Bernoulli **25**(1), 2019, 310–340
https://doi.org/10.3150/17-BEJ988

Error bounds for sequential Monte Carlo samplers for multimodal distributions

DANIEL PAULIN[*], AJAY JASRA[**] and ALEXANDRE THIERY[†]

Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546, Singapore. E-mail: [*] paulindani@gmail.com; [**] staja@nus.edu.sg; [†] a.h.thiery@nus.edu.sg

In this paper, we provide bounds on the asymptotic variance for a class of sequential Monte Carlo (SMC) samplers designed for approximating multimodal distributions. Such methods combine standard SMC methods and Markov chain Monte Carlo (MCMC) kernels. Our bounds improve upon previous results, and unlike some earlier work, they also apply in the case when the MCMC kernels can move between the modes. We apply our results to the Potts model from statistical physics. In this case, the problem of sharp peaks is encountered. Earlier methods, such as parallel tempering, are only able to sample from it at an exponential (in an important parameter of the model) cost. We propose a sequence of interpolating distributions called *interpolation to independence*, and show that the SMC sampler based on it is able to sample from this target distribution at a polynomial cost. We believe that our method is generally applicable to many other distributions as well.

Keywords: asymptotic variance bound; central limit theorem; metastability; Potts model; scale invariance; sequential Monte Carlo

1. Introduction

Sequential Monte Carlo sampling [13,20,29] is a method designed to approximate a sequence of probability distributions $\{\boldsymbol{\mu}_k\}_{0\leq k\leq n}$ defined upon a common measurable space $(E,\mathcal{E})$. The method uses $N\geq 1$ samples (or particles) that are generated in parallel and propagated via importance sampling and resampling methods. In the context of this article, we are concerned with the class of algorithms where $\boldsymbol{\mu}_0$ is an easy to sample distribution and $\boldsymbol{\mu}_n$ is a complex distribution and $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_{n-1}$ interpolate (in some sense) between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_n$. In addition, the particles are moved/mutated through Markov kernels of invariant measure $\boldsymbol{\mu}_k$ at time k. The SMC methodology has proven itself to be a very efficient tools for inference in a wide variety of statistical models and applications including stochastic volatility [22], regression models [8] and approximate Bayesian computation [14]. In this article, we develop theoretical tools for analysing SMC samplers [13], a particular class of SMC algorithms, and introduce a new type of interpolating sequences of distributions that enjoys, in many situations, better convergence properties than more standard tempering sequences that are usually used in practice. The SMC methodology is by now fairly well understood; for example, high-dimensional asymptotic results are obtained in [1,2], the study of the long-time behaviour is presented in [21,36] and its performances for exploring multimodal distributions are described in [16,34]. For a book-length treatment of the subject, the reader is referred to [12]. [17] has shown concentration inequalities and moment bounds that take into account global mixing properties of the Markov kernels. These results are

formulated in terms of the so-called Dobrushin coefficients, that is, the contraction rate of the Markov kernels in total variational distance. The authors also generalise their results to stochastic optimization algorithms. The main results in this paper, Theorems 3.1 and 3.2 only bound the asymptotic variance. Theorem 3.1 is using global mixing properties (spectral gap) of the Markov kernels, while Theorem 3.2 is using local mixing properties (thus it is more suited to multimodal distributions). Proving non-asymptotic bounds under similar conditions is an interesting problem for further research.

Multimodal distributions appear in a wide variety of applications in statistics, physics, economics and many more. However, sampling from such distributions is a challenging problem. In the context of interest, one well-known advantage of SMC samplers over traditional Markov Chain Monte Carlo (MCMC) methods is their ability to work relatively well for multimodal distributions. Although this phenomenon is known to practitioners, there have been only very few attempts to rigorously explain and quantify this behaviour. [16] and [34] were the first to show error bounds (moment bounds) for SMC samplers when applied to explore multimodal distributions. These results are extremely interesting from a conceptual perspective; unfortunately, the applications of these results require very stringent assumptions that are rarely met in practical scenarios of interest to practitioners. One of the main purposes of this article is to develop widely applicable tools for studying the asymptotic properties of SMC samplers when applied to probe multimodal distributions. To this end, we leverage a metastable approximation to obtain new bounds on the asymptotic variance of the SMC estimates; see [30] for a comprehensive monograph on metastability. These bounds show that if the time scale it takes for the Markov kernels to approximate a mixture of local equilibrium distributions sufficiently well is polynomial in some size parameter of the system, then the SMC sampler can sample from the multimodal target distribution in polynomial time. In other words, this shows that Markov kernels with good metastability properties can be leveraged to construct SMC samplers that can explore multimodal target distributions in polynomial time.

We demonstrate the applicability of our results by analysing a model from statistical physics, the Potts model with three colours at critical temperature. Earlier methods, such as parallel tempering, are only able to sample configurations from the Potts model at an exponential cost [4] when using standard tempering bridging distributions; this is mainly caused by the appearance of both wide and narrow peaks in the target distribution. Indeed, [39] have shown that, in general, parallel and simulated tempering using tempering distributions are torpidly mixing for such target distributions. The recent paper [5] has introduced model specific interpolating distributions for the Potts model called *entropy dampening distributions* and proven (Theorem 7.7 of [5]) that the simulated tempering algorithm mixes in polynomial time when using these distributions.

The other main contribution of this article is the introduction of a new general interpolating sequence of distributions, coined *interpolation to independence sequence*. We rigorously prove that an SMC sampler utilizing this newly developed interpolating sequence can generate configurations of the Potts model at a computational cost that only scales cubicly in the system size; this improves improving upon the earlier polynomial rate obtained in [5]. The *interpolation to independence sequence* is not model specific; we believe that it has a wide range of potential applications to many systems that display scale invariance properties.

The paper is organised as follows. In Section 2, we introduce the basic tools required, such as Feynman-Kac semigroups. In Section 3, we state and prove our general asymptotic bounds. In

Section 4, we introduce the interpolation to independence sequence of distributions. Section 5 states our results for the Potts model, and Section 6 contains the proofs of these results.

1.1. Notations

For a function $\varphi : E \to \mathbb{R}$, the supremum norm is written as $\|\varphi\|_\infty = \sup_{x\in E} |\varphi(x)|$. Consider a probability measure $\boldsymbol{\mu}$ on E and $\varphi \in L^1(\boldsymbol{\mu})$; we repeatedly use the shorthand notation $\boldsymbol{\mu}(\varphi) = \int_E \varphi(x)\boldsymbol{\mu}(dx)$ and write $\mathrm{Var}_{\boldsymbol{\mu}}(\varphi)$ for the variance of φ under $\boldsymbol{\mu}$. The Hilbert space $L^2(\boldsymbol{\mu})$ has scalar product

$$\langle f, g\rangle_{\boldsymbol{\mu}} = \int_E f(x)g(x)\boldsymbol{\mu}(dx)$$

and associated norm $\|\cdot\|_{L^2(\boldsymbol{\mu})}$. We sometimes identify $\boldsymbol{\mu}$ with the linear operator from $L^2(\boldsymbol{\mu})$ to itself that maps the function φ to the constant function that equals $\boldsymbol{\mu}(\varphi)$ everywhere. A Markov kernel $\boldsymbol{K}$ that lets $\boldsymbol{\mu}$ invariant is identified with the linear operator $\boldsymbol{K} : L^2(\boldsymbol{\mu}) \to L^2(\boldsymbol{\mu})$

$$\boldsymbol{K}\varphi(x) = \int \boldsymbol{K}(x, dy)\varphi(y).$$

For an operator $\boldsymbol{L} : L^2(\boldsymbol{\mu}) \to L^2(\boldsymbol{\mu})$, its triple norm equals

$$|||\boldsymbol{L}|||_{L^2(\boldsymbol{\mu})} = \sup\big\{\|\boldsymbol{L}\varphi\|_{L^2(\boldsymbol{\mu})} : \boldsymbol{\mu}\big(\varphi^2\big) \le 1\big\}.$$

Similarly, the quantity $|||\boldsymbol{L}|||_\infty$ equals the supremum of $\|\boldsymbol{L}\varphi\|_\infty$ over the set of test functions such that $\|\varphi\|_\infty \le 1$. The notation $\mathcal{N}(m, \sigma^2)$ designates the Gaussian distribution with mean m and variance σ^2. We use the notation $A \sqcup B$ to denote the union of the disjoint subsets $A, B \subset E$. Finally, for a function $\varphi : E \to \mathbb{R}$ and a subset $S \subset E$, the function $\varphi_{|S} : S \to \mathbb{R}$ is the restriction of φ to S; for $x \in S$, we have $\varphi_{|S}(x) = \varphi(x)$.

2. Preliminaries

Suppose that we are interested in inference from some distribution $\boldsymbol{\mu}$ on some Polish state space $(E, \mathcal{E})$. We define an interpolating sequence $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ of distributions with $\boldsymbol{\mu}_n = \boldsymbol{\mu}$; the distribution $\boldsymbol{\mu}_0$ is chosen so that it is straightforward to generate independent samples from it. In this article, we assume that for any index $0 \le k \le n-1$ the distribution $\boldsymbol{\mu}_{k+1}$ is absolutely continuous with respect to $\boldsymbol{\mu}_k$ and denote by $g_{k,k+1}$ the Radon–Nykodym derivative

$$g_{k,k+1} = \frac{d\boldsymbol{\mu}_{k+1}}{d\boldsymbol{\mu}_k}.$$

We work under the standing assumption that these Radon–Nikodym derivatives are bounded and set

$$\Gamma_g = \max\big\{\|g_{k,k+1}\|_\infty : 0 \le k \le n-1\big\} < \infty.$$

We will make extensive use of the linear operator $\boldsymbol{G}_{k,k+1} : L^2(\boldsymbol{\mu}_{k+1}) \to L^2(\boldsymbol{\mu}_k)$ defined as

$$\boldsymbol{G}_{k,k+1}\varphi = g_{k,k+1}\varphi. \tag{2.1}$$

For a test function $\varphi : E \to \mathbb{R}$, our goal is to estimate the performances of the Sequential Monte Carlo (SMC) algorithm for estimating the expectation $\boldsymbol{\mu}(\varphi)$. Recall that the SMC algorithm with N particles proceeds as follows [13]. An initial set $\{\xi_0^1, \ldots, \xi_0^N\}$ of N i.i.d. samples from the probability distribution $\boldsymbol{\mu}_0$ is generated. The empirical distribution

$$\boldsymbol{\mu}_0^N = (1/N) \sum_{i=1}^N \boldsymbol{\delta}_{\xi_0^i},$$

where $\boldsymbol{\delta}_a$ denotes the Dirac mass at $a \in E$, is an approximation of $\boldsymbol{\mu}_0$. In order to produce a particle approximation of $\boldsymbol{\mu}$, the algorithm iterates *mutation* and *resampling* steps. Suppose that a particle approximation $\boldsymbol{\mu}_k^N = (1/N) \sum_{i=1}^N \boldsymbol{\delta}_{\xi_k^i}$ has already been obtained. The mutation steps generates N particles $\{\widetilde{\xi}_k^1, \ldots, \widetilde{\xi}_k^N\}$ distributed as $\widetilde{\xi}_k^i \sim \boldsymbol{K}_k(\xi_k^i, dx)$ where $\boldsymbol{K}_k(x, dy)$ is a Markov kernel that lets the distribution $\boldsymbol{\mu}_k$ invariant; given $\{\xi_k^1, \ldots, \xi_k^N\}$, the particles $\widetilde{\xi}_k^1, \ldots, \widetilde{\xi}_k^N$ are independent. The subsequent particle approximation $\boldsymbol{\mu}_{k+1}^N = (1/N) \sum_{i=1}^N \boldsymbol{\delta}_{\xi_{k+1}^i}$ is obtained through a multinomial resampling step; the particles $\{\xi_{k+1}^1, \ldots, \xi_{k+1}^N\}$ are N i.i.d. samples from the $\{\xi_k^1, \ldots, \xi_k^N\}$-valued random variable that equals ξ_k^i with probability $g_{k,k+1}(\xi_k^i)/[g_{k,k+1}(\xi_k^1) + \cdots + g_{k,k+1}(\xi_k^N)]$. This procedures can be iterated to produce a sequence of particle approximations $\boldsymbol{\mu}_0^N, \ldots, \boldsymbol{\mu}_n^N$. The output of the SMC algorithm employing $N \geq 1$ particles is an empirical approximation $\boldsymbol{\mu}_n^N$ to $\boldsymbol{\mu}_n = \boldsymbol{\mu}$,

$$\boldsymbol{\mu}_n^N = (1/N) \sum_{i=1}^N \boldsymbol{\delta}_{\xi_n^i}.$$

The mutation and resampling steps are also frequently used in genetic optimization algorithms, and some of these algorithms can be analysed in terms of the same Feynman-Kac formulation, we refer the reader to [11] for more details. Asymptotic properties of the SMC algorithm are by now well understood (see, e.g., [15], and [12] for a comprehensive overview). For ease of presentation, we will often present our results for functions with mean zero and finite moment of order $(1 + \varepsilon)$ for some $\varepsilon > 0$; in other words, for a probability distribution $\boldsymbol{\pi}$, we consider the linear subspace

$$L_0^{2+}(\boldsymbol{\pi}) := \big\{\varphi : E \to \mathbb{R} \text{ such that } \boldsymbol{\pi}\big(|\varphi|^{2+\varepsilon}\big) < \infty \text{ for some } \varepsilon > 0 \text{ and } \boldsymbol{\pi}(\varphi) = 0\big\}.$$

Theorem 1 of [9] implies that for a test function $\varphi \in L_0^{2+}(\boldsymbol{\mu})$, the following limit holds in distribution,

$$\lim_{N\to\infty} N^{1/2}\big[\boldsymbol{\mu}_n^N - \boldsymbol{\mu}\big](\varphi) = \mathcal{N}\big(0, V_n(\varphi)\big), \tag{2.2}$$

with asymptotic variance $V_n(\varphi)$ that can be expressed as

$$V_n(\varphi) = \sum_{k=0}^{n} V_{k,n}(\varphi) \qquad \text{with } V_{k,n}(\varphi) := \|\boldsymbol{G}_{k,k+1}\boldsymbol{K}_{k+1}\cdots\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_k)} \tag{2.3}$$

and $V_{n,n}(\varphi) := \|f\|^2_{L^2(\boldsymbol{\mu}_k)} = \mathrm{Var}_{\boldsymbol{\mu}}(\varphi)$. We note that the CLT and the expression (2.3) was first proven for multivariate processes and more general Feynman-Kac models in the case of bounded functions in [15]. In the next section, we establish bounds on the asymptotic variance $V_n(\varphi)$ under various natural conditions.

3. Bounds on the asymptotic variance

In this section, we state and prove new asymptotic variance bounds for SMC empirical averages. Section 3.1 considers bounds under global mixing assumptions. Section A.1 (in the supplementary material [32]) considers the multimodal case, under the assumption that there is no mixing between the modes. In Section 3.2, we obtain general results for multimodal distributions. To lighten the notations, for a positive operator $\boldsymbol{M} : L^2(\boldsymbol{\mu}) \to L^2(\boldsymbol{\mu})$ and test functions $f, g \in L^2(\boldsymbol{\mu})$, we set $\langle f, g\rangle_{\boldsymbol{\mu},\boldsymbol{M}} := \langle f, \boldsymbol{M}g\rangle_{\mu}$ and $\|\varphi\|^2_{L^2(\boldsymbol{\mu}),\boldsymbol{M}} := \langle\varphi, \boldsymbol{M}\varphi\rangle_{\boldsymbol{\mu}}$. In particular, we have that $\|\varphi\|_{L^2(\boldsymbol{\mu}_k),\boldsymbol{G}_{k,k+1}} = \|\varphi\|_{L^2(\boldsymbol{\mu}_{k+1})}$, with $\boldsymbol{G}_{k,k+1}$ defined as in Equation (2.1).

3.1. Bound under global mixing assumptions

The following theorem bounds the asymptotic variance $V_n(\varphi)$ in terms of the "global" mixing properties of the Markov kernels $\boldsymbol{K}_k$ and the size of the relative density $g_{k,k+1}$. Before stating our result, we need to introduce some notations. Recall that $\boldsymbol{\mu}_k : L^2(\boldsymbol{\mu}_k) \mapsto L^2(\boldsymbol{\mu}_k)$ denotes the orthogonal projection operator that maps a function φ to the constant function that equals $\boldsymbol{\mu}_k(\varphi)$ everywhere and that the Markov operator $\boldsymbol{K}_k$ lets $\boldsymbol{\mu}_k$ invariant. We define the quantity

$$\gamma_{\boldsymbol{K}} := 1 - \max\left\{\|\boldsymbol{K}_k - \boldsymbol{\mu}_k\|_{L^2(\boldsymbol{\mu}_k)} : 1 \le k \le n\right\}; \tag{3.1}$$

For any test function $\varphi \in L^2(\boldsymbol{\mu}_k)$, we thus have that $\|(\boldsymbol{K}_k - \boldsymbol{\mu}_k)\varphi\|_{L^2(\boldsymbol{\mu}_k)} \le (1-\gamma_{\boldsymbol{K}})\|\varphi\|_{L^2(\boldsymbol{\mu}_k)}$. In the case where the Markov kernels $\boldsymbol{K}_k$ are reversible, the quantity $\gamma_{\boldsymbol{K}}$ is a uniform lower bound on their absolute spectral gap. The larger $\gamma_{\boldsymbol{K}}$, the better the mixing properties of these Markov kernels.

Theorem 3.1 (Variance bound under a global mixing assumption). *Let $\varphi \in L_0^{2+}(\boldsymbol{\mu})$ be a test function. Assume that*

$$\Gamma_g < \frac{1}{(1-\gamma_{\boldsymbol{K}})^2}. \tag{3.2}$$

The CLT (2.2) holds with asymptotic variance $V_n(\varphi)$ such that

$$V_n(\varphi) \le \frac{1}{1-(1-\gamma_{\boldsymbol{K}})^2\cdot\Gamma_g}\,\mathrm{Var}_{\boldsymbol{\mu}_n}(\varphi).$$

Proof. Recall the formula (2.3) for the asymptotic variance. First, note that $V_{n,n}(\varphi) = \mathrm{Var}_{\boldsymbol{\mu}_n}(\varphi)$. By definition of the upper bound Γ_g and the operators $\boldsymbol{G}_{k,k+1}$, it follows that

$$\begin{aligned} V_{k,n}(\varphi) &= \|\boldsymbol{G}_{k,k+1}\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_k)} \\ &\leq \Gamma_g\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_k),\boldsymbol{G}_{k,k+1}} \\ &= \Gamma_g\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_{k+1})}. \end{aligned}$$

Also, since the Markov kernel $\boldsymbol{K}_j$ let $\boldsymbol{\mu}_j$ invariant and $\boldsymbol{\mu}(\varphi) = 0$, we have that

$$\boldsymbol{\mu}_{k+1}\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi(x) = 0$$

for any $x \in E$. Consequently, the quantity $\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_{k+1})}$ can also be expressed as

$$\left\|(\boldsymbol{K}_{k+1} - \boldsymbol{\mu}_{k+1})\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\right\|^2_{L^2(\boldsymbol{\mu}_{k+1})}. \tag{3.3}$$

The definition (3.1) of $\gamma_{\boldsymbol{K}}$ further yields that (3.3) is less than

$$(1-\gamma_{\boldsymbol{K}})^2\|\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_{k+1})} \tag{3.4}$$

so that $V_{k,n}(\varphi) \leq \Gamma_g(1-\gamma_{\boldsymbol{K}})^2\|\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2}\cdot\ldots\cdot\boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_{k+1})}$. Keeping in mind that $\|\varphi\|^2_{L^2(\boldsymbol{\mu}_k)} = \mathrm{Var}_{\boldsymbol{\mu}_k}(\varphi)$, iterating the same arguments shows that

$$V_{k,n}(\varphi) \leq \left(\Gamma_g(1-\gamma_{\boldsymbol{K}})^2\right)^{n-k}\mathrm{Var}_{\boldsymbol{\mu}}(\varphi).$$

Since $V_n(\varphi) = \sum_{k=0}^n V_{k,n}(\varphi)$ and $\Gamma_g(1-\gamma_{\boldsymbol{K}})^2 < 1$, the conclusion follows. □

Theorem 3.1 gives an improvement over the quadratic error bounds provided by Theorem 1.2 of [34]; indeed, contrarily to their result, our bound on the asymptotic variance does not depend on the number $n \geq 1$ of resampling stages. Note that the required assumption (3.2) can easily be enforced by including a sufficient number $n \geq 1$ of resampling stages and/or by increasing the amount of MCMC steps at each stage. It is important to note that Theorem 3.1 does not assume that the target distribution $\boldsymbol{\mu}$ is unimodal in any sense; instead, assumptions on the global mixing properties of the Markov kernels $\boldsymbol{K}_k$ are leveraged. However, and as is widely acknowledged in the Markov Chain Monte-Carlo literature, it is generally difficult to design Markov kernels with good global mixing properties for multimodal distributions. This is remark is one of main motivations for our work; in the next sections, we describe results that do not require the Markov kernels $\boldsymbol{K}_k$ to possess good global mixing properties.

3.2. Bound for the multimodal case

In this section, we examine the case of multiple modes. Mixing between modes is allowed. We look at partitions of the state space E that may vary with the algorithm index k,

$$E := \bigsqcup_{j=1}^{m(k)} F_k^{(j)}.$$

This means that one allows different modes, and potentially a different number of modes, for each intermediate distribution $\boldsymbol{\mu}_k$. This extra generality allows us to analyse a wider range of interpolating distributions. In particular, we will use this property in the analysis of the Potts model (see Section 5). We define the *growth-within-mode constant* as

$$B_{k,k+1} := \max\left\{\boldsymbol{\mu}_{k+1}\left(F_k^{(r)}\right)/\boldsymbol{\mu}_k\left(F_k^{(r)}\right) : 1 \le r \le m(k)\right\}. \tag{3.5}$$

The restriction of $\boldsymbol{\mu}_k$ to $F_k^{(r)}$, denoted by $\boldsymbol{\mu}_{k,r}$, is defined as

$$\boldsymbol{\mu}_{k,r}(S) := \frac{\boldsymbol{\mu}_k(S \cap F_k^{(r)})}{\boldsymbol{\mu}_k(F_k^{(r)})} \qquad \text{for every measurable } S \subset E.$$

Consider the situation where, as is common in practice, the Markov kernel $\boldsymbol{K}_k$ is of the form $\boldsymbol{K}_k = \boldsymbol{P}_k^{t_k}$ for some Markov kernel $\boldsymbol{P}_k$; in words, the kernel $\boldsymbol{K}_k$ corresponds to iterating t_k steps of the Markov kernel $\boldsymbol{P}_k$. We introduce an approximation called *metastable state*, which is a kernel $\widehat{\boldsymbol{\mu}}_k$ defined as

$$\widehat{\boldsymbol{\mu}}_k(x, \varphi) = \sum_{r=1}^{m(k)} \alpha_{k,r}(x) \boldsymbol{\mu}_{k,r}(\varphi_{|F_k^{(r)}}), \tag{3.6}$$

where for every $x \in E$ and index $1 \le k \le n$ the family $\{\alpha_{k,r}(x)\}_{r=1}^{m(k)}$ is a sequence of non-negative real numbers that are such that

$$\sum_{r=1}^{m(k)} \alpha_{k,r}(x) \le 1. \tag{3.7}$$

Since $\sum_{r=1}^{m(k)} \alpha_{k,r}(x)$ can be strictly smaller than one, the metastable operator $\widehat{\boldsymbol{\mu}}_k$ is not necessarily a Markov kernel. A natural choice is $\alpha_{k,r}(x) = \boldsymbol{K}_k(x, F_k^{(r)})$ (the probability of ending up in mode $F_k^{(r)}$ when started from x). Another possibility, useful when the chain mixes well globally, consists in setting $\alpha_{k,r}(x) = \boldsymbol{\mu}_k(F_k^{(r)})$; this approximation results in $\widehat{\boldsymbol{\mu}}_k(x, dy) = \boldsymbol{\mu}_k(dy)$. As will become clear in Section 3.3, for a suitable choice of coefficients $\alpha_{k,r}(x)$, the approximation $\boldsymbol{K}_k \approx \widehat{\boldsymbol{\mu}}_k$ is often accurate, even for reasonably small values of t_k. The following result is our variance bound in this setting.

Theorem 3.2 (Variance bound for multimodal case with mixing). *Assume that* $\Gamma_g < \infty$. *For a bounded and measurable test function* φ, *the CLT* (2.2) *holds with an asymptotic variance* $V_n(\varphi) = \sum_{k=0}^{n} V_{k,n}(\varphi)$ *where, for any index* $0 \le k \le n-1$, *we have*

$$V_{k,n}(\varphi) \le \Gamma_g \prod_{j=k+1}^{n-1} \left\{ B_{j,j+1} + \Gamma_g \|\boldsymbol{K}_j - \widehat{\boldsymbol{\mu}}_j\|_\infty \right\} \|\varphi\|_\infty^2. \tag{3.8}$$

Proof. Since $\|\boldsymbol{G}_{k,k+1}\|_\infty \le \Gamma_g$, we have

$$V_{k,n}(\varphi) \le \Gamma_g \|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_{k+1})}.$$

Moreover, $\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|^2_{L^2(\boldsymbol{\mu}_{k+1})}$ is less than

$$\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|_\infty \times \|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|_{L^1(\boldsymbol{\mu}_{k+1})}.$$

Since the Markov kernel $\boldsymbol{K}_{k+1}$ lets $\boldsymbol{\mu}_{k+1}$ invariant, this is a contraction in $L^1(\boldsymbol{\mu}_{k+1})$; consequently

$$\begin{aligned}
&\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|_{L^1(\boldsymbol{\mu}_{k+1})} \\
&\quad \le \|\boldsymbol{G}_{k+1,k+2}\boldsymbol{K}_{k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|_{L^1(\boldsymbol{\mu}_{k+1})} \\
&\quad = \|\boldsymbol{K}_{k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|_{L^1(\boldsymbol{\mu}_{k+2})} \le \cdots \le \|\varphi\|_{L^1(\boldsymbol{\mu})} \le \|\varphi\|_\infty.
\end{aligned}$$

Also, since $\|\boldsymbol{K}_n\varphi\|_\infty \le \|\varphi\|_\infty$, we have that

$$\|\boldsymbol{K}_{k+1}\boldsymbol{G}_{k+1,k+2} \cdot \ldots \cdot \boldsymbol{G}_{n-1,n}\boldsymbol{K}_n\varphi\|_\infty \le \left\{ \prod_{j=k+1}^{n-1} |||\boldsymbol{K}_j\boldsymbol{G}_{j+1}|||_\infty \right\} \|\varphi\|_\infty.$$

Furthermore, $|||\boldsymbol{K}_j\boldsymbol{G}_{j+1}|||_\infty \le |||\widehat{\boldsymbol{\mu}}_j\boldsymbol{G}_{j+1}|||_\infty + |||(\boldsymbol{K}_j - \widehat{\boldsymbol{\mu}}_j)\boldsymbol{G}_{j+1}|||_\infty$. Definition (3.5) yields that $|||\widehat{\boldsymbol{\mu}}_j\boldsymbol{G}_{j+1}|||_\infty$ is less than $B_{j,j+1}$; similarly, $|||(\boldsymbol{K}_j - \widehat{\boldsymbol{\mu}}_j)\boldsymbol{G}_{j+1}|||_\infty \le \Gamma_g |||\boldsymbol{K}_j - \widehat{\boldsymbol{\mu}}_j|||_\infty$. It follows that

$$|||\boldsymbol{K}_j\boldsymbol{G}_{j+1}|||_\infty \le B_{j,j+1} + \Gamma_g |||\boldsymbol{K}_j - \widehat{\boldsymbol{\mu}}_j|||_\infty,$$

as required. □

Note that unlike Theorem 3.1, here we use the supremum norm (thus our result is restricted to bounded functions), because we have encountered some technical difficulties when using $\|\cdot\|_{L^2(\boldsymbol{\mu}_k)}$ norms in this setting. The main improvement in this theorem over the results of [34] is that mixing is allowed between the modes. For completeness, a variant of the asymptotic variance bound of [34] (when no mixing is allowed between the modes) is presented in Section A.1 (in the supplementary material [32]).

3.3. Framework for metastable approximation

In order to apply Theorem 3.2, one needs to bound the norm $\|\boldsymbol{K}_k - \widehat{\boldsymbol{\pi}}_k\|_{L^\infty}$ for $1 \le k \le n$. This section provides with a framework for establishing such bounds. Suppose that the state space E is partitioned into modes $\{\mathcal{F}^{(j)}\}_{j=1}^m$ and that each mode is comprised of an *inner region* $\mathcal{I}^{(j)}$ and a *border region* $\mathcal{B}^{(j)}$; in other words, we have the following decomposition of the state space,

$$E = \bigsqcup_{i=1}^{m} \mathcal{F}^{(j)} = \bigsqcup_{i=1}^{m} \{\mathcal{I}^{(j)} \sqcup \mathcal{B}^{(j)}\}.$$

It will reveal useful to set

$$\mathcal{I} = \bigsqcup_{j=1}^{m} \mathcal{I}^{(j)} \quad \text{and} \quad \mathcal{B} = \bigsqcup_{j=1}^{m} \mathcal{B}^{(j)}.$$

For every $1 \le j \le m$, we denote restrictions of $\boldsymbol{\mu}$ to $\mathcal{F}^{(j)}$ by $\boldsymbol{\mu}^{(j)}$, defined by the relation

$$\boldsymbol{\mu}^{(j)}(\varphi) = \frac{\boldsymbol{\mu}(\varphi_{|\mathcal{F}^{(j)}})}{\boldsymbol{\mu}(\mathcal{F}^{(j)})}. \tag{3.9}$$

For $x \in E$ and an integer $t \ge 1$, consider the quantity

$$q^{(i)}(x,t) := \mathbb{P}\big(X_{\tau_{\mathcal{B}}} \in \mathcal{I}^{(i)}, \tau_{\mathcal{B}} \le t | X_0 = x\big),$$

where $\{X_k\}_{k\ge 0}$ is a Markov chain with Markov transition kernel $\boldsymbol{P}$, and $\tau_{\mathcal{B}}$ is the time of exit from $\mathcal{B}$, $\tau_{\mathcal{B}} := \inf\{t \ge 0 : X_t \notin \mathcal{B}\}$. This expresses the probability that we exit the border regions in one of the first t steps, and the first step outside $\mathcal{B}$ is in the inner region $\mathcal{I}^{(i)}$. Our main result in this section, Theorem 3.3 quantifies the approximation $\boldsymbol{P}^t \approx \widehat{\boldsymbol{\pi}}^{(t)}$, where the kernel $\widehat{\pi}^{(t)}$ is defined as

$$\widehat{\pi}^{(t)}(x, dy) := \begin{cases} \boldsymbol{\mu}^{(j)}(dy) & \text{for } x \in \mathcal{I}^{(j)}, \\ \sum_j q^{(j)}\big(x, \lfloor t/2 \rfloor\big) \cdot \boldsymbol{\mu}^{(j)}(dy) & \text{for } x \in \mathcal{B}. \end{cases} \tag{3.10}$$

Note that $\sum_{i=1}^{m(k)} q^{(i)}(x,t) = \mathbb{P}(\tau_{\mathcal{B}} \le t) \le 1$ so condition (3.7) is satisfied. The bound on the discrepancy $\|\!|\boldsymbol{P}^t - \widehat{\boldsymbol{\pi}}^{(t)}|\!\|_\infty$ is expressed in terms of the event $\mathrm{S}_{\mathcal{B}}(x,t)$ (stay in the border region) defined as

$$\mathrm{S}_{\mathcal{B}}(x,t) := \{\text{start at } x \in \mathcal{B} \text{ and stay inside } \mathcal{B} \text{ for } t \text{ steps}\}.$$

When starting in the inner region of a mode and after a number of steps slightly larger than the local mixing time, the Markov chain is typically approximately distributed according to the restriction of the stationary distribution to that mode; nevertheless, the Markov chain is still not likely to escape from that mode. When starting from a border region, the Markov chain typically enters the inner region rapidly then stay there in the rest of the steps, and mix well within that mode. The number of steps thus needs to be chosen carefully to make sure that we exit the border

regions and mix well within the modes, but do not exit the modes once we have entered an inner region (due to the "potential well" effect).

Theorem 3.3 (Quantifying the quality of metastable approximation). *Let $\widehat{\boldsymbol{\pi}}^{(t)}$ be defined as in* (3.10)*, then the following bound holds,*

$$\begin{aligned}\left|\!\left|\!\left| \boldsymbol{P}^t - \widehat{\boldsymbol{\pi}}^{(t)} \right|\!\right|\!\right|_\infty \leq & \max_{x\in\mathcal{B}} \mathbb{P}\big(\mathrm{S}_{\mathcal{B}}\big(x, \lfloor t/2\rfloor\big)\big) \\ & + 2 \max_{1\leq i\leq m} \max_{\lceil t/2\rceil \leq r\leq t} \sup_{x\in\mathcal{I}^{(i)}} d_{\mathrm{TV}}\big(\boldsymbol{P}^r(x,\cdot), \boldsymbol{\mu}^{(i)}\big).\end{aligned} \tag{3.11}$$

Proof. For two distributions η_1, η_2 on E (which are not necessarily probability distributions), we define their total variational distance as

$$d_{\mathrm{TV}}(\eta_1,\eta_2) := \frac{1}{2} \sup_{f:E\to[-1,1]} \big|\eta_1(f) - \eta_2(f)\big|,$$

where the supremum is taken among Borel-measurable functions from $\mathbb{E}$ to $[-1,1]$. By the definition of $|\!|\!|\boldsymbol{P}^t - \widehat{\boldsymbol{\pi}}^{(t)}|\!|\!|_\infty$, we can rewrite it as

$$\left|\!\left|\!\left| \boldsymbol{P}^t - \widehat{\boldsymbol{\pi}}^{(t)} \right|\!\right|\!\right|_\infty = 2 \sup_{x\in E} d_{\mathrm{TV}}\big(\boldsymbol{P}^t(x,\cdot), \widehat{\boldsymbol{\pi}}^{(t)}(x,\cdot)\big),$$

so we need to bound this total variational distance for every $x \in E$. We consider two separate cases.

Fist, assume that $X_0 = x \in \mathcal{I}^{(i)}$ for some $1 \leq i \leq m$. In this case, $q^{(i)}(x, \lfloor t/2\rfloor) = 1$, so $\widehat{\boldsymbol{\pi}}^{(t)}(x,\cdot) = \boldsymbol{\mu}^{(i)}(\cdot)$, and thus

$$2d_{\mathrm{TV}}\big(\boldsymbol{P}^t(x,\cdot), \widehat{\boldsymbol{\pi}}^{(t)}(x,\cdot)\big) = 2d_{\mathrm{TV}}\big(\boldsymbol{P}^t(x,\cdot), \boldsymbol{\mu}^{(i)}\big),$$

which is bounded by the right-hand side of (3.11).

Alternatively, assume that $X_0 = x \in \mathcal{B}$. For $1 \leq i \leq m$, define the events

$$E^{(i)} := \big\{X_0 = x, X_{\tau_{\mathcal{B}}} \in \mathcal{I}^{(i)}, \tau_{\mathcal{B}} \leq \lfloor t/2\rfloor\big\}.$$

Let $E^{\cup} := \cup_{1\leq i\leq m} E^{(i)}$, and $E^c = (E^{\cup})^c$ (the complement of $E^{\cup}$). Then $\mathbb{P}(E^{(i)}) = q^{(i)}(x, \lfloor t/2\rfloor)$, and we can write the two kernels $\boldsymbol{P}^t(x,dy)$ and $\widehat{\boldsymbol{\pi}}^{(t)}(x,dy)$ as

$$\boldsymbol{P}^t(x,dy) = \mathbb{P}\big(X_t \in dy | E^c\big)\cdot\mathbb{P}\big(E^c\big) + \sum_{i=1}^m \mathbb{P}\big(X_t \in dy | E^{(i)}\big)\cdot\mathbb{P}\big(E^{(i)}\big),$$

$$\widehat{\boldsymbol{\pi}}^{(t)}(x,dy) = \sum_{i=1}^m \boldsymbol{\mu}^{(i)}(dy)\cdot\mathbb{P}\big(E^{(i)}\big).$$

Based on this, by the triangle inequality, we have

$$2d_{\mathrm{TV}}\big(\boldsymbol{P}^t(x,\cdot), \widehat{\boldsymbol{\pi}}^{(t)}(x,\cdot)\big) \leq \mathbb{P}\big(E^c\big) + 2\sum_{i=1}^m \mathbb{P}\big(E^{(i)}\big)\cdot d_{\mathrm{TV}}\big(\mathcal{L}\big(X_t|E^{(i)}\big), \boldsymbol{\mu}^{(i)}\big),$$

where $\mathcal{L}(X_t|E^{(i)})$ denotes the law of X_t conditioned on $E^{(i)}$. The result now follows from the fact that $\mathbb{P}(E^c) \leq \max_{x\in\mathcal{B}} \mathbb{P}(\mathrm{S}_{\mathcal{B}}(x, \lfloor t/2 \rfloor))$ and

$$d_{\mathrm{TV}}\big(\mathcal{L}\big(X_t|E^{(i)}\big), \boldsymbol{\mu}^{(i)}\big) \leq \max_{1\leq i\leq m} \max_{\lceil t/2\rceil \leq r\leq t} \sup_{x\in\mathcal{I}^{(i)}} d_{\mathrm{TV}}\big(\boldsymbol{P}^r(x,\cdot), \boldsymbol{\mu}^{(i)}\big). \qquad \square$$

To use Theorem 3.3, one does not need to know the exact probabilities $\boldsymbol{P}^t(x, \mathcal{F}^{(i)})$ of ending up in the different modes. Instead, one only needs to estimate the probability of staying in the border regions for many steps, and the total variational distance of $\boldsymbol{P}^t(x,\cdot)$ from the local restriction $\boldsymbol{\mu}^{(i)}$ when started from a point x in an inner region $\mathcal{I}^{(i)}$. As shall be seen in the application, these quantities can typically be bounded using concentration inequalities and drift arguments.

4. Interpolation to independence sequence

Suppose that we have a sequence of probability distributions $(\eta_k)_{k\in\mathbb{Z}_+}$ defined on $(\Lambda^k)_{k\in\mathbb{Z}_+}$, respectively (which are product spaces of increasing dimension), and that our target is η_d. Suppose that these distributions satisfy some sort of *scale invariance property*. By this, we mean that for sufficiently high k, the number, position, and probability mass of the modes are essentially constant in some appropriate coordinate system (i.e., the positions of the modes have approximately reached a limit). For such systems, we define the *interpolation to independence sequence* as a sequence of distributions on Λ^d, denoted by $\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_d$ such that $\boldsymbol{\mu}_d := \eta_d$ (i.e., the target measure in dimension d), and $\boldsymbol{\mu}_k$ corresponds to the distribution when the first k coordinates are distributed according to η_k, and the rest of the coordinates are i.i.d. uniformly distributed on Λ (independently of the first k coordinates). For the SMC sampler based on this sequence, we use some appropriate MCMC kernels $\boldsymbol{K}_k$ that first change the first k coordinates (such as versions of the Glauber dynamics), and then replace the rest of the coordinates by independent copies.

The interpolation to independence sequence consists of miniaturised versions of the original system (and some independent coordinates to keep the state space invariant). As we are going to see, if the system satisfies the scale invariance property, then the number and location of the modes is essentially the same across all the distributions. This ensures that the growth-within-mode constants $B_{k,k+1}$ are small, and thus the method is efficient if the MCMC moves are chosen appropriately (see Theorems 3.2 and 3.3). This is a key difference with the standard tempering sequence, because the change of the temperature parameter might alter the number and location of the modes drastically, so the growth-within-mode constants might be very large (see [4]).

The idea of interpolating to independence have appeared in the literature of Stein's method, see Section 3.4 of [7]. At this point, we note that somewhat similar ideas have appeared for SMC methods in [3], where a gradual coarsening of the grid is used for the solution of PDEs, and in [26], where graphical models are studied, and the interpolating sequence is chosen by breaking them into smaller blocks gradually. Multigrid methods have been fruitful for solving challenging problems in numerical analysis, and MCMC samplers based on this idea have been also proposed in [18]. In addition, some variations to the standard tempering distributions in have been proposed in the literature to parallel tempering, such as truncating the peaks [23] and moving across different dimensional spaces [27].

However, most of these works are lacking in theoretical explanation of the efficiency of the proposed methodology. Moreover, it is not clear to us whether they can overcome the type of problem we encounter for the Potts model (where some of the modes have exponentially small probability according the initial distribution, and thus they might take exponentially long time to be discovered). We note that Theorem 7.7 of [5] has shown that for some model specific interpolating distributions called entropy dampening distributions, the simulated tempering MCMC sampler for the Potts model mixes in polynomial time.

The following section states our application in this paper, a bound on the asymptotic variance of the SMC method using the interpolation to independence sequence of distributions applied to the Potts model. We would like to emphasise the fact that the interpolation to independence methodology is not limited to this single example. A natural generalisation is to use it for sampling from exponential random graph models. For such models, MCMC sampling has been shown to mix very slowly, and the location of the modes depends of the temperature parameter (see [6]). However, they are believed to be scale invariant, so our method could be useful for sampling from them.

5. Application to the Potts model

5.1. Introduction

In this section, we are going to study the Potts model introduced in [33]. Let G be a simple graph with M vertices. The model consists of M spins $\sigma := (\sigma_1, \ldots, \sigma_M)$ taking values in $\Omega := \{1, \ldots, q\}^M$ for some $q \geq 2$ (these are called "colours"). The Hamiltonian of the model is defined as

$$H(\sigma) = -\sum_{(i,j)\in G} J \cdot 1_{[\sigma_i = \sigma_j]}, \tag{5.1}$$

where the summation is over all the edges in G. The sign of J determines whether the neighbors prefer the same colour (ferromagnetic case) or different colour (antiferromagnetic case). The Gibbs distribution on configurations is given by

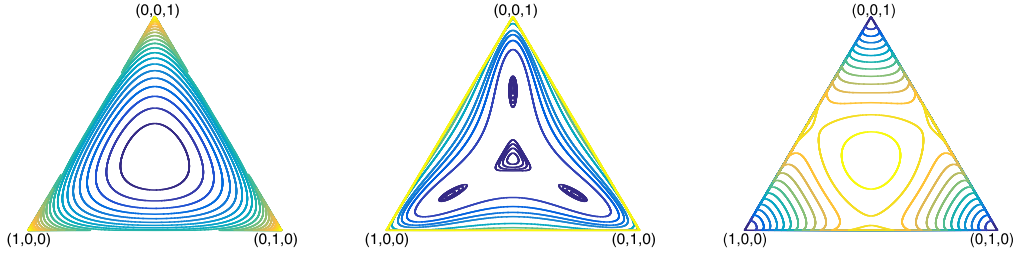
$$\boldsymbol{\mu}^{\text{Potts}}_{\tilde{\beta},G}(\sigma) := \frac{\exp(-\beta \cdot H(\sigma))}{Z(\beta)}, \tag{5.2}$$

where β is the inverse temperature parameter (a constant independent of σ), and $Z(\beta)$ is the normalising constant.

We will consider the 3 colour mean-field case, when G is the complete graph, and $q = 3$. In this case, a simple rearrangement (see [4]) shows that we can equivalently write the model as

$$\boldsymbol{\mu}^{\text{Potts}}_{\tilde{\beta},M}(\sigma) := \frac{\exp(M\tilde{\beta}(s_1(\sigma)^2 + s_2(\sigma)^2 + s_2(\sigma)^2))}{\tilde{Z}(\tilde{\beta})}, \tag{5.3}$$

where $s_k(\sigma)$ is the ratio of spins of colour k for $k = 1, 2, 3$, $\tilde{\beta} := \frac{\beta J}{2} M$, and $\tilde{Z}(\tilde{\beta})$ is the new normalising constant. We call the triple $s_1(\sigma)$, $s_2(\sigma)$, $s_3(\sigma)$ the magnetisation vector.



(a) $\tilde{\beta} < \tilde{\beta}_c$ (b) $\tilde{\beta} = \tilde{\beta}_c$ (c) $\tilde{\beta} > \tilde{\beta}_c$

Figure 1. Contour plots of the log-likelihood as a function of (s_1, s_2, s_3).

This model is known to undergo a first order phase transition at $\tilde{\beta}_c = 2\log(2)$. We denote the distribution of the magnetisation vector as

$$\boldsymbol{\mu}^{\text{mag}}_{\tilde{\beta},M}(s_1, s_2, s_3) := \boldsymbol{\mu}^{\text{Potts}}_{\tilde{\beta},M}\big(\{\sigma : s_1(\sigma) = s_1, s_2(\sigma) = s_2, s_3(\sigma) = s_3\}\big).$$

To show the difference between the different phases, we include 3 contour plots of

$$\log \boldsymbol{\mu}^{\text{mag}}_{\tilde{\beta},M}(s_1, s_2, s_3) = \log\left(\binom{M}{Ms_1}\binom{M - Ms_1}{Ms_2} \cdot \frac{\exp(M\tilde{\beta}(s_1^2 + s_2^2 + s_3^2))}{\tilde{Z}(\tilde{\beta})}\right)$$

for $\tilde{\beta} = \tilde{\beta}_c/2$, $\tilde{\beta} = \tilde{\beta}_c$ and $\tilde{\beta} = 2\tilde{\beta}_c$, for $M = 1000$. The plots show s_1, s_2 and s_3 in a barycentric coordinate system, the darker colours correspond to areas of higher probability (see Figure 1). As we can see, for $\tilde{\beta} < \tilde{\beta}_c$, there is a single local maximum centered at $(s_1, s_2, s_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. At $\tilde{\beta} = \tilde{\beta}_c$, there are 4 local maximums, centered at

$$\begin{aligned} C_1 &:= \left(\frac{2}{3}, \frac{1}{6}, \frac{1}{6}\right), \qquad C_2 := \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right), \\ C_3 &:= \left(\frac{1}{6}, \frac{1}{6}, \frac{2}{3}\right) \quad \text{and} \quad C_4 := \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right). \end{aligned} \tag{5.4}$$

Finally, for $\tilde{\beta} > \tilde{\beta}_c$, there are 3 local maximums, centered at $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

The Glauber dynamics Markov chain updates a randomly chosen spin conditioned on the rest of the spins in each step. [10] has shown that this chain is fast mixing in part of the region $\tilde{\beta} < \tilde{\beta}_c$ (called the *high temperature* region), but its mixing time increases exponentially in the number of spins for $\tilde{\beta} \geq \tilde{\beta}_c$. This phenomenon is caused by the existence of multiple modes for $\tilde{\beta} \geq \tilde{\beta}_c$.

Parallel tempering (also called Metropolis coupled MCMC) is a popular method that has been shown to work well for some multimodal distributions (see [28,38]). However, it was shown in [4] that parallel and simulated tempering will have exponentially slow mixing time for $\tilde{\beta} \geq \tilde{\beta}_c$ if the tempering distributions with the standard temperature ladder are used. The reason for this is that the 3 new modes that appear for $\tilde{\beta} \geq \tilde{\beta}_c$ have very little probability according to the uniform

distribution on all configurations. This means that even if theoretically we could move between the modes in the levels of high temperature, practically we will almost never move to the 3 new modes.

The goal of this section is to show that this difficulty can be overcome by SMC methods using the *interpolation to independence* sequence of distributions defined in Section 4. For simplicity, we will only consider the case $\tilde{\beta} = \tilde{\beta}_c$ (but we believe that similar arguments work for any temperature and number of colours). The MCMC moves that we are going to use for $\boldsymbol{\mu}_k$ do one step in the Glauber dynamics for $(\sigma_1, \ldots, \sigma_k)$, and replace the rest of the spins by independent copies. We denote the Markov kernel corresponding to this move by $\boldsymbol{P}_k$, and the kernel combining t_k such steps by $\boldsymbol{K}_k = \boldsymbol{P}_k^{t_k}$. Then it is easy to see that this kernel is a reversible with respect to $\boldsymbol{\mu}_k$.

It is not difficult to see heuristically that if we applying the SMC algorithm with Glauber dynamics steps, and choose the interpolating distributions $\boldsymbol{\mu}_i$ as $\boldsymbol{\mu}^{\text{Potts}}_{\tilde{\beta}_c \cdot (i/n), M}$ for $0 \le i \le n$ (tempering distributions), then most of the particles would stay close to the central mode, and they would never discover the 3 other modes (we have some numerical evidence for this). For such a sequence of distributions, the product of the growth-within-mode constants, $\prod_{j=0}^{M-1} B_{j,j+1}$, grows exponentially with M.

5.2. Main result

Theorem 5.1 (SMC variance bound for the Potts model). *Suppose that $\tilde{\beta} = \tilde{\beta}_c$. There is a constant $C_1 \in \mathbb{R}_+$ such that for the SMC sampler described above, assuming that the number of MCMC steps in stage k is chosen as $t_k = \lceil C_1 k \log(k)^2 \rceil$, the asymptotic variance of the SMC empirical average of any bounded function $f : \Omega \to \mathbb{R}$ satisfies that*

$$V_M(f) \le C_2 M \|f\|_\infty^2, \tag{5.5}$$

for some absolute constant $C_2 \in \mathbb{R}_+$.

Due to space considerations we only prove this result for $\tilde{\beta} = \tilde{\beta}_c$, but we believe that a similar result holds for the Potts model with any number of colours, and any temperature parameter $\tilde{\beta}$. Here we note that the overall amount of computational effort needed to obtain a sample of unit variance by this algorithm is $\mathcal{O}(M^3 \log^2(M))$. This is significantly better than the mixing rate obtained in [5]. We think that this could be improved to a smaller power of M by resampling only when the effective sample size parameter (ESS) is below a certain threshold (see [13]), since resampling is not necessary in each stage because the ratio of particles in the separate modes is converging quickly to a limit.

6. Proofs for the Potts model

The proof of Theorem 5.1, based on our theoretical results, is rather complex. To make the presentation clear, in Section 6.1 we first state 5 key propositions bounding the maximal density ratio (Γ_g) and growth-within-mode ($B_{j,j+1}$) constants in Theorem 3.2 and the 3 terms in Theorem 3.3 (the probability of escaping from the inner regions, the probability of staying in the border regions, and the total variational distance to the local restriction of the stationary distribution to the

mode when started from a place in one of the inner regions). Based on these propositions, we then prove Theorem 5.1.

In Section 6.2, we show some preliminary results that will be used in the proof of the key propositions. This is followed by Section 6.3, where we show a lemma bounding the drift of the Glauber dynamics chain. Finally, in Sections 6.4–6.6, we prove the key propositions used in the proof of Theorem 5.1.

In the rest of this paragraph, we introduce some notations that will be used through the proof. Let $\sigma(0) \in \Omega = \{1,2,3\}^M$ be a fixed starting point, and $\sigma(0), \sigma(1), \sigma(2), \ldots$ be a realisation of the Glauber dynamics chain started at $\sigma(0)$. Let

$$S(k) := \big(s_1\big(\sigma(k)\big), s_2\big(\sigma(k)\big), s_3\big(\sigma(k)\big)\big)$$

be the vector of the ratios of different colours in $\sigma(k)$. It is shown in [10] that $S(k)$ is also a Markov chain, called the *magnetisation chain*. We call the state space of this chain Ω^S. In order to understand the geometry of Ω^S, we will think of each point of it in barycentric coordinates. We will call by $T := \{(s_1, s_2, s_3) : 0 \le s_1, s_2, s_3 \le 1, s_1 + s_2 + s_3 = 1\}$ the *main triangle*. On Figure 2, we illustrate this triangle with an equilateral triangle with side length 1. The centroid of this triangle is point C_4, and we denote the 3 vectors pointing from C_4 to the three corners of an equilateral triangle by e_1, e_2 and e_3 (each of them has length $\frac{\sqrt{3}}{3}$). Then to each point $s = (s_1, s_2, s_3) \in T$, the sum $C(s) := s_1 e_1 + s_2 e_2 + s_3 e_3$ is the corresponding two dimensional vector. We define the distance of two points $s, s' \in T$, denoted by $d(s, s')$, the Euclidean distance of $C(s)$ and $C(s')$, which can be rewritten as

$$\begin{aligned} d(s,s') &:= \big\|(s_1 - s_1')e_1 + (s_2 - s_2')e_2 + (s_3 - s_3')e_3\big\| \\ &= \sqrt{\big\langle (s_1 - s_1')e_1 + (s_2 - s_2')e_2 + (s_3 - s_3')e_3, (s_1 - s_1')e_1 + (s_2 - s_2')e_2 + (s_3 - s_3')e_3 \big\rangle} \quad (6.1) \\ &= \frac{1}{\sqrt{2}} \sqrt{(s_1 - s_1')^2 + (s_2 - s_2')^2 + (s_3 - s_3')^2}. \end{aligned}$$

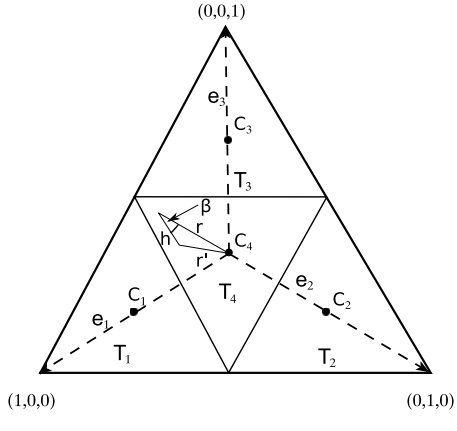


Figure 2. Position of modes in barycentric coordinates.

The main triangle T is divided into 4 *subtriangles*, $T_i := \{(s_1, s_2, s_3) : s_i \in (1/2, 1], 0 \leq s_1, s_2, s_3 \leq 1, s_1 + s_2 + s_3 = 1\}$ for $i = 1, 2, 3$, and $T_4 := \{(s_1, s_2, s_3) : 0 \leq s_1, s_2, s_3 \leq 1/2, s_1 + s_2 + s_3 = 1\}$. Then it is easy to see that the these are equilateral triangles, with centroid C_i for T_i, for $1 \leq i \leq 4$. We define the distance between a point $s \in T$ to the closes of the centres $C_1, \ldots, C_4$ as

$$d_C(s) := \min\big(d(s, C_1), d(s, C_2), d(s, C_3), d(s, C_4)\big). \tag{6.2}$$

Then one can see that for every $1 \leq i \leq 4$, $s \in T_i$, we have $d_C(s) = d(s, C_i)$.

We define the modes in the following way. For $0 \leq j \leq 10^7$, we set $m(j) := 1$, and $F_j^{(1)} := \Omega$. For $10^7 \leq j \leq M$, we set $m(k) := 4$, and

$$F_j^{(i)} := \left\{ x \in \Omega : \frac{\sum_{k=1}^{j} \mathbb{1}[x_k = i]}{j} > \frac{1}{2} \right\} \qquad \text{for } i = 1, 2, 3 \quad \text{and} \tag{6.3}$$

$$F_j^{(4)} := \Omega \setminus \big(F_j^{(1)} \cup F_j^{(2)} \cup F_j^{(3)}\big). \tag{6.4}$$

Thus we compute the ratio of the spins of each colour among the first k spins, look at which triangle this ratio vector falls into on Figure 2, and assign them to the corresponding mode. Now we proceed with the definition of the inner regions. For $j \leq 10^7$, choose a single inner region as $\mathcal{I}_j^{(1)} := \Omega$, and the border region is an empty set. For $10^7 < j \leq M$, and $1 \leq i \leq 4$, we define the inner regions as

$$\mathcal{I}_j^{(i)} := \left\{ \sigma \in \Omega : -\frac{\rho}{4} \leq s_k(\sigma_{1:j}) - C_{i,k} \leq \frac{\rho}{2} \text{ for } k = 1, 2, 3 \right\} \qquad \text{with } \rho := 10^{-6}, \tag{6.5}$$

where $s_k(\sigma_{1:j}) := (\sum_{l=1}^{j} 1_{[\sigma_l = k]})/j$ is the ratio of spins of colour k among the first j spins. The points $(s_1(\sigma_{1:j}), s_2(\sigma_{1:j}), s_3(\sigma_{1:j}))$ for $\sigma \in \mathcal{I}_j^{(i)}$ fall in a small equilateral triangle centered at C_i with sides parallel to the sides of T. The fact that $j > 10^7$ ensures that the inner regions are non-empty. The border regions are defined as $\mathcal{B}_j^{(i)} := F_j^{(i)} \setminus \mathcal{I}_j^{(i)}$ for $1 \leq i \leq 4$.

6.1. Key propositions and proof of Theorem 5.1

In this section, we state 5 key propositions bounding the various terms in Theorems 3.2 and 3.3, and then prove Theorem 5.1 based on them. Our first proposition bounds the maximal density ratio constant Γ_g.

Proposition 6.1. *The maximal density constant* $\Gamma_g := \max_{0 \leq k \leq n-1} \max_{x \in \Omega} \frac{\boldsymbol{\mu}_{k+1}(x)}{\boldsymbol{\mu}_k(x)}$ *satisfies that* $\Gamma_g \leq \exp(2\tilde{\beta}_c)$.

Proof. Let us denote the number of spins of colours $1, 2, 3$ among $x_1, \dots, x_k$ by $n_1(x_{1:k})$, $n_2(x_{1:k})$ and $n_3(x_{1:k})$, respectively. Then

$$\frac{\boldsymbol{\mu}_{k+1}(x)}{\boldsymbol{\mu}_k(x)} = \frac{Z_k(\tilde{\beta}_c)}{Z_{k+1}(\tilde{\beta}_c)} \cdot \exp\Bigg(\tilde{\beta}_c\Bigg(\frac{n_1^2(x_{1:k+1}) + n_2^2(x_{1:k+1}) + n_3^2(x_{1:k+1})}{k+1} - \frac{n_1^2(x_{1:k}) + n_2^2(x_{1:k}) + n_3^2(x_{1:k})}{k}\Bigg)\Bigg).$$

Now it is easy to show that

$$\left|\frac{n_1^2(x_{1:k+1}) + n_2^2(x_{1:k+1}) + n_3^2(x_{1:k+1})}{j+1} - \frac{n_1^2(x_{1:k}) + n_2^2(x_{1:k}) + n_3^2(x_{1:k})}{k}\right| \leq 1,$$

thus

$$\frac{Z_k(\tilde{\beta}_c)}{Z_{k+1}(\tilde{\beta}_c)} \exp(-\tilde{\beta}_c) \leq \frac{\boldsymbol{\mu}_{k+1}(x)}{\boldsymbol{\mu}_k(x)} \leq \frac{Z_k(\tilde{\beta}_c)}{Z_{k+1}(\tilde{\beta}_c)} \exp(\tilde{\beta}_c).$$

The fact that $\sum_{x\in\Omega} \boldsymbol{\mu}_k(x) = \sum_{x\in\Omega} \boldsymbol{\mu}_{k+1}(x) = 1$ implies that $\frac{Z_k(\tilde{\beta}_c)}{Z_{k+1}(\tilde{\beta}_c)} \leq \exp(\tilde{\beta}_c)$, and thus $\Gamma_g \leq \exp(2\tilde{\beta}_c)$. □

Our second result bounds the growth-within-mode constants $B_{j,j+1}$ defined in (3.5).

Proposition 6.2 (Bounds on the growth-within-mode constants). *We have* $B_{0,1} = B_{1,2} = 1$, *and for any* $2 \leq j \leq M-1$,

$$B_{j,j+1} \leq 1 + C \cdot \frac{\log(j)^5}{j^{3/2}}$$

for some absolute constant $C > 0$.

The proof of this result, based on Taylor expansions, is quite technical, with no probabilistic ingredient, so it is included in Section A.5 (in the supplementary material [32]). The third proposition bounds the time needed to approach one of the centers. The proof is included in Section 6.5.

Proposition 6.3 (Time to get to a central region). *Let* $\{S(t)\}_{t\geq 0}$ *be the magnetisation chain on* Ω^S. *Let*

$$\tau := \inf\left\{k \in \mathbb{N} : d_C\big(S(k)\big) \leq \frac{\rho}{8}\right\},$$

that is, the first time we get closer than $\frac{\rho}{8}$ *to one of the centers. Then for any initial position* $s \in \Omega^S$, *any* $r \in \mathbb{R}_+$,

$$\mathbb{P}\big(\tau > r \cdot CM\log(M) | S(0) = s\big) \leq \exp\big(-\lfloor r \rfloor\big),$$

where $C > 0$ *is an absolute constant.*

The fourth proposition bounds the total variational distance from the local distribution in the modes. The proof is included in Section 6.6.

Proposition 6.4. *Suppose that $k \geq 330M\log(M)$. Let P denote the Markov kernel of the Glauber dynamics on $\boldsymbol{\mu}_M$. Then for M larger than some absolute constant, for any $\sigma \in \mathcal{I}_M^{(i)}$ for some $1 \leq i \leq 4$, we have*

$$d_{\mathrm{TV}}\big(\boldsymbol{P}^k(\sigma,\cdot),\boldsymbol{\mu}_M^{(i)}\big) \leq \frac{C}{M^2} + 3k^2\exp\big(-C'\cdot M\big),$$

where $C, C' > 0$ are some absolute constants.

Now we state the proof of our variance bound based on these key propositions.

Proof of Theorem 5.1. By (2.3), the asymptotic variance satisfies that $V_M(f) = \sum_{j=0}^{M} V_{j,M}(f)$, with the terms $V_{j,M}(f)$ can be bounding by Theorem 3.2 as

$$\begin{aligned} V_{j,M}(f) &\leq \|f\|_\infty^2 \Gamma_g \prod_{i=j+1}^{M-1} \big\{B_{i,i+1} + \Gamma_g |||\boldsymbol{K}_i - \widehat{\boldsymbol{\pi}}_i|||_\infty\big\} \\ &\leq \|f\|_\infty^2 \Gamma_g \prod_{i=0}^{M-1} B_{i,i+1} \cdot \prod_{i=j+1}^{M-1} \big\{1 + \Gamma_g |||\boldsymbol{K}_i - \widehat{\boldsymbol{\pi}}_i|||_\infty\big\}. \end{aligned} \tag{6.6}$$

By Proposition 6.1, we have $\Gamma_g \leq \exp(2\tilde{\beta}_c)$. Proposition 6.2 implies that $\prod_{i=0}^{M-1} B_{i,i+1} \leq C_B$ for some absolute constant $C_B < \infty$.

Let us choose $t_i = \lceil Ri\log(i)^2 \rceil$ for some constant R, and $\boldsymbol{K}_i := \boldsymbol{P}_i^{t_i}$, where $\boldsymbol{P}_i$ is the Markov kernel described in Section 4 (combining a Glauber dynamics step in the first i coordinates $(\sigma_1,\ldots,\sigma_i)$ with respect to $\boldsymbol{\mu}_{\tilde{\beta}_c,k}^{\mathrm{Potts}}$ and replacing the rest of the coordinates $(\sigma_{i+1},\ldots,\sigma_M)$ by independent copies). Based on Theorem 3.3, we have

$$\begin{aligned} |||\boldsymbol{K}_i - \widehat{\boldsymbol{\pi}}_i|||_\infty = \big|\big|\big|\boldsymbol{P}_i^{t_i} - \widehat{\boldsymbol{\pi}}^{(t_i)}\big|\big|\big|_\infty &\leq \max_{x\in\mathcal{B}_i} \mathbb{P}\big(\mathrm{S}_{\mathcal{B}}\big(x,\lfloor t_i/2\rfloor\big)\big) \\ &\quad + 2\max_{1\leq l\leq 4}\max_{\lceil t_i/2\rceil\leq r\leq t_i}\sup_{x\in\mathcal{I}^{(l)}} d_{\mathrm{TV}}\big(\boldsymbol{P}_i^r(x,\cdot),\boldsymbol{\mu}^{(l)}\big), \end{aligned} \tag{6.7}$$

where $\mathcal{B}_i := \cup_l \mathcal{B}_i^{(l)}$ is the union of the border regions. By using Propositions 6.3 and 6.4 (applied by substituting $M = i$ and using the fact that the rest of the spins are independent), by choosing R sufficiently large, for i larger than some absolute constant, we have

$$\max_{x\in\mathcal{B}_i} \mathbb{P}\big(\mathrm{S}_{\mathcal{B}}\big(x,\lfloor t_i/2\rfloor\big)\big) \leq \frac{1}{i^2}, \quad \text{and} \tag{6.8}$$

$$2\max_{1\leq l\leq 4}\max_{\lceil t_i/2\rceil\leq r\leq t_i}\sup_{x\in\mathcal{I}^{(l)}} d_{\mathrm{TV}}\big(\boldsymbol{P}_i^r(x,\cdot),\boldsymbol{\mu}^{(l)}\big) \leq \frac{C'}{i^2}, \tag{6.9}$$

for some absolute constant $C' < \infty$. The result now follows by (6.6). □

6.2. Preliminary results

In this section, we will first prove anti-concentration and concentration results for sequences of random variables satisfying certain drift conditions. After these, we show a coupling argument for showing how good mixing in the magnetisation chain can be used to show good mixing in in the original Glauber dynamics chain. The proofs are included in Section A.3 (in the supplementary material [32]).

Lemma 6.5 (A lower bound on the exit time). *Assume that $(X_l)_{l\geq 0}$ is a sequence of random variables adapted to some filtration $(\mathcal{F}_l)_{l\geq 0}$. Suppose that*

(1) $|X_{l+1} - X_l| \leq R$ *almost surely for every* $l \geq 0$ *for some absolute constant* R,
(2) $\mathbb{E}(X_{l+1} - X_l|\mathcal{F}_l) \leq \delta$ *for some* $\delta > 0$,
(3) $X_0 = x_0$ *for some fixed constant* $x_0 \in \mathbb{R}$, *and*
(4) $\mathrm{Var}(X_{l+1} - X_l|\mathcal{F}_l) \geq v$ *for every* $t \geq 0$ *for some absolute constant* $v > 0$.

Suppose that $z \geq 12\frac{R}{\sqrt{v}}$, *then*

$$\mathbb{P}\Big[\min_{0\leq l\leq 4(z+2R/\sqrt{v})^2}(X_l - x_0) \leq -z\sqrt{v} + 4(z + 2R/\sqrt{v})^2\delta\Big] \geq \frac{1}{6}.$$

Remark 6.6. This lemma quantifies the fact that if the variance of the jumps is always at least an absolute constant greater than 0, and the drift δ towards to right is sufficiently small, then after $\mathcal{O}(r)$ steps, we will move to the left $\mathcal{O}(\sqrt{r})$ with reasonable probability.

Lemma 6.7 (Moving in a region of negative drift). *Assume that $(X_l)_{l\geq 0}$ is a sequence of random variables adapted to some filtration $(\mathcal{F}_l)_{l\geq 0}$. Suppose that*

(1) $|X_{l+1} - X_l| \leq R$ *almost surely for every* $l \in \mathbb{N}$ *for some absolute constant* R,
(2) $\mathbb{E}(X_{l+1} - X_l|\mathcal{F}_l) \leq -\delta$ *for some* $\delta > 0$,
(3) $X_0 = x_0$ *for some fixed constant* $x_0 \in \mathbb{R}$.

Then the probabilities of moving backward, and forward, respectively, can be bounded as

$$\mathbb{P}(X_{\lceil(1+c)T/\delta\rceil} - x_0 \geq -T) \leq \exp\left(-\frac{cT\delta}{4R^2}\right) \qquad \text{for any } c \geq 1, T > 0, \tag{6.10}$$

$$\mathbb{P}\Big(\max_{0\leq k\leq l}(X_k - x_0) \geq T\Big) \leq l\exp\left(-\frac{T\delta}{R^2}\right) \qquad \text{for any } l \in \mathbb{N}, T > 0. \tag{6.11}$$

The following lemma shows that once two chains have met in magnetisation, they can be coupled together in $\mathcal{O}(M\log(M))$ time with high probability.

Lemma 6.8 (From coupling in magnetisation to coupling in spins). *Let $\sigma(0), \widetilde{\sigma}(0) \in \Omega$ such that $s(\sigma(0)) = s(\widetilde{\sigma}(0))$. Let $(\sigma(t))_{t\geq 0}$ and $(\widetilde{\sigma}(t))_{t\geq 0}$ be two Glauber dynamics chains with temperature parameter $\tilde{\beta}_c$ started at $\sigma(0)$ and $\widetilde{\sigma}(0)$, respectively. Let $\tau := \inf\{t \geq 0 : \sigma(t) = \widetilde{\sigma}(t)\}$.*

Then there is a coupling $((\sigma(t))_{t\geq 0}, (\widetilde{\sigma}(t))_{t\geq 0})$ *such that for this coupling,*

$$\mathbb{P}(\tau > t) \leq \frac{M}{2} \exp\left(-\frac{t}{9M}\right).$$

6.3. Bounding the drift towards the centers

The following lemma shows bounds the drift towards the centers at a given distance from the centers.

Lemma 6.9 (Drift bound). *Let* $\{S(k)\}_{k\geq 0}$ *be the magnetisation chain on* Ω^S. *Then for any* $s \in \Omega^S$ *with* $d_C(s) > 1/M$, *we have*

$$\begin{aligned}&\mathbb{E}\big(d_C\big(S(1)\big) - d_C\big(S(0)\big) | S(0) = s\big)\\ &\quad\leq -\frac{\varphi(d_C(s))}{M} + \frac{1}{M^2} \cdot \left(8 + \frac{1}{2(d_C(s) - 1/M)}\right),\end{aligned}$$

with $\varphi : [0, \sqrt{3}/6] \to \mathbb{R}$ *defined as*

$$\varphi(t) := \begin{cases} 0.002 \cdot \dfrac{t}{\sqrt{3}/24} & \text{for } 0 \leq t \leq \sqrt{3}/24, \\ 0.002 \cdot \dfrac{\sqrt{3}/12 - t}{\sqrt{3}/24} & \text{for } \sqrt{3}/24 \leq t \leq \sqrt{3}/12 \quad \text{and} \\ 0.002 \cdot \dfrac{t - \sqrt{3}/12}{\sqrt{3}/24} & \text{for } \sqrt{3}/12 \leq t \leq \sqrt{3}/6. \end{cases}$$

The proof of Lemma 6.9 is included in Section A.2 (in the supplementary material [32]). Figure 3 plots $\varphi(t)$ for $0 \leq t \leq \sqrt{3}/6$. Note that $\varphi(t) = 0$ for $t = 0$ and $t = \sqrt{3}/12$.

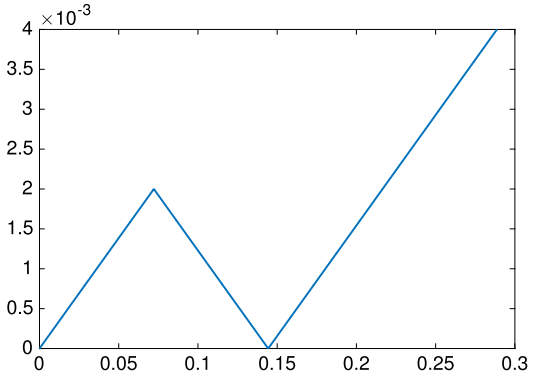


Figure 3. $\varphi(t)$ for $0 \leq t \leq \sqrt{3}/6$.

6.4. Bounds on escaping from the center

In this section, we prove the following proposition bounding the probability of escaping from the region near one of the centers.

Proposition 6.10 (Escaping from a central region). *Let $\{S(t)\}_{t\geq 0}$ be the magnetisation chain on Ω^S. Assuming that $S(0)=s$ with $d_C(s)\leq \rho\cdot\frac{\sqrt{3}}{2}$, the probability of getting further away than $\rho\sqrt{3}$ in l steps is bounded as*

$$\mathbb{P}\Big(\max_{0\leq k\leq l} d_C\big(S(k)\big)>\rho\sqrt{3}\big|S(0)=s\Big)\leq l^2\exp(-C_{\mathrm{esc}}M), \tag{6.12}$$

for M larger than some absolute constant, where C_{esc} is an absolute constant that can be chosen as $C_{\mathrm{esc}}=0.005\rho^2$.

Remark 6.11. By the inner mode sets $\mathcal{I}_M^{(i)}$ and the sets $\tilde{\Lambda}^{(i)}:=\{\sigma\in\Omega:-\rho\leq s_l(\sigma)-C_{i,l}\leq 2\rho \text{ for } l=1,2,3\}$ satisfy that in the magnetisation space, their defining equations restrict them into equilateral triangles, with edge lengths $\frac{\sqrt{3}}{2}\cdot\rho$ for $\mathcal{I}_M^{(i)}$, and $2\sqrt{3}\cdot\rho$ for $\tilde{\Lambda}^{(i)}$. Therefore, in particular, this proposition implies that if we start from $\sigma(0)=\sigma\in\mathcal{I}_M^{(i)}$, then the probability of the Glauber dynamics chain $\{\sigma(k)\}_{k\geq 0}$ exiting from $\tilde{\Lambda}^{(i)}$ in the first l steps can be bounded as

$$\mathbb{P}\big(\sigma(k)\notin\tilde{\Lambda}^{(i)}\text{ for some }1\leq k\leq l|\sigma(0)=\sigma\big)\leq l^2\exp(-C_{\mathrm{esc}}M), \tag{6.13}$$

for M larger than some absolute constant.

Proof of Proposition 6.10. Using Lemma 6.9, we can see that for $s\in\Omega^S$ satisfying that $d_C(s)\in[\rho\frac{\sqrt{3}}{2},\rho\sqrt{3}]$, for M larger than some absolute constant,

$$\mathbb{E}\big(d_C\big(S(1)\big)-d_C\big(S(0)\big)|S(0)=s\big)\leq-\frac{\varphi(\frac{\rho\sqrt{3}}{2})}{2M}\leq\frac{-0.012\rho}{M}.$$

In order to get further away than ρ from one of the centers, we need to spend a period of time when the distance is between $\rho\frac{\sqrt{3}}{2}$ and $\rho\sqrt{3}$, and then exceed $\rho\sqrt{3}$.

Notice that we cannot apply the martingale-type inequalities of Lemma 6.7 directly to $\{d_C(S(k))\}_{k\geq 0}$ because the drift does not hold uniformly in every k. Instead of direct application, we use a coupling argument. For every $0\leq k\leq l-1$, we define a sequence of random variables $D_0^{(k)},D_1^{(k)},\ldots,$ as follows. First, we set $D_0^{(k)}:=d_C(S(k))$. After this, we define the rest of the sequence such that it satisfies that for every $j\in\mathbb{N}$,

$$\begin{aligned}D_{j+1}^{(k)}-D_j^{(k)}:=&\,1_{[d_C(S(k+j))\in[\rho\frac{\sqrt{3}}{2},\rho\sqrt{3}][\rho\frac{\sqrt{3}}{2},\rho\sqrt{3}]]}\cdot\big[d_C\big(S(k+j+1)\big)-d_C\big(S(k+j)\big)\big]\\&-1_{[d_C(S(k+j))\notin[\rho\frac{\sqrt{3}}{2},\rho\sqrt{3}]]}\cdot\frac{0.012\rho}{M}.\end{aligned}$$

From the above definitions it follows that for $s \in \Omega^S$ satisfying that $d_C(s) \in \left[\rho\frac{\sqrt{3}}{2}, \rho\sqrt{3}\right]$,

$$\mathbb{P}\Big(\max_{0\le k\le l} d_C\big(S(k)\big) > \rho \big| S(0) = s\Big) \le \sum_{k=0}^{l-1} \mathbb{P}\Big(\max_{0\le i\le l-k}\big(D_i^{(k)}\big) \ge \sqrt{3}\rho \Big| S(0) = s\Big).$$

Then from their definition, we can see that the random variables $(D_i^{(k)})_{0\le i\le l-k}$ satisfy the conditions of Lemma 6.7 with $R = \frac{1}{M}$ and $\delta := \frac{0.012\rho}{M}$, and the claim of the proposition follows by applying (6.11) with $T = \frac{\sqrt{3}}{2}\rho - \frac{1}{M}$ and $\delta = \frac{0.012\rho}{M}$ on $\max_{0\le i\le l-k}(D_i^{(k)})$, and then summing up. □

6.5. Getting to one of the centers

In this section, we prove Proposition 6.3 based on drift arguments and concentration inequalities. The following lemma will be used for the proof.

Lemma 6.12 (Minimum variance of jumps). *Let $\{S(t)\}_{t\ge 0}$ be the magnetisation chain on Ω^S. Then for M larger than some absolute constant, for any starting point $s \in \Omega^S$,*

$$\mathrm{Var}\big(d_C\big(S(1)\big) | S(0) = s\big) \ge \frac{v_{\min}}{M^2} \qquad \textit{with } v_{\min} := 0.001.$$

Proof. Using the notations of the proof of Lemma 6.9, it is straightforward to show that

$$P_{i\to j}(s_1, s_2, s_3) \ge s_i \cdot \frac{1}{2 + \exp(\tilde{\beta}_c)} = \frac{s_i}{18}.$$

Moreover, it is also easy to check that

$$P_{\circlearrowleft}(s_1, s_2, s_3) \ge \sum_{1\le i\le 3} \frac{s_i}{1 + \exp[\frac{2}{M} + 2\tilde{\beta}_c(s_j - s_i)] + \exp[\frac{2}{M} + 2\tilde{\beta}_c(s_k - s_i)]} \ge \frac{1}{54}.$$

The proof is based on the fact that for an equilateral triangle of unit edge length, and a point on the plane outside the triangle, the difference between the distances of the point and the closest and furthest away corners of the triangle is at least $\frac{\sqrt{3}}{2} - \frac{1}{2}$.

Assume first that (s_1, s_2, s_3) more than $\frac{1}{M}$ away from the edges of the central triangle T_4 (in d distance). Then without loss of generality, assume that $s_1 \ge \frac{1}{3}$. Then $P_{1\to 2}(s_1, s_2, s_3) \ge \frac{1}{54}$ and $P_{1\to 3}(s_1, s_2, s_3) \ge \frac{1}{54}$. Since the three positions s, $s^{1\to 2}$ and $s^{1\to 3}$ form an equilateral triangle of side length $\frac{1}{M}$, unless a central point is included in this triangle, there is at least $\frac{1}{M}(\frac{\sqrt{3}}{2} - \frac{1}{2})$ difference between two of $d_C(s)$, $d_C(s^{1\to 2})$ and $d_C(s^{1\to 3})$. Therefore, the variance is lower bounded as

$$\mathrm{Var}\big(d_C\big(S(1)\big) | S(0) = s\big) \ge \frac{2}{54}\left(\frac{1}{2}\cdot\left(\frac{\sqrt{3}}{2} - \frac{1}{2}\right)\cdot\frac{1}{M}\right)^2 > \frac{0.001}{M^2}.$$

If the triangle formed by s, $s^{1\to 2}$ and $s^{1\to 3}$ contains a central point, then one can check that the same bound still holds for M greater than some absolute constant by suitably choosing the direction $i \to j$.

Finally, if (s_1, s_2, s_3) is no more than $\frac{1}{M}$ away from the edges of the central triangle T_4, then one can still choose $i \to j$ in a way that we do not exit from the triangle that we are in (T_1, T_2, T_3 or T_4), for M larger than some absolute constant, $|d_C(s^{i\to j}) - d_C(s)| > \frac{0.4}{M}$, and $s_i \geq \frac{1}{4}$. For this choice, we have $P_{i\to j}(s_1, s_2, s_3) \geq \frac{1}{72}$, and thus

$$\mathrm{Var}\big(d_C\big(S(1)\big)|S(0) = s\big) \geq \frac{2}{72}\left(\frac{1}{2}\cdot\frac{0.4}{M}\right)^2 > \frac{0.001}{M^2}. \qquad \square$$

Now we are ready to prove the main result of this section.

Proof of Proposition 6.3. Let $c_1 := 1000$ and $c_2 := 1$. Let

$$r := \left\lceil \left(\frac{\sqrt{3}}{12} - c_1\sqrt{\frac{\log(M)}{M}}\right) \Big/ \left(\frac{c_2}{\sqrt{M}}\right) \right\rceil + 1 \quad \text{and}$$

$$m := r + 1 + \left\lceil \left(\frac{\sqrt{3}}{12} - c_1\sqrt{\frac{\log(M)}{M}} - \frac{\rho}{8}\right) \Big/ \left(\frac{c_2}{\sqrt{M}}\right) \right\rceil.$$

Define a sequence of distances $d_0 > d_1 > \cdots > d_m$ as follows,

$$d_0 := \frac{\sqrt{3}}{6}, \qquad d_{r-1} := \frac{\sqrt{3}}{12} + c_1\sqrt{\frac{\log(M)}{M}}, \qquad d_r := \frac{\sqrt{3}}{12}, \qquad d_{r+1} := \frac{\sqrt{3}}{12} - c_1\sqrt{\frac{\log(M)}{M}},$$

$$d_m := \frac{\rho}{8}, \qquad d_k - d_{k+1} := \frac{c_2}{\sqrt{M}} \qquad \text{for } k \in \{1, \ldots, m-2\} \setminus \{r-1, r\}.$$

For $1 \leq j \leq m$, we define the arrival times $\tau_j := \inf\{k \geq 0 : d_C(S(k)) \leq d_j\}$. The proof will consist of subsequently estimating the differences $\tau_{j+1} - \tau_j$ for $0 \leq j \leq m-1$. Figure 4 illustrates the position of the distances $(d_j)_{0\leq j\leq m}$, and the function $\varphi(t)$.

Based on Lemma 6.9, we obtain the following simplified drift bounds. For M larger than some absolute constant, for every $s \in \Omega^S$, we have

$$\mathbb{E}\big(d_C\big(S(1)\big) - d_C\big(S(0)\big)|S(0) = s\big) \leq \frac{20}{M^2}. \tag{6.14}$$

Based on the definition of the distances d_j and the function $\varphi(t)$, it follows that for M larger than some absolute constant, for every $j \in \{0, \ldots, m-1\} \setminus \{r, r-1\}$, for every $s \in \Omega^S$ such that $d_C(s) \in [d_{j+1} - \frac{1}{3}c_1\sqrt{\frac{\log(M)}{M}}, d_j + \frac{1}{3}c_1\sqrt{\frac{\log(M)}{M}}]$, we have

$$\mathbb{E}\big(d_C\big(S(1)\big) - d_C\big(S(0)\big)|S(0) = s\big) \leq -\frac{\varphi(d_j)}{2M}. \tag{6.15}$$

First, we are going to estimate $\tau_{r+1} - \tau_{r-1}$. By applying Lemma 6.5 to $\{d_C(S(\tau_{r-1} + k))\}_{k\geq 0}$ with $\delta = \frac{20}{M^2}$, $R = \frac{1}{M}$, $v = \frac{v_{\min}}{M^2}$ (based on Lemma 6.12), and $z = 4c_1\sqrt{\frac{\log(M)}{M}}/\sqrt{v} =$

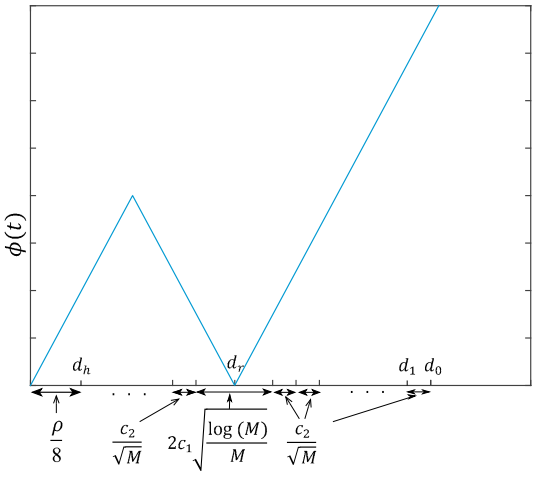


Figure 4. The distances $d_0, d_1, \ldots, d_m$.

$\frac{4c_1}{\sqrt{v_{\min}}}\sqrt{M\log(M)}$, it follows that for M larger than some absolute constant, and any $s \in \Omega^S$,

$$\mathbb{P}\big(\tau_{r+1} - \tau_{r-1} \leq R_{r-1} + R_r | S(0) = s\big) > c_4, \tag{6.16}$$

where $R_{r-1} = R_r := c_3 M\log(M)$, $c_3 := \frac{70c_1^2}{v_{\min}}$ and $c_4 := \frac{1}{6}$. For $j \in \{0, \ldots, m-1\} \setminus \{r-1, r\}$, let

$$R_j := \left\lceil \left(2 + \frac{32\log(M)}{c_2 M^{1/2}\varphi(d_j)}\right) \cdot \frac{2c_2\sqrt{M}}{\varphi(d_j)} \right\rceil.$$

We are going to show that the probability of $\tau_{j+1} - \tau_j$ being greater than R_j is small. Let us define the interval I_j as

$$I_j := \left[d_{j+1} - \frac{1}{3}c_1\sqrt{\frac{\log(M)}{M}}, d_j + \frac{1}{3}c_1\sqrt{\frac{\log(M)}{M}} \right].$$

Notice that the drift bound (6.15) only holds for positions s for which $d_C(s)$ is within interval I_j. Since we can possibly exit from this interval within R_j steps from time τ_j, we cannot apply the martingale inequalities directly to $\{d_C(S(\tau_j + k))\}_{k\geq 0}$ as previously. This difficulty can be resolved by a coupling argument similar to the one in the proof of Proposition 6.10.

For every $j \in \{0, \ldots, m-1\} \setminus \{r-1, r\}$, we define a sequence of random variables $D_0^{(j)}, D_1^{(j)}, \ldots$, as follows. First, we set $D_0^{(j)} := d_C(S(\tau_j))$. After this, we define the rest of the sequence based on the condition that for every $l \in \mathbb{N}$,

$$\begin{aligned} D_{l+1}^{(j)} - D_l^{(j)} := {} & 1_{[d_C(S(\tau_j+l))\in I_j]} \cdot \big[d_C\big(S(\tau_j + l + 1)\big) - d_C\big(S(\tau_j + l)\big)\big] \\ & - 1_{[d_C(S(\tau_j+l))\notin I_j]} \cdot \frac{\varphi(d_j)}{2M}. \end{aligned}$$

Due to the definition of this sequence, for every $s \in \Omega^S$, we have

$$\begin{aligned}&\mathbb{P}\big(\tau_j - \tau_{j-1} > R_j | S(0) = s\big)\\ &\quad\le \mathbb{P}\bigg(\max_{0\le i\le R_j} D_0^{(j)} > d_j + \frac{1}{3}c_1\sqrt{\frac{\log(M)}{M}} | S(0) = s\bigg) + \mathbb{P}\big(D_{R_j}^{(j)} > d_{j+1} | S(0) = s\big).\end{aligned} \tag{6.17}$$

Based on the definition, it follows that $\{D_i^{(j)}\}_{i\ge 0}$ satisfies the conditions of Lemma 6.7 with filtration $(\mathcal{F}_l)_{l\ge 0} := (\sigma(S(\tau_j), \dots, S(\tau_j + l)))_{l\ge 0}$, $R = \frac{1}{M}$, and $\delta = \frac{\varphi(d_j)}{2M}$. Therefore by (6.10) and (6.11), for M larger than some absolute constant, for every $j \in \{0, \dots, m-1\} \setminus \{r-1, r\}$, for every $s \in \Omega^S$, we have $\mathbb{P}(D_{R_j}^{(j)} > d_{j+1} | S(0) = s) \le \frac{1}{2M^2}$, and

$$\mathbb{P}\bigg(\max_{0\le i\le R_j} D_0^{(j)} > d_j + \frac{1}{3}c_1\sqrt{\frac{\log(M)}{M}} | S(0) = s\bigg) \le \frac{1}{2M^2},$$

therefore

$$\mathbb{P}\big(\tau_j - \tau_{j-1} > R_j | S(0) = s\big) \le \frac{1}{M^2}. \tag{6.18}$$

Now it is easy to show that for M larger than some absolute constant,

$$\sum_{j=0}^{m} R_j \le c_5 M \log(M), \tag{6.19}$$

for some absolute constant c_5. Moreover, based on (6.16) and (6.18), by the union bound, it follows that for M larger than some absolute constant, for every $s \in \Omega^S$,

$$\mathbb{P}\big(\tau_m < c_5 M \log(M) | S(0) = s\big) > \frac{c_4}{2}. \tag{6.20}$$

Since this result holds for any $s \in \Omega^S$, therefore by repeated application, we obtain that for any $k \in \mathbb{Z}_+$,

$$\mathbb{P}\big(\tau_m > k(c_5+1) M \log(M) | S(0) = s\big) \le \bigg(1 - \frac{c_4}{2}\bigg)^k, \tag{6.21}$$

which implies the claim of the theorem. □

6.6. Fast mixing at the center via curvature

In this section, we are going to show that once we get near the center of one of the modes, we will quickly approach the stationary distribution restricted to that mode, proving Proposition 6.4. For this, we will use a curvature (i.e., path-coupling) argument.

Let (Λ, d_Λ) be a Polish metric space, and $P(x, y)$ a Markov kernel on Λ. For two probability measures $\boldsymbol{\mu}, \boldsymbol{\eta}$ on Λ, we define $\Pi(\boldsymbol{\mu}, \boldsymbol{\eta})$ as the set of measures ν on $\Lambda \times \Lambda$ with first and second

marginals $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ (that is, $\int_{y\in\Lambda} \nu(dx,dy) = \boldsymbol{\mu}(dx)$ and $\int_{x\in\Lambda} \nu(dx,dy) = \boldsymbol{\eta}(dy)$), and we define the Wasserstein distance of $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$, denoted by $W_1(\boldsymbol{\mu},\boldsymbol{\eta})$, as

$$W_1(\boldsymbol{\mu},\boldsymbol{\eta}) := \sup_{\nu\in\Pi(\boldsymbol{\mu},\boldsymbol{\eta})} \int_{x,y\in\Lambda} d(x,y)\nu(dx,dy).$$

Then for $x, y \in \Lambda$, $x \neq y$, [31] defines

$$\kappa(x,y) := \frac{d_\Lambda(x,y) - W_1(\boldsymbol{P}(x,\cdot),\boldsymbol{P}(y,\cdot))}{d_\Lambda(x,y)},$$

and calls $\kappa := \inf_{x,y\in\Lambda,x\neq y} \kappa(x,y)$ the *coarse Ricci curvature*. This quantity can be used to estimate the speed of convergence to the stationary distribution of the Markov chain.

[31] says that (Λ, d_Λ) satisfies the ε-geodesic property if for any two points $x, y \in \Lambda$, there exists a set of points $x_0 := x, x_1, \ldots, x_l := y$ such that $d_\Lambda(x,y) = d_\Lambda(x_0,x_1) + \cdots + d_\Lambda(x_{l-1},x_l)$ and $d_\Lambda(x_i,x_{i+1}) \leq \varepsilon$ for every $0 \leq i < l$. Proposition 19 of [31] shows that for such spaces, $\kappa = \inf_{x,y\in\Lambda,x\neq y,d_\Lambda(x,y)\leq\varepsilon} \kappa(x,y)$, that is, it suffices to estimate $\kappa(x,y)$ for pairs of points whose distance is at most ε. We will choose Λ as a subset of Ω^S (the state space of the colour ratio vector S), and define the distance

$$d_\Lambda(x,y) := \frac{M}{2}\big(|x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3|\big),$$

corresponding to the amount of edges one needs to traverse on the hexagonal lattice to get from x to y. This distance is 1-geodesic, thus it will suffice to bound $\kappa(x,y)$ for neighbouring points x, y. It turns out that if we choose $\boldsymbol{P}$ as the Markov kernel of the magnetisation chain, $\kappa(x,y)$ will not be positive through the whole state space Ω^S. This negative curvature still persists even for the local restrictions of the Markov kernel to the 4 modes (the 4 triangles on Figure 2). This is caused by the non-convexity of the distribution at the regions separating the modes. For $1 \leq i \leq 4$, we define 4 regions near the center of the modes as

$$\Lambda^{(i)} := \big\{s \in \Omega^S : -\rho \leq s_l - C_{i,l} \leq 2\rho \text{ for } l = 1,2,3\big\}. \tag{6.22}$$

The following proposition shows that the curvature is positive in these regions. The proof of this result is included in Section A.4 (in the supplementary material [32]).

Proposition 6.13 (Curvature bound). *Consider the state space $(\Lambda^{(m)}, d_\Lambda)$ as above for some $1 \leq m \leq 4$, and let $\boldsymbol{P}^{\mathrm{mag}}_{(m)}$ be the restriction of the kernel of the magnetisation chain $\boldsymbol{P}^{\mathrm{mag}}$ to $\Lambda^{(i)}$, i.e. for $x, y \in \Lambda^{(m)}$, we let*

$$\boldsymbol{P}^{\mathrm{mag}}_{(m)}(x,dy) := \boldsymbol{P}^{\mathrm{mag}}\big(x, (\Lambda^{(m)})^c\big)\cdot\delta_x(dy) + 1_{[y\in\Lambda^{(m)}]}\cdot\boldsymbol{P}(x,dy),$$

where $\delta_x(dy)$ corresponds to the Dirac-δ distribution at x. Let κ denote the coarse Ricci curvature of the kernel $\boldsymbol{P}^{\mathrm{mag}}_{(m)}$ on the metric space $(\Lambda^{(m)}, d_\Lambda)$. Then for M larger than some absolute constant, we have $\kappa \geq \frac{0.01}{M}$.

Let $\tilde{\Lambda}^{(i)} := \{\sigma \in \Omega : s(\sigma) \in \Lambda^{(i)}\}$. Let $\boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}$ denote the restriction of $\boldsymbol{\mu}_M$ to $\tilde{\Lambda}^{(i)}$ (i.e. $\boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}(f) = \frac{\boldsymbol{\mu}_M(f \cdot 1_{\tilde{\Lambda}^{(i)}})}{\boldsymbol{\mu}_M(\tilde{\Lambda}^{(i)})}$, and $\boldsymbol{\mu}_M^{(i)}$ denote the restriction of $\boldsymbol{\mu}_M$ to $F_M^{(i)}$ (i.e. $\boldsymbol{\mu}_M^{(i)}(f) = \frac{\boldsymbol{\mu}_M(f \cdot 1_{F_M^{(i)}})}{\boldsymbol{\mu}_M(F_M^{(i)})}$). The following lemma bounds the total variational distance of these two distributions. It will be used in the proof of Proposition 6.4.

Lemma 6.14. *For M larger than some absolute constant, for every $1 \le i \le 4$, we have*

$$d_{\mathrm{TV}}\big(\boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}, \boldsymbol{\mu}_M^{(i)}\big) \le \frac{C_\Lambda}{M^2} \quad \text{and} \quad \boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}\big(\tilde{\Lambda}^{(i)} \setminus \mathcal{I}_M^{(i)}\big) \le \frac{C_\Lambda}{M^2}, \tag{6.23}$$

for some absolute constant $C_\Lambda < \infty$ (the inner regions $\mathcal{I}_M^{(i)}$ are defined as in (6.5)).

This lemma essentially states that most of the mass of the distribution $\boldsymbol{\mu}_M$ is contained near the centers. The proof of is included in Section A.6 (in the supplementary material [32]). Now we are ready to prove Proposition 6.4.

Proof of Proposition 6.4. Let $\boldsymbol{P}_{\Lambda^{(i)}}$ be a Markov kernel that is the restriction of $\boldsymbol{P}$ to $\Lambda^{(i)}$, i.e., for every $x, y \in \Lambda^{(i)}$,

$$\boldsymbol{P}_{\tilde{\Lambda}^{(i)}}(x, dy) := \boldsymbol{P}\big(x, \big(\Lambda^{(i)}\big)^c\big) \cdot \delta_x(dy) + 1_{[y \in \tilde{\Lambda}^{(i)}]} \cdot \boldsymbol{P}(x, dy).$$

Let $\sigma(0)$ be a fixed element of $\tilde{\Lambda}^{(i)}$, $\sigma'(0) \sim \boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}$, and define copies of them as $\Sigma(0) := \sigma(0)$ and $\Sigma'(0) = \sigma'(0)$. Let $\{\sigma(i)\}_{i \ge 0}$ and $\{\sigma'(i)\}_{i \ge 0}$ be two Markov chains evolving according to the kernel P, and let $\{\Sigma(i)\}_{i \ge 0}$ and $\{\Sigma'(i)\}_{i \ge 0}$ be two Markov chains evolving according to the kernel $P_{\tilde{\Lambda}^{(i)}}$. We are going to obtain the total variational distance bound by creating a coupling $\{\sigma(i), \sigma'(i), \Sigma(i), \Sigma'(i)\}_{0 \le i \le k}$ of these four chains. Let $\boldsymbol{\mu}_{(i)}^{\mathrm{mag}}$ denote the restriction of the magnetisation distribution $\boldsymbol{\mu}_{\tilde{\beta}_c, M}^{\mathrm{mag}}$ to $\Lambda^{(i)}$. First, we note that for $k_1 := \lfloor 300 M \log(M) \rfloor$, based on Proposition 6.13, and Corollary 21 of [31], we have

$$\begin{aligned} d_{\mathrm{TV}}\big(\big(\boldsymbol{P}_{(i)}^{\mathrm{mag}}\big)^{k_1}\big(s\big(\sigma(0)\big), \cdot\big), \boldsymbol{\mu}_{(i)}^{\mathrm{mag}}\big) &\le W_1\big(\big(\boldsymbol{P}_{(i)}^{\mathrm{mag}}\big)^{k_1}\big(s\big(\sigma(0)\big), \cdot\big), \boldsymbol{\mu}_{(i)}^{\mathrm{mag}}\big) \\ &\le 3\rho M \cdot \left(1 - \frac{0.01}{M}\right)^{k_1} \le \frac{1}{2M^2}, \end{aligned} \tag{6.24}$$

for M larger than some absolute constant. In the first step, we have used fact that the minimum distance between disjoint two points in our metric d_Λ is 1.

Let $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ be two probability measures on a finite space W. Proposition 4.7 of [25] shows the existence of an optimal coupling, that is, a coupling (X, Y) of two random variables $X \sim \boldsymbol{\nu}$ and $Y \sim \boldsymbol{\eta}$ such that $\mathbb{P}(X \ne Y) = d_{\mathrm{TV}}(\boldsymbol{\nu}, \boldsymbol{\eta})$.

We choose the coupling $(s(\Sigma(k_1)), s(\Sigma'(k_1)))$ as an optimal coupling. By (6.24) means that they satisfy that

$$\mathbb{P}\big(s\big(\Sigma(k_1)\big) \ne s\big(\Sigma'(k_1)\big)\big) \le \frac{1}{M^2} \tag{6.25}$$

for M larger than some absolute constant. Given the joint distribution of $(s(\Sigma(k_1)), s(\Sigma'(k_1)))$, we choose the joint distribution $(\Sigma(k_1), \Sigma'(k_1))$ arbitrarily among the possibilities. After this, we define $\{\Sigma(i)\}_{1\le i\le k_1-1}$ and $\{\Sigma'(i)\}_{1\le i\le k_1-1}$ based on their conditional distribution given $\Sigma(0)$, $\Sigma(k_1)$, and $\Sigma'(0)$, $\Sigma'(k_1)$, respectively (their joint distribution can be chosen arbitrarily among the possibilities).

Now that $\{\Sigma(i), \Sigma'(i)\}_{0\le i\le k_1}$ is defined, we define $\{\sigma(i)\}_{0\le i\le k_1}$ and $\sigma'(i)\}_{0\le i\le k_1}$ recursively, based on the optimal coupling with $\{\Sigma(i)\}_{0\le i\le k_1}$ and $\{\Sigma'(i)\}_{0\le i\le k_1}$, respectively. That is, if we have already defined $\{\sigma(j)\}_{0\le j\le i}$ for some $0 \le i \le k_1 - 1$, then we define $\sigma(i+1)$ such that $\sigma(i+1)$ and $\Sigma(i+1)$ are optimally coupled, and similarly for $\sigma'(i+1)$ and $\Sigma'(i+1)$. Due to the definition of the Markov kernels $\boldsymbol{P}_{\tilde{\Lambda}^{(i)}}$ and $\boldsymbol{P}$, we have

$$\mathbb{P}\big(\sigma(k_1) \ne \Sigma(k_1)\big) \le \mathbb{P}\big(\sigma(i) \notin \Lambda^{(i)} \text{ for some } 1 \le i \le k_1\big),$$

and since $\sigma(0) \in \mathcal{I}_M^{(i)}$, by Proposition 6.10, we have

$$\mathbb{P}\big(\sigma(k_1) \ne \Sigma(k_1)\big) \le k_1^2 \exp(-C_{\mathrm{esc}} M). \tag{6.26}$$

Based on Lemma 6.14, we have $\mathbb{P}(\sigma'(0) \notin \mathcal{I}_M^{(i)}) \le \frac{C_\Lambda}{M^2}$, and therefore by the same argument, we have

$$\mathbb{P}\big(\sigma'(k_1) \ne \Sigma'(k_1)\big) \le k_1^2 \exp(-C_{\mathrm{esc}} M) + \frac{C_\Lambda}{M^2}. \tag{6.27}$$

At this point, by combining (6.25), (6.26) and (6.27), we can see that

$$\mathbb{P}\big(s(\sigma_{k_1}) \ne s\big(\sigma'_{k_1}\big)\big) \le 2k_1^2 \exp(-C_{\mathrm{esc}} M) + \frac{C_\Lambda + 1}{M^2}. \tag{6.28}$$

From this point onwards, whenever $s(\sigma_{k_1}) = s(\sigma'_{k_1})$, we define the joint distribution $\{\sigma_i, \sigma'_i\}_{k_1\le i\le k}$ conditioned on $(\sigma_{k_1}, \sigma'_{k_1})$ as the coupling given by Lemma 6.8. When $s(\sigma_{k_1}) \ne s(\sigma'_{k_1})$, the joint distribution $\{\sigma_i, \sigma'_i\}_{k_1\le i\le k}$ is chosen arbitrarily. Then based on Lemma 6.8, for M larger than some absolute constant, we have

$$\begin{aligned}\mathbb{P}\big(\sigma_k \ne \sigma'_k\big) &\le 2k_1^2 \exp(-C_{\mathrm{esc}} M) + \frac{C_\Lambda + 1}{M^2} + \frac{M}{2}\exp\left(-\frac{(k-k_1)}{9M}\right) \\ &\le 2k_1^2 \exp(-C_{\mathrm{esc}} M) + \frac{C_\Lambda + 2}{M^2}.\end{aligned} \tag{6.29}$$

Finally, we define the joint distribution of $\{\Sigma'(i), \sigma'(i)\}_{k_1\le i\le k}$ as the optimal coupling in each step as previously. With the same argument as in (6.27), we have

$$\mathbb{P}\big(\sigma'(k) \ne \Sigma'(k)\big) \le k^2 \exp(-C_{\mathrm{esc}} M) + \frac{C_\Lambda}{M^2}, \tag{6.30}$$

for M larger than some absolute constant. By combining (6.29) and (6.30), we obtain that for M larger than some absolute constant,

$$\mathbb{P}\big(\sigma_k \ne \Sigma'_k\big) \le 3k^2 \exp(-C_{\mathrm{esc}} M) + \frac{2C_\Lambda + 2}{M^2}, \tag{6.31}$$

and the result follows by noticing that Σ_k' is distributed according to $\boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}$ and that by Lemma (6.14), $d_{\mathrm{TV}}(\boldsymbol{\mu}_{M|\tilde{\Lambda}^{(i)}}, \boldsymbol{\mu}_M^{(i)}) \leq \frac{C_\Lambda}{M^2}$. □

Acknowledgements

AJ & DP were supported by AcRF Tier 2 grant R-155-000-143-112. AT and DP were supported by AcRF Tier 1 grant R-155-000-150-133. AJ is affiliated with the Risk Management Institute and the Centre for Quantitative Finance at the National University of Singapore. We thank Pierre Del Moral for his encouragement and insightful comments. We thank Tobias Terzer for finding several typos and one error in the paper. We thank the anonymous referees for their insightful remarks.

Supplementary Material

Supplement to "Error bounds for sequential Monte Carlo samplers for multimodal distributions" (DOI: 10.3150/17-BEJ988SUPP; .pdf). Some technical proofs from the paper are given here.

References

[1] Beskos, A., Crisan, D. and Jasra, A. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.* **24** 1396–1445. MR3211000

[2] Beskos, A., Crisan, D.O., Jasra, A. and Whiteley, N. (2014). Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions. *Adv. in Appl. Probab.* **46** 279–306. MR3189059

[3] Beskos, A., Jasra, A., Law, K., Tempone, R. and Zhou, Y. (2017). Multilevel sequential Monte Carlo samplers. *Stochastic Process. Appl.* **127** 1417–1440. MR3630230

[4] Bhatnagar, N. and Randall, D. (2004). Torpid mixing of simulated tempering on the Potts model. In *Proceedings of the Fifteenth Annual ACM–SIAM Symposium on Discrete Algorithms* 478–487. New York: ACM.

[5] Bhatnagar, N. and Randall, D. (2015). Simulated tempering and swapping on mean-field models. Preprint. Available at 1508.04521.

[6] Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461. MR3127871

[7] Chen, L.H.Y. and Röllin, A. (2010). Stein couplings for normal approximation. Preprint. Available at 1003.6039.

[8] Chopin, N. (2002). A sequential particle filter for static models. *Biometrika* **89** 539–551.

[9] Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* **32** 2385–2411. MR2153989

[10] Cuff, P., Ding, J., Louidor, O., Lubetzky, E., Peres, Y. and Sly, A. (2012). Glauber dynamics for the mean-field Potts model. *J. Stat. Phys.* **149** 432–477.

[11] Del Moral, P. (2004). *Feynman–Kac Formulae*: *Genealogical and Interacting Particle Systems with Applications*. *Probability and Its Applications* (*New York*). New York: Springer. MR2044973

[12] Del Moral, P. (2013). *Mean Field Simulation for Monte Carlo Integration*. *Monographs on Statistics and Applied Probability* **126**. Boca Raton, FL: CRC Press. MR3060209
[13] Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333
[14] Del Moral, P., Doucet, A. and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22** 1009–1020. MR2950081
[15] Del Moral, P. and Miclo, L. (2000). Branching and interacting particle systems approximations of Feynman–Kac formulae with applications to non-linear filtering. In *Séminaire de Probabilités, XXXIV*. *Lecture Notes in Math.* **1729** 1–145. Berlin: Springer.
[16] Eberle, A. and Marinelli, C. (2013). Quantitative approximations of evolving probability measures and sequential Markov chain Monte Carlo methods. *Probab. Theory Related Fields* **155** 665–701.
[17] Giraud, F. and Del Moral, P. (2017). Nonasymptotic analysis of adaptive and annealed Feynman–Kac particle models. *Bernoulli* **23** 670–709. MR3556789
[18] Goodman, J. and Sokal, A.D. (1989). Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D* **40** 2035–2071.
[19] Grafakos, L. (2004). *Classical and Modern Fourier Analysis*. Upper Saddle River, NJ: Pearson Education, Inc.
[20] Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **78** 2690–2693.
[21] Jasra, A. and Doucet, A. (2008). Stability of sequential Monte Carlo samplers via the Foster–Lyapunov condition. *Statist. Probab. Lett.* **78** 3062–3069.
[22] Jasra, A., Stephens, D.A., Doucet, A. and Tsagaris, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Stat.* **38** 1–22. MR2760137
[23] Kou, S.C., Zhou, Q. and Wong, W.H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.* **34** 1581–1652. With discussions and a rejoinder by the authors.
[24] Levin, D.A., Luczak, M.J. and Peres, Y. (2010). Glauber dynamics for the mean-field Ising model: Cut-off, critical power law, and metastability. *Probab. Theory Related Fields* **146** 223–265.
[25] Levin, D.A., Peres, Y. and Wilmer, E.L. (2009). *Markov Chains and Mixing Times*. Providence, RI: Amer. Math. Soc. With a chapter by James G. Propp and David B. Wilson. MR2466937
[26] Lindsten, F., Johansen, A.M., Naesseth, C.A., Kirkpatrick, B. Schön, T.B., Aston, J. and Bouchard-Côté, A. (2015). Divide-and-conquer with sequential Monte Carlo. Preprint. Available at 1406.4993.
[27] Liu, J.S. and Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87** 353–369.
[28] Madras, N. and Zheng, Z. (2003). On the swapping algorithm. *Random Structures Algorithms* **22** 66–97.
[29] Neal, R.M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. MR1837132
[30] Olivieri, E. and Vares, M.E. (2005). *Large Deviations and Metastability*. *Encyclopedia of Mathematics and Its Applications* **100**. Cambridge: Cambridge Univ. Press.
[31] Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **256** 810–864.
[32] Paulin, D., Jasra, A. and Thiery, A. (2017). Supplement to "Error bounds for sequential Monte Carlo samplers for multimodal distributions." DOI:10.3150/17-BEJ988SUPP.
[33] Potts, R.B. (1952). Some generalized order-disorder transformations. *Proc. Cambridge Philos. Soc.* **48** 106–109.
[34] Schweizer, N. (2012). Non-asymptotic error bounds for Sequential MCMC methods Ph.D. thesis, Universitäts-und Landesbibliothek Bonn. Available at http://hss.ulb.uni-bonn.de/2012/2906/2906.pdf.
[35] Warwick, T. (2011). *Validated Numerics*: *A Short Introduction to Rigorous Computations*. Princeton, NJ: Princeton Univ. Press.

[36] Whiteley, N. (2012). Sequential Monte Carlo samplers: Error bounds and insensitivity to initial conditions. *Stoch. Anal. Appl.* **30** 774–798. MR2966098

[37] Whittaker, E.T. and Watson, G.N. (1962). *A Course of Modern Analysis*, 4th ed. Cambridge: Cambridge Univ. Press.

[38] Woodard, D.B., Schmidler, S.C. and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.* **19** 617–640.

[39] Woodard, D.B., Schmidler, S.C. and Huber, M. (2009). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electron. J. Probab.* **14** 780–804.

Received February 2017 and revised July 2017