigroup associated with $X_t^{x,k}$ fulfills a Logarithmic Sobolev inequality with constant c provided that the Hamiltonian H associated with the stationary measure is C^2 and fulfills the inequality

$$x^T (\text{Hess } H)(x)\, x \geq c\, x^T x \tag{3.31}$$

for all $x \in E$. Corollary 3.2 of Wang (1997) extends this result to diffusions reflected at the boundary of a manifold with convex boundary such as our Langevin diffusion on $B_{2r}(0)$. For our families of measures μ_k we have

$$(\text{Hess } H)(x) = Q_k^T A_k Q_k$$

for all $x \in E$ and thus (3.31) holds with

$$c = \min_i a_k^i =: a_k^*.$$

We thus have (see e.g. Deuschel and Stroock (1990) or Ané et al (2000)) the hypercontractivity inequality

$$\|K_k(f)\|_{L_{q(p,t_k)}(\mu_k)} \leq \|f\|_{L_p(\mu_k)} \tag{3.32}$$

for $f \in B(E)$ and $q(p, t_k) = 1 + (p-1)\exp(2a_k^* t_k)$. Since the spectral gap can be bounded from below by the Logarithmic Sobolev constant (see Deuschel and Stroock (1990) or Chen and Wang (1997)) we furthermore have the L_2-L_2 inequality

$$\|K_k(f) - \mu_k(f)\|_{L_2(\mu_k)}^2 \leq \exp(-2a_k^* t_k)\ \|f - \mu_k(f)\|_{L_2(\mu_k)}^2. \tag{3.33}$$

These are the contractivity inequalities (3.14) and (3.15) needed for our error bounds of Corollary 3.7. Combining these observations with the bound on relative densities from Proposition 3.4 we have thus shown how Corollary 3.7 can be applied to moving Gaussian distributions restricted to the ball $B_{2r}(0)$.

4 Sequential MCMC on Trees

In this section we study the ability of our Sequential MCMC algorithm to explore a multimodal state space by abstracting from the problem of mixing within modes: We consider the algorithm on a simple tree structure. We assume that our sequence of probability distributions $(\mu_k)_k$ lives on a sequence of state spaces $(I_k)_k$ where the states in I_{k+1} have unique predecessors in I_k. Particle movements in the MCMC steps are restricted to moving from a state in I_k to one of its successors in I_{k+1}.

Section 4.1 introduces the model including the notation for the tree structure. Section 4.2 states the algorithm and the error bounds for this setting. While the algorithm considered here should be viewed as a stylized version of the one introduced in Section 1.2, it nevertheless fits into the framework of Section 2. Section 4.3 introduces an alternative algorithm, Sequential Importance Sampling, which is based on weighting particles instead of resampling them. In Section 4.4 we provide an extensive discussion of an elementary example where the error of our Sequential MCMC algorithm grows polynomially in the number of levels n while the error of Sequential Importance Sampling increases exponentially fast.

4.1 The Model

Consider a sequence of probability distributions $\mu_0, \dots, \mu_n$ on a sequence of finite state spaces $I_0, \dots, I_n$. Assume that each μ_k gives positive mass to each point in its state space I_k. Denote by $B(I_k)$ the bounded measurable functions from I_k to $\mathbb{R}$. We define a tree structure on the sequence of state spaces by introducing for $k \in \{0, \dots, n-1\}$ the predecessor function $p_k : I_{k+1} \cup \dots \cup I_n \to I_k$ which maps $x \in I_l$ to its predecessor in I_k for $l > k$. We assume transitivity of the functions p_k, i.e., for $j < k < l$ and $x \in I_l$ we assume that

$$p_j(p_k(x)) = p_j(x).$$

Denote by $\mathcal{P}(I_k)$ the collection of subsets of I_k. Conversely to p_k, we define the successor function $s_k : I_0 \cup \dots \cup I_{k-1} \to \mathcal{P}(I_k)$ as follows: For $x \in I_l$ with $0 \le l < k \le n$ the successors in I_k of x are given by

$$s_k(x) = \{y \in I_k | p_l(y) = x\}.$$

We assume that no branches die out, i.e., for all $x \in I_0 \cup \dots \cup I_{n-1}$

$$s_n(x) \neq \emptyset.$$

In order to obtain a genuine tree structure we could additionally assume that $|I_0| = 1$ but this assumption is not needed in the following (and is thus not made). Additionally, for $0 \leq k < l \leq n$, define for a probability distribution μ on I_l the probability distribution $\mu^{\rightarrow k}$ on I_k as the projection of μ to I_k: For $x \in I_k$,

$$\mu^{\rightarrow k}(x) = \mu(s_l(x)).$$

For $0 \leq k < n$, denote by $g_{k,\,k+1} \in B(I_k)$ an unnormalized relative density between μ_k and $\mu_{k+1}^{\rightarrow k}$: For all $f \in B(I_k)$

$$\mu_{k+1}^{\rightarrow k}(f) = \frac{\mu_k(f g_{k,\,k+1})}{\mu_k(g_{k,\,k+1})}.$$

Denote by $K_{k+1} : I_k \times I_{k+1} \to [0, 1]$ a Markov transition kernel for which

$$\mu_{k+1}(f) = \mu_{k+1}^{\rightarrow k}(K_{k+1}(f))$$

for all $f \in B(I_{k+1})$. Any pair of probability distributions μ_k and μ_{k+1} with full support on, respectively, I_k and I_{k+1} can be related through such a pair $(g_{k,k+1}, K_{k+1})$. Moreover K_{k+1} is unique and $g_{k,k+1}$ is unique up to a normalizing constant. For $x \in I_k$ and $y \in I_{k+1}$, K_{k+1} is given explicitly by

$$K_{k+1}(x, y) = \begin{cases} \frac{\mu_{k+1}(y)}{\mu_{k+1}(s_{k+1}(x))} & \text{if } y \in s_{k+1}(x) \\ 0 & \text{otherwise.} \end{cases}$$

The tree structure, concretely, the fact that the states in I_k are not connected by K_k, is a simple model of a multimodal state space: The elements of I_k stand for components of a continuous state space which are separated by regions of very low probability. For the particle dynamics we consider subsequently, the consequence is that particles can move between different branches only through the resampling step but not through the mutation step: This is consistent with our aim of studying, how helpful the resampling step is in overcoming problems associated with multimodality. Accordingly, μ_n is not necessarily thought to be a severely multimodal distribution on I_n – the problems associated with multimodality are captured by the tree structure.

Now define $q_{k,k+1} : B(I_{k+1}) \to B(I_k)$ by

$$q_{k,\,k+1}(f) = \frac{g_{k,\,k+1} K_{k+1}(f)}{\mu_k(g_{k,\,k+1})}$$

for all $f \in B(I_{k+1})$. Furthermore, define for $0 \leq j \leq k \leq n$ the mapping $q_{j,k} : B(I_k) \to B(I_j)$ by

$$q_{j,k}(f) = q_{j,j+1}(q_{j+1,j+2}(\ldots q_{k-1,k}(f))) \qquad \text{for } j < k$$

and $q_{k,k}(f) = f$. We then have the relation

$$\mu_j(q_{j,k}(f)) = \mu_k(f) \quad \text{for } 0 \leq j \leq k \leq n \text{ and } f \in B(I_k)$$

and the semigroup property

$$q_{j,l}(q_{l,k}(f)) = q_{j,k}(f) \quad \text{for } 0 \leq j \leq l \leq k \leq n.$$

This model is a special case of the framework of Section 2.1.2. The following lemma gives an explicit expression for $q_{j,k}(f)$:

Lemma 4.1. *For $0 \leq j < k \leq n$, $f \in B(I_k)$ and $x \in I_j$ we have*

$$q_{j,k}(f)(x) = \frac{\mu_k(f\,1_{\{s_k(x)\}})}{\mu_j(x)}. \tag{4.1}$$

In particular,

$$\begin{aligned} |q_{j,k}(f)(x)| &\leq \left(\max_{y \in I_k} |f(y)|\right) q_{j,k}(1)(x) \qquad (4.2)\\ &\leq \left(\max_{y \in I_k} |f(y)|\right) \left(\max_{z \in I_j} \frac{\mu_k^{\to j}(z)}{\mu_j(z)}\right). \end{aligned}$$

Proof of Lemma 4.1. Observe that for $x \in I_j$ and $f \in B(I_k)$ we have

$$q_{j,k}(f\,1_{\{s_k(y)\}})(x) = 0$$

for $x \neq y \in I_j$. Thus we can write

$$q_{j,k}(f)(x) = \sum_{y \in I_j} q_{j,k}(f1_{\{s_k(y)\}})(x) = q_{j,k}(f\,1_{\{s_k(x)\}})(x).$$

since $q_{j,k}(f)$ is linear in f. Therefore we have

$$\begin{aligned} &\mu_k(f1_{\{s_k(x)\}}) = \mu_j(q_{j,k}(f1_{\{s_k(x)\}}))\\ = \;&\mu_j(x)q_{j,k}(f1_{\{s_k(x)\}})(x) = \mu_j(x)q_{j,k}(f)(x) \end{aligned}$$

which can be rearranged into (4.1). (4.2) follows from

$$|q_{j,k}(f\,1_{\{s_k(x)\}})(x)| \leq \left(\max_{y \in I_k} |f(y)|\right) q_{j,k}(1_{\{s_k(x)\}})(x) = \left(\max_{y \in I_k} |f(y)|\right) \frac{\mu_k(s_k(x))}{\mu_j(x)}$$

and the definition of $\mu_k^{\to j}$ □

4.2 Sequential MCMC

We now introduce the interacting particle system associated with the Sequential MCMC algorithm for the tree model and derive non-asymptotic bounds on the approximation error. This algorithm corresponds to the particle system of Section 2.1.3 in the special case of our tree model.

4.2.1 The Interacting Particle System

We construct an interacting particle system approximating the sequence of measures μ_k. We start with N independent samples $\xi_0 = (\xi_0^1, \dots, \xi_0^N)$ from μ_0. The particle dynamics alternates two steps: Importance Sampling Resampling and Mutation: A vector of particles ξ_k approximating μ_k is transformed into a vector $\hat{\xi}_{k+1}$ approximating $\mu_{k+1}^{\to k}$ by drawing N conditionally independent samples from the empirical distribution of ξ_k weighted with the functions $g_{k,k+1}$. Afterwards, $\hat{\xi}_{k+1}$ is transformed into a vector ξ_{k+1} approximating μ_{k+1} by moving the particles $\hat{\xi}_{k+1}^i$ independently with the transition kernel K_{k+1}: A particle $\hat{\xi}_{k+1}^i \in I_k$ is moved to a position in $s_{k+1}(\hat{\xi}_{k+1}^i) \subseteq I_{k+1}$ with probabilities proportional to $\mu_{k+1}(\cdot|s_{k+1}(\hat{\xi}_{k+1}^i))$.

We thus have two arrays of random variables $(\xi_k^j)_{0\le k\le n, 1\le j\le N}$ and $(\hat{\xi}_k^j)_{1\le k\le n,1\le j\le N}$ where ξ_k^j and $\hat{\xi}_{k+1}^j$ take values in I_k. The random variables $\xi_0^1, \dots, \xi_0^N$ are independent and distributed according to μ_0. The distributions of the remaining $\hat{\xi}_k^j$ and ξ_k^j are pinned down by the transition probabilities

$$\mathbb{P}[\hat{\xi}_{k+1} \in dx | \xi_k = z] = \prod_{j=1}^N \sum_{i=1}^N \frac{g_{k,k+1}(z^i)}{\sum_{l=1}^N g_{k,k+1}(z^l)} \delta_{z^i}(dx^j)$$

and

$$\mathbb{P}[\xi_{k+1} \in dx | \hat{\xi}_{k+1} = z] = \prod_{j=1}^N K_{k+1}(z^j, dx^j).$$

4.2.2 Error Bounds

Denote by $\mathcal{F}_k$ the σ-algebra generated by $\xi_0, \dots \xi_k$ and $\hat{\xi}_1, \dots \hat{\xi}_k$ and denote the empirical measure of ξ_k by η_k^N, i.e.

$$\eta_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{\xi_k^i}.$$

Recall that by Lemma 2.1 we have for $f \in B(E)$ and $1 \le k \le n$ that

$$\mathbb{E}[\eta_k^N(f) | \mathcal{F}_{k-1}] = \frac{\eta_{k-1}^N(q_{k-1,k}(f))}{\eta_{k-1}^N(q_{k-1,k}(1))}.$$

We are interested in the question how well η_k^N approximates μ_k.

Following the analysis of Section 2.2 we define for $0 \leq k \leq n$ the sequence of measures

$$\nu_k^N(f) = \varphi_k \, \eta_k^N(f)$$

on I_k where φ_k is given by

$$\varphi_k = \prod_{j=0}^{k-1} \eta_j^N(q_{j,j+1}(1)).$$

Recall that we have for $f \in B(I_k)$

$$\mathbb{E}[\nu_k^N(f)|\mathcal{F}_{k-1}] = \nu_{k-1}^N(q_{k-1,k}(f)),$$

and that, by Proposition 2.1, $\nu_k^N(f)$ is an unbiased estimator for $\mu_k(f)$ with quadratic error given in that proposition. Moreover we can control the approximation error of η_k^N through the approximation error of ν_k^N by Lemma 2.2.

We next apply to our model the error bounds of Theorem 2.2. To achieve this we need to define a series of norms $\|\cdot\|_j$ on $B(I_j)$ and find constants $d_{j,k}$ such that the inequality

$$\max\left(\|1\|_j \|q_{j,k}(f)^2\|_j, \|q_{j,k}(f)\|_j^2\right) \leq d_{j,k} \; \|f\|_k^2. \tag{4.3}$$

is satisfied. We choose $\|\cdot\|_j$ to be the maximum-norm on $B(I_j)$, i.e. for $f \in B(I_j)$

$$\|f\|_j = \max_{x \in I_j} |f(x)|.$$

Next we derive constants $d_{j,k}$ which guarantee that (4.3) is satisfied. Observe that we have $\|f^2\|_j = \|f\|_j^2$, $\|1\|_j = 1$ and by Lemma 4.1

$$\|q_{j,k}(f)\|_j \leq \|q_{j,k}(1)\|_j \|f\|_n.$$

Moreover by the same lemma we have

$$\|q_{j,k}(1)\|_j = \max_{x \in I_j} \frac{\mu_k^{\to j}(x)}{\mu_j(x)} \geq 1 \tag{4.4}$$

Thus we can choose

$$d_{j,k} = \left(\max_{x \in I_j} \frac{\mu_k^{\to j}(x)}{\mu_j(x)}\right)^2$$

$d_{j,k}$ is large when a node in the tree which is unimportant at level j has offspring which carries considerably more probability mass at level k. Notably, the constant $d_{j,k}$ does not take into account any further branching of the state space which occurs at levels $j+1, \ldots, n$. In order to state our error bound we define another series of constants following the definitions of Section 2.4: Define

$$\widehat{d}_k = 2 \sum_{j=0}^{k} d_{j,k},$$

and

$$\widehat{v}_k = \sup\left\{\sum_{j=0}^{k} \mathrm{Var}_{\mu_j}(q_{j,k}(f)) \middle| f \in B(I_k), \|f\|_k \leq 1\right\}.$$

and

$$\varepsilon_k^N = \sup\left\{\mathbb{E}[|\nu_k^N(f) - \mu_k(f)|^2] \middle| f \in B(I_k), |f|_k \leq 1\right\}.$$

Moreover define

$$\overline{d}_k = \max_{j \leq k} \widehat{d}_j, \quad \overline{v}_k = \max_{j \leq k} \widehat{v}_j \quad \text{and} \quad \overline{\varepsilon}_k^N = \max_{j \leq k} \varepsilon_j^N.$$

Then the following error bound is an immediate consequence of Theorem 2.2:

Corollary 4.1. *Let $N \geq 2\overline{d}_n$. Then for $f \in B(I_n)$ we have*

$$N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + \|f\|_n^2 \widehat{d}_n \, \overline{\varepsilon}_n^N$$

and

$$\overline{\varepsilon}_n^N \leq 2\frac{\overline{v}_n}{N}$$

Finally, observe that we can bound $\mathrm{Var}_{\mu_j}(q_{j,n}(f))$ through $d_{j,n}$ by

$$\mathrm{Var}_{\mu_j}(q_{j,n}(f)) \leq \mu_j(q_{j,n}(f)^2) \leq \|f\|_n^2 \, \|q_{j,n}(1)\|_j \, \mu_j(q_{j,n}(1)) \leq \sqrt{d_{j,n}} \, \|f\|_n^2. \tag{4.5}$$

This also implies

$$\widehat{v}_k \leq \sum_{j=0}^{k} \sqrt{d_{j,k}}.$$

Thus our error bounds depend on the measures μ_j through the maxima of the relative densities between μ_j and $\mu_k^{\to j}$. This is the maximal importance gain of a component of the partition at level j between levels j and k. We next set these bounds into perspective by deriving a lower bound on $\widehat{v}_k$ by calculating the asymptotic variance

$$\mathrm{Var}_k^{\mathrm{ass}}(f) = \sum_{j=0}^{k} \mathrm{Var}_{\mu_j}(q_{j,k}(f))$$

for the test function $f \equiv 1 \in B(I_k)$:

Proposition 4.1.

$$\mathrm{Var}_k^{ass}(1) = \sum_{j=0}^{k} \sum_{x \in I_j} \mu_k^{\to j}(x) \left(\frac{\mu_k^{\to j}(x)}{\mu_j(x)} - 1\right) = \sum_{j=0}^{k} \mu_k^{\to j}(q_{j,k}(1) - 1)$$

Proof of Proposition 4.1. Observe that

$$\mathrm{Var}_{\mu_j}(q_{j,k}(1)) = \left[\sum_{x \in I_j} \mu_j(x)\left(q_{j,k}(1)(x)\right)^2\right] - \mu_j(q_{j,k}(1))^2$$

By Lemma 4.1 we have

$$(q_{j,k}(1)(x))^2 = \left(\frac{\mu_k^{\to j}(x)}{\mu_j(x)}\right)^2.$$

By the fact that

$$\mu_j(q_{j,k}(1))^2 = \mu_k(1)^2 = 1 = \sum_{x \in I_j} \mu_k^{\to j}(x)$$

we can thus write

$$\mathrm{Var}_{\mu_j}(q_{j,k}(1)) = \sum_{x \in I_j} \mu_k^{\to j}(x)\left(\frac{\mu_k^{\to j}(x)}{\mu_j(x)} - 1\right) = \mu_k^{\to j}(q_{j,k}(1) - 1).$$

Summing over j completes the proof. □

Denote the expression for $\mathrm{Var}_{\mu_j}(q_{j,k}(1))$ from the proposition by $v_{j,k}$, i.e.,

$$v_{j,k} = \sum_{x \in I_j} \mu_k^{\to j}(x)\left(\frac{\mu_k^{\to j}(x)}{\mu_j(x)} - 1\right).$$

$d_{j,k}$ may be large even when $v_{j,k}$ is small: $d_{j,k}$ is large if the successors at level k of $x \in I_j$ are – relatively – much more important under μ_k than x is under μ_j. In this case $v_{j,k}$ may still be small if the absolute importance of the successors of x is small under μ_k. In short, $v_{j,k}$ may be much smaller than $d_{j,k}$ if the largest (relative) gains in importance are made by regions of the state space that remain (absolutely) unimportant.

As a by-product, note that from the proof of Proposition 4.1 we immediately get an upper bound on $\mathrm{Var}_{\mu_j}(q_{j,k}(f))$ which is sharper than (4.5):

$$\mathrm{Var}_{\mu_j}(q_{j,k}(f)) \leq \mu_j(q_{j,k}(1)^2)\, \|f\|_k^2 = \widetilde{d}_{j,k}\|f\|_k^2$$

where $\widetilde{d}_{j,k}$ is defined as

$$\widetilde{d}_{j,k} = \sum_{x \in I_j} \frac{\mu_k^{\to j}(x)^2}{\mu_j(x)} = \mu_k^{\to j}(q_{j,k}(1)).$$

We obtain corresponding sharper upper bounds on $\widehat{v}_k$ and $\overline{v}_k$. This allows to bound the leading term in the error bounds of Corollary 4.1 by $\widetilde{d}_{j,k}$ instead of $\sqrt{d_{j,k}}$.

4.3 Sequential Importance Sampling

For the purpose of comparison, we also introduce the Sequential Importance Sampling algorithm for the tree model and give an explicit expression for the approximation error for a class of test functions.

In Sequential Importance Sampling, particles are moved independently according to the kernels K_k. Afterwards, importance weights ω are calculated for the particles which allow to obtain an estimator for μ_n through a weighted empirical measure of the particles, see Section 1.3.5 for further discussion. In the present framework, Sequential Importance Sampling is equivalent to simple Importance Sampling between the probability distribution π_n on I_n given by

$$\pi_n = \mu_0 K_1 \dots K_n$$

and μ_n. For simplicity, we consider only unnormalized Importance Sampling, i.e., we assume that we can calculate the weights exactly (and not only up to a normalizing constant). This has the advantage that we do not have to consider a bias introduced by normalizing the particle weights through their sum.

Instead of a system of particles, it is thus sufficient to consider only the vector of particles $(\widetilde{\xi}_n^i)_{1\leq i\leq N}$ which are distributed independently according to π_n. We define the importance weight function $\omega_n \in B(I_n)$ by

$$\omega_n(x) = \frac{\mu_n(x)}{\pi_n(x)}$$

for all $x \in I_n$. Then for $f \in B(I_n)$ our Sequential Importance Sampling estimator $\widetilde{\eta}_n(f)$ is given by

$$\widetilde{\eta}_n(f) = \frac{1}{N}\sum_{i=1}^{N} f\left(\widetilde{\xi}_n^i\right)\omega_n\left(\widetilde{\xi}_n^i\right).$$

$\widetilde{\eta}_n(f)$ is an unbiased estimator for $\mu_n(f)$, i.e.,

$$\mathbb{E}[\widetilde{\eta}_n(f)] = \mu_n(f).$$

We next calculate a formula for the quadratic approximation error for test functions of the form $f = 1_{\{x\}}$ where $x \in I_n$:

Lemma 4.2. *For $x \in I_n$ and $f = 1_{\{x\}}$ we have*

$$\mathbb{E}[|\widetilde{\eta}_n(f) - \mu_n(f)|^2] = \frac{\mu_n(x)^2}{N}\left(\frac{1}{\pi_n(x)} - 1\right)$$

Proof of Lemma 4.2. To prove the lemma we only need the following calculation based on the unbiasedness of $\widetilde{\eta}_n(f)$:

$$\begin{aligned}
\mathbb{E}[|\widetilde{\eta}_n(f) - \mu_n(f)|^2] &= \mathbb{E}\left[\left(\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\mu_n(x)}{\pi_n(x)}1_{\{x\}}\left(\widetilde{\xi}_n^i\right)\right) - \mu_n(x)\right)^2\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\mu_n(x)}{\pi_n(x)}1_{\{x\}}\left(\widetilde{\xi}_n^i\right) - \mu_n(x)\right)^2\right] \\
&= \frac{1}{N}\left(\pi_n(x)\left(\frac{\mu_n(x)}{\pi_n(x)} - \mu_n(x)\right)^2 + (1-\pi_n(x))\mu_n(x)^2\right) \\
&= \frac{\mu_n(x)^2}{N}\left(\frac{1}{\pi_n(x)} - 1\right)
\end{aligned}$$

□

We thus see that Sequential Importance Sampling can only perform well if the distribution π_n is sufficiently close to μ_n, more precisely, if no state which is unimportant under π_n is important under μ_n.

4.4 Example: Weighting or Resampling?

We now apply the error bounds we just developed to a concrete example depicted in Figure 4.1. Our aim is to show that in this case Sequential MCMC, notably, its Resampling step, succeeds in a multimodal setting in which Sequential Importance Sampling severely suffers from weight degeneracy. Section 4.4.1 introduces the setting of the example. Section 4.4.2 derives upper bounds on $q_{j,k}(1)$. Sections 4.4.3 and 4.4.4 contain the error analysis for, respectively, Sequential MCMC and Sequential Importance Sampling. Section 4.4.5 closes our comparison of Sequential MCMC and Sequential Importance Sampling by discussing some further examples.

4.4.1 The Model

We consider the sequence of state spaces $I_0, \ldots, I_n$ given by

$$I_k = \{0_k, \ldots, k_k\}.$$

Thus the elements of I_k are the natural numbers from 0 to k, indexed by k in order to keep the notation clearer.

For $l > k$, the predecessor in I_k of $j_l \in I_l$ is given by j_k if $j \leq k$, otherwise it is k_k:

$$p_k(j_l) = \begin{cases} j_k & \text{if } j \leq k \\ k_k & \text{if } j > k. \end{cases}$$

We thus have a simple tree structure where from level k to level $k+1$ the "largest" node k_k has two successors, k_{k+1} and $(k+1)_{k+1}$, while all other nodes j_k have only

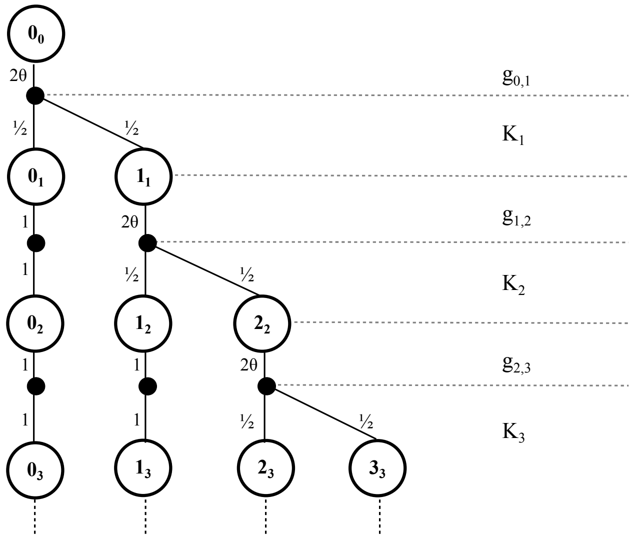

Figure 4.1: Weighting or Resampling?

one successor j_{k+1}. Accordingly, for $l > k$ and $j_k \in I_k$, the successor function is given by

$$s_l(j_k) = \begin{cases} \{j_l\} & \text{if } j < k \\ \{k_l, \dots, l_l\} & \text{if } j = k. \end{cases}$$

We define the sequence $\mu_0, \dots, \mu_n$ implicitly through $g_{k,k+1}$ and K_{k+1}. We choose the unnormalized density $g_{k,k+1} \in B(I_k)$ such that only the mass of k_k is modified while the relative masses of the other nodes remain the same:

$$g_{k,k+1}(j_k) = \begin{cases} 1 & \text{if } j < k \\ 2\theta & \text{with } \theta > 0 \text{ if } j = k. \end{cases}$$

The transition kernel $K_{k+1} : I_k \times I_{k+1} \to [0,1]$ is chosen such that $K_{k+1}(j_k, \cdot)$ is the uniform distribution on the successors of j_k:

$$K_{k+1}(j_k, i_{k+1}) = \begin{cases} 1 & \text{if } i = j < k \\ \frac{1}{2} & \text{if } j = k \text{ and } i \in \{k, k+1\} \\ 0 & \text{otherwise.} \end{cases}$$

Observe that for $\theta > \frac{1}{2}$ we have two countervailing effects, one from the kernels K_k and one from the functions $g_{k,k+1}$: On the one hand, the kernels K_k favor that mass is concentrated on j_k with small j. If we had a constant function $g_{k,k+1}$ (i.e. $\theta = \frac{1}{2}$), μ_k would be a geometric distribution with parameter $\frac{1}{2}$ and maximum in 0_k . On the other hand, the weight functions $g_{k,k+1}$ move mass to the largest node k_k. As becomes clear from the explicit formula for μ_k calculated next, the case of $\theta > 1$

which we mainly consider is the case where the second effect is sufficiently strong in the sense that $\mu_k(k_k) > \mu_k(j_k)$ for $j < k-1$. As θ approaches 1, μ_k converges to the uniform distribution on I_k. The cases where $\theta < 1$ are largely omitted in our error bounds, not because they are more difficult, but because they are less interesting and would need a largely separate analysis.

Corollary 4.2. *For $j_k \in I_k$ we have*

$$\mu_k(j_k) = \begin{cases} \frac{\theta^{j+1}}{Z_k} & \text{if } j < k \\ \\ \frac{\theta^k}{Z_k} & \text{if } j = k \end{cases}$$

where the normalizing constant Z_k is given by

$$Z_k \;=\; \theta^k + \sum_{j=0}^{k-1} \theta^{j+1}. \qquad (4.6)$$

Moreover for $\theta \neq 1$,

$$Z_k \;=\; \theta^k + \frac{\theta}{\theta - 1}(\theta^k - 1). \qquad (4.7)$$

The corollary is an immediate consequence of our choices of $g_{k,k+1}$ and K_{k+1}. Thus for $\theta > 1$, μ_k can be characterized as follows: It is a geometric distribution with maximum in $(k-1)_k$ on $0_k, \dots, (k-1)_k$. Additionally we have $\mu_k((k-1)_k) = \mu_k(k_k)$.

4.4.2 Controlling the Semigroup

From here on we mostly focus on the case $\theta \geq 1$. In order to apply the error bounds of Section 4.2.2 we have to study the expressions $q_{j,k}(1)$ for this example. This is begun in the following lemma:

Lemma 4.3. *For $0 \leq k < l \leq n$, we have*

$$q_{k,l}(1)(j_k) = \begin{cases} \frac{Z_k}{Z_l} & \text{if } j < k \\ \\ \frac{Z_k Z_{l-k}}{Z_l} & \text{if } j = k. \end{cases}$$

Furthermore for $\theta \geq 1$,

$$\max_{j_k \in I_k} q_{k,l}(1)(j_k) = \frac{Z_k Z_{l-k}}{Z_l}.$$

Proof of Lemma 4.3. Recall from Lemma 4.1 that

$$q_{k,l}(1)(j_k) = \frac{\mu_l(s_l(j_k))}{\mu_k(j_k)}.$$

Thus for $j_k \neq k_k$ Corollary 4.2 immediately implies

$$q_{k,l}(1)(j_k) = \frac{\mu_l(j_l)}{\mu_k(j_k)} = \frac{Z_k}{Z_l}.$$

For $j_k = k_k$ we have

$$\begin{aligned} q_{k,l}(1)(k_k) &= \frac{\mu_l(\{k_l, \ldots, l_l\})}{\mu_k(k_k)} \\ &= \frac{Z_k}{Z_l}\left(\frac{\theta^l + \sum_{i=k}^{l-1}\theta^{i+1}}{\theta^k}\right) \\ &= \frac{Z_k}{Z_l}\left(\theta^{l-k} + \sum_{i=0}^{l-k-1}\theta^{i+1}\right) \\ &= \frac{Z_k Z_{l-k}}{Z_l}. \end{aligned}$$

Observe from (4.6) that $Z_k < Z_l$ and thus for $j_k \neq k_k$

$$q_{k,l}(1)(j_k) < 1.$$

Since both μ_k and μ_l are probability measures and since

$$\mu_k(q_{k,l}(1)) = \mu_l(1) = 1$$

this implies

$$\max_{j_k \in I_k} q_{k,l}(1)(j_k) = q_{k,l}(1)(k_k) > 1.$$

□

Thus in order to control $q_{k,l}(1)$ we need bounds on the constants Z_k. The following lemma gives two pairs of bounds on Z_k. The bounds in (4.8) get sharp as θ approaches 1 while the bounds in (4.9) get sharp as θ gets large.

Lemma 4.4. *We have for $\theta \geq 1$*

$$(k+1)\theta \leq Z_k \leq (k+1)\theta^k \tag{4.8}$$

and

$$2\theta^k \leq Z_k \qquad \text{and, if } \theta > 1, \qquad Z_k \leq \rho(\theta)\theta^k \tag{4.9}$$

where we define

$$\rho(\theta) = 2 + \frac{1}{\theta - 1}. \tag{4.10}$$

Proof of Lemma 4.4. The bounds in (4.8) and the lower bound in (4.9) follow immediately from (4.6) and from the fact that for $k > i$ we have $\theta^k > \theta^i$. The upper bound in (4.9) follows from (4.7) since

$$Z_k = \theta^k + \frac{\theta}{\theta-1}(\theta^k - 1) < \left(1 + \frac{\theta}{\theta-1}\right)\theta^k = \left(2 + \frac{1}{\theta-1}\right)\theta^k.$$

□

We thus arrive at the following upper bound on $\|q_{k,l}(1)\|_k$ (where as before $\|\cdot\|_k$ denotes the maximum norm on $B(I_k)$):

Corollary 4.3. *For $k < l$ and $\theta > 1$ we have*

$$\|q_{k,l}(1)\|_k \leq \min\left(\frac{\rho(\theta)^2}{2}, \frac{\rho(\theta)^2}{l+1}\theta^{l-1}, \frac{(l+2)^2}{8}, \frac{l+2}{2}\theta^{l-1}\right)$$

Proof of Corollary 4.3. By combining each time one lower bound and one upper bound from Lemma 4.4 we obtain four upper bounds on

$$\|q_{k,l}(1)\|_k = \frac{Z_k Z_{l-k}}{Z_l}.$$

Applying the inequalities $(k+1)(l-k+1) \leq \frac{1}{4}(l+2)^2$ and

$$\frac{(k+1)(l-k+1)}{l+1} \leq \frac{l+2}{2}$$

completes the proof. □

For θ sufficiently close to 1 the upper bound

$$\|q_{k,l}(1)\|_k \leq \frac{l+2}{2}\theta^{l-1} \tag{4.11}$$

which is obtained from using both directions of (4.8) is the sharpest one. For sufficiently large θ the bound

$$\|q_{k,l}(1)\|_k \leq \frac{\rho(\theta)^2}{2} \tag{4.12}$$

obtained from (4.9) is best. Depending on the values of k and l, one of the two other bounds may be even better for intermediate values of θ. Finally, note that the third and fourth bounds also apply to $\theta = 1$ since they do not rely on the upper bound from (4.9).

It is quite intuitive, that for $\theta \approx 1$ our bounds on $q_{j,k}(1)$ depend more sensitively on k. With a large value of θ mass is concentrated quickly in the highest branch of the tree such that the sequence $a_k = \mu_k(k_k)$ varies relatively little in k. For $\theta \approx 1$, mass is accumulated only slowly in k_k as k increases such that the same sequence a_k is increasing substantially in k at least for small values of k. This increase is reflected in the fact that our upper bound on $q_{j,k}(1)$ is increasing with k in that case. Put differently, for θ close to 1 and k not large, the distributions μ_k are not very concentrated (i.e. close to the uniform distribution) and thus more costly to approximate. As we will see, the approximation error of our algorithm is indeed of worse order in n at $\theta = 1$ than for $\theta > 1$ (or $\theta < 1$). This can also be seen as an elementary manifestation of the critical slowing down phenomenon.

4.4.3 Error Bounds for Sequential MCMC

In the following we give two error bounds, both based on Corollary 4.1: one which degenerates as θ approaches 1, and one which does not degenerate but which is worse for θ sufficiently greater than 1. Before we begin, note that a dependence on the parameter n enters the error bound from two sources: While the two terms of the error bound of Corollary 4.1 are, respectively, linear and quadratic in n, we obtain a stronger dependence on n in Proposition 4.3 below since n is also the size of the state space I_n and a parameter of the distribution μ_n. To confirm that this difference between the results is not an artefact of our upper bounds, we calculate the asymptotic variance in the case $\theta = 1$ explicitly in Lemma 4.5 at the end of this section.

The first result, for θ sufficiently greater than 1, is based on the bound (4.12), i.e. we choose

$$\|q_{k,l}(1)\|_k^2 \leq d_{k,l} = \frac{\rho(\theta)^4}{4}.$$

with $\rho(\theta)$ as defined in (4.10)

Proposition 4.2. *Consider $\theta > 1$, $N > \rho(\theta)^4 n$ and $f \in B(I_n)$. Then we have*

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \|f\|_n^2 \left(\frac{\rho(\theta)^2}{2} \frac{n+1}{N} + \rho(\theta)^6 \frac{(n+1)^2}{N^2} \right).$$

Proof of Proposition 4.2. In order to apply Corollary 4.1 we have to control the constants introduced in Section 4.2.2. By our choice of $d_{j,k}$, we get

$$\widehat{d}_k \leq \frac{\rho(\theta)^4(k+1)}{2}$$

Since this bound is increasing in k we also have

$$\overline{d}_k \leq \left(\frac{1}{2}\rho(\theta)^4 + \frac{1}{4}\rho(\theta)^5 \right) k.$$

Furthermore by (4.5) we have

$$\sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) \leq \frac{\rho(\theta)^2}{2}(n+1)\,\|f\|_n^2,$$

and

$$\widehat{v}_k \leq \frac{\rho(\theta)^2}{2}(k+1).$$

Inserting these definitions into the bound of Corollary 4.1 gives the desired bound.
□

These bounds degenerate quickly as θ approaches 1 since $\rho(\theta)$ gets arbitrarily large then. To demonstrate that we obtain reasonable constants in our bounds for sufficiently large θ, we give the following result derived from the special case $\theta = 2$ (and thus $\rho(2) = 3$):

Corollary 4.4. *Consider $\theta \geq 2$, $N > 81n$ and $f \in B(I_n)$. Then we have*

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \|f\|_n^2 \left(\frac{9}{2} \frac{n+1}{N} + 729 \frac{(n+1)^2}{N^2} \right)$$

We now turn to a bound which does not degenerate at $\theta = 1$. For the sake of simplicity we rely on the bound

$$\|q_{k,l}(1)\|_k^2 \leq d_{k,l} = \frac{(l+2)^4}{64} \tag{4.13}$$

from Corollary 4.3 instead of the bound (4.11) which, for small θ, is sharper and has a better order in k but which degenerates quickly as θ increases.

Proposition 4.3. *Consider $\theta \geq 1$, $N > \frac{1}{16}(n+2)^5$ and $f \in B(I_n)$. Then we have*

$$\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \|f\|_n^2 \left(\frac{1}{8} \frac{(n+2)^3}{N} + \frac{1}{128} \frac{(n+2)^8}{N^2} \right)$$

Proof of Proposition 4.3. By our choice of $d_{j,k}$, we get

$$\widehat{d}_k \leq \frac{(k+2)^5}{32}$$

Since this bound is increasing in k we also have

$$\overline{c}_k \leq \frac{1+\theta}{32}(k+2)^5.$$

Furthermore by (4.5) we have

$$\sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) \leq \frac{(n+2)^3}{8} \|f\|_n^2,$$

and

$$\widehat{v}_k \leq \frac{(k+2)^3}{8}.$$

Inserting these definitions into the bound of Corollary 4.1 gives the desired bound.
□

As noted above we used in Proposition 4.3 a bound of order n^4 on $\|q_{k,n}(1)\|_k^2$ instead of relying on (4.11) which may have led to a better order at least for θ close to 1. Thus we expect that the error bound of Proposition 4.3 can be improved concerning the order in n. In Section 4.4.4, we show however that the approximation error of Sequential Importance Sampling is growing exponentially in n in this example. Thus Proposition 4.3 is strong enough to make our point that the resampling step in our Sequential MCMC algorithm overcomes the problem of weight degeneracy.

To close our analysis of the error bound for θ close to 1, we explicitly calculate the asymptotic variance – and thus the leading coefficient in the error bound of Corollary 4.1 – for the case $\theta = 1$ and $f \equiv 1 \in B(I_n)$. This asymptotic variance is quadratic in n which proves that it is no artifact of our upper bounds, that we do not achieve as good an order in n in Proposition 4.3 as in Proposition 4.2.

Lemma 4.5. *For $\theta = 1$ we have*

$$\mathrm{Var}_n^{ass}(1) = \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(1)) = \frac{n^2(n-1)}{12(n+1)}$$

Proof of Lemma 4.5. By Proposition 4.1 we have

$$\mathrm{Var}_n^{\mathrm{ass}}(1) = \sum_{j=0}^{n} w_j$$

where

$$w_j = \sum_{x \in I_j} \mu_n^{\to j}(x) \left(\frac{\mu_n^{\to j}(x)}{\mu_j(x)} - 1 \right).$$

Now observe that for $\theta = 1$ we have

$$\mu_j(x) = \frac{1}{j+1}$$

for all $x \in I_j$ and

$$\mu_n^{\to j}(x) = \begin{cases} \frac{n-j+1}{n+1} & \text{for } x = j_j \\ \\ \frac{1}{n+1} & \text{otherwise.} \end{cases}$$

Thus we have

$$\begin{aligned} w_j &= \sum_{x \in I_j} \mu_n^{\to j}(x)((j+1)\mu_n^{\to j}(x) - 1) \\ &= -1 + (j+1) \sum_{x \in I_j} \mu_n^{\to j}(x)^2 \\ &= -1 + \frac{j+1}{(n+1)^2} \left(j + (n-j+1)^2 \right) \end{aligned}$$

It is then straightforward to calculate that

$$\mathrm{Var}_n^{\mathrm{ass}}(1) = \sum_{j=0}^{n} w_j = \frac{n^2(n-1)}{12(n+1)}$$

which completes the proof. □

4.4.4 Weight Degeneracy of Sequential Importance Sampling

We now turn to the analysis of Sequential Importance Sampling as introduced in Section 4.3 for our example. In the present setting, the distribution π_n is given by

$$\pi_n(j_n) = \begin{cases} 2^{-j+1} & \text{for } j < n \\ 2^{-n} & \text{for } j = n \end{cases}$$

To prove that depending on the value of θ the approximation error of $\widetilde{\eta}_n(f)$ may grow exponentially in n, we consider the approximation error for the test function $f = 1_{\{n_n\}}$. We have the following explicit formula for the approximation error:

Corollary 4.5. *For $f = 1_{\{n_n\}}$ and $\theta > 0$ we have*

$$\mathbb{E}[|\widetilde{\eta}_n(f) - \mu_n(f)|^2] = \begin{cases} \frac{2^n - 1}{N\left(1+\frac{\theta}{\theta-1}(1-\theta^{-n})\right)^2} & \textit{for } \theta \neq 1 \\ \\ \frac{2^n-1}{N(n+1)^2} & \textit{for } \theta = 1 \end{cases}$$

Moreover, $\mathbb{E}[|\widetilde{\eta}_n(f) - \mu_n(f)|^2]$ grows exponentially in n whenever $\theta > 2^{-\frac{1}{2}}$.

Proof of Corollary 4.5. The explicit formula for the error is a direct consequence of Lemma 4.2, the fact that $\pi_n(n_n) = 2^{-n}$, and the representation of μ_n given in Corollary 4.2 which yields

$$\mu_n(n_n) = \frac{1}{1 + \frac{\theta}{\theta-1}(1-\theta^{-n})}$$

for $\theta \neq 1$ and

$$\mu_n(n_n) = \frac{1}{n+1}$$

for $\theta = 1$. The error grows exponentially in n whenever

$$\frac{2^n}{\theta^{-2n}}$$

tends to infinity in n which is the case for $\theta > 2^{-\frac{1}{2}}$. □

Notably, we see that Sequential Importance Sampling suffers from weight degeneracy when approximating $f = 1_{\{n_n\}}$ even in some cases (i.e. $2^{-\frac{1}{2}} < \theta < 1$) where $\mu_n(n_n)$ is decreasing exponentially itself.

4.4.5 Further Examples

The poor performance of Sequential Importance Sampling in the previous example stems from the fact that the particles' movements only depend on the kernels K_k and do not take into account the reweighting through the functions $g_{k,k+1}$. It is easy to construct a (somewhat artificial) example where this turns out to be an advantage and where accordingly Sequential Importance Sampling outperforms Sequential MCMC.[1] This is done in the following. The notation of the previous example is retained unless otherwise noted.

Consider the sequence of state spaces $I_0 = \{0_0\}$ and $I_k = \{0_k, 1_k\}$ for $1 \leq k \leq 3$. Define a sequence of probability measures μ_k on I_k through $\mu_0(0_0) = 1$,

$$\mu_1(0_1) = \mu_1(1_1) = \mu_3(0_3) = \mu_3(1_3) = \frac{1}{2}$$

and $\mu_2(0_2) = \alpha$, $\mu_2(1_2) = 1 - \alpha$ where $0 < \alpha < 1$.

The tree structure is given by $p_k(0_{k+1}) = 0_k$, $p_0(1_1) = 0_0$ and, for $k > 0$, $p_k(1_{k+1}) = 1_k$. This implies that

$$K_1(0_0, 0_1) = K_1(0_0, 1_1) = \frac{1}{2}$$

while all other transition kernel are trivial, i.e., for $k > 1$ and $j \in \{0, 1\}$

$$K_k(j_k, j_{k+1}) = 1.$$

We first consider the approximation error of Sequential Importance Sampling as defined in the previous section: In this example the Importance Sampling proposal distribution π_3 coincides with μ_3. Thus from Lemma 4.2, we obtain the following: For $f = 1_{\{0_3\}}$

$$\mathbb{E}[|\widetilde{\eta}_3(f) - \mu_3(f)|^2] = \frac{\mu_3(0_3)^2}{N}\left(\frac{1}{\pi_3(0_3)} - 1\right) = \frac{1}{4N} \tag{4.14}$$

Observe that this error is independent of α: When moving from μ_1 to μ_2, the weights are changed, but this change is removed when moving (back) to μ_3 and throughout the particles' movements are unaffected. So to say, the particles "accidentally" do the right thing when moving from μ_0 to μ_1. To see this, we replace μ_1 by μ_1' which is essentially the same as μ_2, $\mu_1'(0_1) = \alpha$ and $\mu_1'(1_1) = 1 - \alpha$. Intuitively, this might make the problem easier, because it leads to a "smoother" sequence μ_k. The opposite is the case however: The proposal distribution π_3' is now given by $\pi_3'(0_3) = \alpha$ and

[1] Neal (1996) discusses numerical results for a more natural example that follows a similar logic. Instead of Sequential MCMC and Sequential Importance Sampling, he considers the Simulated Tempering and Tempered Transitions algorithms whose respective ways of discovering the state space are intuitively similar to our two algorithms.

$\pi'_3(1_3) = 1 - \alpha$. Accordingly we get the error bound

$$\mathbb{E}[|\widetilde{\eta}_3(f) - \mu_3(f)|^2] = \frac{1}{4N}\left(\frac{1}{\alpha} - 1\right)$$

which gets arbitrarily large for small α.

Now we consider the asymptotic variance of Sequential MCMC for the same example, again with the original μ_1 and with the test function $f = 1_{\{0_3\}}$. We thus have to evaluate

$$\mathrm{Var}_3^{\mathrm{ass}}(f) = \sum_{j=0}^{3} \mathrm{Var}_{\mu_j}(q_{j,3}(f)).$$

Using the formula (4.1) for $q_{j,k}(f)$ it is straightforward to calculate that

$$q_{0,3}(1_{\{0_3\}}) = \frac{1}{2}, \qquad q_{1,3}(1_{\{0_3\}}) = 1_{\{0_1\}},$$
$$q_{2,3}(1_{\{0_3\}}) = \frac{1}{2\alpha}1_{\{0_2\}} \quad \text{and} \quad q_{3,3}(1_{\{0_3\}}) = 1_{\{0_3\}}.$$

Accordingly we have $\mathrm{Var}_{\mu_0}(q_{0,3}(f)) = 0$,

$$\mathrm{Var}_{\mu_1}(q_{1,3}(f)) = \mathrm{Var}_{\mu_3}(q_{3,3}(f)) = \frac{1}{4}$$

and

$$\mathrm{Var}_{\mu_2}(q_{2,3}(f)) = \frac{1}{4}\left(\frac{1}{\alpha} - 1\right).$$

Thus the asymptotic variance is given by

$$\mathrm{Var}_3^{\mathrm{ass}}(f) = \frac{1}{4}\left(\frac{1}{\alpha} + 1\right).$$

Recall that the asymptotic variance also coincides with the coefficient of the leading term in our error bound of Corollary 4.1. Thus we observe that the approximation error gets arbitrarily large for small values of α. This is in contrast to the error (4.14) of Sequential Importance Sampling for the same example which is independent of α.

Changing μ_1 to μ'_1 with $\mu'_1(0_1) = \alpha$ and $\mu'_1(1_1) = 1 - \alpha$ does not lead to a qualitative change of the error bound: We then get

$$q'_{1,3}(1_{\{0_3\}}) = \frac{1}{2\alpha}1_{\{0_1\}} \text{ and } \mathrm{Var}_{\mu'_1}(q'_{1,3}(f)) = \frac{1}{4}\left(\frac{1}{\alpha} + 1\right)$$

which leads to an asymptotic variance of

$$\mathrm{Var}_3^{\mathrm{ass}\,\prime}(f) = \frac{1}{4}\left(\frac{2}{\alpha} - 1\right).$$

Observe that again – despite the fact that the sequence $\mu_0, \mu_1, \mu_2, \mu_3$ varies more strongly than $\mu_0, \mu_1', \mu_2, \mu_3$ – the asymptotic variance for small values of α is larger under the second sequence than under the first sequence. The reason for this lies in the fact μ_1 is a better approximation of $\mu_3^{\to 1}$ than μ_1'.

For $\alpha > \frac{1}{2}$, the asymptotic variance under μ_1' is smaller than the one under μ_1 and both are well-behaved. But in this case the asymptotic variance for $f' = 1_{\{1_3\}}$ increases more quickly under μ_1' than under μ_1 as α approaches 1. In this sense, the asymptotic variance is more stable under μ_1 than under μ_1'.

We thus close our comparison of Sequential Importance Sampling and Sequential MCMC on trees with the following conclusion: Sequential Importance Sampling works well if the proposal distribution π_n constructed from μ_0 and the transition kernels K_k is sufficiently close to the target distribution μ_n. Sequential MCMC works well if the distributions μ_j are sufficiently close to the projected distributions $\mu_n^{\to j}$. While there is no obvious relationship between these two properties, it seems clear that Sequential MCMC is more suited to applications where the relative densities $g_{k,k+1}$ play a significant role. Furthermore for both algorithms it is easy to construct examples where they perform arbitrarily bad. Finally note that for the last example we only considered the asymptotic variance of Sequential MCMC. In order to obtain good constants in our error bounds, we also need that μ_j is sufficiently close to $\mu_k^{\to j}$ for $j < k < n$.

5 L_p-bounds under Local Mixing

This chapter brings together the perspectives of Chapters 3 and 4. We return to the basic framework of Chapter 3 and thus to the algorithm as introduced in Section 1.2. However, instead of deriving stability from global mixing properties of the MCMC dynamics, we assume that the MCMC dynamics mixes well only *within* the elements of increasingly finer partitions of the state space. The latter assumption takes the role of the tree structure of Chapter 4. Sections 5.1 and 5.2 introduce the setting and restate our error bounds in it. Section 5.3, the main part of this chapter, derives stability of the Feynman-Kac semigroup $q_{j,k}$ from local mixing properties, concretely, from good mixing within each disconnected component of the state space and from a condition that disconnected components do not gain too much weight. The latter condition is similar to the one that appeared in Chapter 4, see the discussion at the end of Section 5.3.

5.1 The Model

Recall the measure-valued model and interacting particle system introduced in Sections 3.1 and 3.2.1: Let (E, r) be a Polish space and let $\mathcal{B}(E)$ be the σ-algebra of Borel subsets of E. Denote by $M(E)$ the space of finite signed Borel measures on E. Let $M_1(E) \subset M(E)$ be the subset of all probability measures. Let $B(E)$ be the space of bounded, measurable real-valued functions on E. Consider the sequence of probability distributions $(\mu_k)_{k=0}^n$, $\mu_k \in M_1(E)$. The μ_k are related through

$$\mu_k(f) = \frac{\mu_{k-1}(g_{k-1,k}f)}{\mu_{k-1}(g_{k-1,k})}$$

for strictly positive (unnormalized) relative densities $g_{k-1,k} \in B(E)$. For $1 \leq k \leq n$, let $K_k(x, A)$ be an integral operator with $K_k(\cdot, f) \in B(E)$ for all $f \in B(E)$ and with $K_k(x, \cdot) \in M_1(E)$ for all $x \in E$. Assume that K_k is stationary with respect to μ_k. Define the mapping $q_{k-1,k} : B(E) \to B(E)$ by

$$q_{k-1,k}(f) = \frac{g_{k-1,k}K_k(f)}{\mu_{k-1}(g_{k-1,k})}$$

Furthermore, let the semigroup $q_{k,l}$, the interacting particle system $(\xi_k^j)_{0\leq k\leq n, 1\leq j\leq N}$ and $(\hat{\xi}_k^j)_{1\leq k\leq n, 1\leq j\leq N}$, the empirical distribution η_k^N and the weighted empirical distribution ν_k be defined as in Chapter 3.

In place of the tree structure of Chapter 4 we now introduce a sequence of partitions of E: Let I_0,...,I_n be a collection of finite index sets. Define $I = I_0 \cup \ldots \cup I_n$ and for $0 \leq k \leq n$

$$I_{>k} = I_{k+1} \cup \ldots \cup I_n \text{ and } I_{<k} = I_0 \cup \ldots \cup I_{k-1}.$$

For all $j \in I$ there is a set $F_j \in \mathcal{B}(E)$ with $\mu_0(F_j) > 0$. Moreover, we assume that for all $0 \leq k \leq n$ the collection $(F_j)_{j \in I_k}$ is a disjoint partition of E. We assume that partitions successively get finer: For $1 \leq k \leq n$, assume that for all $j \in I_k$ there exists an $i \in I_{k-1}$ with $F_j \subseteq F_i$. Thus for $0 \leq k \leq n-1$, a well-defined predecessor function $p_k : I_{>k} \to I_k$ is characterized as follows: For $1 \leq k < l \leq n$, $j \in I_k$ and $i \in I_l$ define

$$p_k(i) = j \quad \text{if} \quad F_i \subseteq F_j.$$

Conversely, define a successor function $s_k : I_{<k} \to \mathcal{P}(I_k)$ via

$$s_k(i) = \{j \in I_k | p_l(j) = i\} \text{ for } i \in I_l \text{ with } 0 \leq l < k.$$

Thus, for $l < k$ and $i \in I_l$ the collection $(I_j)_{j \in s_k(i)}$ is a disjoint partition of F_i.

We make the simplifying assumption that particles move between partition elements only through the resampling step: For $1 \leq k \leq n$ and $j \in I_k$ assume

$$K_k(1_{F_j})(x) = 0 \text{ for all } x \in E \setminus F_j. \tag{5.1}$$

This assumption ensures that if f has support only in F_j, $j \in I_k$, then $K_k(f)$ has support only in F_j as well. While this technical assumption will not be literally fulfilled in most applications of interest, it can be seen as an approximation of the fact that particles will move between different modes only rarely through the MCMC dynamics.

5.2 Error Bounds for Sequential MCMC

In order to apply the error bounds of Section 2.3 we need to introduce a sequence of norms on E. Unlike in Chapter 3 we want to rely only on local mixing properties. Thus we replace the L_p-norms of Chapter 3 by stronger norms which are composed of local L_p-norms. To introduce these norms we need a few additional definitions. For $0 \leq k \leq n$ and $j \in I$, denote by $\mu_{k,j} \in M_1(E)$ the restriction of μ_k to F_j: For $f \in B(E)$,

$$\mu_{k,j}(f) = \frac{\mu_k(f 1_{F_j})}{\mu_k(F_j)}.$$

It will prove to be convenient to view $\mu_{k,j}$ as a probability distribution on E (and not on F_j). Note that we define $\mu_{k,j}$ for all $j \in I$ (and not only for $j \in I_k$). Furthermore, by assumption (5.1), K_k is stationary with respect to $\mu_{k,j}$ for all $j \in I_k$. Now for $0 \leq k \leq n$, $j \in I$ and $p \geq 1$ denote by $\| \cdot \|_{k,j,p}$ the L_p-norm with respect to $\mu_{k,j}$: For $f \in B(E)$,

$$\|f\|_{k,j,p} = \mu_{k,j}(|f|^p)^{\frac{1}{p}}.$$

Next define the norm $\|\cdot\|_{k,p}$ to be the maximum over the L_p-norms with respect to $\mu_{k,j}$ with $j \in I_k$: For $f \in B(E)$ and $0 \leq k \leq n$,

$$\|f\|_{k,p} = \max_{j \in I_k} \|f\|_{k,j,p}.$$

With this choice of norm we have

$$\|f\|_{L_p(\mu_k)} \leq \|f\|_{k,p}.$$

Now define $\widetilde{c}_{j,k}(p,q)$ to be the constant in an L_p-L_q bound for the semigroup $q_{j,k}$: For $p > q > 1$ and $0 \leq j < k \leq n$ we have

$$\|q_{j,k}(f)\|_{j,p} \leq \widetilde{c}_{j,k}(p,q)\|f\|_{k,q} \quad \text{for all} \quad f \in B(E)$$

Such constants will be studied in Section 5.3 below. Fix $p > 2$ and define

$$c_{j,k}(p) = \max\left(\widetilde{c}_{j,k}\left(p, \frac{p}{2}\right),\ \widetilde{c}_{j,k}(2p,p)^2\right).$$

This choice of $c_{j,k}$ satisfies (3.1), i.e., for $p > 2$, $0 \leq j < k \leq n$ and $f \in B(E)$ we have

$$\max\left(\|1\|_{j,p}\|q_{j,k}(f)^2\|_{j,p}, \|q_{j,k}(f)\|_{j,p}^2, \|q_{j,k}(f^2)\|_{j,p}\right) \leq c_{j,k}(p)\ \|f\|_{k,p}^2.$$

This follows with the same reasoning as in the proof of Lemma 3.1.

Now define another series of constants following the definitions of Section 2.3: Define

$$\widehat{c}_k(p) = \sum_{j=0}^{k-1} c_{j,k}(p)\,(2 + \|q_{j,j+1}(1) - 1\|_{j,p})$$

and

$$\widehat{v}_k(p) = \sup\left\{ \sum_{j=0}^{k} \mathrm{Var}_{\mu_j}(q_{j,k}(f)) \,\middle|\, f \in B(E), \|f\|_{k,p} \leq 1 \right\}$$

and

$$\varepsilon_k^N(p) = \sup\left\{ \mathbb{E}[|\nu_k^N(f) - \mu_k(f)|^2] \,\middle|\, f \in B(E), \|f\|_{k,p} \leq 1 \right\}.$$

Moreover define

$$\overline{c}_k(p) = \max_{j \leq k} \widehat{c}_j(p), \quad \overline{v}_k(p) = \max_{j \leq k} \widehat{v}_j(p) \quad \text{and} \quad \overline{\varepsilon}_k^N(p) = \max_{j \leq k} \varepsilon_j^N(p).$$

Then the following error bound is an immediate consequence of Theorem 2.1:

Corollary 5.1. *Let $p > 2$ and $N \geq 2\overline{c}_n(p)$. Then for $f \in B(E)$ we have*

$$N\mathbb{E}[|\nu_n^N(f) - \mu_n(f)|^2] \leq \sum_{j=0}^{n} \mathrm{Var}_{\mu_j}(q_{j,n}(f)) + \|f\|_{n,p}^2 \widehat{c}_n(p)\ \overline{\varepsilon}_n^N(p)$$

and

$$\overline{\varepsilon}_n^N(p) \leq 2\frac{\overline{v}_n(p)}{N}$$

5.3 Stability of Feynman-Kac Semigroups under Local Mixing

In this section, we generalize the analysis of Section 3.3 to the present setting, weakening the global mixing assumptions to local ones. We begin with a few more definitions. For $j \in I$ let $m_{k,k+1}(j)$ be the relative change in the mass of F_j between μ_k and μ_{k+1},

$$m_{k,k+1}(j) = \frac{\mu_{k+1}(F_j)}{\mu_k(F_j)}.$$

Furthermore for $0 \leq k \leq n-1$, denote by $\overline{g}_{k,k+1}$ the normalized relative density between μ_k and μ_{k+1},

$$\overline{g}_{k,k+1}(x) = \frac{g_{k,k+1}(x)}{\mu_k(g_{k,k+1})} \qquad \text{for} \qquad x \in E.$$

Next we define restricted relative densities: For $0 \leq k \leq n-1$, $j \in I$ and $x \in E$,

$$\overline{g}_{k,k+1,j}(x) = \frac{1}{m_{k,k+1}(j)}\overline{g}_{k,k+1}(x)\ 1_{F_j}(x).$$

Observe that with this choice of $\overline{g}_{k,k+1,j}$ we have for $f \in B(E)$, $0 \leq k \leq n-1$ and $j \in I$ that

$$\mu_{k+1,j}(f) = \frac{\mu_{k+1}(f1_{F_j})}{\mu_{k+1}(F_j)} = \frac{1}{m_{k,k+1}(j)}\frac{\mu_k(f\overline{g}_{k,k+1}1_{F_j})}{\mu_k(F_j)} = \mu_{k,j}(\overline{g}_{k,k+1,j}f),$$

i.e., $\overline{g}_{k,k+1,j}$ is a relative density between $\mu_{k,j}$ and $\mu_{k+1,j}$.

Like in Section 3.3 we postulate a uniform upper bound on relative densities, this time on restricted relative densities: We assume that for some $\gamma > 1$ we have for every $0 \leq k \leq n-1$, every $j \in I_k$ and every $x \in F_j$

$$\overline{g}_{k,k+1,j}(x) = \frac{\mu_k(F_j)}{\mu_{k+1}(F_j)}\overline{g}_{k,k+1}(x) \leq \gamma.$$

This assumption is neither stronger nor weaker than the corresponding bound on $\overline{g}_{k,k+1}$ assumed in Section 3.3. In many cases it will however be weaker in the sense of being fulfilled with a smaller constant γ. Roughly, this is the case when the largest values of $\overline{g}_{k,k+1}$ occur in components F_j which gain importance. For instance, in the extreme case where $\overline{g}_{k,k+1}$ is constant on each component F_j with $j \in I_k$ we can choose $\gamma = 1$.

Again, it proves to be convenient not to work with $q_{j,k}$ directly but to work with the semigroup $\hat{q}_{j,k}$ defined as follows: For $1 \leq k \leq n-1$ define $\hat{q}_{k,k+1} : B(E) \to B(E)$

by

$$\hat{q}_{k,k+1}(f) = K_k \left(\overline{g}_{k,k+1} f\right)$$

Furthermore, for $1 \leq j < k \leq n$ the mapping $\hat{q}_{j,k} : B(E) \to B(E)$ is given by

$$\hat{q}_{j,k}(f) = \hat{q}_{j,j+1}(\hat{q}_{j+1,j+2}(\ldots \hat{q}_{k-1,k}(f))) \quad \text{and} \quad \hat{q}_{k,k}(f) = f,$$

so that $\hat{q}_{j,k}$ is a semigroup. $q_{j,k}$ and $\hat{q}_{j+1,k}$ are related through

$$q_{j,k}(f) = \overline{g}_{j,j+1} \hat{q}_{j+1,k}(K_k(f)).$$

In Lemma 5.5 below, we show how L_p-L_q-bounds for $\hat{q}_{j,k}$ can be used to obtain L_p-L_q-bounds for $q_{j,k}$.

Like in Section 3.3, we proceed by considering first L_2-bounds for one time-step and then iterated L_2-bounds. From these we conclude one-step L_p-bounds and then, in Proposition 5.1, iterated L_p-bounds for $\hat{q}_{j,k}$. Afterwards, we show how to extend this result to L_p-L_q-bounds, using local hyperboundedness, and to the semigroup $q_{j,k}$. Corollary 5.5 concludes the bound for $q_{j,k}$ needed in order to make the constants in the error bound for Sequential MCMC in Corollary 5.1 explicit.

It proves to be useful, to consider mostly inequalities which bound, for $i \in I_j$, $\|\hat{q}_{j,k}(f)\|_{j,i,p}$ against $\max_{l \in s_k(i)} \|f\|_{k,l,p}$. The inequalities which bound $\|\hat{q}_{j,k}(f)\|_{j,p}$ against $\|f\|_{k,p}$ can then be concluded by taking the maximum over $i \in I_j$. So to say, the latter inequalities are the final results while the former are more useful tools in proving further results.

In order to keep track of how mass is shifted between different components, two more definitions are needed: For $0 \leq j < k \leq n$ and $i \in I_j$ define by $M_{j,k}(i)$ the following iterated version of $m_{j,j+1}(i)$:

$$M_{j,k}(i) = \max_{l \in s_k(i)} \prod_{r=j}^{k-1} m_{r,r+1}(p_r(l)).$$

This is the maximal product of relative mass changes one has to go through when moving from F_i, $i \in I_j$ to one of its successors F_l, $l \in s_k(i) \subseteq I_k$. For the transition from r to $r+1$ the relative mass change of the predecessor of F_l at level r is taken into account. Observe that for $i \in I_j$ we have the relation

$$M_{j,k}(i) = m_{j,j+1}(i) \max_{l \in s_{j+1}(i)} M_{j+1,k}(l). \tag{5.2}$$

Furthermore, we define for $0 \leq j < k \leq n$ the constant $A_{j,k}$ by

$$A_{j,k} = \max_{i \in I_j} M_{j,k}(i).$$

Before we come to local mixing properties and L_p-bounds, we briefly look at the L_1-case:

Lemma 5.1. *For $0 \leq j < k \leq n$, $f \in B(E)$ and $i \in I_j$ we have*

$$\|\hat{q}_{j,k}(f)\|_{j,i,1} \leq M_{j,k}(i) \max_{l \in s_k(i)} \|f\|_{k,l,1}. \tag{5.3}$$

Moreover,

$$\|\hat{q}_{j,k}(f)\|_{j,1} \leq A_{j,k} \|f\|_{k,1}. \tag{5.4}$$

Proof. We can write

$$\begin{aligned} \|\hat{q}_{j,k}(f)\|_{j,i,1} &= \mu_{j,i}(|K_j(\overline{g}_{j,j+1}\hat{q}_{j+1,k}(f))|) \\ &\leq m_{j,j+1}(i)\mu_{j,i}(\overline{g}_{j,j+1,i}|\hat{q}_{j+1,k}(f)|) \\ &\leq m_{j,j+1}(i) \max_{l \in s_{j+1}(i)} \mu_{j+1,l}(|\hat{q}_{j+1,k}(f)|) \\ &\leq m_{j,j+1}(i) \max_{l \in s_{j+1}(i)} \|\hat{q}_{j+1,k}(f)\|_{j+1,l,1}. \end{aligned} \tag{5.5}$$

Iterating this bound yields

$$\|\hat{q}_{j,k}(f)\|_{j,i,1} \leq m_{j,j+1}(i) \max_{l_{j+1} \in s_{j+1}(i)} m_{j+1,j+2}(l_{j+1}) \dots \max_{l_{k-1} \in s_{k-1}(l_{k-2})} m_{k-1,k}(l_{k-1}) \|f\|_{k,l_{k-1},1}.$$

Note that by iterating (5.2) we obtain

$$M_{j,k}(i) = m_{j,j+1}(i) \max_{l_{j+1} \in s_{j+1}(i)} m_{j+1,j+2}(l_{j+1}) \dots \max_{l_{k-1} \in s_{k-1}(l_{k-2})} m_{k-1,k}(l_{k-1}).$$

Thus applying

$$\|f\|_{k,l_{k-1},1} \leq \max_{l \in s_k(i)} \|f\|_{k,l,1}$$

in (5.5) yields (5.3). Taking the maximum over $i \in I_j$ gives (5.4). □

The proof illustrates how the constants $M_{j,k}(i)$ and $A_{j,k}$ come into play in our bounds. The same arguments appear – in less detail and alongside further complications – in our proofs for $p > 1$. In fact, this didactic purpose is the main motivation behind Lemma 5.1: The argument (3.4) from Section 3.3 still goes through, showing that

$$\|\hat{q}_{j,k}(f)\|_{L_1(\mu_j)} \leq \|f\|_{L_1(\mu_k)}.$$

Moreover, in a similar fashion one can show that for all $i \in I$

$$\|\hat{q}_{j,k}(f)\|_{j,i,1} \leq \frac{\mu_k(F_i)}{\mu_j(F_i)} \|f\|_{k,i,1}.$$

This implies

$$\|\hat{q}_{j,k}(f)\|_{j,1} \leq \left(\max_{i \in I_j} \frac{\mu_k(F_i)}{\mu_j(F_i)} \right) \|f\|_{k,1} \tag{5.6}$$

which is generally an improvement over Lemma 5.1.

We now state the local mixing conditions behind our L_p-bounds for the case $p \geq 2$: We assume that we have uniform constants $\alpha > 0$ and $\beta \in [0,1]$ such that for all

$1 \leq k < n$, for all $f \in B(E)$ and for all $i \in I_k$

$$\|\hat{q}_{k,k+1}(f)\|_{k,i,2}^2 \leq m_{k,k+1}(i)^2 \left(\alpha \|f\|_{k+1,i,2}^2 + \beta \mu_{k+1,i}(f)^2\right). \tag{5.7}$$

One way to ensure that (5.7) holds is to assume that the kernels K_k possess the following contraction property: There exists $\rho \in (0,1)$ such that for all $1 \leq k < n$ for all $f \in B(E)$ and for all $i \in I_k$

$$\mu_{k,i}(K_k(f - \mu_{k,i}(f))^2) \leq (1-\rho)\mathrm{Var}_{\mu_{k,i}}(f). \tag{5.8}$$

Then it can be shown with the same reasoning as in Lemma 3.6 that (5.7) holds with $\alpha = (1-\rho)\gamma$ and $\beta = \rho$. Moreover, with the same arguments as in Lemma 3.7 it follows that (5.8) holding with a sufficiently large ρ is equivalent to a local Poincaré inequality with a sufficiently large spectral gap being satisfied.

We now turn to proving an L_2 inequality for $\hat{q}_{j,k}$. Note first that (5.7) immediately implies the following one-step L_2-bounds:

Corollary 5.2. *For $1 \leq k < n$, for all $f \in B(E)$ and for all $i \in I_k$ we have*

$$\|\hat{q}_{k,k+1}(f)\|_{k,i,2}^2 \leq m_{k,k+1}(i)^2 \left(\alpha \left(\max_{l \in s_{k+1}(i)} \|f\|_{k+1,l,2}^2\right) + \beta \left(\max_{l \in s_{k+1}(i)} \mu_{k+1,l}(f)^2\right)\right). \tag{5.9}$$

and

$$\begin{aligned} \|\hat{q}_{k,k+1}(f)\|_{k,2}^2 &\leq A_{k,k+1}^2 \left(\alpha \|f\|_{k+1,2}^2 + \max_{l \in I_{k+1}} \beta \mu_{k+1,l}(f)^2\right) \\ &\leq A_{k,k+1}^2 (\alpha + \beta) \|f\|_{k+1,2}^2. \end{aligned}$$

Next we iterate (5.9) to obtain an L_2-bound for more than one step:

Lemma 5.2. *Assume $\alpha < 1$. Then for $1 \leq j < k \leq n$ and $f \in B(E)$ and for $i \in I_j$ we have the bounds*

$$\|\hat{q}_{j,k}(f)\|_{j,i,2}^2 \leq M_{j,k}(i)^2 \left(\alpha^{k-j} \left(\max_{l \in s_k(i)} \|f\|_{k,l,2}^2\right) + \frac{\beta}{1-\alpha} \left(\max_{l \in s_k(i)} \mu_{k,l}(f)^2\right)\right), \tag{5.10}$$

and

$$\|\hat{q}_{j,k}(f)\|_{j,i,2} \leq M_{j,k}(i) \frac{1}{(1-\alpha)^{\frac{1}{2}}} \max_{l \in s_k(i)} \|f\|_{k,l,2}, \tag{5.11}$$

and

$$\|\hat{q}_{j,k}(f)\|_{j,2} \leq A_{j,k} \frac{1}{(1-\alpha)^{\frac{1}{2}}} \|f\|_{k,2}. \tag{5.12}$$

Proof. Applying (5.9) yields

$$\begin{aligned}\|\hat{q}_{j,k}(f)\|^2_{j,i,2} &\leq m_{j,j+1}(i)^2 \\ &\qquad \left(\alpha \max_{l\in s_{j+1}(i)} \|\hat{q}_{j+1,k}(f)\|^2_{j+1,l,2} + \beta \max_{l\in s_{j+1}(i)} \mu_{j+1,l}(\hat{q}_{j+1,k}(f))^2\right).\end{aligned} \tag{5.13}$$

Arguing as in the proof of Lemma 5.1 yields the inequality

$$\max_{l\in s_{j+1}(i)} \mu_{j+1,l}(\hat{q}_{j+1,k}(f))^2 \leq M_{j+1,k}(i)^2 \max_{l\in s_k(i)} \mu_{k,l}(f)^2 \tag{5.14}$$

which can be used to bound the second term on the right hand side of (5.13). To the first term in (5.13) we can apply again (5.9) which yields again two terms, one which can be bounded through (5.9) and one which can be bounded through (5.14). Iterating this reasoning and collecting the factors $m_{r,r+1}$ into terms $M_{j,k}$ gives us

$$\|\hat{q}_{j,k}(f)\|^2_{j,i,2} \leq M_{j,k}(i)^2 \left(\alpha^{k-j}\left(\max_{l\in s_k(i)} \|f\|^2_{k,l,2}\right) + \beta \sum_{r=0}^{k-j-1} \alpha^r \max_{l\in s_k(i)} \mu_{k,l}(f)^2\right). \tag{5.15}$$

Applying to this the geometric series inequality yields (5.10). Since we have

$$\max_{l\in s_k(i)} \mu_{k,l}(f)^2 \leq \max_{l\in s_k(i)} \|f\|^2_{k,l,2}$$

and since we assumed $\beta \leq 1$ we can conclude from (5.15) that

$$\|\hat{q}_{j,k}(f)\|^2_{j,i,2} \leq M_{j,k}(i)^2 \sum_{r=0}^{k-j} \alpha^r \max_{l\in s_k(i)} \|f\|^2_{k,l,2}$$

which implies (5.11) by the geometric series inequality. Taking the maximum over $i \in I_j$ in (5.11) gives (5.12). □

Our next step is the following one-step L_p-bound:

Lemma 5.3. *For $1 \leq k < n$, for all $f \in B(E)$, for all $i \in I_k$ and for all $p \geq 1$ we have*

$$\begin{aligned}&\|\hat{q}_{k,k+1}(f)\|^{2p}_{k,i,2p} \\ \leq\; & m_{k,k+1}(i)^{2p}\gamma^{2p-2}\left(\alpha\left(\max_{l\in s_{k+1}(i)} \|f\|^{2p}_{k+1,l,2p}\right) + \beta\left(\max_{l\in s_{k+1}(i)} \|f\|^{2p}_{k+1,l,p}\right)\right)\end{aligned} \tag{5.16}$$

and

$$\begin{aligned}\|\hat{q}_{k,k+1}(f)\|^{2p}_{k,2p} &\leq A^{2p}_{k,k+1}\gamma^{2p-2}\left(\alpha\|f\|^{2p}_{k+1,2p} + \beta\|f\|^{2p}_{k+1,p}\right) \\ &\leq A^{2p}_{k,k+1}\gamma^{2p-2}(\alpha+\beta)\|f\|^{2p}_{k+1,2p}.\end{aligned} \tag{5.17}$$

Proof. We can write

$$\begin{aligned}\|\hat{q}_{k,k+1}(f)\|_{k,i,2p}^{2p} &= \mu_{k,i}(|K_k(\overline{g}_{k,k+1}f)|^{2p}) \\ &\leq \mu_{k,i}(K_k(\overline{g}_{k,k+1}^{p}|f|^p)^2) \\ &= \|\hat{q}_{k,k+1}(\overline{g}_{k,k+1}^{p-1}|f|^p)\|_{k,i,2}^2.\end{aligned}$$

This expression we can bound using (5.7) to obtain

$$\begin{aligned}\|\hat{q}_{k,k+1}(f)\|_{k,i,2p}^{2p} &\leq m_{k,k+1}(i)^2\left(\alpha\|\overline{g}_{k,k+1}^{p-1}|f|^p\|_{k+1,i,2}^2+\beta\|\overline{g}_{k,k+1}^{p-1}|f|^p\|_{k+1,i,1}^2\right) \\ &\leq m_{k,k+1}(i)^{2p}\gamma^{2p-2}\left(\alpha\||f|^p\|_{k+1,i,2}^2+\beta\||f|^p\|_{k+1,i,1}^2\right) \\ &\leq m_{k,k+1}(i)^{2p}\gamma^{2p-2}\left(\alpha\|f\|_{k+1,i,2p}^{2p}+\beta\|f\|_{k+1,i,p}^{2p}\right)\end{aligned}$$

which immediately implies (5.16). (5.17) follows by taking the maximum over $i \in I_k$. □

Next we iterate the bound of Lemma 5.3 to show how an L_p-bound for $\hat{q}_{j,k}$ implies an L_{2p}-bound:

Lemma 5.4. *Assume that $\alpha\gamma^{2p-2} < 1$ and that for some $\delta(p) \geq 1$ we have for all $1 \leq j < k \leq n$, $i \in I_j$ and $f \in B(E)$ the inequality*

$$\|\hat{q}_{j,k}(f)\|_{j,i,p} \leq M_{j,k}(i)\delta(p)\max_{l\in s_k(i)}\|f\|_{k,l,p} \tag{5.18}$$

is fulfilled. Then we have

$$\|\hat{q}_{j,k}(f)\|_{j,i,2p} \leq M_{j,k}(i)\delta(2p)\max_{l\in s_k(i)}\|f\|_{k,l,2p} \tag{5.19}$$

with

$$\delta(2p) = \delta(p)\frac{\gamma^{1-\frac{1}{p}}}{(1-\alpha\gamma^{2p-2})^{\frac{1}{2p}}}.$$

Moreover, we have

$$\|\hat{q}_{j,k}(f)\|_{j,2p} \leq A_{j,k}\delta(2p)\|f\|_{k,2p}. \tag{5.20}$$

Proof. Define $\theta = \alpha\gamma^{2p-2}$. Iterating the inequality of Lemma 5.3 and utilizing that $\beta \leq 1$, we get

$$\begin{aligned}\|\hat{q}_{j,k}(f)\|_{j,i,2p}^{2p} &\leq M_{j,k}(i)^{2p}\theta^{k-j}\left(\max_{l\in s_k(i)}\|f\|_{k,l,2p}^{2p}\right) \\ &+ \gamma^{2p-2}\sum_{r=j+1}^{k}\theta^{r-1-j}\max_{l\in s_r(i)}R_{j,r}(l)^{2p}\|\hat{q}_{r,k}(f)\|_{r,l,p}^{2p}, \end{aligned}\tag{5.21}$$

where for $l \in I_r$, $R_{j,r}(l)$ is defined by

$$R_{j,r}(l) = \prod_{t=j}^{r-1} m_{t,t+1}(p_t(l)).$$

Observe that for $i \in I_j$ we have

$$M_{j,k}(i) = \max_{l \in s_k(i)} R_{j,k}(l)$$

and moreover for $j < r < k$ and $i \in I_j$

$$M_{j,k}(i) = \max_{l \in s_r(i)} R_{j,r}(l) M_{r,k}(l) \tag{5.22}$$

Thus applying (5.18) and then (5.22) to bound the factors $\|\hat{q}_{r,k}(f)\|_{r,l,p}$, we obtain from (5.21) the inequality

$$\begin{aligned} \|\hat{q}_{j,k}(f)\|_{j,i,2p}^{2p} &\leq M_{j,k}(i)^{2p} \theta^{k-j} \left(\max_{l \in s_k(i)} \|f\|_{k,l,2p}^{2p} \right) \\ &+ \gamma^{2p-2} M_{j,k}(i)^{2p} \delta(p)^{2p} \sum_{r=j+1}^{k} \theta^{r-1-j} \max_{l \in s_k(i)} \|f\|_{k,l,p}^{2p}. \end{aligned}$$

Since we assumed $\gamma \geq 1$ and $\delta(p) \geq 1$ and since we have

$$\max_{l \in s_k(i)} \|f\|_{k,l,p}^{2p} \leq \max_{l \in s_k(i)} \|f\|_{k,l,2p}^{2p}$$

we thus have

$$\|\hat{q}_{j,k}(f)\|_{j,i,2p}^{2p} \leq \gamma^{2p-2} M_{j,k}(i)^{2p} \delta(p)^{2p} \left(\max_{l \in s_k(i)} \|f\|_{k,l,p}^{2p} \right) \sum_{r=j+1}^{k} \theta^{r-1-j}.$$

By the geometric series inequality and our assumption of $\theta < 1$, we thus get

$$\|\hat{q}_{j,k}(f)\|_{j,i,2p} \leq M_{j,k}(i)^{2p} \delta(2p) \max_{l \in s_k(i)} \|f\|_{k,l,p},$$

with

$$\delta(2p) = \delta(p) \frac{\gamma^{1-\frac{1}{p}}}{(1-\theta)^{\frac{1}{2p}}}.$$

This shows (5.19). (5.20) follows by taking the maximum over $i \in I_j$. □

Combining Lemmas 5.2 and 5.4 we can now state the key result of this section as follows:

Proposition 5.1. *For $r \in \mathbb{N}$, consider $p = 2^r$ and assume that $\alpha\gamma^{p-2} < 1$. Then we have for $1 \leq j < k \leq n$, for $i \in I_j$ and $f \in B(E)$ the inequality*

$$\|\hat{q}_{j,k}(f)\|_{j,i,p} \leq M_{j,k}(i) \delta(p) \max_{l \in s_k(i)} \|f\|_{k,l,p} \tag{5.23}$$

with

$$\delta(p) = \prod_{j=1}^{r} \frac{\gamma^{1-2^{-(j-1)}}}{(1-\alpha\gamma^{2^j-2})^{2^{-j}}} < \frac{\gamma^{r-2+2^{-(r-1)}}}{1-\alpha\gamma^{2^r-2}}.$$

Moreover,

$$\|\hat{q}_{j,k}(f)\|_{j,p} \leq A_{j,k}\delta(p)\|f\|_{k,p} \tag{5.24}$$

Proof. We proceed by induction over r. The cases $r = 0$ and $r = 1$ follow from Lemmas 5.1 and 5.2, respectively. The inequalities for $r > 1$ follow because Lemma 5.4 implies that we can choose

$$\delta(2^r) = \delta(2) \prod_{j=2}^{r} \frac{\gamma^{1-2^{-(j-1)}}}{(1-\alpha\gamma^{2^j-2})^{2^{-j}}}.$$

We can apply Lemma 5.4 iteratively, since $\alpha\gamma^{p-2} < 1$ implies $\alpha\gamma^{q-2} < 1$ for all $q \leq p$. For the upper bound on $\delta(p)$, we apply the geometric series equality in the nominator, bound the term in brackets under the exponent in the denominator by $1 - \alpha\gamma^{p-2}$ and apply the geometric series inequality to the product. This shows (5.23). (5.24) follows by taking the maximum over $i \in I_j$. □

Since the constants $\delta(2^r)$ are monotonically increasing in r, we can immediately extend the bounds of Proposition 5.1 to general $p \geq 1$ using the Riesz-Thorin interpolation theorem (see Davies (1990), §1.1.5):

Corollary 5.3. *Consider $p \in [2^r, 2^{r+1}]$ for $r \in \mathbb{N}$ and assume $\alpha\gamma^{2^{r+1}-2} < 1$. Then for $1 \leq j < k \leq n$ and $f \in B(E)$ and $i \in I_j$ we have*

$$\|\hat{q}_{j,k}(f)\|_{j,i,p} \leq M_{j,k}(i)\delta(p) \max_{l \in s_k(i)} \|f\|_{k,l,p}$$

and

$$\|\hat{q}_{j,k}(f)\|_{j,p} \leq A_{j,k}\delta(p)\|f\|_{k,p}$$

with $\delta(p)$ given by

$$\delta(p) = \delta(2^{r+1})$$

where $\delta(2^{r+1})$ is defined as in Proposition 5.1.

We still need two more results: one which shows how to translate L_p-stability into L_p-L_q-stability using local hyper-boundedness, and one which relates our bounds for $\hat{q}_{j+1,k}$ to corresponding bounds for $q_{j,k}$. We first show how L_p-L_q-inequalities follow from our L_p-inequalities and a local hypercontractivity assumption on the kernels K_j.

Corollary 5.4. *Consider $p \geq 1$ and $q \geq 1$. Let $q \in [2^r, 2^{r+1}]$ for $r \in \mathbb{N}$ and assume $\alpha\gamma^{2^{r+1}-2} < 1$. Assume that for $1 \leq j < n$ we have a constant $\theta_j(p,q) \geq 0$ such that for all $i \in I_j$ and all $f \in B(E)$ we have*

$$\|K_j(f)\|_{j,i,p} \leq \theta_j(p,q)\|f\|_{j,i,q}. \tag{5.25}$$

Then for $j < k \leq n$ we have

$$\|\hat{q}_{j,k}(f)\|_{j,i,p} \leq M_{j,k}(i)\theta_j(p,q)\gamma^{\frac{q-1}{q}}\delta(q) \max_{l\in s_k(i)} \|f\|_{k,l,q} \tag{5.26}$$

and

$$\|\hat{q}_{j,k}(f)\|_{j,p} \leq A_{j,k}\theta_j(p,q)\gamma^{\frac{q-1}{q}}\delta(q)\|f\|_{k,q} \tag{5.27}$$

with $\delta(q)$ as defined in Corollary 5.3.

Proof. By (3.10) we have

$$\|\hat{q}_{j,k}(f)\|_{j,i,p} \leq \theta_j(p,q)m_{j,j+1}(i) \max_{l\in s_{j+1}(i)} \|\overline{g}_{j,j+1,i}\hat{q}_{j+1,k}(f)\|_{j,l,q}$$

and thus by Corollary 5.3

$$\|\hat{q}_{j,k}(f)\|_{j,i,p} \leq \theta_j(p,q)M_{j,k}(i)\gamma^{\frac{q-1}{q}}\delta(q) \max_{l\in s_k(i)} \|f\|_{k,l,q}.$$

This shows (5.26). Taking the maximum over $i \in I_j$ proves (5.27). □

Next, we show how to obtain bounds for $q_{j,k}$ from our bounds for $\hat{q}_{j+1,k}$:

Lemma 5.5. *Assume that for some $p \geq 1$ and $q \geq 1$ and for fixed $1 \leq j+1 < k \leq n$ we have a $\delta \geq 0$ such that for all $l \in I_{j+1}$ and for all $f \in B(E)$*

$$\|\hat{q}_{j+1,k}(f)\|_{j+1,l,p} \leq \delta\, M_{j+1,k}(l) \max_{r\in s_k(l)} \|f\|_{k,r,q}. \tag{5.28}$$

Then we have for all $i \in I_j$

$$\|q_{j,k}(f)\|_{j,i,p} \leq \gamma^{\frac{p-1}{p}}\delta\, M_{j,k}(i) \max_{r\in s_k(i)} \|f\|_{k,r,q}.$$

and

$$\|q_{j,k}(f)\|_{j,p} \leq \gamma^{\frac{p-1}{p}}\delta\, A_{j,k}\, \|f\|_{k,q}.$$

Proof. Note that we have for $i \in I_j$

$$\begin{aligned}
\|q_{j,k}(f)\|_{j,i,p} &= \mu_{j,i}(|\overline{g}_{j,j+1}\hat{q}_{j+1,k}(K_k(f))|^p)^{\frac{1}{p}} \\
&= m_{j,j+1}(i)\,\mu_{j,i}(|\overline{g}_{j,j+1,i}\hat{q}_{j+1,k}(K_k(f))|^p)^{\frac{1}{p}} \\
&\leq \gamma^{\frac{p-1}{p}} m_{j,j+1}(i)\mu_{j+1,i}(|\hat{q}_{j+1,k}(K_k(f))|^p)^{\frac{1}{p}} \\
&\leq \gamma^{\frac{p-1}{p}} m_{j,j+1}(i) \max_{l\in s_{j+1}(i)} \mu_{j+1,l}(|\hat{q}_{j+1,k}(K_k(f))|^p)^{\frac{1}{p}} \\
&= \gamma^{\frac{p-1}{p}} m_{j,j+1}(i) \max_{l\in s_{j+1}(i)} \|\hat{q}_{j+1,k}(K_k(f))\|_{j,l,p}
\end{aligned}$$

and thus by (5.28)

$$\begin{aligned}
\|q_{j,k}(f)\|_{j,i,p} &\leq \gamma^{\frac{p-1}{p}}\, m_{j,j+1}(i)\, \delta \max_{l\in s_{j+1}(i)} M_{j+1,k}(l) \max_{r\in s_k(l)} \|K_k(f)\|_{k,r,q} \\
&\overset{(5.2)}{\leq} \gamma^{\frac{p-1}{p}}\, M_{j,k}(i)\, \delta \max_{r\in s_k(i)} \|K_k(f)\|_{k,r,q} \\
&\leq \gamma^{\frac{p-1}{p}}\, M_{j,k}(i)\, \delta \max_{r\in s_k(i)} \|f\|_{k,r,q}
\end{aligned}$$

where in the last step we used that by Jensen's inequality $|K_k(f)|^q \leq K_k(|f|^q)$ and that K_k is stationary with respect to $\mu_{k,l}$ for all $l \in I$. This shows the first inequality. The second inequality follows by taking the maximum over $i \in I_j$ on both sides. □

Combining Lemma 5.5 and Corollary 5.4 we can conclude the types of inequalities needed in the error bound for Sequential MCMC stated in Corollary 5.1 in the previous section:

Corollary 5.5. *Consider $p \geq 1$ and $q \geq 1$. Let $q \in [2^r, 2^{r+1}]$ for $r \in \mathbb{N}$ and assume $\alpha\gamma^{2^{r+1}-2} < 1$. Assume that for all $1 \leq j < n$ we have a constant $\theta_j(p,q) \geq 0$ such that for all $i \in I_j$ and all $f \in B(E)$ we have*

$$\|K_j(f)\|_{j,i,p} \leq \theta(p,q)\|f\|_{j,i,q}.$$

Then for $1 \leq j < k \leq n$ we have

$$\|q_{j,k}(f)\|_{j,p} \leq \widetilde{c}_{j,k}(p)\|f\|_{k,q} \tag{5.29}$$

with

$$\widetilde{c}_{j,k}(p) = A_{j,k}\theta(p,q)\gamma^{\frac{p-1}{p}}\gamma^{\frac{q-1}{q}}\delta(q)$$

where $\delta(q)$ is as defined in Corollary 5.3.

The stability inequalities in this section thus differ from those of Section 3.3 by containing the factor $A_{j,k}$ on the right-hand side. For the case treated in Section 3.3, i.e., $|I_j| = 1$ for $0 \leq j \leq n$ and $F_i = E$ for all $i \in I$, we obtain $A_{j,k} = 1$. Thus the results of the present section contain those of Section 3.3 as special cases. For the case of invariant partitions treated in Eberle and Marinelli (2010), i.e., $|I_j| = |I_k|$ for $0 \leq j < k \leq n$, we obtain

$$A_{j,k} = \max_{i\in I_j} \frac{\mu_k(F_i)}{\mu_j(F_i)} \tag{5.30}$$

which is a discrete-time analogue of their constant.

All the results of this section can be proved much more easily with $\widetilde{A}_{j,k}$ defined by

$$\widetilde{A}_{j,k} = \prod_{r=j}^{k-1} \max_{i\in I_r} m_{r,r+1}(i)$$

in place of $A_{j,k}$. This constant is typically larger however. For instance, in the case of invariant partitions we obtain

$$\widetilde{A}_{j,k} = \prod_{r=j}^{k-1} \max_{i \in I_r} \frac{\mu_{r+1}(F_i)}{\mu_r(F_i)}$$

which is typically greater than $A_{j,k}$ which is given by (5.30) in this case. In contrast, in the setting of trees in Chapter 4 we had (not only for invariant partitions) constants corresponding to $\widehat{A}_{j,k}$ given by

$$\widehat{A}_{j,k} = \max_{i \in I_j} \frac{\mu_k(F_i)}{\mu_j(F_i)} \tag{5.31}$$

which are typically smaller than these constants $A_{j,k}$. As was shown in (5.6), we can achieve the constants $\widehat{A}_{j,k}$ in the L_1 case. For $p > 1$, in order to apply the local mixing conditions we have to analyze components more separately. This leads to the worse constants $A_{j,k}$.

Compared to the setting of Chapter 4 the present setting is more general in two respects: We take into account local mixing and local variations in relative densities. To disentangle these two factors to some extent, consider the case where we take into account local mixing but assume that the relative densities $g_{k,k+1}$ are constant on each of the sets F_j with $j \in I_k$. In that case, we have $\gamma = 1$ and the inequality (5.29) in Corollary 5.5 becomes

$$\|q_{j,k}(f)\|_{j,p} \leq A_{j,k}\theta(p,q)\frac{1}{1-\alpha}\|f\|_{k,q}.$$

Thus, compared to the results of Chapter 4 we obtain different norms, we obtain the constants $A_{j,k}$ which are similar but worse than the corresponding constants in Chapter 4 and we obtain an additional factor taking into account hyperboundedness and local mixing.

References

[1] C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto and G. Scheffer, *Sur les inégalités de Sobolev logarithmiques*, Panoramas et Synthèses, 10, Société Mathématique de France, Paris, 2000.

[2] C. Andrieu and J. Thoms, *A tutorial on adaptive MCMC*, Statistics and Computing, 18, 343-373, 2008.

[3] Y. Atchadé, G. Fort, E. Moulines and P. Priouret, *Adaptive Markov chain Monte Carlo : Theory and Methods*, In: D. Barber, A. T. Cemgil and S. Chiappia (eds.), Bayesian Time Series Models, Cambridge University Press, 2011.

[4] Y. Atchadé, J. S. Rosenthal and G. O. Roberts, *Optimal scaling of Metropolis-Coupled Markov chain Monte Carlo*, Statistics and Computing, 21, 555-568, 2011.

[5] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*, Springer, New York, 2009.

[6] F. Bassetti and F. Leisen, *Metropolis algorithm and equienergy sampling for two mean field spin system*s, Working Paper, University of Insubria, 2007.

[7] N. Bhatnagar and D. Randall, *Torpid Mixing of Simulated Tempering on the Potts Model*, Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA), 278-287, 2004.

[8] O. Cappé, R. Douc and E. Moulines, *Comparison of Resampling Schemes for Particle Filtering*, Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, 2005.

[9] O. Cappé, E. Moulines and T. Rydén, *Inference in Hidden Markov Models*, Springer, New York, 2005.

[10] S. Caracciolo, A. Pelissetto and A. D. Sokal, *Two Remarks on Simulated Tempering*, Unpublished Manuscript, 1992.

[11] G. Celeux, M. Hurn and C. P. Robert, *Computational and inferential difficulties with mixture posterior distributions*, Journal of the American Statistical Association, 95, 957-970, 2000.

[12] F. Cérou, P. Del Moral and A. Guyader, *A nonasymptotic variance theorem for unnormalized Feynman-Kac particle models*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 47, 629-649, 2011.

[13] M.-F. Chen and F.-Y. Wang, *Estimates of logarithmic Sobolev constant: an improvement of Bakry-Emery criterion*, Journal of Functional Analysis, 144, 287-300, 1997.

[14] N. Chopin, *Central Limit Theorem for Sequential Monte Carlo methods and its application to Bayesian inference*, Annals of Statistics, 32, 2385-2411, 2004.

[15] D. Crisan, J. Gaines and T. Lyons, *Convergence of a Branching Particle Method to the Solution of the Zakai Equation*, SIAM Journal on Applied Mathematics, 58, 1568-1590, 1998.

[16] E. B. Davies, *Heat kernels and spectral theory*, Cambridge University Press, Cambridge, 1990.

[17] P. Del Moral, *Nonlinear filtering: interacting particle solution*, Markov Processes and Related Fields, 2, 555-579, 1996.

[18] P. Del Moral, A. Doucet and A. Jasra, *Sequential Monte Carlo Samplers*, Journal of the Royal Statistical Society B, 68, 411-436, 2006.

[19] P. Del Moral, A. Doucet and A. Jasra, *On Adaptive Resampling Procedures for SMC Methods*, Bernoulli, to appear, 2011.

[20] P. Del Moral and L. Miclo, *Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to nonlinear filtering*, Séminaire de Probabilités XXXIV, Lecture Notes in Mathematics, vol. 1729, Springer, Berlin, 2000, pp. 1-145.

[21] J.-D. Deuschel and D. W. Stroock, *Hypercontractivity and spectral gap of symmetric diffusions with applications to the stochastic Ising models*, Journal of Functional Analysis, 92, 30-48, 1990.

[22] P. Diaconis, *The MCMC Revolution*, Bulletin of the American Mathematical Society, 46, 179-205, 2009.

[23] P. Diaconis and L. Saloff-Coste, *What do we know about the Metropolis algorithm?*, Journal of Computer and System Sciences, 57, 20-36, 1998.

[24] R. Douc and E. Moulines, *Limit theorems for weighted samples with applications to Sequential Monte Carlo Methods*, Annals of Statistics, 36, 2344-2376, 2008.

[25] R. Douc, E. Moulines and J. S. Rosenthal, *Quantitative bounds on convergence of time-inhomogeneous Markov chains*, Annals of Applied Probability, 14, 1643-1665, 2004.

[26] A. Doucet, N. de Freitas and N. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*, Springer, New York, 2001.

[27] D. J. Earl and M. W. Deem, *Parallel Tempering: Theory, Applications, and New Perspectives*, Physical Chemistry Chemical Physics, 7, 3910-3916, 2005.

[28] A. Eberle and C. Marinelli, *L^p estimates for Feynman-Kac propagators with time-dependent reference measures*, Journal of Mathematical Analysis and Applications, 365, 120-134, 2010.

[29] A. Eberle and C. Marinelli, *Quantitative approximations of evolving probability measures and sequential Markov Chain Monte Carlo methods*, Working Paper, University of Bonn, 2011.

[30] L. J. Fogel, A. J. Owens and M. J. Walsh, *Artificial Intelligence through Simulated Evolution*, Wiley Publishing, New York, 1996.

[31] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*, Springer, New York, 2007.

[32] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 609-628, 1984.

[33] C. J. Geyer and E. A. Thompson, *Annealing Markov chain Monte Carlo with Applications to Ancestral Inference*, Journal of the American Statistical Association, 90, 909-920, 1995.

[34] C. J. Geyer, *Markov chain Monte Carlo maximum likelihood*, Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, 156-163, 1991.

[35] N. Gordon, D. Salmond and A. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, IEE Proceedings F Radar and Signal Processing, 140, 107-113, 1993.

[36] R. B. Gramacy, R. J. Samworth and R. King, *Importance Tempering*, Statistics and Computing 20, 1-7, 2010.

[37] Y. Guan and S. M. Krone, *Small-world MCMC and convergence to multi-modal distributions: From slow mixing to fast mixing*, Annals of Applied Probability, 17, 284-304, 2007.

[38] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57, 97-109, 1970.

[39] J. D. Hol, T. B. Schön and F. Gustafsson, *On resampling algorithms for particle filters*, Proceedings of the Nonlinear Statistical Signal Processing Workshop (NSSPW), Cambridge, 2006.

[40] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.

[41] R. A. Holley, S. Kusuoka and D. W. Stroock, *Asymptotics of the spectral gap with applications to the theory of simulated annealing*, Journal of Functional Analysis, 83, 333-347, 1989.

[42] K. Hukushima and K. Nemoto, *Exchange Monte Carlo Method and Application to Spin Glass Simulations*, Journal of the Physical Society of Japan, 65, 1604-1608, 1996.

[43] C. Jarzynski, *Nonequilibrium equality for free energy differences*, Physical Review Letters, 78, 2690-2693, 1997a.

[44] C. Jarzynski, *Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach*, Physical Review E, 56, 5018-5035, 1997b.

[45] M. Jerrum, J.-B. Son, P. Tetali, and E. Vigoda, *Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains*, Annals of Applied Probability, 14, 1741-1765, 2004.

[46] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, *Optimization by Simulated Annealing*, Science, 220, 671-680, 1983.

[47] S. C. Kou, Q. Zhou and W. H. Wong, *Equi-energy sampler with applications in statistical inference and statistical mechanics*, Annals of Statistics, 34, 1581-1619, 2006.

[48] H. R. Künsch, *Recursive Monte-Carlo filters: algorithms and theoretical analysis*, Annals of Statistics, 33, 1983-2021, 2005.

[49] D. A. Levin, Y. Peres and E. L. Wilmer, *Markov Chains and Mixing Times*, AMS, Providence, 2009.

[50] F. Liang, *Determination of normalizing constants for simulated tempering*, Physica A, 356, 468-480, 2005.

[51] J. Liu, *Monte Carlo strategies in scientific computing*, Springer, New York, 2001.

[52] J. S. Liu and C. Sabatti, *Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions*, In: Bayesian Statistics, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), Oxford University Press, New York, 1998.

[53] L. Lovász, R. Kannan and M. Simonovits, *Random walks and an* $O*(n^5)$ *volume algorithm for convex bodies*, Random Structures and Algorithms, 11, 1-50, 1997.

[54] E. Lyman, F. M. Ytreberg and D. M. Zuckerman, *Resolution Exchange Simulation*, Physical Review Letters, 96, 028105, 2006.

[55] N. Madras and M. Piccioni, *Importance sampling for families of distributions*, Annals of Applied Probability, 9, 1202-1225, 1999.

[56] N. Madras and D. Randall, *Markov chain decomposition for convergence rate analysis*, Annals of Applied Probability, 12, 581-606, 2002.

[57] N. Madras and Z. Zheng, *On the swapping algorithm*, Random Structures and Algorithms, 22, 66-97, 2002.

[58] K. F. Man, K. S. Tang and S. Kwong. *Genetic algorithms*, Springer, New York, 1999.

[59] J.-M. Marin, K. Mengersen and C. P. Robert, *Bayesian modelling and inference on mixtures of distributions*, Handbook of Statistics, 25, 459-507, 2005.

[60] E. Marinari and G. Parisi, *Simulated tempering: A new monte carlo scheme*, Europhysics Letters, 19, 451-458, 1992.

[61] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *Equations of state calculations by fast computing machines*, Journal of Chemical Physics, 21, 1087-1092, 1953.

[62] W. Nadler and U. H. E. Hansmann, *Optimizing Replica Exchange Moves For Molecular Dynamics*, Physical Review E, 76, 057102, 2007.

[63] R. M. Neal, *Sampling from multimodal distributions using tempered transitions*, Statistics and Computing, 6, 353-366, 1996.

[64] R. M. Neal, *Annealed importance sampling*, Statistics and Computing, 11, 125-139, 2001.

[65] S. Park and V. Pande, *Choosing weights for simulated tempering*, Physical Review E, 76, 016703, 2007.

[66] C. Predescu, M. Predescu and C. V. Ciobanu, *The incomplete beta function law for parallel tempering sampling of classical canonical systems*, Journal of Chemical Physics, 120, 4119-4128, 2004.

[67] I. Rechenberg, *Evolution Strategy: Optimization of Technical Systems by Means of Biological Evolution*, Fromman-Holzboog, Stuttgart, 1973.

[68] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, Second Edition, Springer, New York, 2004.

[69] G. O. Roberts, *Optimal metropolis algorithms for product measures on the vertices of a hypercube*, Stochastics and Stochastic Reports, 62, 275-283, 1998.

[70] G. O. Roberts and J. S. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*, Statistical Science, 16, 351-367, 2001.

[71] G. O. Roberts and J. S. Rosenthal, *General state space Markov chains and MCMC algorithms*, Probability Surveys, 1, 20-71, 2004.

[72] D. Rudolf, *Explicit error bounds for lazy reversible Markov chain Monte Carlo*, Journal of Complexity, 25, 11-24, 2009.

[73] C. Sherlock and G. O. Roberts, *Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets*, Bernoulli, 15, 774-798, 2009.

[74] G. M. Torrie and J. P. Valleau, *Nonphysical sampling distributions in Monte Carlo free energy estimation: Umbrella sampling*, Journal of Computational Physics, 3, 187-199, 1977.

[75] D. J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, 2003.

[76] F.-Y. Wang, *On estimation of logarithmic Sobolev constant and gradient estimates of heat semigroups*, Probability Theory and Related Fields, 108, 87-101, 1997.

[77] N. Whiteley, *Sequential Monte Carlo samplers: error bounds and insensitivity to initial conditions*, Working Paper, University of Bristol, 2011.

[78] D. B. Woodard, S. C. Schmidler and M. L. Huber, *Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions*, Annals of Applied Probability, 19, 617-640, 2009a.

[79] D. B. Woodard, S. C. Schmidler and M. L. Huber, *Sufficient conditions for torpid mixing of parallel and simulated tempering* Electronic Journal of Probability, 14, 780-804, 2009b.

[80] Z. Zheng, *On swapping and simulated tempering algorithms*, Stochastic Processes and Applications, 104, 131-154, 2002.