Finite Sample Bounds and Path Selection for Sequential Monte Carlo

by

Joseph Marion

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Scott C. Schmidler, Supervisor

_____
Sayan Mukherjee

_____
Robert L. Wolpert

_____
Patrick Charbonneau

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2018

Abstract

Finite Sample Bounds and Path Selection for Sequential Monte Carlo

by

Joseph Marion

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Scott C. Schmidler, Supervisor

_____
Sayan Mukherjee

_____
Robert L. Wolpert

_____
Patrick Charbonneau

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2018

Abstract

Sequential Monte Carlo (SMC) samplers have received attention as an alternative to Markov chain Monte Carlo for Bayesian inference problems due to their strong empirical performance on difficult multimodal problems, natural synergy with parallel computing environments, and accuracy when estimating ratios of normalizing constants. However, while these properties have been demonstrated empirically, the extent of these advantages remain unexplored theoretically. Typical convergence results for SMC are limited to $\mathcal{O}(\sqrt{N})$ results; they obscure the relationship between the algorithmic factors (weights, Markov kernels, target distribution) and the error of the resulting estimator. This limitation makes it difficult to compare SMC to other estimation methods and challenging to design efficient SMC algorithms from a theoretical perspective.

In this thesis, we provide conditions under which SMC provides a randomized approximation scheme, showing how to choose the number of of particles and Markov kernel transitions at each SMC step in order to ensure an accurate approximation with bounded error. These conditions rely on the sequence of SMC interpolating distributions and the warm mixing times of the Markov kernels, explicitly relating the algorithmic choices to the error of the SMC estimate. This allows us to provide finite-sample complexity bounds for SMC in a variety of settings, including finite state-spaces, product spaces, and log-concave target distributions.

A key advantage of this approach is that the bounds provide insight into the

selection of efficient sequences of SMC distributions. When the target distribution is spherical Gaussian or log-concave, we show that judicious selection of interpolating distributions results in an SMC algorithm with a smaller complexity bound than MCMC. These results are used to motivate the use of a well known SMC algorithm that adaptively chooses interpolating distributions. We provide conditions under which the adaptive algorithm gives a randomized approximation scheme, providing theoretical validation for the automatic selection of SMC distributions.

Selecting efficient sequences of distributions is a problem that also arises in the estimation of normalizing constants using path sampling. In the final chapter of this thesis, we develop automatic methods for choosing sequences of distributions that provide low-variance path sampling estimators. These approaches are motived by properties of the theoretically optimal, lowest-variance path, which is given by the geodesic of the Riemann manifold associated with the path sampling family. For one dimensional paths we provide a greedy approach to step size selection that has good empirical performance. For multidimensional paths, we present an approach using Gaussian process emulation that efficiently finds low variance paths in this more complicated setting.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

In general, when an upper case Latin characters are used to the number of items in a set (for example N) the corresponding lower case character is used to index that set (in this case n).

$\mathcal{X}$	State space
λ	Dominating measure on $\mathcal{X}$
$\mathcal{B}$	Borel sets on $\mathcal{X}$
$\mathcal{P}$	Set of probability measures on $\mathcal{X}$ that are absolutely continuous with respect to λ
$\mathcal{F}$	Set of λ measurable functions $f : \mathcal{X} \to \mathcal{R}$
S	Number of SMC steps
N	Number of SMC particles
$\mathcal{N}$	Gaussian distribution

Indexed Symbols

Symbols used to describe SMC algorithm at step s. When present, the superscript n refers to the SMC particle being indexed.

μ_s	SMC interpolating measure in $\mathcal{P}$
p_s	Normalized density of μ_s
q_s	Unnormalized density of μ_s
z_s	Normalizing constant of μ_s

θ_s	Distribution parameters
w_s	Importance sampling weight
K_s	Markov kernel, generally with limiting distribution μ_s
τ_s	The mixing time of K_s
t_s	Number of Markov kernel transitions according to K_s
X_s^n	Particle at the end of step s
$\tilde{X}_s^n$	Particle resulting from re-sampling at step s
$\hat{\mu}_s$	Marginal distribution of X_s^n
$\tilde{\mu}_s$	Marginal distribution of $\tilde{X}_s^n$

Abbreviations

ESS	Effective sample size
GP	Gaussian process
MCMC	Markov chain Monte Carlo
RAS	Randomized approximation scheme
RESS	Relative effective sample size
SMC	Sequential Monte Carlo

Acknowledgements

I would like to thank my advisor Scott Schmidler for his support and guidance during my time at Duke; I would never have been successful without him. Scott pushed me beyond my limits, making my work far better than I thought possible. I will miss our meandering, multi-hour meetings, though Scott may appreciate having his time returned! It has been an honor to be Scott's student.

I would also like to thank my committee members Sayan Mukherjee, Robert L. Wolpert and Patrick Charbonneau for their insightful feedback and helpful discussion. Thank you also to David Banks, who provided me with guidance and conversation during my first year at Duke; many first year students do not have such an attentive mentor. Thanks are also due to the National Science Foundation (research traineeship grant DMS-1045153) and the Statistical and Applied Mathematical Sciences Institute (NSF grant DMS-1638521) for supporting my research.

While at Duke, I had the pleasure of some excellent summer internships that deepened my appreciation for statistics and broadened my versatility. Special thanks for Chaitanya Chemudugunta, John Harer and Alice Broadhead; I learned valuable lessons from all of you and greatly benefited from your mentorship and experience. My time at Blizzard was especially memorable and I'll forever have feelings of fondness for the moments we spent playing games in the office. Thanks to all of you for welcoming me into your workplaces and making me feel impactful.

My time in graduate school would not have been the same without my roommates

Jake, Sayak and Derek. Together we enjoyed many board games, spicy meals, and good ciders. They entertained my math questions at all hours of the day and dutifully left me alone when I needed space to work. Living with them was a highlight of my experience at Duke and I could not have asked for better roommates. Thanks too to my climbing buddies Christoph and John (and sometimes Dave!);climbing kept me sane and weekly pilgrimages to Smashburger kept me sated. Who knew that such a mundane restaurant could provide so many interesting experiences? I was lucky enough to have my friends from Cornell (Greg, Sara, Brittany and Robert) living in D.C. while I was in Durham. They acted as my support group when I left the Army; visiting them helped kept me centered. Thanks to to my friends from El Paso who kept me distracted from the rigours of the PhD through various gaming activities; they provided a much needed outlet. Special thanks to Parker for experiencing the weirdness of the military-to-school transition with me and keeping in touch. Finally, thanks to my dearest friend Cody for his years of patient friendship; through thick and thin he has always stood by me.

I am deeply thankful to my family for their love, support and sacrifices. Mom and Dad, I know it hasn't always been clear what I've been up to, but I appreciate you bearing with it and providing my foundation. I wouldn't be where I am without you. Becca, it has been interesting doing a PhD alongside you, I think you've taken to it better than I have! Finally, thank you to my best friend and the love of my life Brianna. Being with you has made the last few years particularly special and I look forward to spending the rest of my days with you. I couldn't have asked for a better motivation to finish.

1

Introduction

Sequential Monte Carlo (SMC) samplers are a broad class of simulation methods used to estimate expectations of probability distributions. These methods were originally developed for use in Bayesian state space problems where the dimension of the posterior distribution grows with time; in this context they may be reffered to as bootstrap particle filters [1], sequential importance samplers or resample-move particle filters [2]. Chopin [3] first observed that these methods could be also be applied to Bayesian inference problems of fixed/static dimension by establishing artificial dynamics on the posterior distribution, broadening the class of problems that could be tackled by the approach. The class of SMC methods has continued to grow and now constitutes a wide range of algorithms including Annealed Importance Sampling [4], population Monte Carlo [5] and SMC^2 [6]; a unifying view of the field can be found in [7]. The key similarity connecting these algorithms is the repeated transformation of a collection of N particles, via combinations of weighting, resampling and Markov kernel transitions.

SMC has received attention as an alternative to more popular Markov chain Monte Carlo (MCMC) methods due to several potential advantages. The first is

that SMC is more amenable to parallelization than MCMC, which is an inherently serial algorithm. The computation of weights and application of Markov kernels can be done in parallel for each particle, making the algorithm well suited to parallel computing implmentations [8]. Modifications to the algorithm can be made which further adapt the algorithm to these environments [9, 10]. A second advantage of SMC is that it may exhibit properties similar to parallel tempering, making it suitable for difficult multimodal problems where many MCMC methods fail [7, 11, 12]. Finally, SMC samplers provide approximations of the interpolating distributions at each intermediate stage, allowing for the development of powerful methods that automatically adapt to the problem at hand. This includes adaptively selecting weights [13], Markov kernels [3, 14, 15] and sequences of distributions [16, 15].

While many of these advantages have been demonstrated empirically, in general they have not been confirmed theoretically. Convergence results for SMC are typically of the $\mathcal{O}(\sqrt{N})$ variety, proving the validity and stability of the algorithm but providing limited insight into how the weights, Markov kernels, and properties of the target distribution affect the error of an SMC estimator. Results of this kind include SMC central limit theorems, which have been proven in a variety of settings [17, 18, 19, 20], finite-sample bounds [21, 12, 18] and other kinds of stability results [11, 22, 23, 24]; an overview of this literature is given in Chapter 2. This limited theoretical understanding of SMC makes it difficult to design SMC algorithms, as the relationship between the components and the error of SMC estimates remains unknown. In addition, these results allow for only limited comparison with other methods, as the current bounds are too imprecise to be useful for this task.

The first contribution of this thesis is to improve the theoretical understanding of the SMC algorithm for static target distributions. In Chapter 2, we develop conditions under which an SMC estimator is a randomized approximation scheme (RAS). Precisely, for a target distribution π, test function f with $|f| \leqslant 1$, error

tolerance $\epsilon > 0$, and probability $1 - \delta > 0$, we show how to choose the number of samples N and Markov kernel transitions t to ensure:

$$\Pr(|\pi f - \hat{f}| \leqslant \epsilon) \geqslant 1 - \delta$$

The choice of N and t depend explicitly on the sequence of SMC interpolating distributions, an upper bound on the weights, and the mixing times of the Markov kernels. This facilitates the full exploration of the SMC algorithm, demonstrating how algorithmic choices affect the performance of the estimator. Our approach differs from the traditional Feynman-Kac semi-group technique described in [18]. Instead, we focus on the marginal distributions of the particles, developing an inductive method for controlling the error of the particle approximation at each step of the algorithm. We demonstrate our bound in a variety of settings, proving new convergence results for SMC and comparing the results to MCMC. These examples include finite spaces, product distributions, and log-concave target distributions such as Bayesian logistic regression.

Chapter 3 extends these results, replacing the assumption of bounded weights with a bound on the L_2 norm between adjacent distributions, showing more precisely how the selection of interpolating distributions or *path* affects the complexity of the SMC estimator. This is the first finite-sample convergence result for SMC that does not require an upper bound on the importance sampling weights. We use this bound to construct SMC algorithms that have lower complexity than those for MCMC when the target distribution is Gaussian or log-concave, providing the first scenarios where SMC gives a provably faster algorithm. These results are then used to motivate the use of a well-known adaptive SMC algorithm that chooses a path using the relative effective sample size (RESS) [16, 25, 26]. We prove conditions under which this adaptive SMC algorithm provides a RAS and show empirically that it chooses nearly optimal sequences of distributions.

In the final chapter, we show how path selection can be used to provide low-variance estimates of ratios of normalizing constants using an integration technique known as path sampling [27, 28]. First, we show that for sequences of distributions index by a single parameter, an adaptive step size approach (similar to the technique in Chapter 3) can be used to automatically select a path with controlled variance and bias. Then, we develop a two stage approach suitable for multidimensional path selection. This approach is motivated by the connection between path sampling and Riemannian geometry. Our technique uses a small number of samples to tain a Gaussian process (GP) emulator, which is used to evalute the cost of potential paths. We then select a low-variance path from amongst a large set of paths specified by a directed, acyclic graph (DAG) using topological sorting. We validate these algorithms empirically on several examples, including the mean-field Ising model, the Gaussian example from [28], and a random effects model. In each case, the proposed approach provides almost theoretically-optimal performance.

2

Finite Sample Complexity of Sequential Monte Carlo Estimators

2.1 Introduction

Sequential Monte Carlo samplers (SMC) [3, 7] have recently received attention as an alternative to Markov chain Monte Carlo (MCMC) for Bayesian inference problems. Practitioners cite a variety of reasons for using SMC over MCMC. One reason is that it provides a natural estimate of the normalizing constant and may be the preferred method for estimating marginal likelihoods or Bayes factors [4, 25, 29]. SMC is also believed to outperform MCMC in parallel computing environments, and a variety of methods have been developed to facilitate its implementation on graphics processing units or clusters of computers [8, 10, 9, 15]. Finally, SMC may exhibit similar properties to tempering, making it well suited for difficult or multimodal problems [4, 3, 7]. While these properties could make SMC a competitive alternative to MCMC, they have rarely been verified theoretically.

The preponderance of SMC theory focuses on the asymptotic regime, where the number of particles approaches infinity. The existence of a central limit theorem for

the SMC estimator was established by Del Moral and Guionnet [17] and extended by Chopin [19]. A similar CLT was shown to hold for adaptive resampling methods by Douc and Moulines [30] and later by Beskos et al. [20]. Other asymptotic theory includes the work of Jasra et al. [11], who proved a bound on the asymptotic variance under local mixing assumptions. Beskos et al. [24], showed non-degeneracy of the particle approximation as the dimension increases for problems with product measures. Eberle and Marinelli [22, 23] developed asymptotic error bounds for the continuous time analogue of SMC. Finite sample results have been largely concerned with the L_p stability of SMC. This includes Whiteley [21], who developed L_p error bounds on non-compact spaces using drift and minorization conditions, and Schweizer [12] who demonstrated L_p stability for finite-sample SMC on compact spaces using global and local mixing conditions. While these finite sample results are useful for establishing general characteristics of SMC, they depend on expectations and norms of the associated Feynman-Kac measures, making them difficult to evaluate in practice.

In this chapter we develop finite sample bounds which enable the characterization of SMC as a randomized approximation scheme. Let π be a target measure on $\mathcal{X}$ and $f : \mathcal{X} \to \mathcal{R}$ a bounded measurable function. Our main result is to provide, for any error tolerance $\epsilon > 0$ and error probability $\delta \in (0, 1/4]$, a choice of the number of particles N and the number of Markov chain transitions t at each step of the algorithm to ensure

$$Pr(|\hat{\pi} f - \pi f| < \epsilon) \geqslant 1 - \delta$$

where πf denotes the expectation of f with respect to π and $\hat{\pi} f$ is the SMC estimator of πf. In contrast to other finite sample SMC bounds, we make explicit the dependence of N and t on an upper bound to the weights, an upper bound on the ratio of normalizing constants between adjacent interpolating distributions, the mixing times of the Markov kernels, and the specified ϵ and δ. The primary advantage of

such bounds is that they allow for the interrogation of the algorithm, identifying how changes in the sequence of distributions and Markov kernels affect the computational cost of the estimator. The bound provided here also facilitates explicit comparison with other methods such as MCMC, potentially identifying situations where one method may be preferred over another. Our approach differs from previous analyses by focusing on the marginal distribution of individual particles rather than following the Feynman-Kac semi-group approach popularized by Del Moral [18]. We use an inductive approach to controlling the error at each step of the algorithm, developing sufficient conditions for propagating forward accurate particle approximations with high probability.

The chapter is structured as follows. Section 2.2 introduces some notation and describes the general form of the SMC algorithm studied in this chapter and concludes with a statement of our main result. Section 2.3 presents the proof of our error bound, developing conditions for inductively controlling the error. Section 2.4 uses our bound to compare the performance of SMC with MCMC on sequences of distributions obtain via geometric mixtures with application to finite state spaces. This comparison highlights important differences between the algorithms and provides some guidance on how to select the interpolating distributions. Section 2.5 uses our bounds to explore the scaling of the SMC with dimension on product measures, and compares our results to those obtained previously in asymptotic and continuous time settings. This example also demonstrates the utility of our bounds in comparing SMC behavior under distinct choices of distribution sequences, showing that when the target is Gaussian with precision ϕ a careful choice of intermediate distributions can decrease the complexity from exponential in ϕ to logarithmic. Section 2.6 considers the case of log-concave target distributions and provides an application to Bayesian logistic regression. To the best of our knowledge this represents the first non-asymptotic SMC bound on a problem of direct interest to Bayesian statistical

practice.

2.2 Sequential Monte Carlo

Let π be a target probability measure on a space $\mathcal{X}$ with σ-algebra $\mathcal{B}$ and dominating measure $\lambda(dx)$. Consider a test function $f : \mathcal{X} \to \mathcal{R}$. Our goal is to quantify the finite sample error arising from estimating $\mathbb{E}_\pi f$ using sequential Monte Carlo. In this section, we introduce our probabilistic setting and the SMC algorithm studied in this chapter.

2.2.1 Notation

Let $\mathcal{P}$ be the set of probability measures on $\mathcal{X}$ that are absolutely continuous with respect to λ and $\mathcal{F}$ the set of measurable functions $f : \mathcal{X} \to \mathcal{R}$. Each measure acts on functions $f \in \mathcal{F}$ from the left by $\mu f = \int f(x)\mu(dx) = \mathbb{E}_\mu f$. We say that a measure $\nu \in \mathcal{P}$ is ω-warm with respect to μ if $\sup_{B\in\mathcal{B}}\ \nu(B) \leqslant \omega \cdot \mu(B)$ [31, 32]. Let $\mathcal{P}_\omega(\mu)$ be the set of all such measures.

Let $K : \mathcal{X} \times \mathcal{B} \to [0, 1]$ be an ergodic Markov kernel with limiting distribution μ. Markov kernels operate on functions from the left $Kf(x) = \int K(x, dy)f(y)$ and probability distributions from the right $\mu K(dx) = \int \mu(dy)K(y, dx)$. Define the mixing time of K by

$$\tau_K(\epsilon, \omega) = \min\ \left\{t : \sup_{\nu\in\mathcal{P}_\omega(\mu)} ||\nu K^t - \mu||_{\mathrm{TV}} \leqslant \epsilon\right\}$$

where $||\cdot||_{\mathrm{TV}}$ is the total variation distance. Note that this is a somewhat weaker notion of mixing time than commonly used. In particular, obtaining samples from μ in polynomial time requires not only that τ_K grows at most polynomially in $1/\epsilon$ and ω, but also the ability to draw an initial state from an ω-warm distribution. Part of our result will be to show that SMC with appropriately chosen parameters guarantees an ω-warm starting distribution.

2.2.2 Sequential Monte Carlo

In sequential Monte Carlo a collection of particles transition through a sequence of measures $\mu_0, ..., \mu_S \in \mathcal{P}$ where $\mu_S = \pi$. Denote the density of each intermediate measure by $q_s(x)/z_s$. The ratios $w_s(x) = q_s(x)/q_{s-1}(x)$ are assumed to be bounded so we have $\frac{q_s(x)/z_s}{q_{s-1}(x)/z_{s-1}} \leqslant W \cdot Z$, where W and Z are the maximum values of $\sup w_s(x)$ and z_{s-1}/z, respectively. This assumption requires that the tails of μ_{s-1} are suitably heavy relative to μ_s. In addition to the sequence of measures, we are given a collection of Markov transition kernels $K_1, ..., K_S$. Each kernel K_s is assumed to be irreducible, aperiodic, μ_s-reversible, and has mixing time $\tau_s(\epsilon, \omega)$.

In this chapter we consider the following sequential Monte Carlo algorithm. Initialize by drawing N independent samples $X_0^{1:N} = \big(X_0^1, ..., X_0^N\big)$ from μ_0. The realizations of these particles are denoted by $x_0^{1:N} = \big(x_0^0, ..., x_0^N\big)$. For $s = 1, ..., S$ perform the following:

1. Assign each particle an importance sampling weight equal to the unnormalized density ratio.

$$w_s\big(x_{s-1}^n\big) = \frac{q_s\big(x_{s-1}^n\big)}{q_{s-1}\big(x_{s-1}^n\big)}$$

2. Sample a new set of particles with replacement according to the weights (multinomial resampling).

$$Pr\Big(\tilde{X}_s^n = x \;\Big|\; X_{s-1}^{1:N} = x_{s-1}^{1:N}\Big) \propto \sum_{n=1}^{N} w_s\big(x_{s-1}^n\big) \cdot \delta_{x_{s-1}^n}(x)$$

3. Apply t steps of the kernel K_s to each re-sampled particle, producing $X_s^{1:N}$.

$$X_s^n \sim K^t\big(\tilde{X}_s^n, \cdot\big)$$

The final step of SMC produces empirical measure $\hat{\pi} = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_S^n}$ and corresponding estimate $\hat{\pi} f = \frac{1}{N}\sum_{n=1}^{N} f\left(x_S^n\right)$ of πf. Intuitively, the weighting step identifies particles in regions of high relative density, while the resampling step oversamples particles in underrepresented regions while removing particles with low weights, so that computation is not wasted on particles in low density areas. This comes at the cost of increased dependence among the particles, often referred to as particle degeneracy, and characterized by multiple particles sharing the same value immediately after resampling. The last step combats this degeneracy by evolving the resampled particles under the Markov kernel. This contracts the marginal distribution $\tilde{\mu}_s$ towards the desired distribution μ_s and reduces dependence between the particles.

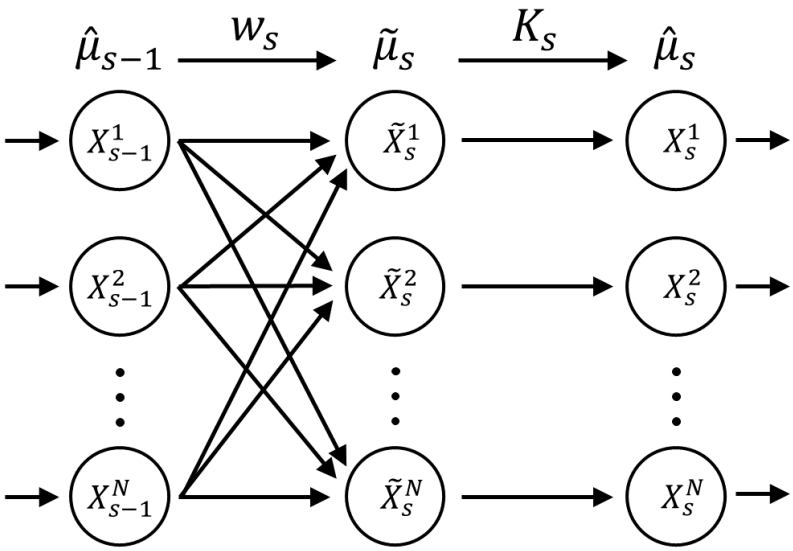

FIGURE 2.1: Marginal distributions and dependence structure of the particles.

Our approach to controlling the error of SMC depends on relating the marginal distributions of the particles to the prespecified interpolating distributions. At the beginning of each step, the particles $X_{s-1}^{1:N}$ are identically distributed according to $\hat{\mu}_{s-1}$. After resampling, the particles remain identically distributed and have marginal distribution $\tilde{X}_s^n \sim \tilde{\mu}_s$. Applying t steps of K_s to each particle changes

the marginal distribution to $\hat{\mu}_s = \tilde{\mu}_s K_s^t$. The dependence structure and marginal distributions of the particles are displayed visually in Figure 2.1. At each step of the algorithm, we show that the marginal distributions $\tilde{\mu}_s$ and $\hat{\mu}_s$ remain close to the desired distribution μ_s.

2.2.3 Main result

In the next section, we prove the following theorem which bounds the probability of error of the SMC estimator as a function of N and t. This allows us to establish SMC as a *randomized approximation scheme*, i.e. an algorithm which guarantees $|\hat{\pi}f - \pi f| < \epsilon$ with probability at least $1 - \delta$ (see e.g. [33]). It is standard to show this bound for $\delta \geqslant 3/4$; this can then be improved to $1 - \delta$ in $\mathcal{O}\big(\log(1/\delta)\big)$ time [34].

Theorem 1 (Total error bound for SMC)**.**
Fix $\epsilon > 0$ and assume $X_0^{1:N}$ are sampled independently from μ_0. Let

1. $N \geqslant 2\log\big(16S\big) \cdot \max\big\{9W^2Z^2, \frac{1}{\epsilon^2}\big\}$
2. $t \geqslant \max_s\ \tau_s\Big(\frac{1}{8NS},\ 2\Big)$

Then for any $f \in \mathcal{F}$ with $|f| \leqslant 1$,

$$\big|\hat{\pi}f - \pi f\big| \leqslant \epsilon.$$

with probability at least $3/4$.

We also present a second error bound that may be tighter in some settings. This bound can be applied when $\mathcal{X}$ is either a finite space or R^d and a lower bound on the spectral gap ρ_s of K_s is available. This bound uses a number of Markov kernel transitions t_S at the final step which may be larger than the number required at intermediate steps, in order to obtain the desired final accuracy.

Theorem 2 (Alternative bound for SMC)**.**

Assume that $\mathcal{X}$ is either a finite space or R^d and that $X_0^{1:N}$ are sampled independently from μ_0. If $\mathcal{X} = \mathcal{R}^d$ further assume that q_π is a continuous distribution. Let $\rho = \min_s \rho_s$ and choose:

1. $N \geqslant 8 \log\left(8S + 8\right) \cdot \max\left\{9W^2Z^2, \frac{1}{\epsilon^2}\right\}$
2. $t \geqslant \frac{4+\log\left(W^2Z^2\right)}{\rho}$
3. $t_S \geqslant \frac{3+\log\left(\frac{WZ}{\epsilon}\right)}{\rho}$

Then with probability at least $3/4$ we have

$$||\hat{\mu}_S - \pi||_{1,\pi} \leqslant \epsilon$$

and consequently for any $f \in \mathcal{F}$ with $|f| \leqslant 1$,

$$\left|\hat{\pi} f - \pi f\right| \leqslant \epsilon.$$

The primary advantage of Theorem 2 is that the choice of t does not depend on S, though it requires a more restrictive setting. The proof of Theorem 1 is presented in Section 2.3, which first develops two key lemmas before presenting the proof. The proof of Theorem 2 is similar and delayed to Section 2.8.

2.3 Error bounds

Our proof uses an inductive approach to bounding the error of SMC, bounding the one-step error conditional on the bound holding at the previous step. Define $\mathbf{C}_s$ to be the event that the following conditions hold:

$$\begin{aligned} &\mathbf{C}_s(\text{i}) \quad X_s^n \sim \mu_s \text{ for } n = 1, \ldots, N \\ &\mathbf{C}_s(\text{ii}) \quad \bar{w}_{s+1} \geqslant \mu_s w_{s+1} \cdot \frac{2}{3} \end{aligned} \tag{2.1}$$

where $\bar{w}_{s+1} = N^{-1} \sum_{n=1}^{N} w_{s+1}\big(x_s^n\big)$ is the average sample weight at the beginning of step s. Condition (i) requires the marginal distribution of each particle at the beginning of the step be the chosen interpolating distribution. Condition (ii) ensures that the average weight does not significantly underestimate $\mu_s w_{s+1} = z_{s+1}/z_s$. We will show that, given these conditions, the marginal distribution $\tilde{\mu}_s$ of the resampled particles is 2-warm with respect to μ_{s+1}. This in turn is sufficient to ensure the event $\mathbf{C}_{s+1}$ with high probability by choosing N and t as functions of τ_s, W, and Z. Applying these conditions inductively enables us to establish $\mathbf{C}_{S-1}$ with high probability, and therefore to bound the error of the final particle approximation $\hat{\pi} f$ with high probability.

First, we show that $\mathbf{C}_s$(i) holds, assuming $\mathbf{C}_{s-1}$ is true. We begin with particles $X_{s-1}^1, \ldots, X_{s-1}^n$ each marginally distributed according to μ_{s-1}. Each particle is weighted according to the density ratio w_s and new particles are drawn using multinomial resampling. The resulting particles $\tilde{X}_s^1, \ldots, \tilde{X}_s^n$ are identically distributed with marginal distribution $\tilde{\mu}_s$. The following lemma proves that $\tilde{\mu}_s \,|\, \mathbf{C}_{s-1}$ is 2-warm with respect to μ_s.

Lemma 3 (Error of the resampling distribution)**.**
Assume $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geqslant 3/4$. *Then* $\tilde{\mu}_s | \boldsymbol{C}_{s-1} \in \mathcal{P}_2(\mu_s)$.

Proof. Let $E_{m,n}$ be the event that particle $\tilde{X}_s^n$ inherits from particle X_{s-1}^m. Fix a set $B \in \mathcal{B}$. Then

$$\Pr\Big(\tilde{X}_s^n \in B \;\Big|\; \mathbf{C}_{s-1}\Big) = \sum_{n=1}^{N} \Pr\Big(X_{s-1}^m \in B,\, E_{m,n} \;\Big|\; \mathbf{C}_{s-1}\Big) \tag{2.2}$$

Each term of the right hand side is identical and can be bounded above:

$$\begin{aligned}
\Pr\Big(X_{s-1}^n \in B,\, E_{m,n} \;\Big|\; \mathbf{C}_{s-1}\Big) &= \int_B \Pr\Big(E_{m,n} \;\Big|\; X_{s-1}^n = x_{s-1}^m,\, \mathbf{C}_{s-1}\Big)\cdot \\
&\qquad \hat{\mu}_{s-1}\big(dx_{s-1}^m \,\big|\, \mathbf{C}_{s-1}\big) \\
&\leqslant \int_B \frac{w_s\big(x_{s-1}^m\big)}{N\cdot \mu_{s-1}w_s\cdot \frac{2}{3}} \cdot \hat{\mu}_{s-1}\big(dx_{s-1}^m \,\big|\, \mathbf{C}_{s-1}\big) \qquad (2.3)\\
&\leqslant \frac{3}{2N}\int_B \frac{\mu_s\big(dx_{s-1}^m\big)}{\mu_{s-1}\big(dx_{s-1}^m\big)} \cdot \frac{\mu_{s-1}\big(dx_{s-1}^m\big)}{\Pr\big(\mathbf{C}_{s-1}(\text{ii})\big)} \\
&\leqslant \frac{2}{N}\cdot \mu_s(B)
\end{aligned}$$

The second line follows from $\mathbf{C}_{s-1}$(ii) and the third line follows from $\mathbf{C}_{s-1}$(i) and Bayes' rule. Combining equations (2.2) and (2.3) proves the result. ■

For the remainder of this section we omit the dependence of $\tilde{\mu}_s$ on the event $\mathbf{C}_{s-1}$ for readability. After resampling, t steps of K_s are applied to each sample to obtain $X_s^1, ..., X_s^N$. The marginal distribution of each particle becomes $\hat{\mu}_s = \tilde{\mu}_s K_s^t$. The following corollary shows how to choose t to ensure that $\hat{\mu}_s = \mu_s$ with high probability (conditional on $\mathbf{C}_{s-1}$).

Corollary 4 (Correctness of the predictive distribution)**.**
Assume $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geqslant 3/4$. *Then for any* $0 < \delta_0 < 1$ *and* $t \geqslant \tau_s\big(\frac{\delta_0}{2N}, 2\big)$

$$Pr\big(\boldsymbol{C}_s(i) \;\big|\; \boldsymbol{C}_{s-1}\big) \geqslant 1 - \delta_0/2$$

Proof. By the choice of t and Lemma 3:

$$||\hat{\mu}_s - \mu_s||_{\text{TV}} \leqslant \frac{\delta_0}{2N}$$

Applying the standard coupling argument "divine intervention" (see for example Lemma 4.2 of [35]) gives $\Pr\big(X_s^m \sim \mu_s\big) \geqslant 1 - \frac{\delta_0}{2N}$. Using a union bound over all the particles gives the result. ■

Having shown how to ensure condition (i) of $\mathbf{C}_s$ holds with high probability, we turn our attention to condition (ii). First, we show that the average weight $\bar{w}_{s+1}$ concentrates around it's mean.

Lemma 5 (Lower bound on the average weights)**.**
Fix any $0 < \delta_0 < 1$ *and choose* $N \geqslant 18\log(4/\delta_0) \cdot W^2 Z^2$*. Then*

$$Pr\Big(|\bar{w}_{s+1} - \hat{\mu}_s w_{s+1}| \leqslant \mu_s w_{s+1}/3\Big) \geqslant 1 - \delta_0/2$$

Proof. Let $\mathcal{W}_{s+1}^n = \sum_{m=1}^n \Big(w_{s+1}\big(X_s^m\big) - \hat{\mu}_s[w_{s+1}]\Big)$ be the partial sum of residuals and $X_s^{1:n} = X_s^1, ..., X_s^n$ denote the first n particles. Azuma's inequality requires that $\mathcal{W}_{s+1}^n$ is a martingale with bounded increments. The increments are bounded by W and it is straightforward to show that $\mathcal{W}_{s+1}^n$ is a martingale:

$$\begin{aligned}
E\Big(\mathcal{W}_{s+1}^{n+1}\Big|\mathcal{W}_{s+1}^n\Big) &= E\Big(\mathcal{W}_{s+1}^{n+1}\Big|X_s^{1:n}\Big) \\
&= \mathcal{W}_{s+1}^n + E\Big(w_{s+1}\big(X_s^{n+1}\big)\Big|X_s^{1:n}\Big) - \hat{\mu}_s[w_{s+1}] \\
&= \mathcal{W}_{s+1}^n + E\bigg(E\Big(w_{s+1}\big(X_s^{n+1}\big)\Big|X_s^{1:n},\ \tilde{X}_s^{n+1}\Big)\bigg) - \hat{\mu}_s[w_{s+1}] \qquad (2.4) \\
&= \mathcal{W}_{s+1}^n + E\bigg(E\Big(w_{s+1}\big(X_s^{n+1}\big)\Big|\tilde{X}_s^{n+1}\Big)\bigg) - \hat{\mu}_s[w_{s+1}] \\
&= \mathcal{W}_{s+1}^n
\end{aligned}$$

With these conditions verified, Azuma's inequality yields the result. ■

We now combine Corollary 4 and Lemma 5 to give the one step induction condition.

Corollary 6 (One step induction condition)**.**
Assume $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geqslant 3/4$*. Fix* $0 < \delta_0 < 1$*. Choose* $N \geqslant 18\log(4/\delta_0) \cdot W^2 Z^2$ *and* $t \geqslant \tau_s\big(\frac{\delta_0}{2N}, 2\big)$*. Then the inductive condition* $\boldsymbol{C}_s|\boldsymbol{C}_{s-1}$ *holds with probability at least* $1 - \delta_0$

Proof. Observe that Lemma 5 implies $\mathbf{C}_s(\text{ii})$ conditional on $\mathbf{C}_s(\text{i})$. Then by Lemma 5 and Corollary 4

$$\begin{aligned} \Pr\big(\mathbf{C}_s \,|\, \mathbf{C}_{s-1}\big) &= \Pr\big(\mathbf{C}_s(\text{ii}) \,|\, \mathbf{C}_{s-1}(\text{i})\big) \cdot \Pr\big(\mathbf{C}_s(\text{i}) \,|\, \mathbf{C}_{s-1}\big) \\ &\geqslant 1 - \delta_0 \end{aligned} \tag{2.5}$$

■

To finish the proof we apply the one step iteration condition and obtain conditions for controlling the bounds on the error of the full SMC algorithm. The error of the final estimator can be controlled using ideas similar to Lemma 5 supposing that $\mathbf{C}_{S-1}$ holds. We then show that N and t can be chosen such that the event $\mathbf{C}_{S-1}$ holds with high probability. The proof of Theorem 1 follows.

Proof of Theorem 1 *Proof.* Fix $\delta_0 = \frac{1}{4S}$. Then

$$\begin{aligned} \Pr\big(|\hat{\pi}f - \pi f| \leqslant \epsilon\big) &\geqslant \Pr(|\hat{\pi}f - \pi f| \leqslant \epsilon \;\big|\; \mathbf{C}_S(\text{i})\big) \cdot \Pr\big(\mathbf{C}_S(\text{i})\big) \\ &\geqslant \Pr(|\hat{\pi}f - \pi f| \leqslant \epsilon \;\big|\; \mathbf{C}_S(\text{i})\big) \cdot \Pr\big(\mathbf{C}_S(\text{i}) \,\big|\, \mathbf{C}_{S-1}\big) \cdot \\ &\quad \prod_{s=1}^{S-1} \Pr\big(\mathbf{C}_s \,\big|\, \mathbf{C}_{s-1}\big) \cdot \Pr\big(\mathbf{C}_0(\text{ii})\big) \\ &\geqslant (1 - \delta_0)^S \\ &\geqslant 1 - \frac{1}{4} \end{aligned} \tag{2.6}$$

The third line follows from the induction condition of Corollary 6. The error probability $\Pr\big(|\hat{\pi}f - \pi f| \leqslant \epsilon \;\big|\; \mathbf{C}_S(\text{i})\big)$ is controlled in the same manner as Lemma 5. Condition (ii) of event $\mathbf{C}_0$ can be shown to hold with probability at least $1 - \delta_0/2$ via Höeffding's inequality. ■

This proves the main result. The requirement that $X_0^{1:N}$ are identically distributed according to μ_0 can be relaxed as long as $\mathbf{C}_0$ holds with high probability. This could be the case, for example, when the initial particles are drawn using a

rapidly mixing Markov chain. In the following sections we use this bound to compare SMC to MCMC in a variety of settings.

2.4 SMC with geometric mixtures

Geometric mixtures are a common and straightforward way of specifying a sequence of SMC distributions. Consider the problem of sampling from π with density $q_\pi(x)/z_\pi$ known up to z_π. Suppose we can draw independent samples from an initial distribution ν with density $p_\nu(x) = q_\nu(x)/z_\nu$. In Bayesian inference, ν is often chosen to be the prior distribution for the posterior π of interest [15, 3, 25]. When $\mathcal{X}$ is finite or compact ν may be chosen to be uniform on $\mathcal{X}$. If the uniform distribution is improper or is not easy to sample from, ν may be chosen to be a tempered version of π which is accessible via MCMC. Choosing the initial distribution to be either uniform or tempered is analogous to simulated annealing, starting from a relatively diffuse distribution and moving towards a more concentrated distribution of interest. Define the geometric mixture distribution μ_β with mixture term $\beta \in [0, 1]$ by the density

$$q_\beta(x) = q_\pi(x)^\beta \cdot q_\nu(x)^{(1-\beta)}/z_\beta. \tag{2.7}$$

The parameter β controls the rate at which our initial distribution is changed to the distribution of interest. In the case of tempering, β is called the inverse temperature. For simplicity re-scale the density ratio $q_\pi(x)/q_\nu(x)$ by it's supremum so that $w_s(x) \leqslant 1$. Define $\Gamma = z_\nu/z_\pi$. We will assume that for each $\beta \in (0, 1]$ we can construct an ergodic Markov kernel K_β with spectral gap ρ_β.

We consider the computational complexity of SMC using geometric mixtures, measured in terms of the number of total Markov kernel transitions SNt required to obtain a (δ, ϵ) randomized approximation scheme. This serves as a measure of overall computational complexity as the Markov kernel transitions tend to domi-

nate the computational cost of the SMC algorithm. If parallel computing resources are available, the performance of SMC may be improved by a constant factor via parallelization, however the overall complexity of the bounds does not change.

2.4.1 Finite sample bounds for SMC

To specify the SMC algorithm, we need to choose a sequence of inverse temperatures $\beta_0, ..., \beta_S$. We choose $S = \lceil \log \Gamma \rceil$ and $\beta_s = s/S$. While Γ is often unknown, in the examples that follow we show that a bound on Γ is sufficient to apply the following corollary. Using this sequence of distributions we can apply Theorem 1.

Corollary 7 (Complexity of SMC with geometric mixtures)**.**
Let $Z = \max_{s=1,...,S} z_{s-1}/z_s$, $\rho = \min_s \rho_{\beta_s}$, *and fix* $\epsilon > 0$. *Then for any* $f \in \mathcal{F}$ *with* $|f| \leqslant 1$, *the number of Markov kernel transitions required to ensure* $|\hat{\pi} f - \pi f| \leqslant \epsilon$ *with probability at least* $3/4$ *is bounded above by*

$$\mathcal{O}^* \left(\frac{1}{\rho} \cdot \log \Gamma \cdot \frac{1}{\epsilon^2 \wedge Z^{-2}} \cdot \log^2 \log \Gamma \right)$$

The notation $\mathcal{O}^*$ indicates that lower order terms ($\log \log \log \Gamma$, $\log Z$ and $\log 1/\epsilon$) have been omitted for readability. The $\mathcal{O}\big(\frac{1}{\rho} \cdot \log \Gamma\big)$ term is the number of Markov chain transitions required to ensure that particles have the correct marginal distribution after the final step. The $\mathcal{O}\big(\epsilon^{-2} \vee Z^2\big)$ term includes both the number of samples from this distribution required to estimate πf with sufficient accuracy, and the number required at intermediate steps to ensure the one step iteration condition. The final $\mathcal{O}\big(\log^2 \log \Gamma\big)$ term is the additional factor required to ensure that the iteration conditions hold simultaneously across all steps of the algorithm. In settings where Theorem 2 can be applied, this term is reduced to $\mathcal{O}\big(\log \log \Gamma\big)$.

The quantity $\epsilon^2 \wedge Z^{-2}$ plays an important role in the way that our bounds characterize the error of SMC. When the desired accuracy is undemanding (ϵ is large),

the number of particles required to approximate z_s/z_{s-1} (and therefore μ_s after resampling) with small relative error remains bounded by $O(Z^2)$. Thus our bounds do not decay monotonically with N; we believe that in practice this behaviour manifests as particle degeneracy. One way to mitigate this effect is to choose a large number of steps S, ensuring that Z is is $O(\epsilon^{-2})$. This is in accordance with SMC folklore, which suggests large numbers of steps with smaller numbers of particles are preferable. Unfortunately, using this method for ensuring Z sufficiently large comes at the cost of a factor of $\log(S)$ in our bounds.

2.4.2 Comparison of SMC and MCMC

We compare the bound for SMC with a similar bound for MCMC. We assume that the MCMC approximation is created by drawing N independent samples according to ν and applying t' transitions of K_1 to each sample. We write $\bar{\pi}$ to denote the analogous empirical measure constructed from the resulting samples.

Corollary 8 (Complexity of MCMC with independent chains)**.**
Fix $\epsilon > 0$. Then for any function $f \in \mathcal{F}$ with $|f| \leqslant 1$, the number of Markov kernel transitions required to ensure $|\bar{\pi}f - \pi f| \leqslant \epsilon$ with probability at least $3/4$ is bounded above by

$$\mathcal{O}^*\left(\frac{1}{\rho_1} \cdot \log \Gamma \cdot \frac{1}{\epsilon^2}\right)$$

Proof. By assumption $\nu \in \mathcal{P}_\Gamma(\pi)$, so choosing $t' = \mathcal{O}\left(\frac{\log(\Gamma/\epsilon)}{\rho_1}\right)$ ensures that $||\nu K^{t'} - \pi||_{\mathrm{TV}} \leqslant \epsilon/2$ and therefore $|\nu K^{t'} f - \pi f| \leqslant \epsilon/2$. Choosing $N = \mathcal{O}\left(\epsilon^{-2}\right)$ ensures that $|\bar{\pi}f - \nu K^{t'} f| \leqslant \epsilon/2$ with probability at least $3/4$ by Höeffding's inequality. The result follows from the triangle inequality. ■

An alternative bound can be obtained by running a single Markov chain to near stationarity, then taking samples every $\mathcal{O}(\rho^{-1})$ transitions to obtain a sequence of

nearly independent particles [36, 37]. This does not change the overall complexity of the method in terms of ρ_1 and Γ, as it is dominated by the mixing time.

For simplicity we assume $\epsilon \in (0, Z^{-1}/3]$ when comparing the bounds presented in Corollaries 7 and 8, though as noted above the SMC bound will not decrease for larger ϵ. We see that the bound for SMC requires an additional factor of $\mathcal{O}\big(\log^2 \log \Gamma\big)$ to ensure the induction condition at each step. Note also that the complexity of MCMC depends only on ρ_1 rather than ρ. If an intermediate Markov kernel mixes substantially slower than K_1, more transitions will be required to achieve comparable accuracy. On the other hand, if ρ_1 is much smaller than ρ_{β_s} for $s < S$, the SMC bound may be unnecessarily pessimistic and SMC may outperform MCMC practice. Finally, while both bounds depend on $\Gamma = z_\pi / z_\nu$, the SMC bound also depends on the smallest ratio z_{s-1}/z_s between any pair of neighboring distributions. When the upper bound Z is close to 1, the complexities differ only by a factor of $\log^2 \log \Gamma$, but if one ratio z_{s-1}/z_s is much smaller than 1, a large increase in N is required to control the error at that step. Choosing the β's so that the ratios are close to one and approximately equal provides the smallest upper bound. This agrees with heuristics for the selection of inverse temperatures found in the simulated tempering literature [38], which aim to space distributions so that the ratios of normalizing constants between adjacent distributions are approximately constant across temperature. If small, evenly sized z_{s-1}/z_s are difficult to ensure *a priori*, S may be chosen to be large to increase $\min_s \Gamma_s$, at the cost of a logarithmic increase in complexity.

2.4.3 Comparison with importance sampling

It is instructive to use our bound to to demonstrate the advantages of SMC over standard importance sampling. When Γ is unknown, the importance sampling estimator is $\sum_{n=1}^N \frac{q_\pi(x_n)}{q_\nu(x_n)} f(x_n) / \sum_{n=1}^N \frac{q_\pi(x_n)}{q_\nu(x_n)}$. To ensure that the absolute error of the estimator is less than ϵ, both the numerator $\frac{1}{N} \sum_{n=1}^N \frac{q_\pi(x_n)}{q_\nu(x_n)} f(x_n)$ and it's normaliza-

tion $\frac{1}{N}\sum_{n=1}^{N}\frac{q_\pi(x_n)}{q_\nu(x_n)}$ need to be accurately estimated. The numerator is relatively easy to estimate and requires $\mathcal{O}\big(1/\epsilon^2\big)$ samples. On the other hand, the normalization needs to have small error relative to Γ^{-1} which requires $\mathcal{O}\big(\Gamma^2/\epsilon^2\big)$ samples (Hoeffding), giving an overall complexity of $\mathcal{O}\big(\Gamma^2/\epsilon^2\big)$. The complexity remains the same in the case where Γ is known. So while the complexity of importance sampling is quadratic in Γ, SMC decreases this dependence to logarithmic, at the cost of a factor of $\mathcal{O}\big(1/\rho\big)$. For many problems of interest Γ may be exponentially large and SMC can be expected to substantially outperform importance sampling.

2.4.4 Example: finite spaces

Let $\mathcal{X}$ be a finite space with $\pi(x)\propto q(x)$ and $0 < q(x) \leqslant 1$. Let $\pi_0 = \min \pi(x)$ and let $x_0 = \arg\min \pi(x)$ be a state at which this is attained. Let initial distribution $\nu(x) = \mathbb{1}_{x=x_0}$ assign mass one to x_0, yielding bound $\Gamma \leqslant \frac{1}{\pi_0}$. The complexity of Markov chain Monte Carlo estimator is bounded above by

$$\mathcal{O}^*\left(\frac{1}{\rho_1}\cdot\log\left(\frac{1}{\pi_0}\right)\cdot\frac{1}{\epsilon^2}\right)$$

A comparable bound can be obtained for SMC using our results. Let $\mu_0\propto\pi^{\beta_0}$ with $\beta_0 = 1/\log\frac{1}{\pi_0}$; samples from μ_0 can be drawn in $\mathcal{O}\big(\frac{1}{\rho}\big)$ time using independent Markov chains beginning at x_0. Set $S = \log\frac{1}{\pi_0} - 1$ and choose $\mu_s\propto q(x)^{\beta_s}$ with $\beta_s = \frac{s+1}{\log\frac{1}{\pi_0}}$ giving $Z \leqslant e$. Applying Theorem 2, the complexity of SMC is bounded above by

$$\mathcal{O}^*\left(\frac{1}{\rho_1}\cdot\log\left(\frac{1}{\pi_0}\right)\cdot\frac{1}{\epsilon^2}\cdot\log\log\left(\frac{1}{\pi_0}\right)\right)$$

On many problems the increase in complexity of $\log\log\frac{1}{\pi_0}$ may be negligible and can be offset by SMC's advantages such as parallelization.

2.5 SMC on product measures

Product measures have previously been used to assess the dimension dependence of SMC [24, 12, 22]. Consider initial distribution ν and target distribution π, with corresponding weight $w(x) = q_\pi(x)/q_\nu(x) \in (0, 1]$, ratio of normalizing constants $\Gamma = z_\nu/z_\pi$, and let K be a π-reversible, geometrically ergodic Markov kernel with spectral gap ρ. Define π_d and ν_d to be independent product measures on $\mathcal{X}^d$ with weight $w_d = \prod_{i=1}^d \frac{q_\pi(x_i)}{q_\nu(x_i)}$, and define the product kernel $K_d = \prod_{i=1}^d K(x_i, dx_i)$; the spectral gap of K_d is independent of the dimension for a product kernel [22, 12], though the computational cost of each kernel transition increases linearly in d. Using a sequence of geometric mixtures and choosing $S = \mathcal{O}(d)$, we can find a sequence of β_s so that $\frac{z_{s-1}}{z_s} = \mathcal{O}(\Gamma)$. Note that in practice, choosing the inverse temperatures in order to ensure this condition may be hard to achieve, so this is really an idealized SMC algorithm. However, the difficulty in specifying the inverse temperatures is addressed via a similar assumption in [22] and [12] and dealt with via asymptotics in [24], making it a relevant assumption for the purposes of comparison. Applying Theorem 1 gives a bound on the computational complexity:

$$\mathcal{O}(d^2 \log^2 d)$$

Under the additional assumptions required by Theorem 2 we obtain:

$$\mathcal{O}(d^2 \log d)$$

This improves upon the $\mathcal{O}(d^3)$ finite sample results of Schweizer [12] and Eberle and Marinelli [22], though it falls short of the $\mathcal{O}(d^2)$ rate obtained by Beskos et al. [24] for the situation of infinite particles and dimensions.

In the next section we apply these bounds to a Gaussian measure where we can explicitly specify the inverse temperatures. This allows us to investigate the effect of inverse temperature selection on the computational complexity.

2.5.1 Example: spherical Gaussian in d-dimensions

Let π be d-dimensional spherical Gaussian centered at the origin with precision $\phi > 1$ and unnormalized pdf $q(x|\phi) = \exp\big(-\frac{1}{2}\phi \cdot x^T x\big)$ for $x \in \mathcal{R}^d$. As many posterior distributions arising from Bayesian analyses are well approximated by normal distributions as the number of observations grows, this may also lend some insight into the performance of SMC more generally.

Let ν be the d-dimensional standard normal distribution and construct interpolating distributions using geometric mixtures with $S = d$ and $\beta_s = s/d$. Then μ_s is also spherical normal, characterized by precision $\phi_s = 1 + \frac{s}{d}(\phi - 1)$. This choice yields $z_0/z_1 = \big(1 + \frac{\phi-1}{d}\big)^{d/2} \geqslant \exp\big(\frac{\phi-1}{2}\big)$ with $Z \approx \exp\big(\frac{\phi-1}{2}\big)$ for d large. Assuming d sufficiently large, the overall complexity of SMC is then bounded above by

$$\mathcal{O}^*\Big(d^2\phi \cdot \max\Big\{\exp(\phi), \frac{1}{\epsilon^2}\Big\} \log d\Big)$$

omitting terms of order ϕ and $\log \epsilon$. Note that while the complexity in d remains $\mathcal{O}^*(d^2 \log d)$, there is an exponential dependence on ϕ. This comes from the first step of the algorithm, where the initial distribution is very flat relative to the first interpolating distribution and z_1/z_0 becomes exponentially small, requiring many samples to estimate with low relative error. A better temperature ladder would ensure that μ_1 is not too peaked relative to μ_0, and then aim for similar spacing at subsequent steps. In fact, a polynomial dependence on ϕ can be made obtained by choosing a log-linear spacing on the precision. Choosing the same number of intermediate distributions S but taking $\phi_s = \phi^{\frac{s}{d}}$ gives $Z \leqslant \sqrt{\phi}$ and yields an SMC bound of

$$\mathcal{O}^*\left(d^2 \log \phi \cdot \max\Big\{\phi, \frac{1}{\epsilon^2}\Big\} \log d\right)$$

a dramatic improvement from the linear spacing. This demonstrates the importance of the choice of interpolating distributions. In fact, the dependence on ϕ can be

further reduced to logarithmic by choosing $S = d\lceil \log \phi \rceil$ and $\beta_s = \exp(s/d) \wedge 1$. Under this choice $Z \leqslant e^{1/2}$ and the complexity is bounded above by

$$\mathcal{O}^*\left(d^2 \cdot \frac{\log \phi \cdot \log d}{\epsilon^2}\right)$$

Our finite sample bounds allow us to see how choosing better sequences of inverse temperatures leads to dramatic improvements in our bounds. This example has important implications for Bayesian inference problems. The situation where ϕ is large is analogous to posterior distributions that that are highly concentrated, suggesting that proper selection of the temperature ladder may be crucial for achieving reasonable performance for large data sets.

2.6 Log-concave distributions

Log-concave target distributions are of interest in many settings. Log-concave sampling problems in statistics include Bayesian analysis of regression and logistic regression problems with priors corresponding to convex penalties, such as the Bayesian ridge or LASSO priors. In this section we apply our bounds to these log-concave problems, utilizing key results from Dwivedi et al. [39].

Let $\pi(x) \propto q(x)$ be a distribution on $\mathcal{R}^d$. We say that q is strongly log-concave if $q^{1-\alpha}(x) \cdot q^{\alpha}(y) < q\big(\alpha x + (1-\alpha)y\big)$ for $x, y \in \mathcal{R}^d$ and $\alpha \in (0,1)$. To be able to use the results of [39], we will also assume that $\log q$ is L-smooth and m-strongly concave, i.e. that

$$-\frac{L}{2}||x-y||_2^2 \leqslant \log \frac{q(x)}{q(y)} - \nabla \log q(x)^T (x-y) \leqslant -\frac{m}{2}||x-y||_2^2$$

for all $x, y \in \mathcal{R}^d$. For x^* the mode of π, this implies that $-L||x-x^*||_2^2 \leqslant 2 \log q(x) \leqslant -m||x-x^*||_2^2$. Let $\kappa = L/k$ denote the condition number of $\log q(x)$. Intuitively, κ is a measure of the curvature of the distribution q and is large when one dimension has a large range relative to the others.

For our analysis, we assume $\epsilon > 2e^{-d}$ to simplify the presentation and we restrict $\mathcal{R}^d$ to a ball B of radius $4\sqrt{d/m}$ centered at x^*; a similar restriction is made in [40]. This restriction ensures that the ratio of normalizing constants is bounded in the first step; since $\pi(B) \leqslant 1 - \epsilon/2$ this assumption has minimal impact on the results of our analysis [39]. We choose $\mu_0 = N(x^*, 1/L)$ and use a tempered sequence of interpolating distributions. Choosing $S = \lceil d\kappa \rceil$ and $\beta_s = s/S$ gives $W = 1$ and $Z = \mathcal{O}(1)$.

We choose the Metropolis-adjusted Langevin algorithm (MALA) Markov kernel, which gives the smallest upper bound. Slightly larger bounds are immediately available for other kernels, e.g the *ball walk* and the *hit-and-run* walk [41]. Dwivedi et al. [39] show that the mixing time of MALA on log-concave problems is $\mathcal{O}\Big(d\kappa \cdot \log \frac{2\omega}{\epsilon} \cdot \max\big\{1, \sqrt{\kappa/d}\big\}\Big)$, when starting from an ω-warm initial distribution. Tempering q does not change the condition number, so the mixing time of K_s is the same for all s. Plugging this mixing time into our SMC bounds gives a complexity of

$$\mathcal{O}^*\big(d^2\kappa^2 \cdot \log^2 d\kappa \cdot \max\big\{1, \sqrt{\kappa/d}\big\}\big)$$

This is larger than the $\mathcal{O}^*\big(d^2\kappa \cdot \log \kappa \cdot \max\big\{1, \sqrt{\frac{\kappa}{d}\log \kappa}\big\}\big)$ obtained for MALA by [39]. Besides the $\log^2 d\kappa$ term, which is the penalty our bound pays to control the worst case error across each step, the SMC bound grows quadratically in κ whereas the MCMC bound grows as $\kappa \log \kappa$. This increased complexity comes from the difficulty in constructing an optimal path for SMC: since the ratio z_ν/z_π is bounded above by $\kappa^{d/2}$, there exists a path of length $d \log \kappa$ which ensures $Z \leqslant e^{\frac{1}{2}}$. Such a path would reduce the dependence on κ from κ^2 to $\kappa \log \kappa$ and eliminate this difference in the bounds. However, constructing an optimal path is non-trivial, therefore we present our bounds for the easily-constructed path above.

2.6.1 Example: Bayesian logistic regression

Consider fitting a logistic regression model to a binary observation vector $Y \in \{0,1\}^n$ and associated matrix of covariates $X \in \mathcal{R}^{n \times p}$, via Bayesian inference. The corresponding likelihood is given by:

$$p(Y|X,\beta) \propto \exp\left(Y^T X\beta - \sum_{i=1}^{n} \log\left(1 + e^{X_i^T \beta}\right)\right)$$

Assign prior $p_0(\beta) = N\left(0, \frac{\alpha}{n}(X^T X)^{-1}\right)$ with the parameter α controlling the strength of the prior shrinkage toward zero. The resulting posterior distribution $q(\beta) \propto p_0(\beta) \cdot p(Y \mid X, \beta)$ is log-concave and satisfies the above assumptions of L-smoothness and m-strong concavity with $L \leqslant (n/4 + \alpha) \cdot \sigma_{\max}$ and $m \geqslant \alpha \cdot \sigma_{\min}$ for $\sigma_{\max}$ and $\sigma_{\min}$ the largest and smallest eigenvalues of $(X^T X)^{-1}/n$, respectively [39]. Inserting into our bounds gives an upper bound on the complexity of sampling via SMC:

$$\mathcal{O}^*\left(\left(\frac{dn}{\alpha} \cdot \frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \cdot \log^2\left(\frac{dn}{\alpha} \cdot \frac{\sigma_{\max}}{\sigma_{\min}}\right) \cdot \max\left\{1, \sqrt{\frac{n}{d\alpha} \cdot \frac{\sigma_{\max}}{\sigma_{\min}}}\right\}\right)$$

This example demonstrates the utility of our approach for practical problems: we are unaware of any previous finite-sample error bounds for non-trivial problems in Bayesian statistics using SMC. The dependence of the bound on $\sigma_{\max}/\sigma_{\min}$ can be removed be improving the condition number via pre-conditioning (see [42]).

2.7 Conclusion

The finite-sample bounds on SMC error provided here enable rigorous analysis of the computational complexity of SMC sampling algorithms on static spaces. As we have demonstrated, this allows for interesting comparisons between the efficiency of various SMC sampling algorithms, including the crucial dependence on the choice of interpolating distributions. However, significant areas remain for potential improvement of these bounds and extensions in future work.

The SMC bounds presented in sections 2.4, 2.5, and 2.6 suffer additional logarithmic complexity in Γ, d, and $\log d\kappa$ respectively in comparison to MCMC. This arises from the requirement that the worst-case error is controlled across all steps (ensuring $\mathbf{C}_s$ for all s). It has been suggested to us that it may be possible to remove this through use of Talagrand's generic chaining method, and we are exploring this approach. Similar techniques could also be used to improve on the union bound used in Corollary 4, making the bound in Theorem 1 closer to Theorem 2.

Another area of interest is target distributions exhibiting multimodality, where Markov kernels may have good local mixing behaviour, yet exhibit poor mixing globally (e.g. [43, 44]). Sequential Monte Carlo has been observed to perform well empirically for some of these target distributions. This also was demonstrated asymptotically by Jasra et al. [11] for some problems studied by [45, 43]. Incorporating local mixing conditions into our methods along the lines of [45, 12, 11] would allow us to obtain results more directly comparable to [45, 43] and answer the interesting question of whether such beneficial behavior persists outside the asymptotic setting.

Finally, our approach is well suited to comparison of the many variations on SMC sampling algorithms, and could be extended to include adaptive SMC methods. Adaptive methods can exhibit substantial performance gains in practice through adaptive selection of distributions and Markov kernels, but theoretical results for these methods to date are limited to adaptive resampling times [30, 20]. The techniques described in this chapter used to prove the stability of this adaptive technique in Chapter 3

2.8 Proof of Theorem 2

In the following we assume that $\mathcal{X}$ is either finite with counting measure $\lambda(dx)$, or $\mathcal{X} = \mathcal{R}^d$ with $\lambda(dx)$ the Lebesgue measure. If $\mathcal{X} = \mathcal{R}^d$ each measure $\mu \in \mathcal{P}$ is assumed to have a continuous density. We begin with some additional setup and

notation.

2.8.1 Additional Notation

Measures $\mu \in \mathcal{P}$ come equipped with an inner product $\langle f, g \rangle_\mu = \mu(f \cdot g)$ and function space $L_p(\mu) = \{f \in \mathcal{F} : \mu|f|^p < \infty\}$ for $p \geqslant 1$. For a signed measure η on $\mathcal{X}$ with $\eta << \mu$, define the p-norm with respect to μ by $||\eta||_{p,\mu} = \mu\left|\frac{\eta(dx)}{\mu(dx)}\right|^p$. Two cases of interest are $p = 1$, where the total variation distance is given by $\frac{1}{2}||\nu - \mu||_{1,\mu}$, and $p = 2$, where $\|\nu - \mu\|_{2,\mu}$ is the chi-squared distance. We will often use the alternate characterization of the total variation distance $\frac{1}{2}||\nu - \mu||_{1,\mu} = \sup_{f \in \mathcal{F}: |f| < M} \; M^{-1} \cdot |\nu f - \mu f|$ for $M > 0$.

Let K be a geometrically ergodic Markov kernel with spectral gap $\rho \in (0, 1)$. For any $\nu \in \mathcal{P}$ with $\nu << \mu$ and any positive integer t we have [46, 47]:

$$||\nu K^t - \mu||_{2,\mu} \leqslant ||\nu - \mu||_{2,\mu} \cdot (1 - \rho)^t \tag{2.8}$$

and the number of steps required to ensure the total variation distance between νK^t and μ is less than ϵ is upper bounded by $\frac{\log(||\nu - \mu||_{2,\mu} - \log(2\epsilon)}{\rho}$.

2.8.2 Proof of Theorem 2

The proof of Theorem 2 uses a modified set of iteration conditions. Define $\mathbf{C}_s^*$ to be the event that the following conditions hold:

$$\begin{aligned} &\mathbf{C}_s^*(\text{i}) \quad ||\hat{\mu}_s - \mu_s||_{2,s} \leqslant e - 1 \\ &\mathbf{C}_s^*(\text{ii}) \quad \bar{w}_{s+1} \geqslant \mu_s w_{s+1} \cdot \frac{2}{3} \end{aligned} \tag{2.9}$$

Condition $\mathbf{C}_s^*$(i) replaces the assumption that the marginal distribution $\hat{\mu}_s$ is exactly μ_s with the assumption that it is close to μ_s. Condition $\mathbf{C}_s^*$(ii) is the same as $\mathbf{C}_s$(ii) and ensures the fidelity of the resampling step. We proceed using the same inductive strategy used to prove Theorem 1. The following Lemma is analogous to Lemma 3 and bounds the chi-squared distance between $\tilde{\mu}_s$ and μ_s conditional on $\mathbf{C}_{s-1}^*$

Lemma 9 (Error of the resampling distribution)**.**

Assume $P(\boldsymbol{C}^*_{s-1}) \geqslant 3/4$. *Then* $\big|\big|\tilde{\mu}_s(\cdot \mid \boldsymbol{C}^*_{s-1}) - \mu_s\big|\big|_{2,s} \leqslant 4e \cdot W \cdot Z$

Proof. Let $E_{m,n}$ be the event that particle $\tilde{X}^n_s$ inherits from particle X^m_{s-1}. For any set $B \in \mathcal{B}$ the probability that $\tilde{X}^n_s \in B$, conditional on $\mathbf{C}^*_{s-1}$, can be bounded as follows:

$$\begin{aligned}
\Pr\Big(\tilde{X}^n_s \in B \;\Big|\; \mathbf{C}^*_{s-1}\Big) &= \sum_{m=1}^{N} \Pr\Big(X^m_{s-1} \in B,\, E_{m,n} \mid \mathbf{C}^*_{s-1}\Big) \\
&\leqslant \sum_{m=1}^{N} \frac{\sup_{x\in B}\ q_s(x)/q_{s-1}(x)}{S \cdot \bar{w}_s} \cdot \Pr\Big(X^m_{s-1} \in B \;\Big|\; \mathbf{C}^*_{s-1}\Big) \\
&\leqslant \frac{\sup_{x\in B}\ p_s(x)/p_{s-1}(x)}{S \cdot 2/3}. \qquad (2.10) \\
&\quad \sum_{m=1}^{N} \Pr\Big(\mathbf{C}^*_{s-1} \;\Big|\; X^m_{s-1} \in B\Big)\Pr\Big(X^m_{s-1} \in B\Big)/\Pr\Big(\mathbf{C}^*_{s-1}\Big) \\
&\leqslant 2 \cdot \sup_{x\in B}\ p_s(x)/p_{s-1}(x) \cdot \hat{\mu}_{s-1}(B)
\end{aligned}$$

In order to bound the chi-squared distance between the re-sampled distribution $\hat{\mu}_s$ and the desired interpolating distribution μ_s, we must convert this upper bound on the probability to an upper bound on the density. This is straightforward when $\mathcal{X}$ is a finite space, since choosing $B = \{x\}$ gives pmf

$$\tilde{p}_s(x|\mathbf{C}^*_{s-1}) \leqslant 2 \cdot p_s(x)/p_{s-1}(x) \cdot \hat{p}_{s-1}(x)$$

The same result holds λ a.e. when $\mathcal{X} = \mathcal{R}^d$. To see this, let $B(x,r)$ denote the open ball of radius r centered at x. Then for λ almost-all $x \in \mathcal{X}$ such that $p_{s-1}(x) > 0$

$$\begin{aligned}
\tilde{p}_s(x|\mathbf{C}_{s-1}) &= \lim_{r\to 0} \frac{\tilde{\mu}_s\big(B(x,r) \mid \mathbf{C}^*_{s-1}\big)}{\lambda\big(B(x,r)\big)} \\
&\leqslant 2 \cdot \lim_{r\to 0} \sup_{x\in B(x,r)}\ p_s(x)/p_{s-1}(x) \cdot \frac{\hat{\mu}_{s-1}(B)}{\lambda\big(B(x,r)\big)} \qquad (2.11) \\
&= 2 \cdot p_s(x)/p_{s-1}(x) \cdot \hat{p}_{s-1}(x)
\end{aligned}$$

The first line is the definition of the Radon-Nikodym derivative on $\mathcal{R}^d$ (see [48] chapter 9 or [49] chapter 5), the second is the upper bound from (2.10), and the last line comes from the continuity of p_s/p_{s-1} and the definition of the derivative. Having upper-bounded the density, we bound the chi-squared distance as follows

$$\begin{aligned}
\left|\left|\tilde{\mu}_s(\cdot \mid \mathbf{C}^*_{s-1}) - \mu_s(\cdot)\right|\right|_{2,s} &\leqslant \int \left[\frac{\tilde{p}_s(x)}{p_s(x)}\right]^2 p_s(x)\lambda(dx) \\
&\leqslant 4 \cdot \int \left[\frac{\hat{p}_{s-1}(x)}{p_{s-1}(x)}\right]^2 \frac{p_s(x)}{p_{s-1}(x)} \cdot p_{s-1}(x)\lambda(dx) \\
&\leqslant 4 \cdot W \cdot Z \cdot \left(||\hat{\mu}_{s-1} - \mu_{s-1}||_{2,s} + 1\right) \\
&\leqslant 4e \cdot W \cdot Z
\end{aligned} \tag{2.12}$$

The third line uses the upper bound on the density ratio and the definition of chi-squared distance and the last is the second condition of $\mathbf{C}^*_{s-1}$. ■

The following corollary shows how to choose t to ensure the first condition of $\mathbf{C}^*_s$ is satisfied (conditional on $\mathbf{C}^*_{s-1}$).

Corollary 10 (Error of the predictive distribution)**.**
Assume $P(\boldsymbol{C}^*_{s-1}) \geqslant 3/4$. Then $\boldsymbol{C}^*_s$(i) is holds conditional on $\boldsymbol{C}^*_{s-1}$ for $t \geqslant \frac{2+\log WZ}{\rho}$.

Proof. Follows directly from Lemma 9 and equation (2.8), using $-\log(1-\rho) \geqslant \rho$

$$\begin{aligned}
||\hat{\mu}_s(\cdot \mid \mathbf{C}^*_{s-1}) - \mu_s||_{2,s} &\leqslant 4e \cdot W \cdot Z \cdot (1-\rho)^t \\
&\leqslant e - 1
\end{aligned} \tag{2.13}$$

■

We know show that condition $\mathbf{C}^*_s$(ii) holds with high probability for appropriately chosen t and N. The proof here differs from the proof presented in Section 2.3 in that $\bar{w}_{s+1}$ may be a biased estimator of $\mu_s w_{s+1}$. We control this deterministic error by choosing t large enough to ensure that the total variation distance between $\hat{\mu}_s$

and μ_s is small. The stochastic variation in $\bar{w}_{s+1}$ is controlled in the same manner as Lemma 5 using Azuma's inequality.

Lemma 11 (Lower bound on the average weights)**.**
Assume $P(\boldsymbol{C}^*_{s-1}) \geqslant 3/4$. *Fix* $\delta_0 \in (0,1]$. *Choose* $N \geqslant 72\log(1/\delta_0) \cdot W^2Z^2$ *and* $t \geqslant \frac{4+\log(W^2Z^2)}{\rho}$. *Then* $\boldsymbol{C}^*_s(ii) \mid \boldsymbol{C}^*_{s-1}$ *holds with probability at least* $1-\delta_0$

Proof. Begin by decomposing the error into a deterministic bias and stochastic component.

$$\left|\bar{w}_{s+1} - z_{s+1}/z_s\right| \leqslant \left|\bar{w}_{s+1} - \hat{\mu}_s w_{s+1}\right| + \left|\hat{\mu}_s w_{s+1} - \mu_s w_{s+1}\right| \tag{2.14}$$

The bias conditional on $\mathbf{C}^*_{s-1}$ can be controlled using Lemma 9

$$\begin{aligned}
\left|\hat{\mu}_s w_{s+1} - \mu_s w_{s+1}\right| &\leqslant \frac{W}{2} \cdot ||\hat{\mu}_s(\cdot \mid \mathbf{C}^*_{s-1}) - \mu_s||_{1,s} \\
&\leqslant \frac{W}{2} \cdot ||\tilde{\mu}_s(\cdot \mid \mathbf{C}^*_{s-1}) - \mu_s||_{2,s} \cdot (1-\rho)^t \\
&\leqslant \frac{4e \cdot W^2 \cdot Z}{2} \cdot (1-\rho)^t \qquad (2.15) \\
&\leqslant \frac{Z^{-1}}{6} \\
&\leqslant \frac{z_{s+1}/z_s}{6}
\end{aligned}$$

The stochastic error can be controlled with Azuma's inequality in the same manner as Lemma 5 yielding

$$\left|\bar{w}_{s+1} - \hat{\mu}_s w_{s+1}\right| \leqslant \frac{z_{s+1}/z_s}{6} \tag{2.16}$$

Combining (2.14), (2.15), and (2.16) gives the desired result. ■

We summarize the one step induction condition for Theorem 2 in the next corollary.

Corollary 12 (Modified one step induction condition)**.**

Assume $P(\boldsymbol{C}^*_{s-1}) > 3/4$. *Fix* $\delta_0 \in (0, 1/4]$. *Choose* $N \geqslant 72 \log(1/\delta_0) \cdot W^2 Z^2$ *and* $t \geqslant \frac{4+\log(W^2Z^2)}{\rho}$. *Then the inductive condition* $\boldsymbol{C}^*_s | \boldsymbol{C}^*_{s-1}$ *holds with probability at least* $1 - \delta_0$

The final step is to apply the modified one step iteration condition to prove Theorem 2 in the same manner as Theorem 1. We allow the number of Markov kernel transitions t_S to vary at the final step which may require more steps than in intermediate stages in order to obtain the desired final accuracy.

Proof of Theorem 2

Proof. Fix $\delta_0 = \frac{1}{4(S+1)}$. First we ensure that $Pr\big(|\hat{\pi} f - \pi f| \leqslant \epsilon \mid \mathbf{C}^*_{S-1}\big) \geqslant 1 - \delta_0$ using the same approach as in Lemma 11. For $t_S \geqslant \frac{3+\log(\frac{WZ}{\epsilon})}{\rho}$ the deterministic bias $|\hat{\pi} f - \pi f|$ is less than or equal to $\epsilon/2$. Similarly, the stochastic error can be made less than $\epsilon/2$ with probability at least $1 - \delta_0$ by choosing $N \geqslant 8 \log\left(\frac{1}{2\delta_0}\right) \cdot \epsilon^{-2}$. Therefore for this N and t_S

$$\Pr\big(|\hat{\pi} f - \pi f| \leqslant \epsilon \big| \mathbf{C}^*_{S-1}\big) \geqslant 1 - \delta_0$$

Lower bounding $\Pr\big(|\hat{\pi} f - \pi f| \leqslant \epsilon\big)$ follows as in the proof of Theorem 1. ■

3

Path Selection for Sequential Monte Carlo

Sequential Monte Carlo (SMC) is a sampling method that moves particles drawn from an initial distribution μ_0 to a target distribution π via a sequence of interpolating distributions $\mu_0, \ldots, \mu_S = \pi$. Choosing an appropriate sequence of distributions, which we refer to as a *path* [28, 26], is critical to obtaining an efficient SMC sampler. Common path selection approaches for static (fixed dimension) SMC problems include batch processing with data [3], tempering with deterministic schedules [7, 4, 25], and tempering with adaptively chosen temperatures [16, 25, 26]. Comparison of paths is generally limited to simulation studies; the theoretical SMC literature treats the sequence of interpolating distributions as given and does not consider the effect of path selection on the accuracy of the resulting SMC estimator [21, 12, 11, 19, 30, 18].

In the first part of this chapter, we directly relate the computational complexity of obtaining a bounded-error SMC estimator to the selection of interpolating distributions. More formally, we demonstrate conditions under which SMC provides a *randomized approximation scheme* for estimating expectations of π. The bound presented here improves on the recent results in [50], relaxing the assumption of

bounded density ratios and requiring only a bound on the L_2 distance between adjacent distributions. This allows us to explicitly relate the distributions in the selected path to the error in the resulting estimator and the computational complexity of the algorithm. This in turn allows us to identify sequences of interpolating distributions (paths) that lead to substantial improvements in efficiency. Unlike other finite sample results for SMC in the literature [50, 21, 12], it also enables us to establish the convergence of SMC in situations where the the importance sampling weights are unbounded.

We use this new bound to illustrate the improvements obtainable by better path selection on two examples. The first is a spherical Gaussian target distribution, where we show that a path using geometric mixtures and tempering has superior complexity to a path using only geometric mixtures. The second example considers general log-concave target distributions. We use the path from [40] in combination with the sampling algorithm from [39] to provide an upper bound for SMC that obtains state of the art complexity for this problem.

In practice, pre-specifying a sequence of distributions that efficiently controls the L_2 distance between steps may be difficult. The second part of the chapter provides a practical scheme for adaptively choosing a sequence of distributions so that the L_2 distance between steps is provably controlled when weights are bounded. This is accomplished through monitoring the relative effective sample size (RESS). Adaptive path selection using the RESS is well known to the SMC community [16, 26, 25]; we provide conditions under which the RESS can be used to estimate the L_2 distance between steps with high accuracy, justifying its use for choosing an SMC path. We then extend our error bounds to this adaptive situation, giving conditions under which SMC using adaptive path selection remains a randomized approximation scheme.

We conclude by demonstrating that this algorithm has good empirical perfor-

mance on two examples. The first is a mean field Ising model, where we show that the adaptive SMC using tempered distributions finds nearly optimal sequences of interpolating distributions. The second is a Bayesian linear regression using a data-tempering approach. We demonstrate that the traditional path may result in steps with large L_2 distances causing significant instability in the resulting estimator. This problem is addressed using a hybrid path that combines the computational advantages of data-tempering with the stability of traditional tempering and provides bounded errors.

3.1 SMC error bounds

Before stating the main result we introduce some notation and describe the SMC algorithm studied in this chapter. Let $\big(\mathcal{X}, \mathcal{B}, \lambda\big)$ be a probability space. Define $\mathcal{P}$ be the set of probability measures on $\mathcal{X}$ that are absolutely continuous with respect to λ and $\mathcal{F}$ the set of measurable functions $f : \mathcal{X} \to \mathcal{R}$. Each measure acts on functions $f \in \mathcal{F}$ from the left by $\mu f = \int f(x)\mu(dx) = \mathsf{E}f$. We say that a measure $\nu \in \mathcal{P}$ is ω-warm with respect to μ if $\sup_{B \in \mathcal{B}(\mathcal{X})} \nu(B) \leqslant \omega \cdot \mu(B)$ [31, 32]. Let $\mathcal{P}_\omega(\mu)$ be the set of all such measures. For an ergodic Markov kernel $K : \mathcal{X} \times \mathcal{B} \to [0, 1]$ with limiting distribution μ, define the ω-warm mixing time of K by $\tau_K(\epsilon, \omega) = \min \big\{t : \sup_{\nu \in \mathcal{P}_\omega(\mu)} \|\nu K^t - \mu\|_{\mathrm{TV}} \leqslant \epsilon\big\}$, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance.

In this chapter we study the following SMC algorithm. Before sampling, the user specifies a path $\mu_0, \ldots, \mu_S$ where $\mu_s \in \mathcal{P}$ and $\mu_{s-1} << \mu_s$. We abuse notation and write the density $\mu_s(x) = q_s(x)/z_s$ where $q_s(x)$ is a known, unnormalized density. The algorithm is initialized by drawing N samples $X_0^{1:N} = X_0^1, \ldots, X_0^N$ independently from μ_0, then proceeds in S steps. At the beginning of step s, each particle is assigned an importance sampling weight $w_s\big(X_{s-1}^n\big) = q_s\big(X_{s-1}^n\big)/q_{s-1}\big(X_{s-1}^n\big)$. Then, a new set of particles $\tilde{X}_s^{1:N}$ is drawn with replacement from the current particles according to

the weights (multinomial resampling); i.e. a copy of X_{s-1}^n is drawn with probability proportional to $w_s\big(X_{s-1}^n\big)$. Finally, each resampled particle evolves independently according to a Markov kernel K_s with stationary distribution μ_s, resulting in a new set of particles $X_s^n \sim K^t\big(\tilde{X}_s^n,\ \cdot\ \big)$. Following step S of the algorithm πf is estimated using the particle average $\hat{\pi} f = \frac{1}{N}\sum_{n=1}^N f\big(X_S^n\big)$. Detailed descriptions of this SMC algorithm can be found in [7, 3, 50].

3.1.1 Convergence of SMC using the L2 distance

The convergence result presented in this section depends on the L_2 distance between interpolating distributions and the mixing times of the Markov kernels. For $\mu, \eta \in \mathcal{P}$ define the L_2 *distance from* η *to* μ by the $L_2(\mu)$ norm of η/μ:

$$\|\eta/\mu\|_{L_2(\mu)} = \int \left(\frac{\eta(dx)}{\mu(dx)}\right)^2 \mu(dx)$$

This quantity is not symmetric and therefore not a true metric, however we follow standard convention and refer to it as the L_2 distance because it provides an appropriate measure of the divergence between μ and η. Note that subtracting one yields the traditional χ^2 "distance" which plays a familiar role in importance sampling, where it provides the variance of the importance weights under instrumental distribution μ and target distribution η. This idea has been used in SMC, where the empirical variance of the weights is used to assess the efficiency of the particle system at each step, using the relative effective sample size (RESS):

$$\hat{E}_s = \frac{\Big(N^{-1}\sum_{s=1}^N w_s\big(X_{s-1}^n\big)\Big)^2}{N^{-1}\sum_{s=1}^N w_s\big(X_{s-1}^n\big)^2} \tag{3.1}$$

When $\hat{E}_s$ is small the particle system is described as degenerate. As $n \to \infty$, $\hat{E}_s$ converges to $\mathcal{E}_s \triangleq \|\mu_s/\mu_{s-1}\|^{-1}_{L_2(\mu_{s-1})}$, the reciprocal of the L_2 distance; we therefore

interpret the RESS as an estimate of $\mathcal{E}_s$ and in Section 3.3 give bounds on its approximation error. Our main result bounds the approximation error of SMC in terms of a bound on the maximal L_2 distance between adjacent distributions:

$$\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant \mathcal{E}^{-1} \tag{3.2}$$

Theorem 13 establishes conditions under which SMC serves as a *randomized approximation scheme*. An algorithm is a randomized approximation scheme if, for any user-specified $\epsilon > 0$ and $\delta \in (0, 1]$, it guarantees $|\hat{\pi} f - \pi f| < \epsilon$ with probability at least $1 - \delta$ [33]. To simplify the presentation, we establish this for $\delta = 1/4$, but this is easily improved to arbitrary $\delta > 0$ at a cost of $\mathcal{O}\big(\log(1/\delta)\big)$ using the median approach [34].

Theorem 13 (Error bound for SMC)**.**
Fix $\epsilon > 0$ and sample $X_0^{1:N}$ independently from μ_0. Let $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant \mathcal{E}^{-1} < \infty$. Choose

1. $N \geqslant 2\log\big(16S\big) \cdot \max\big\{\frac{42}{\mathcal{E}}, \frac{1}{\epsilon^2}\big\}$
2. $t \geqslant \max_s\ \tau_s\big(\frac{1}{8NS},\ 2\big)$

Then for any $f \in \mathcal{F}$ with $|f| \leqslant 1$,

$$\big|\hat{\pi} f - \pi f\big| \leqslant \epsilon.$$

with probability at least 3/4.

The proof of Theorem 13 is given in the appendix and closely follows the proof of Theorem 13 in [50]. The key difference is the use of an improved martingale concentration inequality to ensure concentration of the weights, which replaces Lemma 4 of [50] and results in a modified one-step induction condition yielding Theorem 13 above.

Assumption (3.2) replaces the upper bound W on the weights and a lower bound Z on the ratios of normalizing constants required by [50]. When such bounds are available we immediately have $1/\mathcal{E} \leqslant W^2 Z^2$ to apply Theorem 13. However, requiring a bound on $\frac{1}{\mathcal{E}}$ instead has several advantages. First, the assumption of bounded weights restricts the sequences of interpolating distributions that can be considered and is frequently violated in applications. (Despite this, it is commonly assumed in theoretical results for both asymptotic and finite sample convergence of SMC). Another advantage of assumption (3.2) is that we can compare the resulting SMC bounds directly to similar bounds for MCMC. This advantage is explored in the next section.

3.2 Path selection and complexity

Finite sample bounds such as Theorem 13 facilitate explicit comparison between algorithms. In this section, we compare our bounds on the computational complexity of SMC with existing bounds for MCMC to highlight the advantages of each algorithm. Complexity is given in total number of Markov kernel transitions required to approximate πf. Suppose that $K_1, \ldots, K_S$ are geometrically ergodic and reversible with spectral gaps $\rho_1, \ldots, \rho_S \in (0, 1)$. Then the number of transitions according to K_S required to sample approximately from π using a Markov chain starting with a draw from μ_0 is [46, 47]

$$\mathcal{O}\left(\frac{\log \|\mu_0/\pi\|_{L_2(\pi)}}{\rho_S}\right)$$

In comparison, Theorem (13) gives the following complexity bound for SMC

$$\mathcal{O}\left(\frac{S/\mathcal{E} \cdot \log^2\left(S/\mathcal{E}\right)}{\rho^*}\right)$$

where $\rho^* = \max_s \rho_s$. When the spectral gaps of the Markov kernels K_s are of the same order, the bounds differ primarily by the cost of moving from the initial

distribution to the target distribution. For MCMC, this factor is $\log \|\mu_0/\pi\|_{L_2(\pi)}$, whereas for SMC this factor is an upper bound on $S/\mathcal{E} \geqslant S \cdot \max_s \|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})}$, which we call *the path length*. Note that the L_2 distance is not symmetric and the SMC and MCMC bounds depend on this quantity in opposite directions, and therefore differ even for $S = 1$ (importance sampling). For example when μ_0 is heavy tailed relative to π, $\|\mu_0/\pi\|_{L_2(\pi)}$ may be much larger than $1/\mathcal{E}$. Since specifying the initial distribution μ_0 to be heavier tailed than the target π is generally easier than the reverse, this indicates an advantage for SMC. On the other hand, the amount of computation required by SMC grows linearly in $S/\mathcal{E}$ whereas the bound for MCMC grows logarithmically in $\|\mu_0/\pi\|_{L_2(\pi)}$. This can be advantageous for MCMC when finding a sequence of distributions that ensures $S/\mathcal{E}$ small is difficult.

The remainder of this section compares the relative cost of moving from μ_0 to π for SMC versus MCMC. The first example investigates the problem of sampling from a spherical Gaussian target distribution, studying the path complexity with regards to the target precision, mean, and dimension. The second example considers the problem of sampling a general log-concave target distribution and uses an optimal path identified by [40] to obtain an SMC path with low complexity. This bound improves upon the best existing results for MCMC.

3.2.1 Gaussian example

Consider the problem of approximating expectations with respect to a d-dimensional spherical Gaussian target distribution $\pi(x) = N_d(1_d \cdot \theta,\, I_d/\phi)$, where $\theta \geqslant 2$ and $\phi \geqslant 1$. This problem is representative of many Bayesian inference problems with large sample sizes via the Bernstein-von Mises theorem. A simpler version of this problem (with $\theta = 0$) was studied by [50]; however, results for the more challenging problem when $\theta \neq 0$ are now possible as Theorem 13 allows for unbounded importance sampling weights.

We assume that the initial distributions for both SMC and MCMC are chosen to be standard Gaussian, $\mu_0 = N_d(0, I_d)$. The cost of MCMC, relative to the spectral gap, is given by

$$\mathcal{O}\left(\frac{\theta^2 d}{\phi(2-\phi)}\right) \tag{3.3}$$

assuming $\phi < 2$ (see Appendix). We will consider two different choices of the interpolating distribution sequences for SMC which lead to bounds with improved complexity with respect to ϕ, θ and d. These results are applicable for any $\phi \geqslant 1$.

A standard approach to constructing a sequence of interpolating distributions is a geometric path, with $\mu_\beta(x) \propto \mu_0(x)^{1-\beta}\pi(x)^\beta$ for $\beta \in [0,1]$. Such paths are commonly used to estimate ratios of normalizing constants, where they are sometimes referred to as power paths or tempered paths [28, 51]. The path is specified by a sequence $\beta_0 = 0 \leqslant \beta_1 \leqslant \ldots \leqslant \beta_S = 1$ controlling the rate at which the path moves from μ_0 to π. Choosing $\beta_s = \left(1 + \frac{2}{\theta\sqrt{d}}\right)^{s-1} / \left(\phi \cdot \theta\sqrt{d}\right) \wedge 1$ and $S = 1 + \left\lceil \frac{\theta\sqrt{d}}{2} \log\left(\phi^2 \cdot \theta\sqrt{d}\right)\right\rceil$ ensures $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant \mathcal{O}(1)$ and gives an upper bound on path length $S/\mathcal{E}$ of (see Appendix 3.7.1):

$$\mathcal{O}\left(\theta\sqrt{d} \cdot \log\left(\phi^2 \cdot \theta\sqrt{d}\right)\right) \tag{3.4}$$

This bound improvemes the dimension dependence from $\mathcal{O}(d)$ to $\mathcal{O}(\sqrt{d}\log\sqrt{d})$ relative to the MCMC bound. We also see a super-exponential improvement in dependence on the precision, from $\mathcal{O}\left(\frac{1}{\phi-2}\right)$ to $\mathcal{O}(\log\phi)$, as well as an improvement in the location dependence from $\mathcal{O}(\theta^2)$ to $\mathcal{O}(\theta\log\theta)$.

However there exists an even better path inspired by a result from [28]. First, choose $s_1 = \lceil 3\sqrt{d}\log(d\theta^2)\rceil$ distributions to be $\mu_s = N_d\left(0, I_d/\phi_{1,s}\right)$ with $\phi_{1,s} = (1 - 1/\sqrt{9d})^s \vee \frac{1}{d\theta^2}$. The next distribution changes the location in a single step: $\mu_{s_1+1} = N_d(1_d\theta, I_d \cdot d\theta^2)$. Finally, we take $s_2 = \lceil\sqrt{d}\log(d\theta^2\phi)\rceil$ steps using $\mu_s = N_d\left(1_d\theta, I_d/\phi_{2,s}\right)$ with $\phi_2, s = \frac{1}{d\theta^2}\left(1 + 1/\sqrt{d}\right)^{s-s_1-1} \wedge \phi$. We call this the *precision*

path because it uses the fact that when the precision is sufficiently small it is possible to move between normal distributions with differing locations in a single step. Because precision can be decreased exponentially quickly, this shortens the overall path, yielding an improved complexity in θ compared to varying the mean and precision simultaneously. More precisely, the precision path ensures $1/\mathcal{E} \leqslant 2$, giving a path length bound of (see Appendix 3.7.2):

$$\mathcal{O}\big(\sqrt{d}\log(\phi \cdot \theta^2 d)\big) \tag{3.5}$$

showing an improvement from $\mathcal{O}(\theta \log \theta)$ to $\mathcal{O}(\log \theta)$. This example highlights the potential speedup available using non-geometric paths, though finding such paths may be challenging.

Gelman and Meng [28] derived an optimal path sampling estimator to estimate (log-) ratios of normalizing constants between normal distributions with different means. This optimal path also flattens the intermediate normal distributions by reducing their precisions, resulting in a similar improvement in complexity from $\mathcal{O}(\theta)$ (tempered path) to $\mathcal{O}(\log \theta)$. The similarity between good path-sampling and SMC paths arises due to the necessity of estimating intermediate ratios of normalizing constants to satisfy the one-step induction condition for SMC shown in [50]. In fact, when $d = 1$ and $\phi = 1$, a sufficiently fine discretization of the Gelman and Meng path yields the same complexity bound as the precision path. It is unlikely that this path is optimal for SMC, however, since the optimal path-sampling sequence from π to μ_0 is the reverse of the optimal path from π to μ_0, while this will not generally be true for SMC as the L_2 "distance" is asymmetric and optimal paths should reflect this asymmetry.

3.2.2 Log-concave target distributions

Let $\pi(x) \propto q(x)$ be a log-concave target distribution on $\mathcal{R}^d$. A function q is said to be strongly log-concave if $q^{1-\alpha}(x) \cdot q^{\alpha}(y) < q\big(\alpha x + (1-\alpha)y\big)$ for $x, y \in \mathcal{R}^d$ and $\alpha \in (0,1)$. In general bounds on the ω-warm mixing times of Markov kernels targeting a sequence of distributions obtained by tempering a log-concave distribution will have the same complexity at each step. For example, if we choose K_s to be the Metropolis-adjusted Langevin algorithm (MALA) [52], the complexity of the bound on the ω-warm mixing time is independent of the temperature parameter [39, 50] and a similar result holds for other Markov kernels including the *ball-walk* or *hit-and-run walk* Markov kernels [41]. Therefore, when the target distribution is log-concave, we can again focus on finding interpolating sequences that minimize the path length.

Efficient path selection for log-concave distributions has received substantial attention in the theoretical computer science literature, where the volume of a convex body is estimated by sampling from a sequence of tempered distributions [35, 40]. A key factor in volume computation is the L_2 distance between adjacent distributions, which controls the relative error when estimating the corresponding volume ratios. The following corollary follows from Theorem 13, using the the tempering path from [40] and the bounds on the mixing time from [39].

Corollary 14 (SMC complexity for log-concave target distributions)**.** *Let $\pi(x) \propto q(x)$ be log-concave with mode x^* and define $\kappa = L/m$ where for all $x, y \in \mathcal{R}^d$:*

$$-\frac{L}{2}\|x-y\|_2^2 \leqslant \log\frac{q(x)}{q(y)} - \nabla \log q(x)^T(x-y) \leqslant -\frac{m}{2}\|x-y\|_2^2$$

Restrict π to the ball B of radius $4\sqrt{d/m}$ centered at x^* and assume $\epsilon > 2e^{-d}$. Choose $\mu_0 \propto \mathbb{1}_B(x)$ and $\mu_s(x) \propto \pi^{\beta_s}(x)\mathbb{1}_B(x)$ with $\beta_s = \frac{1}{d\kappa}\big(1+\frac{1}{\sqrt{d}}\big)^s$ for $s = 1, \ldots, S = \lceil \sqrt{d}\log(d\kappa)\rceil$. Let K_s be a MALA kernel with step size given in [39]. Then SMC

provides a randomized approximation scheme in time:

$$\mathcal{O}^*\Big(d^{3/2}\kappa \cdot \max\big\{1, \sqrt{\kappa/d}\big\}\Big)$$

The notation $\mathcal{O}^*$ indicates the omission of logarithmic terms in d and κ. The specified path ensures $1/\mathcal{E} \leqslant e$ and gives $S/\mathcal{E} = \mathcal{O}\big(\sqrt{d}\log(d\kappa)\big)$ [40]. The MALA kernel provides an ω-warm mixing time of $\mathcal{O}\Big(d\kappa \cdot \log\frac{2\omega}{\epsilon} \cdot \max\big\{1, \sqrt{\kappa/d}\big\}\Big)$ [39], and the result then follows from Theorem 13. The restriction to B is used to bound the L_2 distance of the first step and has minimal impact on the results of our analysis as $\pi(B) \geqslant 1 - \epsilon/2$; similar restrictions are common in the log-concave sampling literature. The assumption $\epsilon > 2e^{-d}$ serves only to simplify the presentation.

The result in Corollary 14 improves on the $\mathcal{O}^*\big(d^k\kappa^2 \cdot \max\big\{1, \sqrt{\kappa/d}\big)$ result in [50] and the state-of-the art bound for MCMC of $\mathcal{O}^*\Big(d^2\kappa \cdot \max\big\{1, \sqrt{\kappa/d}\big\}\Big)$ [39]. In both cases, the improvement comes solely from the selection of a superior path, as each bound uses the same Markov kernels. To the best of our knowledge, this is the fastest randomized approximation scheme for a log-concave target distribution. We remark, however, that an ever better bound could be obtained by combining the path in [40] with the MALA mixing times in [39] using a time-inhomogenous Markov chain.

3.3 Adaptive path selection

Selecting a path where a bound on the L_2 distance is known *a priori*, let alone an optimal path, can be difficult in practice. In this section we establish conditions under which adaptively choosing distributions using the RESS, as commonly done in practice [16, 25, 26], automatically selects a path with controlled L_2 distance between steps. Given a pre-specified bound $\mathcal{E} \in (0, 1)$, suppose μ_s is chosen from a (possibly large) discrete set of candidate distributions $\nu_{s,1}, \ldots, \nu_{s,M} \in \mathcal{P}$. (For example, this would be the case in adaptive tempering when considering temperature changes of

increasing size.) For each candidate distribution, we compute the RESS $\hat{E}_{s,m}$ and choose $\mu_s \in \{\nu_{s,m}\}$ so that $\hat{E}_{s,m} \geqslant \mathcal{E}$. We give conditions under which $\frac{1}{\hat{E}_{s,m}}$ accurately estimates $\|\eta_{s,m}/\mu_{s-1}\|_{L_2(\mu_{s-1})}$ and so μ_s can be chosen so that $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant 3/\mathcal{E}$ with high probability. We then apply the same techniques used to prove Theorem 13, ensuring that the chosen step will preserve the approximation accuracy of the algorithm with high probability. The section begins by discussing the specification and selection of candidate distributions. Then, we extend the results of Theorem 13 to the adaptive setting, giving conditions under which the adaptive algorithm constitutes a randomized approximation scheme.

3.3.1 Candidate distributions

At step s of the algorithm, we choose μ_s from a finite set of candidate distributions distributions $\nu_{s,1}, \ldots, \nu_{s,M}$. This set may depend on μ_{s-1} but not on the particle values themselves. We assume these candidates are ordered *a priori*, such that a move to $\nu_{s,m}$ is preferred over a move to $\nu_{s,m-1}$. The algorithm proceeds by computing the RESS of the particle system under each candidate distribution:

$$\hat{E}_{s,m} = \frac{\left(N^{-1}\sum_{s=1}^{N} w_{s,m}\left(X_{s-1}^{n}\right)\right)^2}{N^{-1}\sum_{s=1}^{N} w_{s,m}\left(X_{s-1}^{n}\right)^2} \tag{3.6}$$

where $w_{s,m}(x) \propto \frac{\nu_{s,m}(x)}{\mu_{s-1}(x)}$. In order to obtain bounds for the adaptive algorithm, we require the set of candidate distributions to have bounded weights; i.e. $0 < w_{s,m} \leqslant 1$ for each s and m. This differs from the setting of Theorem (13). The next distribution is chosen to be the candidate distribution with sufficiently large RESS that is closest to π:

$$\mu_s = \max_m \left\{\nu_{s,m} : \hat{E}_{s,m} \geqslant \mathcal{E}\right\} \tag{3.7}$$

When $\mathcal{E}$ is close to 1 the SMC algorithm will take small steps, preserving particle

approximations with high effective sample sizes. On the other hand, when $\mathcal{E}$ is close to 0, the algorithm will take larger steps leading to particle approximations where it is possible for relatively few particles to receive substantial weight.

Example: geometric path. Let μ_0 be an initial distribution on $\mathcal{X}$ and define $\mu(x \mid \beta) \propto \mu_0(x)^{1-\beta} \cdot \pi(x)^{\beta}$ for $\beta \in [0,1]$ to be the geometric mixture of μ_0 and π. At step s of the algorithm, the previous distribution is $\mu_{s-1} = \mu(x|\beta_{s-1})$ and the candidate distributions are $\nu_{s,m}(x) = \mu(x \mid \beta_{s,m})$ with $\beta_{s,m} = \beta_{s-1} + \frac{m}{M}(1 - \beta_{s-1})$.

Computing the weights for each candidate distribution is greatly simplified by pre-computing $w^n \propto \pi\big(X_{s-1}^n\big)/\mu_0\big(X_{s-1}^n\big)$ as $w_{s,m}(X_{s-1}^n) = (w^n)^{\beta_{s,m}}$. The optimal candidate distribution can then be found in $\mathcal{O}(\log M)$ time using binary search and is unique as $1/\hat{E}_{s,m}$ is non-increasing with m.

Example: data tempered path. Let $\pi(x) \propto p(y_{1:K}|x)\pi_0(x)$ be a posterior distribution arising from a Bayesian model with K observations and prior distribution π_0. An alternative to the power path is a sequential posterior obtained by adding observations to the likelihood [3]. Let $\mu_{s-1}(x) \propto p(y_{1:k_{s-1}}|x)\pi_0(x)$ where $y_{1:k_s}$ denotes the first $0 < k_s < K$ observations. The candidate distributions are chosen from $\nu_{s,m}(x) \propto p(y_{1:k_{s,m}}|x)\pi_0(x)$ for $k_{s-1} < k_{s,1} < ... < k_{s,M} \leqslant K$.

When the likelihood is independent, i.e. $p(y_{1:K}|x) = \prod_{k=1}^{K} p(y_k|x)$, the weights are computed incrementally $w_{s,m}\big(X_{s-1}^n\big) = w_{s,m-1}\big(X_{s-1}^n\big) \cdot \prod_{k=1+k_{s,m-1}}^{k_{s,m}} p\big(y_k|X_{s-1}^n\big)$. Rather than choosing the largest m to satisfy condition (3.7) it is generally expedient find the first m such that this condition fails and choose $\mu_s = \nu_{s,m-1}$.

Data tempering may be preferred to tempering in practice as it can substantially reduce the number of likelihood evaluations needed to approximate a posterior distribution while also providing a tempering effect. A drawback of this approach is that the set of candidate distributions may not be sufficiently fine, making it impossible to select a distribution satisfying (3.7). In Section 3.4.2 we introduce a hybrid path,

which combines the advantages of both the geometric and data-tempering paths.

3.3.2 An adaptive path selection RAS

To obtain conditions under which the adaptive step size SMC algorithm provides an RAS, we present a slightly modified version the algorithm. Begin by fixing an initial number of steps S, samples N_1, Markov kernel transitions t_1, and a target RESS $\mathcal{E}$. Select distributions according to the following adjusted criteria:

$$\mu_s = \max_m \left\{ \nu_{s,m} : \hat{E}_{s,m} \geqslant \mathcal{E} \text{ and } \bar{w}_{s,m}^2 \geqslant C \right\} \tag{3.8}$$

The additional requirement that $\bar{w}_{s,m}^2 \geqslant C$ for pre-specified $C \in (0,1)$ is used to ensure that $1/\hat{E}_{s,m}$ approximates $\|\nu_{s,m}/\mu_{s-1}\|_{L_2(\mu_{s-1})}$ with bounded relative error. Small values of C may allow for bigger steps, however, this advantage must be balanced by against the number of particles required which will increase as $\mathcal{O}\big(C^{-2}\big)$.

To establish adaptive step size SMC as an RAS we requires an additional modification of the algorithm, as the proof of Theorem 13 requires that N depend on the number of steps S, which is not known in advance for the adaptive algorithm. To achieve a randomized approximation, the number of particles must be allowed to grow as the number of steps adaptively increases. This can be done in a straightforward way, by choosing N based on S and then increasing N if π is not reached within the first S steps. To increase N, N' new samples are drawn independently from μ_0, evolved through the previously chosen path $\mu_1, \ldots, \mu_S$, and used to supplement the existing particles yielding a system of size $N' + N$. A natural choice is doubling ($N' = N$), ensuring that the particle system is valid for S additional steps before doubling is required again. We call these repetitions *epochs*, and denote the current epoch by p. The number of Markov kernel transitions must also be increased at each epoch as this also depends on the N and S.

A disadvantage of the doubling approach is that the total number of particles

at the end of epoch p is $2^p N$. This exponential dependence can be reduced to sub-quadratic by following a procedure that chooses the number of additional particles at each epoch more carefully. This procedure is summarized in algorithm 1

Algorithm 1: Adaptive SMC algorithm

Result: $\Pr\big(|\hat{\pi} f - \pi f| \leqslant \epsilon\big) \geqslant 3/4$ for any $f \in \mathcal{F}$ with $|f| \leqslant 1$
Fix $\mathcal{E} \in (0,1)$ and $S > 0$;
Set $p = 0, s = 0, t = 0$;
while $\mu_s \neq \pi$ **do**
  $s = s + 1$;
  if $s = p \cdot S$ **then**
    Set $p = e + 1$ and $t = t_p$;
    Apply SMC to N_p new particles using the path $\mu_0, \ldots, \mu_{S(p-1)}$;
    Add the new particles to the set of current particles;
  end
  Select μ_s from $\eta_{s,1}, \ldots, \eta_{s,M}$ according to (3.8);
  Apply a step of SMC targeting μ_s;
end

Under the following conditions this adaptive SMC algorithm provides a randomized approximation scheme

Theorem 15 (Error bound for adaptive SMC)**.**

Choose a set of paths according to Section 3.3.1 with $0 < w_{s,m} \leqslant 1$ *for each candidate distribution. Fix an error tolerance* $\epsilon > 0$*, a target RESS* $0 < \mathcal{E} < 1$*, an epoch length* S*, and a lower bound* $0 < C < 1$*. Let:*

1. $N_1 \geqslant \max\left\{\frac{32}{C^2}\log\big(96SM\big),\ \frac{243}{\mathcal{E}}\log\big(24S\big),\ \frac{2}{\epsilon^2}\log\big(24S\big)\right\}$
2. $N_p \geqslant \max\left\{\frac{243(\log(24S)+p)}{\mathcal{E}},\ \frac{23}{C^2},\ \frac{2}{\epsilon^2}\right\}$
3. $t_e \geqslant \sup_{s,m}\ \tau_{\nu_{s,m}}\left(\frac{1}{2^p 24 \cdot \max\left\{N_p \cdot Sp,\ \sum_{i=1}^{p} N_i\right\}},\ 2\right)$

Then for any $f \in \mathcal{F}$ *with* $|f| \leqslant 1$ *Algorithm 1 ensures*

$$\left|\hat{\pi} f - \pi f\right| \leqslant \epsilon.$$

with probability at least 3/4.

Proof of Theorem 15 is given in the appendix. At the end of epoch p, the total number of particles is $\mathcal{O}\big(p \cdot \max\big\{\frac{\log(24S+p)}{\mathcal{E}},\ C^{-2},\ \epsilon^{-2}\big\}\big)$, a substantial improvement on the exponential dependence of the doubling algorithm. Despite the substantial improvement in p, it is still generally more efficient to choose S large so that p remains small, as the complexity of the algorithm grows only logarithmically in S. In fact, when the candidate distributions are chosen so that there is a maximum possible path length S^*, as in data tempering where $S^* = K$, it may be most efficient to choose $S = S^*$ as this ensures $p = 1$.

To our knowledge Theorem 15 provides the first proof of convergence for SMC with adaptively chosen sequences of distributions. Previously, in order to ensure that a central limit theorem held in the adaptive setting, a two stage approach to SMC was employed [15, 25] in which an adaptive SMC algorithm was run to select a path, followed by a non-adaptive SMC run on the selected path to estimate expectations under π. Our result shows that for appropriately chosen N and t, this two stage procedure is unnecessary. In addition, the two stage procedure provides no information about the properties of the chosen path or the resulting estimation error. Theorem 15 may also be seen as a validation of the use of the RESS for selecting distributions. Other approaches have been considered [53, 54], however, these methods currently lack theoretical support.

Note that we have given no conditions that ensure it is possible to choose an interpolating distribution satisfying (3.8); if this condition is not met the algorithm will terminate prematurely. We can also make no claims at this point about near-optimality of the selected path in terms of path length. We explore this issue in the next section. Finally, the additional requirement of the lower bound C in (3.8) may be unnecessarily restrictive in practice, but it is unclear at this time if selecting

distributions according to (3.7) is sufficient.

3.4 Empirical results

We investigate the empirical performance of the adaptive step-size SMC algorithm on two non-trivial target distributions where the L_2 distance can be evaluated exactly. Our goals are twofold. First, the conditions of Theorem 15 are generally difficult to ensure due to the challenge of bounding the mixing time. Therefore, it is sensible to verify whether a naive implementation of the algorithm might maintain controlled L_2 distances at each step. Second, our result says little about the relative optimality of the adaptively chosen path, guaranteeing accurate estimation but not optimal path length. This empirical study allows us to compare adaptively chosen paths to an ideal path which maintains a fixed step size of $\mathcal{E}$. Finally, we provide an example where data tempering may lead to paths where condition (3.7) cannot be satisfied. We introduce a hybrid path that addresses this problem, ensuring condition (3.7) is met at each stage of the algorithm. Empirically, we find that the hybrid approach decreases the overall path length and leads to paths of near-optimal length for a specified step size. For these reasons, we recommend that the hybrid approach be used over the data tempering approach in practice.

3.4.1 Example: Ising model

Consider the well-known mean field Ising model originally developed as a model of ferromagnetism in statistical physics. The D-dimensional model takes values in $\mathcal{X} = \{-1, 1\}^D$ for binary "spins" x_d with probability

$$\pi(x|\alpha) \propto \exp\left(\frac{\alpha}{2D}\Big(\sum_{d=1}^{D} x_d\Big)^2\right)$$

When $\alpha > 0$, the high probability configurations are those where the spins are mostly the same. The hyperparameter α controls the strength of this effect. Related

models have been used in machine learning for image processing [55] and in Bayesian statistics for modeling spatial dependence [56].

Sampling from the Ising model has received considerable attention [57, 58]. A key characteristic of the model is that π undergoes a phase transition as α approaches the critical temperature α_0. This is exhibited in the distribution of the magnetization $M = \sum_{d=1}^{D} x_d$, which rapidly changes from concentrated about 0 to dispersed to the extremes near $M = D$ and $M = -D$. This rapid change in behaviour makes it challenging to sample from the Ising model when $\alpha > \alpha_0$ as it is difficult for MCMC methods to move between these modes. Tempering approaches have proven successful at sampling from this distribution [45]. The selection of an appropriate temperature ladder is crucial to the success of parallel tempering, and subsequently selecting a temperature ladder has received substantial attention in the tempering literature. In contrast, we demonstrate empirically that temperature selection using an adaptive SMC approach achieves nearly optimal performance in the sense of minimizing the number of steps (temperatures) for a given $\mathcal{E}$.

For $D \in \{10, 50, 250\}$ and $\alpha = 2 \geqslant \alpha_0$ we performed SMC using the geometric path given in Section 3.3.1 with μ_0 uniform distribution on $\mathcal{X}$. Markov transitions are made according to the Glauber dynamics (Gibbs sampling), scanning through each component in a randomly chosen order and drawing a new spin from it's conditional distribution. We set $\mathcal{E} = 0.5$ and used $1,000$ particles for each simulation. This SMC procedure was repeated $1,000$ times to assess variability. We then computed the error of the estimated L_2 distance at each step relative to the exact L_2 distance, which can be evaluated numerically. We also compared the adaptively chosen path to the temperature ladder satisfying $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} = 2$ at each step.

The results of the experiment are displayed in Figure 3.1. In general, the adaptively chosen paths follow closely the optimal path, being of comparable length and displaying similar curvature near the critical temperature where they take small steps

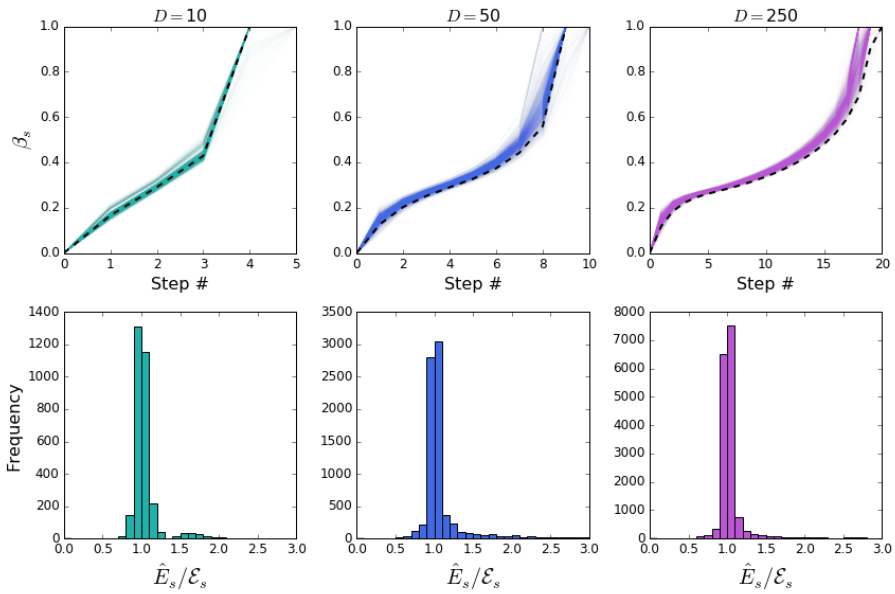

FIGURE 3.1: Results of the empirical path selection for the mean field Ising model. The top row shows the distribution of paths chosen by the adaptive approach, with the optimal path given by the dotted black line. The bottom row shows the distribution of the estimated L_2 distances relative to the true distance.

as the target distribution is changing rapidly. Estimated values of $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})}$ are generally quite accurate. More importantly, the induction condition prescribed by Lemma 18 is achieved at each step of the algorithm and across every step of the simulation. The selection criteria in (3.7) is sufficient to achieve good path selection for this problem, perhaps not requiring the modified criteria (3.8).

3.4.2 Bayesian linear regression

Our second example demonstrates the behaviour of adaptive SMC using data tempering. Consider a Bayesian linear regression model with $Y = X\beta + \epsilon$, where $Y \in \mathcal{R}^K$ is a response vector, $X \in \mathcal{R}^{K\times D}$ is a matrix of covariates, $\beta \in \mathcal{R}^D$ is an unknown coefficient vector and $\epsilon \sim N(0, I_K \cdot \sigma^2)$ is a vector of observation noise. We fit the white wines data set from the UCI machine learning repository [59], which consists of $M = 4898$ observations of wine quality and $D = 11$ physicochemical predictors. Be-

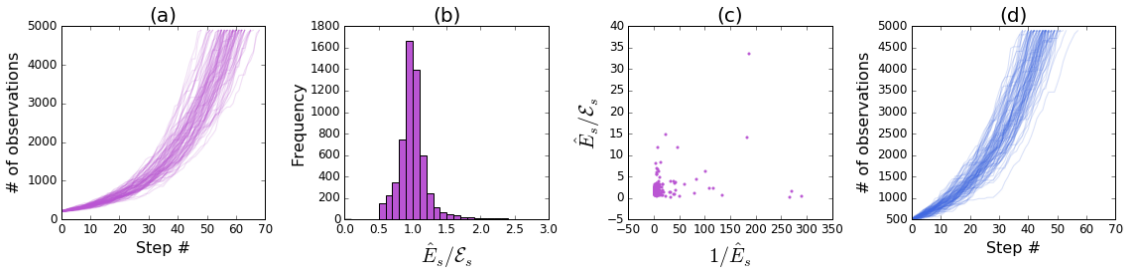

FIGURE 3.2: Adaptive path selection for the linear regression model. (a) Distribution of paths chosen by the adaptive approach. (b) Distribution of the estimated L_2 distance relative to the true L_2 distance for transitions with $1/\hat{E}_s \leqslant 2$. (c) Relative size for steps with bad transitions, i.e. $1/\hat{E}_s \geqslant 2$. (d) Distribution of paths chosen by the hybrid algorithm

fore analysis, the data was centered and scaled. We adopt a normal inverse-gamma prior with $\pi_0(\beta \mid \sigma^2) \propto N\big(0, \sigma^2\big(X^T X\big)^{-1}/K\big)$ and $\pi_0(\sigma^2) \sim$ Inv-Gamma$(4, 4)$. This prior is conjugate, allowing analytic calculation of the L_2 distance between SMC steps for comparison (see appendix 3.9).

We simulated from the posterior distribution of this model using the data tempering approach described in Section 3.3.1. During the initial phase of the algorithm, the target distribution changes rapidly as observations are added, making it difficult to obtain transitions with sufficiently high relative effective sample size. As a result, instead of choosing $\mu_0 = \pi_0$, we let $\mu_0 \propto p(Y_{1:200} \mid \beta, \sigma^2, X_{1:200}) \cdot \pi_0(\beta, \sigma^2)$. This starting point can be easily obtained in practice, either by MCMC or via a geometric path from the prior distribution. The Markov kernels are chosen to be Gibbs samplers, alternating draws of $\beta \mid \sigma^2$ and $\sigma^2 \mid \beta$. We conducted $1,000$ SMC runs, each adaptively choosing a path with $\mathcal{E} = 1/2$. Observation ordering was permuted randomly between each trial to assess the sensitivity of the procedure to the ordering.

The results of the simulation experiment are displayed in Figure 3.2. The number of steps required by the adaptive procedure grows logarithmically with the number of observations, providing evidence of the efficiency of the data tempering approach.

This advantage comes at a cost; a key difference between tempering and data tempering is that the data tempering approach is more likely to fail at controlling the L_2 distance. This occurs when the next observation in the tempering sequence results in a transition with $\hat{E}_{s,1} \leqslant \mathcal{E}$ leading to uncontrolled error. Across all experiments, nearly 5% of the steps resulted in no candidate distributions satisfying $\hat{E}_{s,1} \geqslant \mathcal{E}$; this tends to occur when moving to a high-leverage point, which results in large changes to the posterior. The relative error of these steps is shown in Figure 3.2(c). These steps are characterized by large L_2 distances, which are often catastrophically underestimated.

This problem occurs because sequential introduction of data points has effectively established an SMC step size that is too large, leading to an insufficiently rich set of possible paths. To address this problem, we present a hybrid path that combines the computational advantages of the data tempering with the rich set of paths afforded by tempering. This hybrid path generally ensures that a satisfactory transition can be made at each step of the algorithm and is specified as follows.

Hybrid path for sequential data: Assume the same setting as the data tempered path and suppose $\mu_{s-1}(x) \propto p(y_{1:k_{s-1}}|x)\pi_0(x)$. First, consider the move to $\nu_{s,1}(x) \propto p(y_{1:k_{s-1}+1}|x) \cdot \pi_0(x)$. If $\hat{E}_{s,1} \geqslant \mathcal{E}$, consider additional candidate distributions using the data tempering approach. If not, choose candidate distributions in the same manner as the geometric path, selecting from the family $\eta \propto p(y_{1:k_{s-1}} \mid x)p(y_k \mid x)^{\beta}\pi_0(x)$ for $\beta \in [0, 1]$. Several tempering steps may be required to reach $p(y_{1:k_s+1} \mid x)$, at which time we again consider both data-tempering and tempering moves.

The hybrid path provides a solution to the problem of unacceptably large transitions induced by influential data points. The tempering steps allow for smaller changes in the posterior, effectively allowing for fractional data points in order to

refine the step size. Applying the hybrid approach to the Bayesian linear regression example results in the adaptive criteria being satisfied at each step, with no failed transitions and accurate estimation of the L_2 distances. These paths are shown in Figure 3.2(d). Somewhat surprisingly, these paths tend to be shorter than those from the data tempered approach. This is due to a cascading effect; following a poor transition there is increased error in the estimation of the L_2 distance, resulting in unnecessarily conservative transitions.

3.5 Conclusion

The results presented in this chapter demonstrate the importance of path selection in the design of SMC sampling algorithms. Proper selection of paths has the potential to dramatically improve the efficiency of SMC; for some target distributions the resulting bounds have lower complexity than those for MCMC. When an efficient path is not known prior to running the algorithm, adequate paths can in some cases be obtained during sampling by adaptively choosing steps using the RESS as an estimate of the L_2 distance. In the examples presented here this leads to near-optimal paths, although we currently have no theoretical guarantees of optimality, only sufficiency. Developing conditions under which this algorithm achieves near-optimal paths would lend further weight to the method.

While the adaptive approach provides an efficient method for selecting SMC steps, choosing a good starting distribution is an equally important aspect of path selection. The theoretical and practical examples in this paper used easily available starting distributions that are unnecessarily far from the target distribution. Selecting an initial distribution that closely approximates π could lead to substantial reductions in the number of steps required to reach the target distribution. Approaches like Laplace approximation or variational Bayesian methods may provide efficient starting distributions with minimal additional effort [60].

The final component of path selection is choosing an appropriate family of interpolating distributions. The tempering, data tempering and hybrid approaches presented in this chapter are relatively generic and applicable in many situations, however, as shown in our Gaussian example, meaningful improvements in complexity can be provided by other types of paths. These families, such as the geometric-tempered family, may not be amenable to selection criteria (3.7) when there is no clear preference among the candidate distributions. Developing path selection criteria for these more complicated families could open the door for substantially faster SMC algorithms.

3.6 Proof of Theorem 13

The proof follows closely the approach in [50]. The key is to ensure that after each step of the algorithm the following holds with high probability

$$\begin{aligned} &\mathbf{C}_s(\text{i}) \quad X_s^n \sim \mu_s \text{ for } n = 1, \ldots, N \\ &\mathbf{C}_s(\text{ii}) \quad \bar{w}_{s+1} \geqslant \mu_s(w_{s+1}) \cdot \frac{2}{3} \end{aligned} \tag{3.9}$$

This is called the one-step induction condition. When $\mathbf{C}_s$(i) and $\mathbf{C}_s$(ii) hold, the marginal distribution of the re-sampled particles is 2-warm for μ_{s+1} by Lemma 3 of [50]. Then, using a coupling argument and martingale concentration inequalities, conditions $\mathbf{C}_{s+1}$(i) and $\mathbf{C}_{s+1}$(ii) can be shown to hold using Corollary 3.1 and Lemma 4 of [50], respectively. Inductively applying these steps leads to the proof of Theorem 1 in [50].

We modify this proof by showing an alternate condition under which $\mathbf{C}_s$(ii) holds. This result uses a different kind of martingale concentration and replaces Lemma 4 of [50].

Lemma 16 (Lower bound on the average weights)**.**
Fix any $0 < \delta_0 < 1$ *and choose* $N \geqslant 81 \log(2/\delta_0) \cdot \|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})}$. *Then, conditional*

on $\boldsymbol{C}_{s-1}(i)$,

$$Pr\Big(\bar{w}_s \geqslant \mu_{s-1}(w_s) \cdot 2/3\Big) \geqslant 1 - \delta_0/2$$

Proof. Let $Y^n = \min\{w_s\big(X_s^n\big),\, \alpha\}$ with $\alpha = \mu_{s-1}(w_s^2)/6\mu_{s-1}(w_s)$ and sample mean $\bar{Y}$. Define the residual errors to be $\mathcal{Y}^n = \sum_{i=1}^{n} \big(Y^n - \mathsf{E}Y^n\big)$. Using the same approach as in Lemma 4 of [50] we can show that $\mathcal{Y}^n$ is a zero-mean martingale. The martingale has increments bounded by α as $\mathcal{Y}^{n-1} - \mathcal{Y}^n \leqslant EY_n \leqslant \alpha$. The variance of each increment is bounded as follows:

$$\begin{aligned} \mathrm{Var}(\mathcal{Y}^n|\mathcal{Y}^{n-1}) &= \mathrm{Var}(Y^n - \mathsf{E}Y^n|\mathcal{Y}^{n-1}) \\ &\leqslant \mathrm{Var}(Y^n - \mathsf{E}Y^n) \\ &\leqslant \mu_{s-1}(w_s^2) \end{aligned} \tag{3.10}$$

These conditions verify the requirements of Theorem 22 in [61]. This gives

$$\Pr\Big(\mathcal{Y}^N/N \leqslant -\epsilon\Big) \leqslant \exp\left(\frac{N\epsilon^2}{\mu_{s-1}(w_s^2) \cdot \big(2 + \epsilon/18\mu_{s-1}(w_s)\big)}\right) \tag{3.11}$$

Therefore choosing $\epsilon = \mu_{s-1}(w_s)/6$ and $N \geqslant 81\log(2/\delta_0)/\mathcal{E}$ ensures $\Pr\big(\bar{Y} - \mathsf{E}Y_1 \geqslant -\mu_{s-1}(w_s)/6\big) \geqslant 1 - \delta_0/2$. The condition $\bar{Y} - \mathsf{E}Y_1 \geqslant -\mu_{s-1}(w_s)/6$ is sufficient to show that the claim holds with probability as least $1 - \delta_0$:

$$\begin{aligned} \bar{w}_{s+1} &\geqslant \bar{Y} \\ &\geqslant \mathsf{E}Y_1 - \mu_{s-1}(w_s)/6 \\ &\geqslant \mu_{s-1}(w_s) - \mu_{s-1}(w_s^2)/\alpha - \mu_{s-1}(w_s)/6 \\ &\geqslant \mu_{s-1}(w_s) \cdot 2/3 \end{aligned} \tag{3.12}$$

The third line follows from Lemma 3.9 of [40] and the final line follows from the choice of α. ■

This inequality can be used to prove the following one step induction condition, the proof of which is the same as corollary 4.1 of [50]

Corollary 17 (One step induction condition)**.**

Assume $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geqslant 3/4$. *Fix* $0 < \delta_0 < 1$. *Choose* $t \geqslant \tau_s\big(\frac{\delta_0}{2N}, 2\big)$ *and* $N \geqslant 81\log(2/\delta_0)\cdot\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})}$. *Then the inductive condition* $\boldsymbol{C}_s|\,\boldsymbol{C}_{s-1}$ *holds with probability at least* $1-\delta_0$

The proof of Theorem 13 follows immediately as in [50] using the new one step induction condition.

3.7 Proof of Gaussian example

First, we derive the L_2 distance from $\eta \sim N(\theta_0, \phi_0^{-1})$ to $\mu \sim N(\theta_1, \phi_1^{-1})$ on $\mathcal{R}$, assuming $2\phi_1 \geqslant \phi_0$. We have

$$\|\mu/\eta\|_{L_2(\eta)} = \frac{\phi_1}{\phi_0^{1/2}} \int \frac{1}{\sqrt{2\pi}} \exp\big(-0.5\big[2\phi_1(x-\theta_1)^2 - \phi_0(x-\theta_0)^2\big]\big) \tag{3.13}$$

Let $\phi^* = 2\phi_1 - \phi_0$ and $\theta^* = 2\phi_1\theta_1 - \phi_0\theta_0$ and complete the square inside the exponential function.

$$\begin{aligned} 2\phi_1(x-\theta_1)^2 - \phi_0(x-\theta_0)^2 &= \phi^*\big(x-\theta^*/\phi^*\big)^2 + 2\phi_1\theta_1^2 - \phi_0\theta_0^2 - \frac{\theta^{*2}}{\phi^*} \\ &= \phi^*\big(x-\theta^*/\phi^*\big)^2 - \frac{2\phi_1\phi_0}{\phi^*}(\theta_1-\theta_0)^2 \end{aligned} \tag{3.14}$$

Inserting (3.14) into (3.13) and substituting $\psi = \phi_1/\phi_0$ gives

$$\|\mu/\eta\|_{L_2(\eta)} = \frac{\psi}{\sqrt{2\psi-1}} \exp\Big(\frac{\phi_1}{2\psi-1}(\theta_1-\theta_0)^2\Big)$$

The L_2 distance for spherical, d-dimensional Gaussians follows immediately:

$$\|\mu/\eta\|_{L_2(\eta)} = \Big(\frac{\psi^2}{2\psi-1}\Big)^{d/2} \exp\Big(\frac{d\phi_1}{2\psi-1}(\theta_1-\theta_0)^2\Big) \tag{3.15}$$

3.7.1 Geometric path

The geometric path from section 3.2.1 consists of a sequence of Gaussian distributions with $\mu_{\beta_s}(x) = N\big(1_d\cdot\theta_s\,,\, I_d/\phi_s\big)$ where $\theta_s = \theta\cdot\beta_s/\phi_s$ and $\phi_s = \beta_s(\phi-1)+1$. We

remind the reader that $\phi > 1$ and $\theta \geqslant 2$ and proceed to bound the L_2 distance by separately bounding the factors in (3.15). Define $\psi_s = \phi_s/\phi_{s-1}$. For $s = 1$:d

$$\begin{aligned} 1 < \psi_1 &= 1 + \frac{\phi - 1}{\phi \cdot \theta\sqrt{d}} \\ &\leqslant 1 + \frac{2}{\theta\sqrt{d}} \end{aligned} \tag{3.16}$$

and when $s > 1$:

$$\begin{aligned} 1 \leqslant \psi_s &= \frac{\beta_s(\phi - 1) + 1}{\beta_{s-1}(\phi - 1) + 1} \\ &= 1 + \frac{(\beta_s - \beta_{s-1})(\phi - 1)}{\beta_{s-1}(\phi - 1) + 1} \\ &\leqslant 1 + \frac{(\beta_s - \beta_{s-1})}{\beta_{s-1}} \\ &= 1 + \frac{2}{\theta\sqrt{d}} \end{aligned} \tag{3.17}$$

Plugging this into the factor $\left(\frac{\psi^2}{2\psi - 1}\right)^{d/2}$ in the L_2 distance (3.15) gives

$$\begin{aligned} \left(\frac{\psi_s^2}{2\psi_s - 1}\right)^{d/2} &\leqslant \left(\frac{\left(1 + \frac{2}{\theta\sqrt{d}}\right)^2}{\left(1 + \frac{4}{\theta\sqrt{d}}\right)}\right)^{d/2} \\ &\leqslant \left(1 + \frac{1}{d}\right)^{d/2} \\ &\leqslant 2 \end{aligned} \tag{3.18}$$

where the second line uses $\theta \geqslant 2$. To bound the second factor in (3.15) we bound the difference in means. For $s = 1$, $\theta_1 - \theta_0 = \frac{1}{\phi\sqrt{d} \cdot \phi_1} \leqslant \frac{1}{\sqrt{d}\phi_1}$ and consequently

$\exp\left(\frac{d\phi_1}{2\psi_1-1}(\theta_1-\theta_0)^2\right) \leqslant e$. For $s>1$:

$$\begin{aligned}\theta_s-\theta_{s-1} &= \theta\Big(\frac{\beta_s}{\phi_s}-\frac{\beta_{s-1}}{\phi_{s-1}}\Big)\\ &= \frac{\theta\cdot\beta_{s-1}}{\phi_{s-1}\phi_s}\left(\Big(1+\frac{2}{\theta\sqrt{d}}\Big)\phi_{s-1}-\phi_s\right)\\ &= \frac{2\beta_{s-1}}{\phi_{s-1}\phi_s\sqrt{d}}\end{aligned} \tag{3.19}$$

Inserting this result into the second term in (3.15) gives

$$\begin{aligned}\exp\left(\frac{d\phi_1}{2\psi-1}(\theta_1-\theta_0)^2\right) &= \exp\left(\frac{4\beta_{s-1}^2}{2\phi_s^2\phi_{s-1}-\phi_s\phi_{s-1}^2}\right)\\ &\leqslant \exp(4)\end{aligned} \tag{3.20}$$

The first line follows using $1\leqslant\phi_{s-1}\leqslant\phi_s$ and $\beta_s\leqslant 1$. Inserting (3.18) and (3.19) into (3.15) shows that for the geometric path $1/\mathcal{E}=\mathcal{O}(1)$ proving (3.4).

3.7.2 Precision path

This path is specified by a sequence of normal distributions $\mu_s=N_d(\theta_s,I_d/\phi_s)$. The location parameter is $\theta_s=0$ for $s\leqslant s_1=\lceil 3\sqrt{d}\log(d\theta^2)\rceil$ and $\theta_s=1_d\theta$ otherwise. The precisions are given by

$$\phi_s=\begin{cases}\big(1-\frac{1}{\sqrt{9d}}\big)^s\vee\frac{1}{d\theta^2}, & \text{if } 0\leqslant s\leqslant s_1\\ \frac{1}{d\theta^2}\big(1+\frac{1}{\sqrt{d}}\big)^{s-s_1-1}\wedge\phi, & \text{otherwise}\end{cases} \tag{3.21}$$

Let $\psi_s=\phi_s/\phi_{s-1}$. When $s\leqslant s_1$, $1\geqslant\psi_s\geqslant\big(1-\frac{1}{\sqrt{9d}}\big)$ and therefore

$$\begin{aligned}\|\mu_s\|_{L_2)(\mu_{s-1})} &= \Big(\frac{\psi_s^2}{2\psi_s-1}\Big)^{d/2}\\ &\leqslant \left(\frac{\big(1-\frac{1}{\sqrt{9d}}\big)^2}{\big(1-\frac{2}{\sqrt{9d}}\big)}\right)^{d/2}\\ &\leqslant \Big(1+\frac{1}{d}\Big)^{d/2}\\ &\leqslant 2\end{aligned} \tag{3.22}$$

The same approach shows that $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant 2$ for $s \geqslant s_1+2$. When $s = s_1+1$, $\phi_s = 1$ and therefore $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} = 1$. therefore $1/\mathcal{E} \leqslant 2$, proving (3.5).

3.8 Proof of Theorem 15

The proof of Theorem 2 follows in a similar manner to Theorem 13. To incorporate the adaptive selection of distributions, we add the following condition in addition to those in (3.9)

$$\mathbf{C}_s(\text{iii}) \quad \|\mu_{s+1,n}/\mu_s\|_{L_2(\mu_s)} \leqslant 3/\hat{E}_{s+1,m} \tag{3.23}$$

This condition will be used to ensure that, for any adaptively chosen distribution $\mu_s = \nu_{s,m}$, that $\mathbf{C}_s$(ii) holds with high probability using Lemma (16). This allows the construction of an adaptive version of the one step induction condition, which will then be used to prove Theorem 15. Before proving the inductive condition, we give conditions under which $\mathbf{C}_s$(iii) holds for any $\nu_{s,m}$ with high-probability.

Lemma 18 (Adaptive selection of distributions)**.**
Suppose at the beginning of step s we have m candidate distributions $\nu_{s,1}, \ldots, \nu_{s,M}$ and that condition $\boldsymbol{C}_{s-1}(i)$ holds. Let $w_{s,m} \propto \nu_{s,m}(x)$ and assume that $0 < w_{s,m} \leqslant 1$ for each m and let $\bar{w}^2_{s,m} = \frac{1}{N}\sum_{n=1}^{N} w_{s,m}\big(X^n_{s-1}\big)^2$. Choose any $0 < C < 1$ and $0 < \delta_0 < 1$. Then for $N \geqslant \frac{32}{C^2} \cdot \log\big(\frac{12M}{\delta_0}\big)$ and every m s.t. $\bar{w}^2_{s,m} \geqslant C$:

$$Pr\Big(\|\nu_{s,m}/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant 3/\hat{E}_{s,m}\Big) \geqslant 1 - \delta_0/3$$

Proof. Azuma's inequality ensures that for this choice of N and any $f \in \mathcal{F}$ with $|f| \leqslant 1$ that $\Pr\big(|\bar{f} - \mu_{s-1}f| \geqslant C/4\big) \leqslant \delta_0/6M$. Applying a union bound gives the following

$$\begin{aligned}
&\Pr\Bigg(\bigcup_{m=1}^{M} \big\{|\bar{w}_{s,m} - \mu_{s-1}(w_{s,m})| \geqslant C/4\big\} \cup \big\{|\bar{w}^2_{s,m} - \mu_{s-1}(w^2_{s,m})| \geqslant C/4\Big) \\
&\leqslant \max_m \Pr\Big(|\bar{w}_{s,m} - \mu_{s-1}(w_{s,m})| \geqslant C/4\Big) \bigvee \Pr\Big(|\bar{w}^2_{s,m} - \mu_{s-1}(w^2_{s,m})| \geqslant C/4\Big) \\
&\leqslant \delta_0/3
\end{aligned} \tag{3.24}$$

Now we show that if $\bar{w}_{s,m}^2 \geqslant C$ that it estimates $\mu_{s-1}(w_{s,m}^2)$ with small relative error. Assume that $|\bar{w}_{s,m}^2 - \mu_{s-1}(w_{s,m}^2)| \leqslant C/4$. Then $\mu_{s-1}(w_{s,m}^2) \geqslant \bar{w}_{s,m}^2 - C/4 \geqslant C \cdot 3/4$ and therefore

$$|\bar{w}_{s,m}^2 - \mu_{s-1}(w_{s,m}^2)| \leqslant C/4 \leqslant \mu_{s-1}(w_{s,m}^2)/3 \tag{3.25}$$

A similar argument gives $|\bar{w}_{s,m} - \mu_{s-1}(w_{s,m})| \leqslant \mu_{s-1}(w_{s,m})/3$, noticing that $C \leqslant \bar{w}_{s,m}^2 \leqslant \bar{w}_{s,m}$ as $0 < w_{s,m} < 1$. This shows that

$$\frac{1}{\hat{E}_{s,m}} \geqslant \frac{3\mu_{s-1}(w_{s,m}^2)}{8\big(\mu_{s-1}(w_{s,m})\big)^2} \geqslant \frac{1}{3}\|\nu_{s,m}/\mu_{s-1}\|_{L_2(\mu_{s-1})} \tag{3.26}$$

Combining (3.24) and (3.26) gives the result. ■

The adaptive one step induction condition follows from combining Lemma 18 with Corollary 17.

Lemma 19 (Adaptive one step induction condition)**.**
Suppose at the beginning of step s we have M candidate distributions $\nu_{s,1}, \ldots, \nu_{s,M}$ with $0 < w_{s,m} \leqslant 1$ and that condition $\boldsymbol{C}_{s-1}(i)$ holds. Fix a lower bound $0 < C < 1$, a target effective sample size $0 < \mathcal{E} < 1$, and a probability $0 < \delta_0 < 1$. Choose:

$$\mu_s = \max_m \left\{\nu_{s,m} : \frac{1}{\hat{E}_{s,m}} \leqslant \frac{1}{\mathcal{E}} \text{ and } \bar{w}_{s,m}^2 \geqslant C\right\} \tag{3.27}$$

Then for $N \geqslant \frac{32}{C^2} \cdot \log\left(\frac{12M}{\delta_0}\right) \vee \frac{243}{\mathcal{E}} \cdot \log\left(\frac{3}{\delta_0}\right)$ and $t \geqslant \tau_{s+1}\left(\frac{\delta_0}{3N}, 2\right)$

$$Pr\Big(\boldsymbol{C}_s(i) \mid \boldsymbol{C}_{s-1}(i)\Big) \geqslant 1 - \delta_0$$

Proof. For this choice of μ_s it follows from Lemma 18 that:

$$\Pr\big(\mathbf{C}_s(\text{iii}) \mid \mathbf{C}_{s-1}(\text{i})\big) \geqslant 1 - \delta_0/3$$

This allows us to apply Corollary 17 with $\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant 3/\mathcal{E}$.

$$\begin{aligned}\Pr\Big(\mathbf{C}_s(\mathrm{i}) \mid \mathbf{C}_{s-1}(\mathrm{i})\Big) &\geqslant \Pr\Big(\mathbf{C}_s(\mathrm{i}) \cap \mathbf{C}_{s-1}(\mathrm{ii}) \cap \mathbf{C}_{s-1}(\mathrm{iii}) \mid \mathbf{C}_{s-1}(\mathrm{i})\Big) \\ &\geqslant \Pr\Big(\mathbf{C}_s(\mathrm{i}) \cap \mathbf{C}_{s-1}(\mathrm{ii}) \mid \mathbf{C}_{s-1}(\mathrm{iii}),\ \mathbf{C}_{s-1}(\mathrm{i})\Big) \\ &\quad \cdot \Pr\Big(\mathbf{C}_{s-1}(\mathrm{iii}) \mid \mathbf{C}_{s-1}(\mathrm{i})\Big) \\ &\geqslant 1 - \delta_0\end{aligned} \tag{3.28}$$

■

The proof of Theorem 15 follows by combining Theorem 13 and Corollary 19.

Proof of Theorem 15

Proof. Let $\delta_p = \frac{1}{8}\sum_{i=0}^{p-1} 2^{-i} < \frac{1}{4}$ and define the inductive epoch condition:

$$\Pr\big(\mathbf{C}_{Sp}\big) \geqslant 1 - \delta_p \tag{3.29}$$

This condition will ensure that the result holds for any $s \leqslant Sp$. To show this holds for $p = 1$, use the same approach as in Theorem 13, replacing Corollary 17 with Corollary 19. We now prove that the inductive step holds.

Suppose that at the end of epoch $p-1$ we have not reached the target distribution and that (3.29) holds. First, we show that condition $\tilde{\mathbf{C}}_{S(p-1)}$ holds for the new particle approximation with probability at least $1 - \frac{1}{8}2^{-p}$ for the specified N_p and t_p. The tilde notation is used to distinguish between the induction conditions for the current particle system and the new particle system. This follows from Theorem 13 conditional on (3.29), suitably modified for this probability, noting that (3.29) ensures $\max_{s=1,\ldots,S(p-1)} \|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} \leqslant 3/\mathcal{E}$. Therefore

$$\begin{aligned}\Pr\big(\tilde{\mathbf{C}}_{S(p-1)} \cap \mathbf{C}_{S(p-1)}\big) &= \Pr\big(\tilde{\mathbf{C}}_{S(p-1)} \mid \mathbf{C}_{S(p-1)}\big)\Pr\big(\mathbf{C}_{S(p-1)}\big) \\ &\geqslant 1 - \delta_{p-1} - \frac{1}{8}2^{-p}\end{aligned} \tag{3.30}$$

As a consequence, the combined particle system also satisfies the induction condition. The next step is to show that the combined particle approximation with $N = \sum_{i=1}^{p} N_i$ particles using $t = t_p$ Markov kernel transitions ensures the stability of the adaptive approach for next S steps. The chosen values ensure that Corollary 19 holds with $\delta_0 = \frac{1}{8}2^{-p}/S$. Then for s in $S(p-1)+1, \ldots, Sp$:

$$\begin{aligned}
\Pr\big(\mathbf{C}_s\big) &\geqslant \prod_{r=S(p-1)+1}^{s} \Pr\big(\mathbf{C}_r \mid \mathbf{C}_r\big) \cdot \Pr\big(\mathbf{C}_{S(p-1)+1} \mid \tilde{\mathbf{C}}_{S(p-1)} \cup \mathbf{C}_{S(p-1)}\big) \cdot \\
&\qquad \Pr\big(\tilde{\mathbf{C}}_{S(p-1)} \cap \mathbf{C}_{S(p-1)}\big) \\
&\geqslant \big(1 - \frac{1}{8}2^{-p}/S\big)^s \cdot \big(1 - \delta_{p-1} - \frac{1}{8}2^{-p}\big) \\
&\geqslant 1 - \delta_p
\end{aligned} \tag{3.31}$$

This verifies that (3.29) holds for epoch p, completing the inductive proof. If the algorithm terminates during epoch p, conditional on the induction condition, the chosen $N = \sum_{i=1}^{p} N_i$ and t_p ensure that the result holds (see [50] Theorem 1 for additional details). ■

3.9 L2 distance for linear regression example

The specified Bayesian linear model leads to a Normal Inverse-Gamma posterior distribution with $\mu_s(\beta, \sigma^2 \mid X_{1:k_s}, Y_{1:k_s}) = \mathcal{N}(\beta \mid \theta_s, \sigma_s^2 \Sigma_s) \cdot \text{Inv-Gamma}(\sigma^2 \mid a_s, b_s)$ where

$$\begin{aligned}
\Sigma_s &= \big(\Sigma_0 + X_{1:k_s}^T X_{1:k_s}\big)^{-1} \\
\theta_s &= \Sigma_s X_{1:k_s}^T Y_{1:k_s} \\
a_s &= 4 + k_s/2 \\
b_s &= 4 + \frac{1}{2}\big(Y_{1:k_s}^T Y_{1:k_s} - \theta_s^T \Sigma_S^{-1} \theta_s\big)
\end{aligned} \tag{3.32}$$

and $\Sigma_0 = (X_{1:K}^T X_{1:K})^{-1}/K$. The L_2 distance is:

$$\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} = \int \frac{\mathcal{N}^2(\beta \mid \theta_{s-1}, \sigma^2\Sigma_{s-1}) \cdot \text{Inv-gamma}^2(\sigma^2 \mid a_{s-1}, b_{s-1})}{\mathcal{N}(\beta \mid \theta_s, \sigma^2\Sigma_s) \cdot \text{Inv-gamma}(\sigma^2 \mid a_s, b_s)} d\beta d\sigma^2 \tag{3.33}$$

The conditional normal distribution on β can be integrated out by completing the square:

$$\begin{aligned}\int \frac{\mathcal{N}^2(\beta \mid \theta_{s-1}, \sigma^2\Sigma_{s-1})}{\mathcal{N}(\beta \mid \theta_s, \sigma^2\Sigma_s)} d\beta &= \int \frac{\exp\Big(-\frac{1}{\sigma^2}(\beta-\theta_{s-1})^T\Sigma_{s-1}^{-1}(\beta-\theta_{s-1})\Big)}{\exp\Big(-\frac{1}{2\sigma^2}(\beta-\theta_s)^T\Sigma_s^{-1}(\beta-\theta_s)\Big)} \cdot \\ &\qquad \frac{|2\pi\sigma^2\Sigma_{s-1}|^{-1}}{|2\pi\sigma^2\Sigma_s|^{-1/2}} d\beta \\ &= \frac{|\Sigma_s|^{1/2}|\Sigma_*|^{1/2}}{|\Sigma_{s-1}|} \exp\Big(-\frac{b_*}{\sigma^2}\Big)\end{aligned} \tag{3.34}$$

where:

$$\begin{aligned}\Sigma_* &= \big(2\Sigma_{s-1}^{-1} - \Sigma_s^{-1}\big)^{-1} \\ \mu_* &= \Sigma_*\big(2\Sigma_{s-1}^{-1}\theta_{s-1} - \Sigma_s^{-1}\theta_s\big) \\ b_* &= \frac{1}{2}\big[2\theta_{s-1}^T\Sigma_{s-1}^{-1}\theta_{s-1} - \theta_s^T\Sigma_s^{-1}\theta_s - \theta_*^T\Sigma_*^{-1}\theta_*\big]\end{aligned} \tag{3.35}$$

The L_2 distance can be found by integrating the resulting unnormalized gamma pdf:

$$\begin{aligned}\|\mu_s/\mu_{s-1}\|_{L_2(\mu_{s-1})} &= \int \frac{\text{Inv-gamma}^2(\sigma^2 \mid a_{s-1}, b_{s-1})}{\text{Inv-gamma}(\sigma^2 \mid a_s, b_s)} \cdot \frac{|\Sigma_s|^{1/2}|\Sigma_*|^{1/2}}{|\Sigma_{s-1}|} \exp\Big(-\frac{b_*}{\sigma^2}\Big) d\sigma^2 \\ &= \frac{|\Sigma_s|^{1/2}|\Sigma_*|^{1/2}}{|\Sigma_{s-1}|} \cdot \frac{b_{s-1}^{2a_{s-1}}}{b_s^{a_s}(b_* + 2b_{s-1} - b_s)^{2a_{s-1}-a_s}} \cdot \frac{\Gamma(a_s)\Gamma(2a_{s-1}-a_s)}{\Gamma(a_{s-1})^2}\end{aligned} \tag{3.36}$$

4

Path Selection for Path Sampling

4.1 Introduction

The problem of computing ratios of normalizing contexts arises in a variety of scientific disciplines, including statistics, computational chemistry, and statistical physics. Consider two probability distributions $p_1(x) = q_t(x)/z_t$ for $t = 0, 1$ on a common probability space $\mathcal{X}$ with $q_t(x)$ known. The ratio of normalizing constants is defined as $r = z_1/z_0$ where $z_t = \int q_t(x)dx$. Estimating z_1/z_0 has received considerable attention; typical approaches involve Monte Carlo simulations which may require large amounts of computation in order to obtain estimates with sufficient accuracy. Methods for estimating the ratio, or its logarithm $l = \log r$, include the Bennett acceptance ratio [62], bridge sampling [63, 64], and Chib's method [65, 66]. For particularly difficult problems, these methods are augmented using sequences of distributions connecting p_0 to p_1. These ensemble methods include the multistate Bennett acceptance ratio [67], annealed importance sampling/sequential Monte Carlo [4, 7], and multiple bridge sampling [63, 68]. Comparisons of these approaches are found in [63, 69, 70].

A particularly powerful ensemble method for estimating the ratio is thermodynamic integration [27, 71] which uses a variable augmentation scheme to embed the two distributions in a larger parametric family of distributions $p(x \mid \theta)$ with $\theta \in [0, 1]$ chosen so that $p(x \mid t) = p_t(x)$ for $t = 0, 1$. The thermodynamic identity is given by:

$$l = \int_0^1 \mathsf{E}_\theta \left[\frac{d}{d\theta} \log q(x \mid \theta) \right] d\theta \tag{4.1}$$

where $q(x \mid \theta)$ is the unnormalized density of $p(x \mid \theta)$ and E_θ is it's expectation. This integral is typically estimated by choosing a sequence of quadrature points $0 = \theta_0 < \theta_1 < \ldots < \theta_S = 1$, estimating $\mathsf{E}_{\theta_s} \left[\frac{d}{d\theta} \log q(x \mid \theta_s) \right]$ at each point using Monte Carlo methods, and combining these estimates using the trapezoid rule to approximate the integral over θ. Various augmentations have been suggested to improve the efficacy of this scheme [51, 25, 72].

Gelman and Meng noted that thermodynamic integration can be generalized using an approach called *path sampling* [28]. Define $p(x \mid \theta)$ as before except that now $\theta = (\theta^{(1)}, \ldots, \theta^{(D)})$ is a D dimensional parameter. The parameter is assumed to be chosen so that that $\theta \in \Theta = [0, 1]^D$ and that $p(x \mid \theta)$ is a proper probability distribution throughout Θ. Abusing notation, let $\theta : [0, 1] \to [0, 1]^D$ be a smooth curve, called a *path*, satisfying $p\big(x \mid \theta(t)\big) = p_t(x)$ for $t = 0, 1$. The path sampling integral identity is given by:

$$l = \int_0^1 \mathsf{E}_{\theta(t)} \left[\sum_{d=1}^{D} \dot{\theta}^{(d)}(t) \cdot U^{(d)}\big(x \mid \theta(t)\big) \right] dt \tag{4.2}$$

where $\dot{\theta}_d(t) = \frac{d}{dt}\theta^{(d)}(t)$ and $U^{(d)}\big(x \mid \theta\big) = \frac{d}{d\theta^{(d)}} \log q(x \mid \theta)$. The multidimensional approach offers richer families of distributions, which in turn can lead to substantially more efficient algorithms [28, 73].

Unfortunately, the potential advantages of multidimensional path sampling remain largely unexplored. Two factors contribute to this stagnation. The first is that

the specification of advantageous, multidimensional families can be quite difficult. While there are several well known generic path sampling families, the literature lacks guidance on when to apply these paths and it is unclear when a particular family may be well-suited to a problem. In addition, in many situations it may be necessary to choose a path sampling family tailored to the unique properties of p_1 and p_2, making it difficult to establish general guidelines for the application of path sampling. The second challenge is that for any given family, the space of possible paths may be vast and selecting a good path is a non-trivial task. Even when an appropriate family is chosen, poor path specification may lead to estimates with large errors.

In this chapter, we address some of these issues, providing practical methods for implementing path sampling in practice. For one dimensional paths, we present a simple algorithm that finds low variance paths, simplifying the problem of path selection. For multidimensional paths, we give several example path finding families and discuss scenarios where they might be applied. We then present a two-part technique for finding efficient paths given a path sampling family with modest D. The approach is motivated by the well-known connection between Riemann geometry and path sampling [28, 74], and combines Gaussian process (GP) emulation and path finding algorithms on graphs. Due to the efficiency of the algorithm, the number of considered paths can be vast, allowing the comparison of a rich set of possible path estimators. The chapter concludes with several examples demonstrating the efficacy of these approaches.

4.2 Path sampling

In this section, we introduce three factors that should be considered when implementing path sampling: choosing a path sampling family, specifying a path, and estimating the integral (4.1). First, to illustrate the idea of a path sampling fam-

ily, we present several examples that can be applied in a variety of settings. Then, deferring the issue of path selection to the next section, we present two estimation methods using quadrature and Monte Carlo respectively. Finally, we discuss the relationship between the variance of the path sampling estimator and the chosen path.

4.2.1 Path families

For a particular path sampling problem there are many families of distributions that could be considered. While the best families may be problem specific, several generic path sampling families have received attention in the literature. The first is the geometric mixture, a well-known one dimensional family given by:

$$p(x \mid \theta) \propto q_0(x)^{1-\theta^{(1)}} q_1(x)^{\theta^{(1)}} \tag{4.3}$$

This one dimensional family is easy to construct for most problems and for this reason has received substantial attention [75, 76, 25]. An augmentation to this approach, which we called the geometric-tempered family, was proposed by [28]:

$$p(x \mid \theta) \propto q_0(x)^{(1-\theta^{(1)})\theta^{(2)}} q_1(x)^{\theta^{(1)}\theta^{(2)}} \tag{4.4}$$

This family combines the geometric family's ease of construction with possibility of tempering through the inverse-temperature parameter $\theta^{(2)}$. Geometric-tempered families are often suitable for problems where geometric families encounter large-variance estimates for each $\theta^{(1)} \in [0, 1]$. Lowering the inverse temperature provides a means of reducing this variance and can provide substantial improvements in complexity (see Section 4.4.2.)

The final example is the component family, studied in [77, 78, 73]. Suppose the the log-likelihood of both models can be written in the following form:

$$p_t(x) \propto \exp\Big(\sum_{d=1}^{D} h_t^{(d)}(x)\Big)$$

The components $h_t^{(d)}$ are not necessarily assumed to be similar (i.e. $h_0^{(d)}$ may be quite different from $h_1^{(d)}$) and some of the components may be set to 0 ($h_t^{(d)}(x) = 0$ for some d, t). The component family is defined as:

$$p(x \mid \theta) \propto \exp\Big(\sum_{d=1}^{D} (1 - \theta^{(d)}) \cdot h_0^{(d)}(x) + \theta^{(d)} \cdot h_1^{(d)}(x) \Big) \tag{4.5}$$

Component families can be used to isolate interactions in the log-likelihoods that hinder estimation, and are a natural choice when p_0 and p_1 come from the same exponential family.

4.2.2 Estimation

For most problems, the path sampling integral (4.2) is intractable and must be estimated as both the inner expectations (with respect to x) and the outer integral (with respect to t) are unknown. The inner expectations are often estimated using Monte Carlo methods and we assume that a suitable sampling method is available for each $p(x \mid \theta)$. The outer integral is typically approximated by one of two approaches: quadrature or Monte Carlo integration

Monte Carlo: draw $S+1$ time points $t_0, \ldots, t_S \sim \text{Unif}(0, 1)$ and draw N samples $x_s^1, \ldots, x_s^N \sim p\big(x \mid \theta(t_s)\big)$ at each t_s. Let $\bar{U}_s^{(d)} = \frac{1}{N} \sum_{n=1}^{N} U^{(d)}\big(x_s^n \mid \theta(t_s)\big)$ for $d = 1, \ldots, D$. The Monte Carlo path sampling estimate is:

$$\hat{l}_{MC} = \frac{1}{S+1} \sum_{s=0}^{S} \sum_{d=1}^{D} \dot{\theta}^{(d)}(t_s) \cdot \bar{U}_s^{(d)} \tag{4.6}$$

Quadrature: choose $S+1$ points $\theta_0, \ldots, \theta_S \in \Theta$; these points can be thought of as a discretization of the underlying curve. Estimation proceeds by sampling $x_s^1, \ldots, x_s^N \sim p\big(x \mid \theta_s\big)$ and setting $\bar{U}_s^{(d)} = \frac{1}{N} \sum_{n=1}^{N} U^{(d)}\big(x_s^n \mid \theta_s\big)$. The quadrature path sampling

estimate uses the trapezoid rule:

$$\begin{aligned} \hat{l}_Q &= \sum_{s=1}^{S} \hat{l}_s \\ \hat{l}_s &= \sum_{d=1}^{D} \left(\theta_s^d - \theta_{s-1}^d\right) \frac{\bar{U}_s^{(d)} + \bar{U}_{s-1}^{(d)}}{2} \end{aligned} \tag{4.7}$$

The individual $\hat{l}_s$ are called *increments* and they estimate the ratio of normalizing constants between $p(x \mid \theta_s)$ and $p(x \mid \theta_{s-1})$. More complicated quadrature rules can be used as in [51], however the higher order derivatives required by these methods require additional estimation and in many cases cannot be estimated using samples from $p(x \mid \theta)$.

The Monte Carlo estimate will be unbiased if the individual $\bar{U}_s^{(d)}$ are also unbiased, unlike the quadrature estimate. This advantage is offset by the slower Monte Carlo rate of convergence compared to quadrature for low dimensional problems. Another difference is that the Monte Carlo estimator requires the specification of a complete curve $\theta(t)$, whereas the quadrature estimate requires only the selection of a finite set of points. Choosing quadrature points may be easier than specifying a complete curve, and for this reason we focus primarily on the quadrature approach.

4.2.3 Optimal paths

The variance of both estimators depends on the family of distributions and the chosen path. Before writing the variance of the Monte Carlo path sampling estimate, we introduce some notation. For a vector $z \in \mathcal{R}^D$ and positive definite matrix M, define the matrix norm $\|z\|_M = \sqrt{z^T M z}$. Define $g(\theta)$ to be the $D \times D$ matrix with elements $g^{cd}(\theta) = \mathsf{E}_\theta \, U^{(c)}(x \mid \theta) U^{(d)}(x \mid \theta)$. Under independent sampling, the variance of the Monte Carlo path sampling estimate is:

$$\mathrm{Var}\left(\hat{l}_{MC}\right) = \frac{1}{(S+1)N} \left[\int_{0,1} \|\dot{\theta}(t)\|^2_{g\left(\theta(t)\right)} dt - l^2 \right] \tag{4.8}$$

Gelman and Meng [28] established that, under suitable regularity conditions on $p(x \mid \theta)$, (Θ, g) forms a Riemann manifold and the variance of this estimator is minimized when θ is chosen to be the geodesic connecting p_0 to p_1 [28, 79]. The geodesic has a physical interpretation as the shortest or lowest energy curve connecting p_0 to p_1, motivating the description of θ as a path. Finding geodesics is a problem in the calculus of variations and they can be described by the Euler-Lagrange equations or via Hamiltonian flows[80]. Unfortunately, finding these curves using numerical methods remains out of reach, as the quantities in these differential equations are generally intractable and the optimal initial conditions are unknown.

The variance of the quadrature estimate can be written in a manner similar to (4.8). Define $\mathcal{I}(\theta)$ to be the Fisher information matrix with elements $\mathcal{I}^{c,d}(\theta) = g^{c,d}(\theta) - E_\theta U^{(c)}(x \mid \theta) \cdot E_\theta U^{(d)}(x \mid \theta)$ and let $\Delta_s = \theta_s - \theta_{s-1}$ be the vector of grid spacings/step sizes. Assuming independent sampling, the variance of the quadrature path estimator is:

$$\mathrm{Var}\big(\hat{l}_Q\big) = \frac{1}{N} \sum_{s=1}^{S} \frac{||\Delta_s||^2_{\mathcal{I}(\theta_s)} + ||\Delta_s||^2_{\mathcal{I}(\theta_{s-1})}}{4} \tag{4.9}$$

Shenfield et al. [74] showed that as $S \to \infty$, the variance minimizing path chooses points so that $||\Delta_s||^2_{\mathcal{I}(\theta_s)} + ||\Delta_s||^2_{\mathcal{I}(\theta_{s-1})}$ is equal for each s. In addition, they showed that these points lie along the geodesic of the $(\Theta, \mathcal{I})$ manifold, with points spaced evenly in energy, so it seems sensible to focus on finding the geodesic in both settings.

4.3 Path selection

Having specified a path sampling family, we present two methods for selecting a path motivated by the optimal paths described in section 4.2.3. When $D = 1$, it is clear that any optimal path should be monotonic and the problem of choosing a path can be reduced to a sequence of step size selection problems. For this setting, we

suggest a simple importance sampling approach that greedily chooses a path which maintains a constant energy/variance at each step.

Selecting a path when $D > 1$ presents a greater challenge, as we can no longer rely on the intuition that step sizes are monotonic. For example. when using a geometric-tempered family, the inverse temperature parameter may be initially decrease from 1 to 0 but must eventually return to it's starting place. Another complicating factor for multidimensional path selection is that moves that have small local variance may actually increase the overall variance of the estimator if they move towards a high variance region. Any reasonable multidimensional path selection method must anticipate these potentially dangerous moves and weigh the short term improvements against long term costs.

For this reason, our approach uses Gaussian process emulation to estimate $\mathcal{I}(\theta)$ throughout the Θ. We use this emulator to evaluate the cost of a large number of prospective paths, represented via a directed, acyclic graph (DAG), and find the lowest variance path using topological sorting.

4.3.1 One dimensional path selection

Throughout this section, we drop the superscript on θ and U used to denote the dimension. For one dimensional paths, we follow the asymptotically optimal strategy of maintaining a constant energy/variance at each step. The algorithm begins with N samples from $p_0(x) = p(x \mid \theta_0)$. Assume that at time $s-1$ we have $x_{s-1}^1, \ldots, x_{s-1}^N \sim p(x \mid \theta_{s-1})$. The next value in the path is given by choosing a step size Δ_s and setting $\theta_s = \theta_{s-1} + \Delta_s$. The step size is chosen to preserve a constant unit variance for each increment:

$$\begin{aligned} \mathrm{Var}(\hat{l}_s) \cdot N &= \frac{||\Delta_s||_{\mathcal{I}(\theta_s)} + ||\Delta_s||_{\mathcal{I}(\theta_{s-1})}}{4} \\ &= \gamma \end{aligned} \tag{4.10}$$

where $\gamma > 0$ is a user specified target variance. Generally, larger values of γ lead to shorter paths with larger variances, whereas smaller values of γ will require larger samples sizes in order to estimate $\mathrm{Var}(\hat{l}_s)$ with good relative accuracy. The variance of a potential step can be estimated with particles from $p(x \mid \theta_{s-1})$ using the sample variance of $U(x \mid \theta_{s-1})$ and the importance sampling estimate of the variance at $p(x \mid \theta_{s-1} + \Delta_s)$:

$$\begin{aligned}
\hat{\gamma} &= \frac{\Delta_s^2}{4(N-1)} \sum_{n=1}^{N} \left(U(x_{s-1}^n \mid \theta_{s-1}) - \bar{U}_{s-1}\right)^2 + w_{s-1}^n(\Delta_s) \cdot \left(U(x_{s-1}^n \mid \theta_{s-1} + \Delta_s) - \bar{U}_s\right)^2 \\
\bar{U}_{s-1} &= \frac{1}{N} \sum_{n=1}^{N} U(x_{s-1}^n \mid \theta_{s-1}) \\
\bar{U}_s &= \frac{1}{N} \sum_{n=1}^{N} w_{s-1}^n(\Delta_s) \cdot U(x_{s-1}^n \mid \theta_{s-1} + \Delta_s) \\
w_{s-1}^n(\Delta_s) &= N \cdot \frac{p(x_{s-1}^n \mid \theta_{s-1} + \Delta_s)}{p(x_{s-1}^n \mid \theta_{s-1})} \Big/ \sum_{m=1}^{N} \frac{p(x_{s-1}^m \mid \theta_{s-1} + \Delta_s)}{p(x_{s-1}^m \mid \theta_{s-1})}
\end{aligned} \tag{4.11}$$

A step size can be efficiently chosen to ensure $\hat{\gamma} = \gamma$ using binary search. This greedy stepsize selection is particularly well-suited to sequential Monte Carlo, where a similar procedure is often used to select sequences of interpolating distributions [16, 25, 81]. The efficacy of this approach is demonstrated on a challenging mean-field Ising model example in Section 4.4.1.

Geometric mixtures

One dimensional path selection is particularly well suited to geometric mixtures. First, in this case the potential and important weights have a simple form:

$$\begin{aligned}
U(x \mid \theta) &= \log q_1(x) - \log q_0(x) \\
w_{s-1}^n(\Delta_s) &\propto \exp\left(\Delta_s U(x \mid \theta)\right)
\end{aligned} \tag{4.12}$$

This means that U can be precomputed for each sample before evaluating (4.11) and subsequently $\hat{\gamma}$ can be quickly computed for a broad range of step sizes. In addition, for geometric families, choosing γ sufficiently small will also control the bias of the increments:

$$|\hat{l}_s - l_s| \leqslant \gamma \tag{4.13}$$

Proof is given in section 4.6. The ability to specify target bias and variance through a single parameter is a powerful property, making adaptive step size selection an attractive option when geometric families of distributions are a suitable path sampling choice.

4.3.2 Multidimensional path selection

A different approach is required for selecting multidimensional paths. One possible approach, proposed [74, 82], is to choose a large number of points throughout Θ, sample from $p(x \mid \theta)$ at each of these points, and then estimate $\mathcal{I}(\theta)$ at each location. A path is then selected from among the chosen points using these estimates. In practice, many points are required to specify a sufficiently fine set of paths and the additional cost of estimating $\mathcal{I}(\theta)$ at each point can easily outweigh the computational improvements afforded by an efficient path. For this reason, we propose an approach inspired by the computer emulation literature, where statistical models are used to approximate the output of expensive computer models [83, 84, 85]. The idea is that when if we can estimate $\mathcal{I}(\theta)$ at a small number of carefully chosen points throughout Θ, we can use a model of $\mathcal{I}(\theta)$ to cheaply predict the metric throughout the manifold, enabling the consideration of a much larger set of paths.

Metric estimation

Begin by selecting a small number of design points $\delta_1, \ldots, \delta_M \in \Theta$. Intuitively, the design points should chosen throughout Θ in a "space-filling" fashion in order to min-

imize the predicted error of the estimator [84]. In our examples we use the maximin Latin hypercube design which selects design points to maximize $\min_{i,j} ||\delta_i - \delta_j||_2^2$. Further discussion of space filling designs can be found in [83, 86]. Another consideration when selecting design points is the sampling method used to draw from $p(x \mid \theta)$. Commonly used ensemble sampling methods, such as parallel tempering, population Monte Carlo, or sequential Monte Carlo, require sequences of distributions spaced so that adjacent distributions are close in a suitable sense. When using these methods, it may be necessary to adjust the design points to accommodate this requirement.

Having chosen a set of design points, we estimate $\mathcal{I}(\delta_m)$ using the sample covariance of $\{U^{(d)}(x \mid \delta_m)\}_{d=1}^{D}$ using N draws from $p(x \mid \delta_m)$. These estimates are transformed using a Cholesky decomposition $\hat{\mathcal{I}}(\delta_m) = L(\delta_m)L(\delta_m)^T$, where $L(\delta_m)$ is a lower-triangular matrix. The elements of $L(\theta)$ are be modelled independently as:

$$\begin{aligned} \log L^{d,d}(\theta) &= f^{d,d}(\theta) + \epsilon^{d,d} \\ L^{c,d}(\theta) &= f^{c,d}(\theta) + \epsilon^{c,d} \quad \text{for } c > d \end{aligned} \tag{4.14}$$

where $\mu^{c,d}$ is a mean function and $\epsilon^{c,d}$ is an error term. Other common approaches to modeling collections of covariance matrices include the variance-correlation, spectral, and generalized auto-regressive parameter decompositions (GARP) [87, 88, 89]. Generalizing these approaches for interpolation is complicated; the first two approaches require complex constraints on the parameter spaces and all three approaches induce dependence structure between the elements of $L(\theta)$. The simple Cholesky decomposition considered here allows for independent, unconstrained modeling of the elements of $L(\theta)$ and is sufficient for our purposes.

The mean functions are assigned Gaussian process priors, $f^{c,d}(\theta) \sim \mathcal{GP}\big(0, K\big)$ for $c \geqslant d$. Gaussian process models are a natural prior choice as they are flexible interpolaters, well suited for problems with small numbers of observations. A Gaussian process is defined as a distribution over functions $f(\theta) \sim \mathcal{GP}(\mu, K)$, such that

every finite realization $\boldsymbol{f} = f(\theta_1), \ldots, f(\theta_m)$ has a multivariate normal distribution with mean vector μ and covariance matrix K. The elements of K are denoted by $K_{i,j} = \sigma^2 K(\theta_i, \theta_j \mid \psi)$ depending on hyper-parameters ψ and σ^2. The choice of kernel and hyper-parameters control the correlation structure and smoothness of the covariance structure. Typical kernel choices include seperable squared exponential and Matérn kernels, for more discussion of Gaussian process specification see [90].

The error terms are modeled as Gaussian noise, $\epsilon^{c,d} \sim \mathcal{N}(0, \tau^2_{c,d})$. This can be interpreted as a specification of Wishart-like observation noise via the Bartlett decomposition, though it is somewhat heavier tailed due to log-normal errors on the diagonal parameters rather than chi-squared errors. To fit this model, the hyper parameters $\{\sigma^2_{c,d}, \tau^2_{c,d}, \psi_{c,d}\}$ are chosen by optimizing the marginal likelihood of the observations $\boldsymbol{L}^{c,d}(\delta) = L^{c,d}(\delta_1), \ldots, L^{c,d}(\delta_m)$:

$$\log p\big(\boldsymbol{L}^{c,d}(\delta) \mid \sigma^2_{c,d}, \tau^2_{c,d}, \psi_{c,d}\big) = -\boldsymbol{L}^{c,d}(\delta)^T (K^{c,d} + \tau^2_{c,d} I)^{-1} \boldsymbol{L}^{c,d}(\delta) - \log |K^{c,d} + \tau^2_{c,d} I| + C$$

Given a set of trained models, interpolating $f^{c,d}$ at M^* new points $\boldsymbol{\theta} = \theta^{(d)}, \ldots, \theta^{(M^*)}$ is straightforward. Define $K^{**}_{c,d}$ and $K^*_{c,d}$, to be the matrices of covariances within the interpolation points and between design and interpolation points respectively. The predictive distributions are:

$$\begin{aligned}
\boldsymbol{L}_{d,d}(\boldsymbol{\theta}) \mid \boldsymbol{L}_{d,d}(\delta) &\sim \log \mathcal{N}\big(\mathbf{f}_{d,d}(\boldsymbol{\theta}),\, M_{d,d}(\boldsymbol{\theta}, \boldsymbol{\theta})\big) \\
\mathbf{f}_{d,d}(\boldsymbol{\theta}) &= K^*_{d,d}\big(K_{d,d} + \tau^2_{d,d} I\big)^{-1} \log \boldsymbol{L}_{d,d}(\delta) \\
\\
\boldsymbol{L}_{c,d}(\boldsymbol{\theta}) \mid \boldsymbol{L}_{c,d}(\delta) &\sim \mathcal{N}\big(\mathbf{f}_{c,d}(\boldsymbol{\theta}),\, M_{c,d}(\boldsymbol{\theta}, \boldsymbol{\theta})\big) \quad \text{for } c > d \\
\mathbf{f}_{c,d}(\boldsymbol{\theta}) &= K^*_{c,d}\big(K_{c,d} + \tau^2_{c,d} I\big)^{-1} \boldsymbol{L}_{c,d}(\delta) \\
M_{c,d}(\boldsymbol{\theta}, \boldsymbol{\theta}) &= K^{**}_{c,d} - K^*_{c,d}\big(K_{c,d} + \tau^2_{c,d} I\big)^{-1} \big(K^*_{c,d}\big)^T
\end{aligned} \tag{4.15}$$

We estimate the Riemannian metric using the posterior mode $\bar{\mathcal{I}}(\theta)$, which can be

obtained by taking the elements of the Cholesky matrix $\bar{L}(\theta)$ to be element-wise posterior modes and taking $\bar{\mathcal{I}}(\theta) = \bar{L}(\theta)\bar{L}(\theta)^T$.

Path selection

Armed with an emulator for $\mathcal{I}$ throughout Θ, we can efficiently choose a path with low variance. We select the best path from a large collection of paths, implied by the specification of a directed, acyclic graph (DAG) G with vertices $V \subset \theta$ and edges E. Several example graphs are given in the next section; a key requirement is that $p_0(x)$ is the root of the graph and that each path terminates with $p_1(x)$. Each edge $(v, v^*) \in V$ is assigned the following weight:

$$w(v, v^*) = (v - v^*)^T \Big(\bar{\mathcal{I}}(v) + \bar{\mathcal{I}}(v^*)\Big)(v - v^*) + \alpha \tag{4.16}$$

The first quantity is proportional to the estimated variance of the step (from (4.9)) and the second term $\alpha \geqslant 0$ is a penalty representing the computational cost of each step. The penalty term controls the trade-off between longer, more computationally demanding paths with low variance and shorter, less computationally demanding paths with high variance. Larger values of α typically lead to shorter paths; we recommend small values such as $\alpha = 0.1$ which provide a reasonable trade-off between path length and estimator variance.

Next, we find the shortest path through the G by combining topological sorting and a shortest path algorithm, which can find the lowest variance path in $\mathcal{O}\big(|V|+|E|\big)$ time [91]. The estimates of $\bar{\mathcal{I}}(v)$ at each node should be pre-computed in order to reduce unnecessary computation during the search. The tuning parameter α can be further optimized to yield the path with the smallest variance relative to the path's length using grid search. Finally, the chosen path is used to produce low variance estimates of l according to (4.7). If the chosen path is insufficiently fine, the path can be augmented by linearly interpolating additional steps.

To reiterate, the main advantage of this approach is that the Gaussian process emulator allows us to evaluate the cost of a large number of paths using a relatively small number of additional samples. These paths can then be efficiently compared due to the linear complexity in $|V|$ and $|E|$ of topological sorting and searching. The requirement to fit $\mathcal{O}(D^2)$ Gaussian processes limits this approach to paths of modest dimension; when D is large it may be practical to model only the diagonal elements of $\mathcal{I}$, reducing the number model parameters that need to be estimated.

4.4 Examples

In this section we demonstrate both path selection approaches. First, we show that the adaptive step size approach obtains excellent performance on the challenging mean field Ising model. Then, we validate the multidimensional path selection approach using the Gaussian example from [28]. We show that our path-finding approach approximates the geodesic and estimates $\mathcal{I}(\theta)$ with good accuracy. Finally, we present a new type of path suitable for prior selection in Bayesian statistics which is a competitive alternative to the geometric path.

4.4.1 Ising example

The mean-field Ising model originated in the statistical physics literature, where is used as a model of ferromagnetism. The model assigns each vector of D "spins" $x \in \{-1, 1\}^D$ the following pmf:

$$p(x \mid \beta) \propto \exp\Big(\frac{\alpha}{2D}\big(\sum_{d=1}^{D} x_d\big)^2\Big) \tag{4.17}$$

For β close to 0, the model places high probability on states where the total magnetization $M = \sum_{d=1}^{D} x_d$ is close to 0. However, as α increases, the model undergoes a phase transition and the regions of high probability shift toward the extreme, placing

most mass near $M = D$ and $M = -D$. This phase transition makes an excellent test case for the one dimensional path selection technique, as path sampling approaches based on geometric families will encounter this phase transition and proper spacing of temperatures near the phase transition is the key to accurate estimation.

We tested our algorithm by performing 100 simulations for $D \in \{10, 50, 250\}$ with $\beta = 3$, $\gamma = 0.1$ and $N = 1,000$. Path sampling was performed using a uniform initial distribution and the geometric family. To sample from the Ising model, we used sequential Monte Carlo [7, 25], due to the noted synergy with adaptive step size selection, and made Markov transitions using Gibbs sampling/Glauber dynamics. Results of the experiment are summarized in Figure 4.1 and Table 4.1

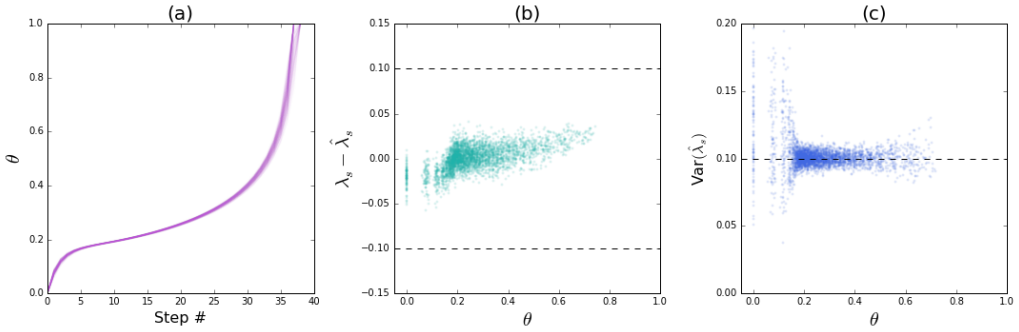

FIGURE 4.1: Adaptive path selection for the mean-field Ising model with $D = 250$. (a) Distribution of paths chosen by adaptive step size selection. (b) Biases of the increments relative to their starting θ and the bound $\gamma = 0.1$. (c) Variances of each increment versus the initial θ and the specified target of $\gamma = 0.1$.

The paths selected by the adaptive step size approach show the desired behaviour; moving quickly towards the critical point, automatically slowing down throughout the phase transition, and taking large steps once the transition is completed. The true variances of the chosen increments tend to be close to the target of $\gamma = 0.1$, with some larger inaccuracies for the initial estimates. This is due to systematic underestimation of the sample variance for $\theta = 0$. The bias of the individual increments were bounded by γ throughout the simulation, demonstrating the ability of the technique to control

both bias and variance simultaneously as suggested in Section 4.3.1. As shown in Table 4.1, adaptive step size selection provides high-accuracy estimates on this challenging problem, with the size of the errors increasing for larger D due to the increased length of the paths.

D	Avg. path length	RMSE	90% CI $(\hat{l} - l)$
10	11	0.06	$(-0.10,\ 0.02)$
50	20	0.07	$(-0.14,\ 0.05)$
250	38	0.09	$(-0.15,\ 0.1)$

Table 4.1: Summaries of the results and estimation errors for the Ising model example

4.4.2 Gaussian example

Consider estimating l when $p_0(x) \sim \mathcal{N}(0, 1)$ and $p_1(x) \sim \mathcal{N}(\mu, 1)$. This problem was studied by[28] and provides an example where embedding the problem in a larger path space and finding optimal paths can dramatically improve the efficiency of the resulting estimator. For this problem, changing the path sampling family from geometric to geometric-tempered and choosing the path to be the implied geodesic improves the sample complexity of the path sampling estimate from $\mathcal{O}(\mu)$ to $\mathcal{O}(\log \mu)$. This example also provides an opportunity to test our shortest path finding algorithm when the optimal path is known and the $\mathcal{I}(\theta)$ has an analytic form.

Before considering the estimation accuracy of the algorithm, we assess the performance of the metric estimation and path selection components. Throughout this section, each run of the algorithm used 144 design points with 100 independent samples drawn at each point. Figure 4.2 shows the errors of an estimated metric for a single run of the algorithm with $\mu = 1000$.

The error of the variance terms (diagonal elements) are generally small throughout Θ, with an average absolute relative error of 0.03 and 0.14 for $\mathcal{I}^{1,1}$ and $\mathcal{I}^{2,2}$ respectively. While the errors of the covariance term $\mathcal{I}^{2,1}$ may seem large, especially

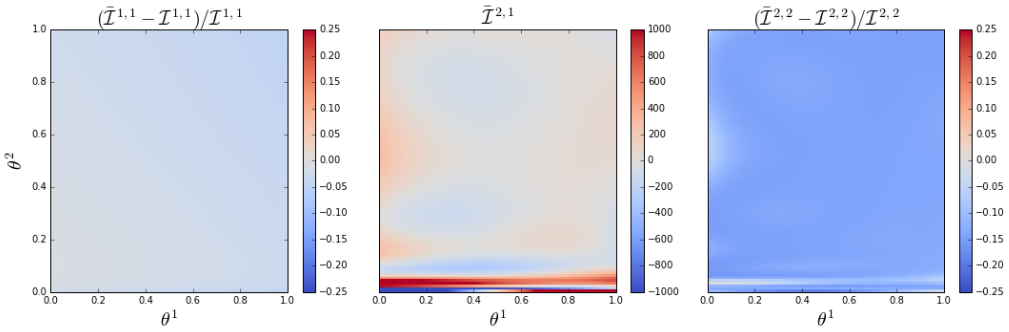

FIGURE 4.2: Errors of the estimated metric on a single run for $\mu = 1000$. For the diagonal elements ($\mathcal{I}^{1,1}$ and $\mathcal{I}^{2,2}$) relative errors are displayed. The middle panel, displaying the covariance term, shows the absolute error as $\mathcal{I}^{1,2} = 0$ throughout Θ.

when $\theta^{(2)}$ is close to 0, they are small relative to the size of the diagonal elements and make little contribution to the estimated cost of a potential path. The model has some difficulty capturing the rate of increase of $\mathcal{I}^{2,2}$ when θ_2 is near ϵ_μ; the interpolated metric has a more gradual rate of change, leading to overestimation near the boundary. The problem of overestimation is most substantial for larger values of μ where the true change is most precipitous. Accurately estimation in this region requires more design points placed near the boundary. Despite this shortcoming, the chosen number of design points was sufficient to achieve our goal of finding efficient paths.

The graph G was specified by vertices $V = \left\{ v_{j,k} \;\middle|\; v_{j,k} = \left(\frac{j}{100},\ 0.5^k \right) \right\}$ and edges $E = \left\{ \left(v_{j,k},\, v_{j',k'} \right) \;\middle|\; j' \in (j,\, j+21),\ k' \in (k-21,\, k+21) \right\}$ for $j = 0, \ldots, 101$ and $k = 0, \ldots, 41$. This leads to a rich set of possible paths, with more than 10^9 possible paths connecting p_0 to p_1. In practice, good results can be obtained with much smaller sets of paths, however, we use this large set to demonstrate that the algorithm can be easily scaled to extremely large path spaces. Figure 4.3 shows the shortest paths from 500 runs of the algorithm for $\mu = 10$ and compares the shortests paths to the optimal path from [28]. In general, the shortest paths found by our

algorithm display similar curvature to the optimal path, however, they generally do not decrease the temperature as much as the optimal path. This is caused by the over estimation of $\mathcal{I}^{2,2}$ when $\theta^{(2)}$ is near 0, which makes paths that approach the lower boundary appear unnecessarily expensive.

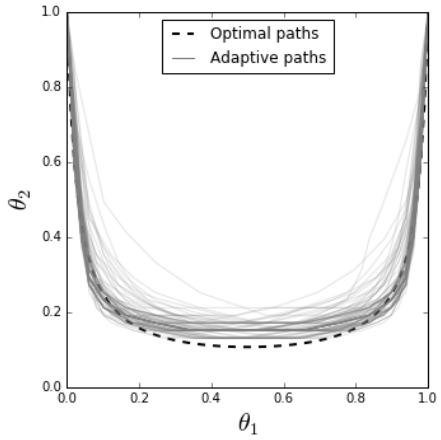

FIGURE 4.3: Distribution of shortest path chosen by the algorithm for $\mu = 10$ using 500 runs of the algorithm. The geodesic/optimal path is shown in black.

To assess the estimation performance of the entire algorithm, estimates using the complete path sampling algorithm were compared to a discrete path sampling estimate using a geometric path with evenly spaced parameters in $[0, 1]$. We performed the metric estimation and path selection algorithm 500 times using the geometric-tempered family for $\mu \in \{10^1, 10^2, 10^3, 10^4\}$. Each run of the algorithm was initialized using 100 independent samples from each of 144 design points spaced according to a Latin hypercube design. These estimates are compared to the path sampling estimate using a geometric path with evenly steps parameters in $[0, 1]$. Even spacing is optimal for this problem using a geometric path, so this comparison serves to highlight the potential advantages of the geometric-tempered path. The number of steps in each linear path was chosen to match the number of steps in a shortest path, and the number of samples was increased above 100 in order to account for the additional

$14,400$ samples used to estimate the $\mathcal{I}(\theta)$. The root mean squared error (RMSE) and 95% confidence intervals (CI) of the errors for both methods are summarized in Table 4.2. The geometric-tempered paths chosen by the path-finding algorithm

	Method	RMSE	90% CI $(\hat{l} - l)$
$\mu = 10$	Geometric-tempered	0.05	$(-0.08,\ 0.08)$
	Geometric	0.05	$(-0.09,\ 0.8)$
$\mu = 10^2$	Geometric-tempered	0.55	$(-0.16,\ 0.16)$
	Geometric	0.35	$(-0.56,\ 0.58)$
$\mu = 10^3$	Geometric-tempered	1.01	$(-0.77,\ 0.71)$
	Geometric	2.89	$(-4.81,\ 4.40)$
$\mu = 10^4$	Geometric-tempered	2.81	$(-4.54,\ 4.32)$
	Geometric	24.34	$(-40.25,\ 38.32)$

Table 4.2: Summary statistics comparing estimation error using the geometric-tempered family with adaptively chosen path and error using geometric paths with even spacing

outperform the standard geometric path in each metric and as μ increases these performance advantages become more pronounced. Unfortunately, the estimates do not exhibit the theoretically optimal logarithmic rate of growth achieved by the optimal path sampling estimator. This results from the overestimation of $\mathcal{I}^{2,2}$ when $\theta^{(2)}$ is near 0, which cause the path finding algorithm to avoid the sufficiently small temperatures used by the optimal path. Regardless, the algorithm still provides a noticeable reduction in error relative to the generic path sampling approach.

4.4.3 Hierarchical model example

Choosing an appropriate prior distribution is a common problem in Bayesian statistics, and ratios of normalizing constants (Bayes factors) can be used to choose between different prior specifications. Often times, data will have a natural hierarchy and a suitable prior should be chosen to reflect the relationship amongst different groups. Two extreme options are the treatment of each group as the same (pooled)

or modelling each group separately (unpooled). Alternatively, a compromise can be specified which allows for information to be shared between models (hierarchical). In this section, we show how multidimensional paths can be used to compare models using these types of prior distributions.

To fix ideas, consider a simple model with observations $y_{i,j} \in \mathcal{R}$, where i is the observation index and j denotes the grouping. The observations are modelled as follows:

$$\begin{aligned} y_{i,j} &= \mu_{i,j} + \epsilon_{i,j} \\ \epsilon_{i,j} &\sim N(0, \sigma^2) \end{aligned} \tag{4.18}$$

The prior choice on $\mu_{i,j}$ determines the structure of the model:

$$\begin{aligned} p(\mu_{i,j}) &= \delta_\mu(\mu_{i,j}) \quad \text{(pooled)} \\ p(\mu_{i,j}) &\sim \mathcal{N}(\mu, \tau^2) \quad \text{(hierarchical)} \\ p(\mu_{i,j}) &\propto 1 \quad \text{(unpooled)} \end{aligned} \tag{4.19}$$

The notation $\delta_\mu(\mu_{i,j})$ indicates that the Dirac delta distribution. The model is completed by placing prior distributions on μ, σ^2 and τ^2.

$$\begin{aligned} p(\mu) &\propto \mathcal{N}(0, \tau_\mu^2) \\ p(\sigma^2) &\propto \sigma^{-2} \\ p(\tau^2) &\sim \mathcal{IG}(a, b) \end{aligned} \tag{4.20}$$

The hyper-parameters τ_μ^2, a and b control the shrinkage of group means towards μ. It is natural to think of the unpooled and pooled models as extreme versions of a single hierarchical model. When $\tau^2 \to 0$, the prior distribution on $\mu_{i,j}$ becomes degenerate leading to the pooled model. On the other hand, as $\tau^2 \to \infty$ the prior distribution on $\mu_{i,j}$ becomes degenerate and the model becomes unpooled. In this section, we use this idea to build a multivariate path-sampling family between these related models.

Geometric-scale families

Consider estimating the ratio of normalizing constants between the unpooled model and the hierarchical model using path sampling with a geometric mixture. In this case, the path sampling family and potential are given by:

$$\begin{aligned} p(\mu, \mu_{i,j}, \sigma^2, \tau^2 \mid y, \theta) &\propto \prod_{i,j} N(y_{i,j} \mid \mu_{i,j}, \sigma^2) \cdot \mathcal{IG}(a, b) \cdot \sigma^{-2} \cdot \Big(\prod_j \mathcal{N}(\mu_{i,j} \mid \mu, \tau^2) \Big)^{\theta} \\ U(\mu, \mu_{i,j}, \sigma^2, \tau^2 \mid \theta) &= \sum_j \log \mathcal{N}(\mu_{i,j} \mid \mu, \tau^2) \end{aligned} \tag{4.21}$$

Estimating l using the geometric family be challenging. As θ increases from 0, the model crosses a high-energy barrier caused by a large discrepancy between the group means $\mu_{i,j}$ and the overall mean μ. While this difference quickly dissipates, the size of the initial change can lead to highly variable estimates of the log ratio of normalizing constants. To address the problem, we propose the *geometric-scale* family:

$$\begin{aligned} p(\mu, \mu_{i,j}, \sigma^2, \tau^2 \mid y, \theta) &\propto \prod_{i,j} N(y_{i,j} \mid \mu_{i,j}, \sigma^2) \cdot \mathcal{IG}(a, b/\theta^{(2)}) \cdot \sigma^{-2} \cdot \\ &\qquad \Big(\prod_j \mathcal{N}(\mu_{i,j} \mid \mu, \tau^2) \Big)^{\theta^{(2)}} \\ U^1(\mu, \mu_{i,j}, \sigma^2, \tau^2 \mid \theta) &= \sum_j \log \mathcal{N}(\mu_{i,j} \mid \mu, \tau^2) \\ U^2(\mu, \mu_{i,j}, \sigma^2, \tau^2 \mid \theta) &= \frac{2\tau^{-2}}{(\theta^{(2)})^2} \end{aligned} \tag{4.22}$$

This family allows the prior scale of the hierarchical variance parameter to vary which we use to modify the cost of the initial path sampling transition. As $\theta^{(2)} \to 0$, the prior mean on τ^2 becomes large, reducing the amount of shrinkage caused by the hierarchical prior and diminishing the energy barrier at the initial transition allowing for low variance moves in $\theta^{(1)}$.

Radon example

We demonstrate the geometric-scale family using the radon data from [92], which consists of 919 log-transformed radon measurements taken from 85 counties. Metric estimation was performed using 25 design points from the latin hypercube and five additional design points evenly spaced between $(0, 10^{-5})$ and $(0, 1)$ in order to accurately capture the behaviour of the metric near the high energy barrier at $\theta^{(1)} = 0$. We drew 100 samples from $p(\mu, \mu_{i,j}, \sigma^2, \tau^2 \mid y, \theta)$ at each design point using the probabilistic programming language Stan [93]. The paths were specified by the graph G in Section 4.4.2.

First, to illustrate the advantages of the geometric-scale family, we show the estimated metric for the Radon problem in Figure 4.4. As discussed, there is a

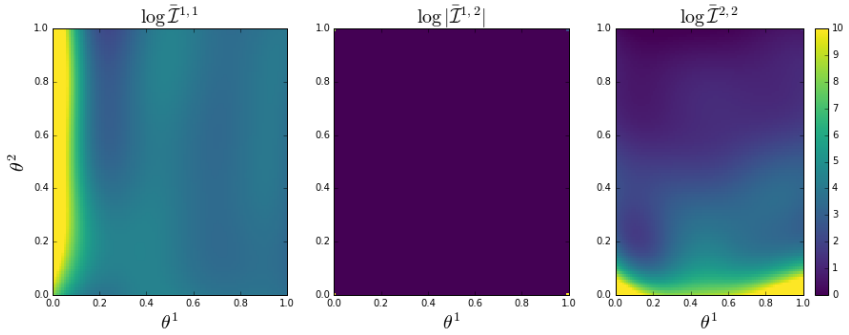

FIGURE 4.4: Estimated metric for the radon example

high energy barrier present for $U^{(1)}$ that can be avoided by appropriately reducing the scale parameter $\theta^{(2)}$. This advantage must be balanced against the increased variability in $U^{(2)}$ for small values of $\theta^{(2)}$. Choosing an appropriate level for $\theta^{(2)}$ is the key to the success when using a geometric-scale family.

Our method was compared to path sampling using a geometric-scale family with three different spacings. The first used linear spacing and the other methods used a geometric spacing with $\theta_s = 1 - (s/S)^p$ for $p = 2$ and $p = 5$ (suggested by [76]).

To assess variability, we ran each method 100 times for $S \in \{11, 51\}$ with $N = 100$. The results are compared to a high quality estimate $l = -67.62$ obtained using a geometric path sampling run with $N = 2,000$, $S = 10,000$ and geometric spacing ($p = 5$). The results of the experiment are shown in Table 4.3. The geometric-scale

	Method	RMSE	90% CI $(\hat{l} - l)$
$S = 11$	Geometric-scale	1.64	$(-5.4,\ -0.1)$
	Geometric (linear)	132.48	$(-431.6,\ -76.3)$
	Geometric (p=2)	10.90	$(-39.5,\ -5.4)$
	Geometric (p=5)	0.79	$(-2.6,\ 0.0)$
$S = 51$	Geometric-scale	0.40	$(-0.9,\ 0.3)$
	Geometric (linear)	16.43	$(-65.1,\ -12.5)$
	Geometric (p=2)	0.53	$(-1.5,\ 0.2)$
	Geometric (p=5)	0.21	$(-0.4,\ 0.3)$
$S = 251$	Geometric-scale	0.15	$(-0.26,\ 0.22)$
	Geometric (linear)	4.38	$(-16.53,\ -1.45)$
	Geometric (p=2)	0.17	$(-0.26,\ 0.27)$
	Geometric (p=5)	0.08	$(-.09,\ 0.15)$

Table 4.3: Summary statistics comparing estimation error using the geometric-scale family with adaptively chosen path and error using geometric paths with linear and geometric spacings

approach is competitive with the best geometric path, though it is not superior to the geometric spacing with $p = 5$. For $S = 11$, the advantage is particularly striking, providing a nearly optimal performance with a small number of samples. Geometric paths can be successful on this problem because the high energy barrier is limited to the region immediately around $\theta = 0$. Tight spacing at the beginning of the path can overcome the initial barrier using a modest number of steps. If the high energy barrier persisted outside of $\theta = 0$, or the high energy barrier occurred away from the edges of the pace, path selection would become more complicated and the geometric-scale path could be more advantageous.

4.5 Conclusion

Future work in this area could follow several directions. First, theoretical gaurantees should be developed for the adaptive step size approach, possibly following the techniques outlined in [81]. Does this approach provide a minimum variance estimator? Does the chosen path approximate the geodesic as $\gamma \to 0$ and $N \to \infty$? How does the complexity of the estimator depend on the characteristics of the path sampling family? Answering these questions could improve our understanding of the algorithm, highlighting it's strengths and weaknesses.

The multidimensional path sampling approach should be applied to more challenging problems in both statistical and scientific settings. Shirts et. al. [70] note that a multidimensional pathway is well suited to the problem of estimating the free energy in the solvation of 3-methylindole and it would be useful to see if our approach could be applied in that setting. Multidimensional paths have also been applied in the machine learning literature [73, 78]; similar problems could also provide a test case for our algorithm.

Finally, it is unclear when the multidimensional approach should be preferred over the one dimensional approach. The development of new path sampling families and guidelines for their application are necessary for making the multidimensional approach generally applicable.

4.6 Proof of equation 4.13

Using the fact that the bias of the increments is the symnetrized Kullback-Leibler divergence [76]:

$$\begin{aligned}
|\hat{l}_s - l_s| &= \frac{1}{2} \cdot \Big| \mathrm{KL}\Big(p(x \mid \theta_s) \,\Big|\Big|\, p(x \mid \theta_{s-1})\Big) - \Big(p(x \mid \theta_{s-1}) \,\Big|\Big|\, p(x \mid \theta_s)\Big) \Big| \\
&\leqslant \frac{1}{2} \cdot \Big| \mathrm{KL}\Big(p(x \mid \theta_s) \,\Big|\Big|\, p(x \mid \theta_{s-1})\Big) + \Big(p(x \mid \theta_{s-1}) \,\Big|\Big|\, p(x \mid \theta_s)\Big) \Big| \\
&\leqslant \frac{\Delta_s}{2} \cdot \Big| E_{\theta_s} U(x \mid \theta) - E_{\theta_{s-1}} U(x \mid \theta) \Big| \\
&\approx \frac{\Delta_s^2}{2} \cdot \Big| \frac{d}{d\theta} E_{\theta_{s-1/2}} U(x \mid \theta) \Big| \quad \text{as } \Delta_s \to 0
\end{aligned} \tag{4.23}$$

where $\theta_{s-1/2} = (\theta_s + \theta_{s-1})/2$. For families of geometric families $\frac{d}{d\theta} E_\theta U(x \mid \theta) = \mathrm{Var}_\theta\big(U(x \mid \theta)\big)$ [51]. If Δ_s is small so that $\mathrm{Var}_\theta\big(U(x \mid \theta)\big)$ is roughly constant, we can further simplify yielding the result:

$$\begin{aligned}
|\hat{l}_s - l_s| &\leqslant \frac{\Delta_s^2}{2} \cdot \Big| \frac{d}{d\theta} E_{\theta_{s-1/2}} U(x \mid \theta) \Big| \\
&= \frac{\Delta_s^2}{2} \cdot \mathrm{Var}_{\theta_{s-1/2}}\big(U(x \mid \theta_{s-1/2})\big) \\
&= \frac{\Delta_s^2}{2} \cdot \frac{\mathrm{Var}_{\theta_s}\big(U(x \mid \theta)\big) + \mathrm{Var}_{\theta_{s-1}}\big(U(x \mid \theta)\big)}{2} \\
&= \mathrm{Var}(\hat{l}_s) \cdot N
\end{aligned} \tag{4.24}$$

5

Conclusion

The majority of this work focused on developing finite sample bounds for SMC algorithms, showing how to obtain an SMC estimator with controlled error. While these bounds may not be easily applicable in practice, due to the difficulty in bounding the mixing time of the Markov kernels, they allowed us to compare the complexity of the algorithm to other methods and to study how the selection of interpolating distributions changes the complexity of the algorithm. In the first part of conclusion, we return to the claims made in the introduction regarding the advantages of SMC over MCMC: scalability via parallelization and performance on multimodal target distributions. Then, we discuss the relationship between the SMC algorithm studied in this paper to other SMC variants. We conclude with a brief discussion of future research directions in this area.

5.0.1 Advantages of SMC

The proofs presented in Chapter 3 are important for validating the scalability of SMC through parallelization. We demonstrated several settings where the order of the SMC path length ($S/\mathcal{E}$) was smaller than the MCMC distance ($\log ||\mu_0/\pi||_{\pi,s}$).

This shows that the SMC algorithm may benefit not only from a linear speedup in the number of available processors, but an additional improvement in complexity due to a reduced path length relative to MCMC. This is important as it shows that SMC has superior performance relative to the algorithm that simply runs many MCMC chains in parallel. Quantifying the extent this improvment more broadly, particularly in the case of the adaptively chosen paths from Chapter 3, could further highlight the advantages of SMC in the parallel setting.

This thesis did not address the performance of SMC on difficult multimodal target problems where Markov kernels mix well locally, but have poor global mixing properties (similar to [11, 45, 43]). While the strategy of proof used in Chapter 2 can be applied more or less directly in this setting, the resulting bound requires $N = \mathcal{O}(2^S)$. This bound seems unrealistically loose; such poor performance would have been easily noticeable in practice. In contrast, a similar asymptotic result requires that $N = \mathcal{O}(S^2)$ [11]. Improvements in our approach are needed to give realistic results for multimodal problems.

The multimodal setting is particularly interesting as we expect SMC to behave somewhat differently than MCMC algorithms, like parallel tempering, on these problems. Efficient parallel tempering algorithms require that the overlap between adjacent distributions is large in order to promote swapping. This concern must be balanced against the overall number of distributions, which should be kept small to allow particles to quickly move from low temperature reasons to high temperature reasons. Balancing these considerations is difficult and is the primary concern when choosing a temperature ladder. In contrast, adaptive SMC algorithms provide a straightforward approach to choosing a temperature ladder, which may make them preferable to parallel tempering in some situations. Unfortunately, like parallel tempering, the performance of SMC seems to degrade as the number of distributions increases. This manifests catastrophically as the loss of modes during sampling, and

unlike parallel tempering, the SMC algorithm is unlikely to recover these modes. Quantifying the extent of this problem could lend important insight into when each algorithm should be used.

5.0.2 Comparison to other SMC algorithms

There are many varieties of SMC and it is natural to wonder whether the algorithm studied here is the best among all the ensemble sampling/SMC flavors. One natural comparison is the SMC algorithm without resampling, known as annealed importance sampling (AIS) [4]. In general, we believe AIS will perform poorly compared to SMC when both methods use rapidly mixing Markov kernels. The repeated application of a Markov kernel is known to decrease the covariance between the initial particle and its end point [35], therefore, we believe that the AIS technique often results in a collection of particles whose weights may be roughly independent of their final locations. This could yield an estimator with artificially increased variance relative to SMC.

Another popular implementation of SMC chooses resampling times adaptively using the RESS. In our frame work, we view this as an alternative method of adaptively choosing distributions, similar to the approach outlined in Chapter 3. The adaptive resampling method could have two potential problems. As a method for choosing distributions, the set of possible steps is usually specified to be unnecessarily coarse and so the adaptive resampling algorithm is prone to taking catastrophically large steps. In addition, the adaptive resampling approach suffers from the same problem as AIS, in that the weights become poor indicators of sample quality as the number of Markov kernel steps increases. While adaptive re-sampling times have been useful in the filtering literature, where the Markov transition kernels do not rejuvinate the samples, for static problems we believe they may actually increase estimation error.

Finally, we consider the time inhomogeneous MCMC approach, which varies

the Markov kernels at each step of the algorithm but uses no weights or resampling. This approach seems to have many theoretical advantages over SMC and in fact will yield bounds with superior complexity to our SMC bounds whenever $||\mu_s/\mu_{s-1}||_{\mu_{s-1},2} = \mathcal{O}(1)$ implies $||\mu_{s-1}/\mu_s||_{\mu_s,2} = \mathcal{O}(1)$. This is the case in the log-concave example explored in Chapter 3, where time inhomogeneous MCMC yields the fastest possible bound. However, this advantage does not seem to be born out in practice, where the time inhomogenous approach seems to yield worse performance than SMC approaches using weights. Understanding this gap in theory could yield new, improved bounds for SMC estimators.

5.0.3 Future work

Upper bounds on the complexity of an algorithm can provide only limited insight into an algorithm's practical performance. Are the bounds of the correct order? How do they relate to the performance of the in algorithm in practice? Addressing these questions, through lower bounds on the error and simulations studies respectively, could further our understanding of SMC algorithms and help identify scenarios where SMC excels.

Bibliography

[1] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[2] Walter R Gilks and Carlo Berzuini. Following a moving targetMonte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146, 2001.

[3] Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

[4] Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[5] Olivier Cappé, Arnauld Guillin, Jeanl-Michel Marin, and Christian Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(3):907–930, 2004.

[6] Nicholas Chopin, Pierre E. Jacob, and Omiros Papaspiliopoulos. SMC^2: an efficient algorithm for sequential analysis of state-space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:397–426, 2012.

[7] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[8] Anthony Lee, Christopher Yau, Michael B Giles, Arnaud Doucet, and Christopher C Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.

[9] Christelle Vergé, Cyrille Dubarry, Pierre Del Moral, and Eric Moulines. On parallel implementation of sequential Monte Carlo methods: the island particle model. *Statistics and Computing*, 25(2):243–260, Mar 2015.

[10] Anthony Lee and Nick Whiteley. Forest resampling for distributed sequential Monte Carlo. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(4):230–248, 2016.

[11] Ajay Jasra, Daniel Paulin, and Alexandre H Thiery. Error bounds for sequential Monte Carlo samplers for multimodal distributions. *arXiv preprint arXiv:1509.08775*, 2015.

[12] Nikolaus Schweizer. *Non-asymptotic error bounds for sequential MCMC methods*. PhD thesis, University of Bonn, 2011.

[13] Jeremy Heng, Adrian N. Bishop, George Deligiannidis, and Arnaud Doucet. Controlled Sequential Monte Carlo. *ArXiv e-prints*, April 2018.

[14] Shixiang Gu, Zoubin Ghahramani, and Richard E. Turner. Neural adaptive sequential Monte Carlo. In *Advances in Neural Information Processing Systems 28*, pages 2629–2637. Curran Associates, Inc., 2015.

[15] Garland Durham and John Geweke. *Adaptive sequential posterior simulators for massively parallel computing environments*, chapter 1, pages 1–44. Emerald Group Publishing Limited, 2014.

[16] Ajay Jasra, David A. Stephens, Arnaud Doucet, and Theodoros Tsagaris. Inference for lvy-driven stochastic volatility models via adaptive sequential monte carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011.

[17] P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Probab.*, 9(2):275–297, 05 1999.

[18] Pierre Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag, 2004.

[19] Nicolas Chopin et al. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.

[20] Alexandros Beskos, Ajay Jasra, Nikolas Kantas, and Alexandre Thiery. On the convergence of adaptive sequential Monte Carlo methods. *Ann. Appl. Probab.*, 26(2):1111–1146, 04 2016.

[21] Nick Whiteley. Sequential Monte Carlo samplers: error bounds and insensitivity to initial conditions. *Stochastic Analysis and Applications*, 30(5):774–798, 2012.

[22] Andreas Eberle and Carlo Marinelli. Quantitative approximations of evolving probability measures and sequential Markov chain Monte Carlo methods. *Probability Theory and Related Fields*, 155(3-4):665–701, 2013.

[23] Andreas Eberle and Carlo Marinelli. Convergence of sequential Markov chain Monte Carlo methods: I. nonlinear flow of probability measures. Technical report, In preparation, 2007.

[24] Alexandros Beskos, Dan Crisan, and Ajay Jasra. On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, 24(4):1396–1445, 08 2014.

[25] Yan Zhou, Adam M. Johansen, and John A.D. Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.

[26] Ming Lin, Rong Chen, and Jun S. Liu. Lookahead strategies for sequential monte carlo. *Statist. Sci.*, 28(1):69–94, 02 2013.

[27] Yosihiko Ogata. A monte carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

[28] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185, 05 1998.

[29] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, 22(3):795–808, May 2012.

[30] Randal Douc and Eric Moulines. Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.*, 36(5):2344–2376, 10 2008.

[31] László Lovász and Santosh Vempala. Hit-and-run from a corner. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 310–314, New York, NY, USA, 2004. ACM.

[32] Santosh Vempala. Geometric random walks: a survey. *Combinatorial and Computational Geometry*, pages 573–612, 2005.

[33] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 1995.

[34] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169 – 188, 1986.

[35] Lszl Lovsz and Santosh Vempala. Simulated annealing in convex bodies and an $\mathcal{O}^*(n^4)$ volume algorithm. *Journal of Computer and System Sciences*, 72(2):392 – 417, 2006. JCSS FOCS 2003 Special Issue.

[36] Ravi Kannan and Guangxing Li. Sampling according to the multivariate normal density. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 204–212. IEEE, 1996.

[37] Ravi Kannan, Lszl Lovsz, and Mikls Simonovits. Random walks and an $\mathcal{O}^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11(1):1–50, 1997.

[38] Sanghyun Park and Vijay S Pande. Choosing weights for simulated tempering. *Physical Review E*, 76(1):016703, 2007.

[39] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! *arXiv preprint arXiv:1509.08775*, 01 2018.

[40] L. Lovasz and S. Vempala. Fast algorithms for logconcave functions: sampling, rounding, integration and optimization. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 57–68, Oct 2006.

[41] Lszl Lovsz and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2006.

[42] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[43] Dawn Woodard, Scott Schmidler, Mark Huber, et al. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780–804, 2009.

[44] D. N. VanDerwerken and S. C. Schmidler. Parallel Markov Chain Monte Carlo. *ArXiv e-prints*, December 2013.

[45] B. Woodard, Scott C. Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Annals of Applied Probability*, pages 617–640, 2009.

[46] Gareth Roberts, Jeffrey Rosenthal, et al. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

[47] Gareth O. Roberts and Richard L. Tweedie. Geometric L^2 and L^1 convergence are equivalent for reversible Markov Chains. *Journal of Applied Probability*, 38:37–41, 2001.

[48] Inder K Rana. *An Introduction to Measure and Integration*. American Mathematical Society, 2017.

[49] Vladimir I. Bogachev. *Measure Theory*. Springer-Verlag, 2007.

[50] Joe Marion and Scott C. Schmidler. Finite Sample Complexity of Sequential Monte Carlo Estimators. *ArXiv e-prints*, March 2018.

[51] Nial Friel, Merrilee Hurn, and Jason Wyse. Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723, Sep 2014.

[52] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.

[53] Luca Martino, Victor Elvira, and Francisco Louzada. Alternative effective sample size measures for importance sampling. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5, June 2016.

[54] Thi Le Thu Nguyen, Francois Septier, Gareth W. Peters, and Yves Delignon. Efficient sequential Monte Carlo samplers for Bayesian inference. *IEEE Transactions on Signal Processing*, 64(5):1305–1319, March 2016.

[55] Ruslan Salakhutdinov. Learning and evaluating Boltzmann machines. Technical report, University of Toronto, 2008.

[56] Sudipto Bannerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman and Hall/CRC, 2014.

[57] Dieter W. Heermann Kurt Binder. *Monte Carlo Simulation in Statistical Physics: an Introduction*. Springer, 2002.

[58] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov Chains and applications to statistical mechanics. *Random Struct. Algorithms*, 9(1-2):223–252, August 1996.

[59] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[60] Nicolas Chopin and James Ridgway. Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Statist. Sci.*, 32(1):64–87, 02 2017.

[61] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Math.*, 3(1):79–127, 2006.

[62] Charles H Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22:245–268, 10 1976.

[63] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

[64] Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002.

[65] Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[66] Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[67] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

[68] Gavin E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Physical Review E*, 61:2361–2366, Mar 2000.

[69] Friel Nial and Wyse Jason. Estimating the evidence a review. *Statistica Neerlandica*, 66(3):288–308, 2011.

[70] Michael R. Shirts and Vijay S. Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *The Journal of Chemical Physics*, 122(14):144107, 2005.

[71] T. P. Straatsma and J. A. McCammon. Multiconfiguration thermodynamic integration. *The Journal of Chemical Physics*, 95(2):1175–1188, 1991.

[72] Chris J Oates, Theodore Papamarkou, and Mark Girolami. The controlled thermodynamic integral for bayesian model comparison. *arXiv preprint arXiv:1404.5053*, 2014.

[73] Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 872–879, New York, NY, USA, 2008. ACM.

[74] Daniel K Shenfeld, Huafeng Xu, Michael P Eastwood, Ron O Dror, and David E Shaw. Minimizing thermodynamic length to select intermediate states for free-energy calculations and replica-exchange simulations. *Physical Review E*, 80(4):046705, 2009.

[75] Nial Friel and Anthony N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.

[76] Ben Calderhead and Mark Girolami. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

[77] Wooseop Kwak and Ulrich H. E. Hansmann. Efficient sampling of protein structures by model hopping. *Phys. Rev. Lett.*, 95:138102, Sep 2005.

[78] Roger B Grosse, Chris J Maddison, and Ruslan R Salakhutdinov. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems 26*, pages 2769–2777. Curran Associates, Inc., 2013.

[79] Colin Atkinson and Ann FS Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 345–365, 1981.

[80] Simon J.A. Malham. *An introduction to Lagrangian and Hamiltonian mechanics*, 2016.

[81] Joe Marion and Scott C. Schmidler. Finite Sample L_2 bounds for sequential Monte Carlo and adaptive path selection. *ArXiv e-prints*, March 2018.

[82] Ryan Muraglia. Path optimization in free energy calculations. Master's thesis, Duke University, 2016.

[83] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[84] Carla Currin, Toby Mitchell, Max Morris, and Don Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.

[85] Dave Higdon, Marc Kennedy, James C. Cavendish, John A. Cafeo, and Robert D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

[86] R. A. Bates, R. J. Buck, E. Riccomagno, and H. P. Wynn. Experimental design and observation for large systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):77–94, 1996.

[87] Michael J. Daniels. Bayesian modeling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors. *Journal of Multivariate Analysis*, 97(5):1185 – 1207, 2006.

[88] Mohsen Pourahmadi, Michael J. Daniels, and Trevor Park. Simultaneous modelling of the cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, 98:568–587, 03 2007.

[89] Andrew Gordon Wilson and Zoubin Ghahramani. Generalised wishart processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 736–744, Arlington, Virginia, United States, 2011. AUAI Press.

[90] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[91] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 2009.

[92] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models, First Edition*. Cambridge University Press, 2006.

[93] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Biography

Joseph Caleb Marion was born on December 3rd, 1987 in Orlando, Florida. He graduated from Cornell University with distinction in Mathematics and Economics in 2010 and was commissioned in the United States Army as a Field Artillery Officer. After deploying to Iraq in 2011 and Afghanistan in 2012 he returned to Duke University to pursue his PhD in Statistical Science, graduating in the summer of 2018. Joe began work as a statistical scientist with Berry Consultants located in Austin, Texas in the Fall of 2018.