Finite Sample Complexity of Sequential Monte Carlo Estimators

Joseph Marion and Scott C. Schmidler
Department of Statistical Science
Duke University

March 28, 2018

Abstract

We present bounds for the finite sample error of sequential Monte Carlo samplers on static spaces. Our approach explicitly relates the performance of the algorithm to properties of the chosen sequence of distributions and mixing properties of the associated Markov kernels. This allows us to give the first finite sample comparison to other Monte Carlo schemes. We obtain bounds for the complexity of sequential Monte Carlo approximations for a variety of target distributions including finite spaces, product measures, and log-concave distributions including Bayesian logistic regression. The bounds obtained are within a logarithmic factor of similar bounds obtainable for Markov chain Monte Carlo.

1 Introduction

Sequential Monte Carlo samplers (SMC) [1,2] 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 [3–5]. 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 [6–9]. Finally, SMC may exhibit similar properties to tempering, making it well suited for difficult or multimodal problems [1–3]. 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 [10] and extended by Chopin [11]. A similar CLT was shown to hold for adaptive resampling methods by Douc and Moulines [12] and later by Beskos et al. [13]. Other asymptotic theory includes the work of Jasra et al. [14], who proved a bound on the asymptotic variance under local mixing assumptions. Beskos et al. [15], showed non-degeneracy of the particle approximation as the dimension increases for problems with product measures. Eberle and Marinelli [16, 17] 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 [18], who developed L_p error bounds on non-compact spaces using drift and minorization conditions, and Schweizer [19] 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 paper 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

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

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 [20]. 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 paper is structured as follows. Section 2 introduces some notation and describes the general form of the SMC algorithm studied in this paper and concludes with a statement of our main result. Section 3 presents the proof of our error bound, developing conditions for inductively controlling the error. Section 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 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 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 Sequential Monte Carlo

Let π be a target probability measure on a space $\mathcal{X}$ with σ-algebra $\mathcal{B}(\mathcal{X})$ 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 paper.

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 $\sup_{B\in\mathcal{B}(\mathcal{X})}\ \nu(B) \leq \omega \cdot \mu(B)$ [21, 22]. Let $\mathcal{P}_\omega(\mu)$ be the set of all such measures.

Let $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ 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

$$\tau_K(\epsilon, \omega) = \min\ \big\{t : \sup_{\nu\in\mathcal{P}_\omega(\mu)} ||\nu K^t - \mu||_{\text{TV}} \leq \epsilon\big\}$$

where $|| \cdot ||_{\text{TV}}$ 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 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 $\frac{q_s(x)/z_s}{q_{s-1}(x)/z_{s-1}} \leq W \cdot Z$, 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 have mixing time $\tau_s(\epsilon, \omega)$.

In this paper 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).

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

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

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

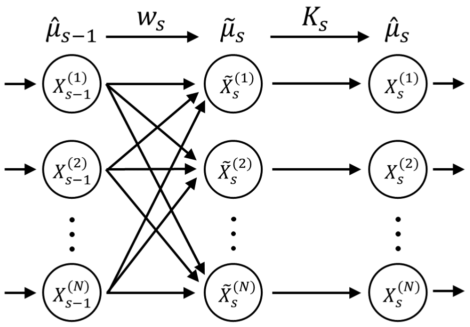

Figure 1: Marginal distributions and dependence structure of the particles.

The final step of SMC produces empirical measure $\hat{\pi} = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_S^{(n)}}$ and corresponding estimate $\hat{\pi}f = \frac{1}{N}\sum_{n=1}^{N} f\big(x_S^{(n)}\big)$ 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 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.

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 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.

2.3 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. [23]). It is standard to show this bound for $\delta \geq 3/4$; this can then be improved to $1-\delta$ in $\mathcal{O}\big(\log(1/\delta)\big)$ time [24].

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

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

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

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

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

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

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

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

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

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 3, which first develops two key lemmas before presenting the proof. The proof of Theorem 2 is similar and delayed to the Appendix.

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:

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

where $\bar{w}_{s+1} = N^{-1}\sum_{n=1}^N w_{s+1}\left(x_s^{(n)}\right)$ 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 $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geq 3/4$. 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 $B \in \mathcal{B}(\mathcal{X})$. Then

$$\Pr\Big(\tilde{X}_s^{(n)} \in B \;\Big|\; \mathbf{C}_{s-1}\Big) = \sum_{n=1}^{N} \Pr\Big(X_{s-1}^{(m)} \in B,\, E_{m,n} \;\Big|\; \mathbf{C}_{s-1}\Big) \tag{2}$$

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

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

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) and (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 3.1 (Correctness of the predictive distribution)**.**
Assume $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geq 3/4$. Then for any $0 < \delta_0 < 1$ and $t \geq \tau_s\big(\frac{\delta_0}{2N}, 2\big)$

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

Proof. By the choice of t and Lemma 3

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

Applying the standard coupling argument "divine intervention" (see for example Lemma 4.2 of [25]) gives $\Pr\big(X_s^{(m)} \sim \mu_s\big) \geq 1 - \frac{\delta_0}{2N}$. 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 4 (Lower bound on the average weights)**.**
Fix any $0 < \delta_0 < 1$ and choose $N \geq 18 \log(4/\delta_0) \cdot W^2 Z^2$. Then

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

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:

$$\begin{aligned} E\Big(\mathcal{W}_{s+1}^{n+1}\Big|\mathcal{W}_{s+1}^{n}\Big) &= E\Big(\mathcal{W}_{s+1}^{n+1}\Big|X_s^{1:n}\Big) \\ &= \mathcal{W}_{s+1}^{n} + E\Big(w_{s+1}\big(X_s^{(n+1)}\big)\Big|X_s^{1:n}\Big) - \hat{\mu}_s[w_{s+1}] \\ &= \mathcal{W}_{s+1}^{n} + E\bigg(E\Big(w_{s+1}\big(X_s^{(n+1)}\big)\Big|X_s^{1:n},\ \tilde{X}_s^{(n+1)}\Big)\bigg) - \hat{\mu}_s[w_{s+1}] \\ &= \mathcal{W}_{s+1}^{n} + E\bigg(E\Big(w_{s+1}\big(X_s^{(n+1)}\big)\Big|\tilde{X}_s^{(n+1)}\Big)\bigg) - \hat{\mu}_s[w_{s+1}] \\ &= \mathcal{W}_{s+1}^{n} \end{aligned} \tag{4}$$

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

We now combine Corollary 3.1 and Lemma 4 to give the one step induction condition.

Corollary 4.1 (One step induction condition).
Assume $P\big(\boldsymbol{C}_{s-1}(ii)\big) \geq 3/4$. Fix $0 < \delta_0 < 1$. Choose $N \geq 18\log(4/\delta_0) \cdot W^2 Z^2$ and $t \geq \tau_s\big(\frac{\delta_0}{2N}, 2\big)$. Then the inductive condition $\boldsymbol{C}_s|\boldsymbol{C}_{s-1}$ holds with probability at least $1-\delta_0$

Proof. Observe that Lemma 4 implies $\mathbf{C}_s$(ii) conditional on $\mathbf{C}_s$(i). Then by Lemma 4 and Corollary 3.1

$$\begin{aligned} \Pr\big(\mathbf{C}_s \,|\, \mathbf{C}_{s-1}\big) &= \Pr\big(\mathbf{C}_s(\text{ii}) \,|\, \mathbf{C}_{s-1}(\text{i})\big) \cdot \Pr\big(\mathbf{C}_s(\text{i}) \,|\, \mathbf{C}_{s-1}\big) \\ &\geq 1 - \delta_0 \end{aligned} \tag{5}$$

■

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 4 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

$$\begin{aligned} \Pr\big(|\hat{\pi}f - \pi f| \leq \epsilon\big) &\geq \Pr(|\hat{\pi}f - \pi f| \leq \epsilon \mid \mathbf{C}_S(\text{i})\big) \cdot \Pr\big(\mathbf{C}_S(\text{i})\big) \\ &\geq \Pr(|\hat{\pi}f - \pi f| \leq \epsilon \mid \mathbf{C}_S(\text{i})\big) \cdot \Pr\big(\mathbf{C}_S(\text{i}) \,|\, \mathbf{C}_{S-1}\big) \cdot \prod_{s=1}^{S-1} \Pr\big(\mathbf{C}_s \,|\, \mathbf{C}_{s-1}\big) \cdot \Pr\big(\mathbf{C}_0(\text{ii})\big) \\ &\geq (1-\delta_0)^S \\ &\geq 1 - \frac{1}{4} \end{aligned} \tag{6}$$

The third line follows from the induction condition of Corollary 4.1. The error probability $\Pr\big(|\hat{\pi}f - \pi f| \leq \epsilon \mid \mathbf{C}_S(\text{i})\big)$ is controlled in the same manner as Lemma 4. 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.

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 [1, 4, 9]. 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

$$q_\beta(x) = q_\pi(x)^\beta \cdot q_\nu(x)^{(1-\beta)}/z_\beta. \tag{7}$$

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 $w_s(x) \leq 1$. 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 dominate 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.

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 4.2 (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 $|f| \leq 1$, the number of Markov kernel transitions required to ensure $|\hat{\pi} f - \pi f| \leq \epsilon$ 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.

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 4.3 (Complexity of MCMC with independent chains).
Fix $\epsilon > 0$. Then for any function $f \in \mathcal{F}$ with $|f| \leq 1$, the number of Markov kernel transitions required to ensure $|\bar{\pi} f - \pi f| \leq \epsilon$ 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 $t' = \mathcal{O}\Big(\frac{\log(\Gamma/\epsilon)}{\rho_1}\Big)$ ensures that $||\nu K^{t'} - \pi||_{\mathrm{TV}} \leq \epsilon/2$ and therefore $|\nu K^{t'} f - \pi f| \leq \epsilon/2$. Choosing $N = \mathcal{O}\big(\epsilon^{-2}\big)$ ensures that $|\bar{\pi} f - \nu K^{t'} f| \leq \epsilon/2$ 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 [26,27]. 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 4.2 and 4.3, 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 [28], 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.

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 normalization $\frac{1}{N}\sum_{n=1}^N \frac{q_\pi(x_n)}{q_\nu(x_n)}$ need to be accurately estimated. The numerator is relatively easy to estimate and requires $\mathcal{O}(1/\epsilon^2)$ samples. On the other hand, the normalization needs to have small error relative to Γ^{-1} which requires $\mathcal{O}(\Gamma^2/\epsilon^2)$ samples (Hoeffding), giving an overall complexity of $\mathcal{O}(\Gamma^2/\epsilon^2)$. 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 $\mathcal{O}(1/\rho)$. For many problems of interest Γ may be exponentially large and SMC can be expected to substantially outperform importance sampling.

4.4 Example: finite spaces

Let $\mathcal{X}$ be a finite space with $\pi(x) \propto q(x)$ and $0 < q(x) \leq 1$. 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 $\Gamma \leq \frac{1}{\pi_0}$. 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 $\mathcal{O}\left(\frac{1}{\rho}\right)$ 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 $Z \leq e$. 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.

5 SMC on product measures

Product measures have previously been used to assess the dimension dependence of SMC [15, 16, 19]. 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 [16, 19], 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 [16] and [19] and dealt with via asymptotics in [15], 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 [19] and Eberle and Marinelli [16], though it falls short of the $\mathcal{O}(d^2)$ rate obtained by Beskos et al. [15] 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.

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 $q(x|\phi) = \exp\left(-\frac{1}{2}\phi \cdot x^T x\right)$ 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 $z_0/z_1 = \left(1 + \frac{\phi-1}{d}\right)^{d/2} \geq \exp\left(\frac{\phi-1}{2}\right)$ with $Z \approx \exp\left(\frac{\phi-1}{2}\right)$ for d large. Assuming d sufficiently large, the overall complexity of SMC is then bounded above by

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

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 $Z \leq \sqrt{\phi}$ and yields an SMC bound of

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

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 $Z \leq e^{1/2}$ 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.

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. [29].

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 [29], we will also assume that $\log q$ is L-smooth and m-strongly concave, i.e. that

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

for all $x, y \in \mathcal{R}^d$. For x^* the mode of π, this implies that $-L||x-x^*||_2^2 \leq 2\log q(x) \leq -m||x-x^*||_2^2$. 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 [30]. This restriction ensures that the ratio of normalizing constants is bounded in the first step; since $\pi(B) \leq 1 - \epsilon/2$ this assumption has minimal impact on the results of our analysis [29]. 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 $Z = \mathcal{O}(1)$.

For our Markov kernel we use the Metropolis-adjusted Langevin algorithm (MALA) 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 [31]. Dwivedi et al. [29] 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 [29]. 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 $Z \leq e^{\frac{1}{2}}$. 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.

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:

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

Assign prior $p_0(\beta) = N\big(0, \frac{\alpha}{n}(X^TX)^{-1}\big)$ with the parameter α controlling the strength of the prior shrinkage toward zero. The resulting posterior distribution $q(\beta) \propto p_0(\beta)p(Y \mid X,\beta)$ is log-concave and satisfies the above assumptions of L-smoothness and m-strong concavity with $L \leq (n/4+\alpha)\cdot\sigma_{\max}$ and $m \geq \alpha\cdot\sigma_{\min}$ for $\sigma_{\max}$ and $\sigma_{\min}$ the largest and smallest eigenvalues of $(X^TX)^{-1}/n$, respectively [29]. Inserting into our bounds gives an upper bound on the complexity of sampling via SMC:

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

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 [32]).

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 4, 5, and 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 3.1, 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. [33, 34]). 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. [14] for some problems studied by [33,35]. Incorporating local mixing conditions into our methods along the lines of [14, 19, 35] would allow us to obtain results more directly comparable to [33, 35] 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 [12, 13]. The techniques described in this paper may be well suited to demonstrating the stability and usefulness of more general adaptive methods.

Appendix 1: 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.

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 $p \geq 1$. For a signed measure η on $\mathcal{X}$ with $\eta << \mu$, define the p-norm with respect to μ by $||\eta||_{p,\mu} = \mu|\frac{\eta(dx)}{\mu(dx)}|^p$. 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 [36, 37]:

$$||\nu K^t - \mu||_{2,\mu} \leq ||\nu - \mu||_{2,\mu} \cdot (1 - \rho)^t \tag{8}$$

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 $\frac{\log(||\nu - \mu||_{2,\mu} - \log(2\epsilon)}{\rho}$.

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:

$$\begin{aligned} &\mathbf{C}_s^*(\text{i}) \quad ||\hat{\mu}_s - \mu_s||_{2,s} \leq e - 1 \\ &\mathbf{C}_s^*(\text{ii}) \quad \bar{w}_{s+1} \geq \mu_s w_{s+1} \cdot \frac{2}{3} \end{aligned} \tag{9}$$

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 5 (Error of the resampling distribution)**.**
Assume $P(\boldsymbol{C}^*_{s-1}) \geq 3/4$. *Then* $\left|\left|\tilde{\mu}_s(\cdot \mid \boldsymbol{C}^*_{s-1}) - \mu_s\right|\right|_{2,s} \leq 4e \cdot W \cdot Z$

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

$$\begin{aligned}
\Pr\Big(\tilde{X}^{(n)}_s \in B \Bigm| \mathbf{C}^*_{s-1}\Big) &= \sum_{m=1}^{N} \Pr\Big(X^{(m)}_{s-1} \in B,\, E_{m,n} \mid \mathbf{C}^*_{s-1}\Big) \\
&\leq \sum_{m=1}^{N} \frac{\sup_{x\in B}\ q_s(x)/q_{s-1}(x)}{S \cdot \bar{w}_s} \cdot \Pr\Big(X^{(m)}_{s-1} \in B \Bigm| \mathbf{C}^*_{s-1}\Big) \\
&\leq \frac{\sup_{x\in B}\ p_s(x)/p_{s-1}(x)}{S \cdot 2/3} \cdot \sum_{m=1}^{N} \Pr\Big(\mathbf{C}^*_{s-1} \Bigm| X^{(m)}_{s-1} \in B\Big)\Pr\Big(X^{(m)}_{s-1} \in B\Big)/\Pr\Big(\mathbf{C}^*_{s-1}\Big) \\
&\leq 2 \cdot \sup_{x\in B}\ p_s(x)/p_{s-1}(x) \cdot \hat{\mu}_{s-1}(B)
\end{aligned} \tag{10}$$

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

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

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 $p_{s-1}(x) > 0$

$$\begin{aligned}
\tilde{p}_s(x|\mathbf{C}_{s-1}) &= \lim_{r\to 0} \frac{\tilde{\mu}_s\big(B(x,r) \mid \mathbf{C}^*_{s-1}\big)}{\lambda\big(B(x,r)\big)} \\
&\leq 2 \cdot \lim_{r\to 0} \sup_{x\in B(x,r)}\ p_s(x)/p_{s-1}(x) \cdot \frac{\hat{\mu}_{s-1}(B)}{\lambda\big(B(x,r)\big)} \\
&= 2 \cdot p_s(x)/p_{s-1}(x) \cdot \hat{p}_{s-1}(x)
\end{aligned} \tag{11}$$

The first line is the definition of the Radon-Nikodym derivative on $\mathcal{R}^d$ (see [38] chapter 9 or [39] chapter 5), the second is the upper bound from (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

$$\begin{aligned}
\left|\left|\tilde{\mu}_s(\cdot \mid \mathbf{C}^*_{s-1}) - \mu_s(\cdot)\right|\right|_{2,s} &\leq \int \left[\frac{\tilde{p}_s(x)}{p_s(x)}\right]^2 p_s(x)\lambda(dx) \\
&\leq 4 \cdot \int \left[\frac{\hat{p}_{s-1}(x)}{p_{s-1}(x)}\right]^2 \frac{p_s(x)}{p_{s-1}(x)} \cdot p_{s-1}(x)\lambda(dx) \\
&\leq 4 \cdot W \cdot Z \cdot \big(||\hat{\mu}_{s-1} - \mu_{s-1}||_{2,s} + 1\big) \\
&\leq 4e \cdot W \cdot Z
\end{aligned} \tag{12}$$

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

Proof. Follows directly from Lemma 5 and equation (8), using the fact that $-\log(1-\rho) \geq \rho$

$$\begin{aligned} ||\hat{\mu}_s(\cdot \,|\, \mathbf{C}_{s-1}^*) - \mu_s||_{2,s} &\leq 4e \cdot W \cdot Z \cdot (1-\rho)^t \\ &\leq e - 1 \end{aligned} \tag{13}$$

■

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 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 4 using Azuma's inequality.

Lemma 6 (Lower bound on the average weights)**.**
Assume $P(\boldsymbol{C}_{s-1}^*) \geq 3/4$. Fix $\delta_0 \in (0,1]$. Choose $N \geq 72\log(1/\delta_0) \cdot W^2Z^2$ and $t \geq \frac{4+\log(W^2Z^2)}{\rho}$. Then $\boldsymbol{C}_s^*$(ii) $|$ $\boldsymbol{C}_{s-1}^*$ holds with probability at least $1-\delta_0$

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

$$\left|\bar{w}_{s+1} - z_{s+1}/z_s\right| \leq \left|\bar{w}_{s+1} - \hat{\mu}_s w_{s+1}\right| + \left|\hat{\mu}_s w_{s+1} - \mu_s w_{s+1}\right| \tag{14}$$

The bias conditional on $\mathbf{C}_{s-1}^*$ can be controlled using Lemma 5

$$\begin{aligned} \left|\hat{\mu}_s w_{s+1} - \mu_s w_{s+1}\right| &\leq \frac{W}{2} \cdot ||\hat{\mu}_s(\cdot \,|\, \mathbf{C}_{s-1}^*) - \mu_s||_{1,s} \\ &\leq \frac{W}{2} \cdot ||\tilde{\mu}_s(\cdot \,|\, \mathbf{C}_{s-1}^*) - \mu_s||_{2,s} \cdot (1-\rho)^t \\ &\leq \frac{4e \cdot W^2 \cdot Z}{2} \cdot (1-\rho)^t \\ &\leq \frac{Z^{-1}}{6} \\ &\leq \frac{z_{s+1}/z_s}{6} \end{aligned} \tag{15}$$

The stochastic error can be controlled with Azuma's inequality in the same manner as Lemma 4 yielding

$$\left|\bar{w}_{s+1} - \hat{\mu}_s w_{s+1}\right| \leq \frac{z_{s+1}/z_s}{6} \tag{16}$$

Combining (14), (15), and (16) gives the desired result. ■

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

Corollary 6.1 (Modified one step induction condition)**.**
Assume $P(\boldsymbol{C}_{s-1}^*) > 3/4$. Fix $\delta_0 \in (0,1/4]$. Choose $N \geq 72\log(1/\delta_0) \cdot W^2Z^2$ and $t \geq \frac{4+\log(W^2Z^2)}{\rho}$. 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 $Pr\big(|\hat{\pi}f - \pi f| \leq \epsilon \mid \mathbf{C}^*_{S-1}\big) \geq 1 - \delta_0$ using the same approach as in Lemma 6. For $t_S \geq \frac{3+\log(\frac{WZ}{\epsilon})}{\rho}$ 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 $N \geq 8\log\big(\frac{1}{2\delta_0}\big) \cdot \epsilon^{-2}$. Therefore for this N and t_S

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

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

References

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

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

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

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

[21] 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.

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

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

[24] 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.

[25] László Lovász 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.

[26] 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.

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

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

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

[35] 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.

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

[37] 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.

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

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