J. Appl. Probab. **62,** 950–971 (2025)
*doi:*10.1017/jpr.2024.102

PERTURBATION ANALYSIS OF MARKOV CHAIN MONTE CARLO FOR GRAPHICAL MODELS

NA LIN,* ** AND
YUANYUAN LIU,* *Central South University*
AARON SMITH,*** *University of Ottawa*

Abstract

The basic question in perturbation analysis of Markov chains is: how do small changes in the *transition kernels* of Markov chains translate to chains in their *stationary distributions*? Many papers on the subject have shown, roughly, that the change in stationary distribution is small as long as the change in the kernel is much less than some measure of the convergence rate. This result is essentially sharp for generic Markov chains. We show that much larger errors, up to size roughly the *square root* of the convergence rate, are permissible for many target distributions associated with graphical models. The main motivation for this work comes from computational statistics, where there is often a tradeoff between the *per-step error* and *per-step cost* of approximate MCMC algorithms. Our results show that larger perturbations (and thus less-expensive chains) still give results with small error.

Keywords: Perturbation error; mixing time; approximate MCMC; Hellinger distance; decay of correlations

2020 Mathematics Subject Classification: Primary 60J20
Secondary 62H22

1. Introduction

Informally, perturbation bounds for Markov chains look like the following: if the distance $d(Q, K)$ between two transition kernels Q, K is sufficiently small, then the distance $d(\mu, \nu)$ between their stationary measures μ, ν is also small. Many results of this form, such as [30, Corollary 3.2] and follow-up work [16, 32, 38], require that the inverse error $d(Q, K)^{-1}$ be much larger than some notion of the 'time to convergence' $\tau(Q)$ of one of the two chains. The main results of this paper show that, in the special case that μ, ν both correspond to graphical models and Q, K 'respect' the same graphical model in a sense made precise in this paper, we can ensure that $d(\mu, \nu)$ remains small even for much larger errors $d(Q, K)$. Our main result, Theorem 2, allows errors of up to size $d(Q, K) \approx \tau(Q)^{-1/2} \gg \tau(Q)^{-1}$. Our main illustrative example, in Section 5, shows how these bounds can be achieved in a simple setting. We also note by example that both the existing bounds and our new bounds are essentially sharp for certain large and natural classes of Markov chains; see Examples 1, 2, and 3.

Received 10 January 2024; accepted 21 October 2024.
* Postal address: School of Mathematics and Statistics, HNP-LAMA, Central South University, Changsha, China.
** Email address: linna929@csu.edu.cn
*** Postal address: Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada.

The main motivation for our work is the analysis of 'approximate' or 'noisy' Markov chain Monte Carlo (MCMC) algorithms. In this setting, we think of Q as an ideal Markov chain that we would like to run, but which is computationally expensive. We think of K as an approximation to Q that is less computationally expensive. As a prototypical example inspired by [21], Q might be the usual Metropolis–Hastings algorithm targeting the posterior distribution associated with a large dataset of size n, while $K = K_m$ might be an algorithm that approximates the posterior distribution using a data subsample of size $m \ll n$. We expect the per-step cost of K_m to increase with m, while we also expect the error $d(Q, K_m)$ to decrease with m. This suggests a tradeoff: we would like to choose a value of m that is large enough for the stationary distributions of Q, K_m to be close, but small enough for K_m to be computationally tractable. Improved perturbation bounds let us confidently decrease the value of m while still obtaining good results.

1.1. Relationship to previous work

There is a long history of analyzing perturbations of matrices and other linear operators – see, e.g., the classic textbook [19]. Many of these results include Markov chains as special cases. The first application of these ideas to MCMC appeared in [6], which used perturbation bounds to show that 'rounding off' numbers to the typical precision of floating-point numbers will usually result in negligible error.

A recent surge of work on applying perturbation bounds to MCMC was inspired by two important new lines of algorithm development. The first was the development of algorithms, such as the unadjusted Langevin algorithm of [33], that are based on discretizations of 'ideal' continuous-time processes. The second was the development of algorithms, such as stochastic gradient Langevin dynamics of [42], the 'minibatch' Metropolis–Hastings of [21], and Monte Carlo within Metropolis of [28, 29], that try to make individual MCMC steps cheaper by using computationally cheap plug-in estimates of quantities appearing in 'ideal' MCMC algorithms. Often, as in [21, 42], the plug-in estimates are obtained by using a small subsample of the original dataset. In both lines, the result was a profusion of useful algorithms that can be viewed as small perturbations of algorithms that have very good theoretical support but are computationally intractable.

We focus now on the second line of algorithms. Perturbation theory has been used to try to show that the new computationally tractable algorithms would inherit many of the good properties of the original computationally intractable algorithms. Some representative theoretical papers on this subject include [2, 16, 32, 38], all of which give generic perturbation bounds that are widely applicable to approximate MCMC algorithms such as [21, 42] under very broad conditions. This generic work has also been applied to more complex algorithms, such as [17, 36].

One of the main questions raised by this work is: how good do the plug-in estimates need to be in order to obtain an algorithm with 'small' error? Theoretical work such as [16, 32, 38] often required that the distance between the 'original' kernel Q and 'approximate' kernel K satisfies a condition along the lines of

$$d(Q, K) \ll \frac{1}{\tau(Q)}, \tag{1}$$

where $d(Q, K)$ is a notion of distance between kernels and $\tau(Q)$ is a notion of time to converge to stationarity. Study of specific algorithms supported the idea that the inequality (1) was often necessary in practice and was *not* satisfied by various naive plug-in estimators [5, 18, 31].

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Since then, a great deal of applied work in the area has focused on developing plug-in estimators that do satisfy this inequality [4, 34].

This work looks at the question from the opposite point of view. Rather than trying to improve control variates to obtain algorithms that satisfy (1), we try to find conditions that are weaker than (1) and hold under conditions of interest in statistics. The *main lesson* in the present paper is that generic perturbation bounds can be *vastly* strengthened by restricting attention to specific statistically interesting settings.

Of course, the point of view in this paper and the point of view in applied work such as [4, 34] do not compete with each other. Improving control variates and improving the condition both allow you to run MCMC algorithms with smaller per-step costs, and these improvements can be combined. The *main remaining question* in this work is: how easy is it to engineer Markov chains for which conditions such as (1) can be weakened?

1.2. Guide to this paper

Section 2 describes the basic notation used throughout this paper. Section 3 is pedagogical, giving simple examples that aim to illustrate when previous results are or are not sharp. Section 4 states and proves the main result, Theorem 2. Finally, Section 5 gives a concrete algorithm and model to which Theorem 2 can be applied.

We note here one question that is raised rather early in the paper but not resolved until near the end. As discussed in Remark 1, there is one main obstacle to constructing an approximate algorithm to which Theore 2 can be applied: the algorithm's stationary distribution must 'respect' the same conditional independence properties as the original target distribution. The main purpose of Section 5 is to show that, while this property is not automatic, it is not terribly difficult to construct an algorithm with this property. In particular, the pseudomarginal framework introduced in [3] provides a powerful general tool for doing this.

2. Notation

We introduce notation used throughout the paper.

2.1. Generic probability notation

For two probability distributions μ, ν on the same Polish space Ω with Borel σ-algebra $(\Omega, \mathcal{F})$, define the total variation (TV) distance as $d_{\mathrm{TV}}(\mu, \nu) = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$.

When μ, ν have densities f, g with respect to a single reference measure λ, the Hellinger distance between μ and ν is defined as

$$d_{\mathrm{H}}(\mu, \nu) = \left(\int_{x \in \Omega} (\sqrt{f(x)} - \sqrt{g(x)})^2 \lambda(\mathrm{d}x) \right)^{1/2},$$

and the $L^2(\lambda)$ distance is defined as

$$d_{L^2(\lambda)}(\mu, \nu) = \left(\int_{x \in \Omega} (f(x) - g(x))^2 \lambda(\mathrm{d}x) \right)^{1/2}.$$

Our argument will rely on the following well-known relationships (see, e.g., [9, p. 135] and [23, Lemma 20.10]):

$$\tfrac{1}{2} d_{\mathrm{H}}^2(\mu, \nu) \leq d_{\mathrm{TV}}(\mu, \nu) \leq d_{\mathrm{H}}(\mu, \nu) \leq d_{L^2}(\mu, \nu). \tag{2}$$

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

By a small abuse of notation, when d is a distance on measures we extend it to a distance on transition kernels via the formula $d(Q, K) = \sup_{x\in\Omega} d(Q(x, \cdot), K(x, \cdot))$.

Let Q be a transition kernel with stationary measure π and let d be a metric on probability measures. Define the associated mixing time, $\tau_{\text{mix}}(Q, d, \varepsilon) = \min\{t : \sup_{x\in\Omega} d(Q^t(x, \cdot), \pi) < \varepsilon\}$. By convention, we write $\tau_{\text{mix}}(Q) = \tau_{\text{mix}}\left(Q, d_{\text{TV}}, \frac{1}{4}\right)$.

When we have two functions f, g on the same state space Ω, we write $f \lesssim g$ as shorthand for: there exists a constant $0 < C < \infty$ such that $f(x) \le Cg(x)$ for all $x \in \Omega$. Similarly, we write $g \gtrsim f$ if $f \lesssim g$, and we write $f \approx g$ if both $f \lesssim g$ and $g \lesssim f$.

2.2. Notation for couplings

For a random variable X, denote by $\mathcal{L}(X)$ the distribution of X. We will use the following standard 'greedy' coupling between two Markov chains on a finite state space. See, e.g., [23, Proposition 4.7] for a proof that the following construction yields well-defined processes.

Definition 1. *(Greedy Markovian coupling.)* Fix a Markov transition kernel K on finite state space Ω. Note that, for points $a, b \in \Omega$, it is possible to write

$$K(a, \cdot) = \delta_{a,b}\, \mu_{a,b} + (1 - \delta_{a,b})\nu_{a,b;1},$$
$$K(b, \cdot) = \delta_{a,b}\, \mu_{a,b} + (1 - \delta_{a,b})\nu_{a,b;2},$$

where $\mu_{a,b}$, $\nu_{a,b;1}$, and $\nu_{a,b;2}$ are probability measures on Ω and $\delta_{a,b} = 1 - d_{\text{TV}}(K(a, \cdot), K(b, \cdot))$.

Next, fix starting points $x, y \in \Omega$ and let $\{X_t\}_{t\ge0}$ (respectively $\{Y_t\}_{t\ge0}$) be a Markov chain evolving through K and starting at $X_0 = x$ (respectively $Y_0 = y$). The *greedy Markovian coupling* of these chains is defined inductively by the following scheme for sampling (X_{t+1}, Y_{t+1}) given X_t, Y_t:

(i) Sample $B_t \in \{0, 1\}$ according to the distribution $\mathbb{P}[B_t = 1] = \delta_{X_t,Y_t}$. Then:

  (a) If $B_t = 1$, sample $Z \sim \mu_{X_t,Y_t}$ and set $X_{t+1} = Y_{t+1} = Z$.

  (b) If $B_t = 0$, sample $X_{t+1} \sim \nu_{X_t,Y_t;1}$ and $Y_{t+1} \sim \nu_{X_t,Y_t;2}$ independently.

We note that the coupling in Definition 1 has the following properties:

(i) The joint process $\{X_t, Y_t\}_{t\ge0}$ is a Markov chain.

(ii) Set $\tau = \min\{t : X_t = Y_t\}$. Then $X_s = Y_s$ for all $s < \tau$ under this coupling.

2.3. Notation for Gibbs samplers and graphical models

Fix $q \in \mathbb{N}$ and define $[q] = \{1, 2, \ldots, q\}$. We denote by the binary tuple $G = (V, \Phi)$ a collection of *vertices* and *factors*. That is, following the notation of [20], we have:

- V is any finite set. We call this set the *vertices*.
- Φ is any collection of functions from $\Omega \equiv [q]^V$ to $\mathbb{R}$. We call this set the *factors*.

We define the Gibbs measure associated with the tuple (V, Φ) to be the probability measure on Ω given by

$$\mu(\sigma) \propto \exp\left(\sum_{\phi\in\Phi} \phi(\sigma)\right), \tag{3}$$

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

where σ represents a configuration in Ω. Factors typically depend on only a few values. To make this precise, we say that a factor $\phi \in \Phi$ *does not depend* on a vertex $v \in V$ if $\phi(\sigma) = \phi(\eta)$ for all $\sigma, \eta \in \Omega$ satisfying $\sigma(u) = \eta(u)$, $u \neq v$. Otherwise, we say ϕ *depends* on v. For $x \in V$, denote by $A[x] = \{\phi \in \Phi : \text{factor } \phi \text{ depends on variable } x\}$ the set of factors that depend on variable x. Similarly, let $S[\phi] = \{x \in V : \text{factor } \phi \text{ depends on variable } x\}$. For fixed (V, Φ), denote by (V, E) the graph with $(i, j) \in E$ if and only if there exists $\phi \in \Phi$ with $i, j \in S[\phi]$. We note that this graph is the usual (minimal) Markov network associated with the data (G, Φ).

For a graph metric d, we abuse notation slightly by defining the distance between two vertex sets A and B as $d(A, B) = \min\{d(x, y) : x \in A, y \in B\}$. For a set $A \subset V$ and integer $r > 0$, we denote by $B_r(A)$ the set of vertices whose distance from A is less than r. That is, $B_r(A) = \{x \in V : d(x, A) < r\}$.

Given $A \subset V$, we denote by $\sigma|_A : A \to [q]^A$ the restriction of the configuration σ to the set A and by $\mu|_A$ the restriction of the measure μ to the set A, i.e. $\sigma|_A(x) = \sigma(x)$ for $x \in A$ and $\mu|_A(\sigma|_A) = \sum_{\eta:\eta|_A = \sigma|_A} \mu(\eta)$. For brevity we write $\mu|_A(\sigma) = \mu|_A(\sigma|_A)$. For $A \subset B \subset V$ and $\sigma \in \Omega$, we denote by $\mu|_A^{\eta|_{B^c}}$ the restriction to A of the conditional distribution of μ given that it takes value η on B^c. That is,

$$\mu|_A^{\eta|_{B^c}}(\sigma) = \frac{\sum_{\theta|_A = \sigma|_A, \theta|_{B^c} = \eta|_{B^c}} \mu(\theta)}{\sum_{\theta|_{B^c} = \eta|_{B^c}} \mu(\theta)}.$$

Let V be partitioned into Γ pairwise disjoint subsets $\{S_i\}$, and $\Pi_1 = \emptyset$ and $\Pi_j \subset \bigcup_{i=1}^{j-1} S_i$, $2 \leq j \leq \Gamma$, correspond to the set of vertices conditioned on which the configuration of S_j is independent from everything else in $\bigcup_{i=1}^{j-1} S_i$. Specifically, let $\{S_j, \Pi_j\}$ satisfy

$$\mu(\sigma) = \prod_{j=1}^{\Gamma} \mu|_{S_j}^{\sigma|_{\Pi_j}}(\sigma). \tag{4}$$

Say that two measures have the same dependence structure if they can both be written in the form (4) with the same lists $\{S_i\}$ and $\{\Pi_i\}$.

Denote by $\sigma_{x \mapsto i}$ be the configuration

$$\sigma_{x \mapsto i}(x) = i,$$
$$\sigma_{x \mapsto i}(y) = \sigma(y), \quad y \neq x.$$

We recall in Algorithm 1 the usual algorithm for taking a single step from a Gibbs sampler targeting (3).

This Gibbs sampler defines a Markov chain with transition matrix

$$Q(\sigma, \sigma_{x \mapsto j}) = \frac{1}{|V|} \cdot \frac{\exp(s_j)}{\sum_{i=1}^{q} \exp(s_i)}.$$

In addition, we denote by $Q|_B$ the Gibbs sampler that only updates labels in B and fixes the value of all labels in B^c. Note that $Q|_B$ with initial state σ has stationary measure $\mu|_B^{\sigma|_{B^c}}$.

We next define the class of algorithms that will be the focus of this paper.

Definition 2. Let $(\mathbb{A}, d)$ be a Polish space and $\mathcal{F}_{\mathbb{A}}$ the associated σ-algebra. We say that a Markov chain K on state space $\Omega \times \mathbb{A}$ is a *perturbed Gibbs sampler* with tuple (V, Φ) if it has the following properties:

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Algorithm 1 Step of standard Gibbs sampler.

Require: Starting state $\sigma \in \Omega$.

1: Sample variable index $x \sim \mathrm{Unif}(V)$.

2: For $i \in [q]$, calculate $s_i = \sum_{\phi \in A[x]} \phi(\sigma_{x \mapsto i})$. Define the distribution p over $[q]$ by $p(i) \propto \exp(s_i)$.

3: Sample $j \sim p$.

4: Return the element $\eta = \sigma_{x \mapsto j}$.

(i) *Single-site updates:* For samples $(\eta, b) \sim K((\sigma, a), \cdot)$, $|\{x \in V : \eta(x) \neq \sigma(x)\}| \leq 1$.

(ii) *Gibbs stationary measure respects factorization structure:* K is ergodic, with stationary measure $\hat{\nu}$ whose marginal on Ω is of the form (3) and has the same dependence structure as (4).

Remark 1. (*Checking that a Gibbs sampler satisfies Definition 2.*) The first part of Definition 2 is fairly innocuous, both in that it is often straightforward to check and in that our arguments are not terribly sensitive to small violations of this assumption. For example, it is straightforward to extend our results from single-site Gibbs updates to the case of Gibbs samplers that update $O(1)$ different entries that are all 'close' in the Markov network.

The second condition is more dangerous, in that it is both harder to check and our arguments do not work as written if it fails. Broadly speaking, we know of two ways to check that a transition kernel satisfies this condition:

(i) In the case that K has the same state space as Q (i.e. $\mathbb{A}$ has one element), the simplest sufficient conditions that we are aware of appear in [15, Proposition 1]. The result in [15] requires that the Markov chain be 'synchronous' (a property that is straightforward to enforce for approximate Gibbs samplers based on subsampling ideas) and reversible (a familiar property). Although this condition is simple and seems strong enough for many purposes, it is far from the strongest possible condition; see, e.g., [22] for sufficient conditions that apply even to non-synchronous Markov chains.

(ii) In the case that K has the form of a pseudomarginal algorithm (see [3] for an introduction to pseudomarginal algorithms and [34] for pseudomarginal algorithms in the context of approximate MCMC), an exact formula for the marginal of the stationary distribution on Ω is available. Thus, as long as the random log-likelihood estimate can be written in the form $\hat{L}(\sigma) = \sum_{\phi \in \Phi} \hat{L}_\phi(\sigma)$, where $\{\hat{L}_\phi(\sigma)\}_{\phi \in \Phi}$ are independent and $\hat{L}_\phi(\sigma)$ depends only on $\{\sigma(i)\}_{i \in S[\phi]}$, the stationary distribution will also be of the form (3). We use this condition in the worked example in Section 5.

In our proof, the second condition of Definition 2 is used to invoke (7) (originally proved as [8, Theorem 1]). The proof in [8] relies on an exact factorization of the associated likelihoods, and it is beyond the scope of this paper to study target distributions that are merely 'close' to such a factorization.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

3. Intuition and simple examples

We give some simple (and far from optimal) calculations to show that simple perturbations are nearly sharp for generic examples, but far from sharp for structured examples. Our simple structured example also shows that our main result, Theorem 2, is nearly sharp.

First, we recall [32, Theorem 1]. To state it, we introduce the following temporary notation. Recall that a kernel Q with stationary distribution μ is said to be geometrically ergodic with factor ρ if there exists a finite function C such that, for every initial distribution ν and for all $n \in \mathbb{N}$, $d_{\mathrm{TV}}(\mu, \nu Q^n) \leq C(\nu)\rho^n$, and is said to be $L^2(\mu)$-geometrically ergodic if

$$d_{L^2(\mu)}(\mu, \nu Q^n) \leq C(\nu)\rho^n. \tag{5}$$

As per [37, Theorem 2.1], if Q is reversible and the initial distribution $\nu \in L^2(\mu)$, then Q is $L^2(\mu)$-geometrically ergodic if and only if it is geometrically ergodic, with this equivalence holding for the same value of ρ (but not necessarily the same value of $C(\nu)$). In this notation, [32, Theorem 1] is given as follows.

Theorem 1. *Let Q,K be two transition kernels with stationary measures μ, ν and assume that Q satisfies (5). Then, for $d_{L^2(\mu)}(Q, K) \ll 1 - \rho$,*

$$d_{L^2(\mu)}(\mu, \nu) \leq \frac{d_{L^2(\mu)}(Q, K)}{\sqrt{(1-\rho)^2 - d^2_{L^2(\mu)}(Q, K)}} \approx \frac{d_{L^2(\mu)}(Q, K)}{1-\rho}.$$

We point out that submultiplicativity of the total variation distance to stationarity, as in [23, Lemma 4.11], implies that a finite state chain with finite mixing time is geometrically ergodic with factor ρ for which $(1-\rho)^{-1} = O(\tau_{\mathrm{mix}})$. In particular, we can see that $d_{L^2(\mu)}(\mu, \nu)$ is small if $d_{L^2(\mu)}(Q, K)$ is much smaller than the mixing rate $\min\{\tau_{\mathrm{mix}}(Q), \tau_{\mathrm{mix}}(K)\}^{-1}$. Qualitatively similar results, of the form

$$d(\mu, \nu) \lesssim \tau_{\mathrm{mix}}(Q, d) d(Q, K), \tag{6}$$

are known to hold for many metrics d, as long as the right-hand side is sufficiently small. For example, see [30] for this result with the choice $d = d_{\mathrm{TV}}$. The following simple example shows that Theorem 1 is close to sharp.

Example 1. For two parameters $p \in (0, 0.5)$ and $C \in (0, p^{-1})$, define the transition kernels Q and K on $\{1, 2\}$ by the following non-holding probabilities:

$$Q(1, 2) = Q(2, 1) = p, \qquad K(1, 2) = Cp, \quad K(2, 1) = p.$$

We view Q as the original chain, and its stationary distribution μ which is uniform serves as the reference measure in the L^2 distance. It is straightforward to verify the following scaling estimates for fixed p and C in the range $C \in (0.5, 2)$:

(i) The distance between kernels $d_{L^2(\mu)}(Q, K) \approx |C - 1|p$.

(ii) The mixing rates of both chains satisfy $\tau_{\mathrm{mix}}(Q, d_{L^2(\mu)}) \approx \tau_{\mathrm{mix}}(K, d_{L^2(\mu)}) \approx 1/p$.

(iii) The distance between stationary distributions $d_{L^2(\mu)}(\mu, \nu) \approx |C - 1|$.

By the above items, we can see that $d_{L^2(\mu)}(\mu, \nu)$ is of the same order as the error

$$\min\{\tau_{\mathrm{mix}}(Q, d_{L^2(\mu)}), \tau_{\mathrm{mix}}(K, d_{L^2(\mu)})\} d_{L^2(\mu)}(Q, K)$$

for fixed p as $C \to 1$.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Example 1 shows that the error upper bound in Theorem 1 scales at best like the product $\min\{\tau_{\text{mix}}(Q), \tau_{\text{mix}}(K)\} d_{L^2(\mu)}(Q, K)$ when this product is small. When the product is large, the error can actually 'blow up'. The following example illustrates this 'blowing up' in the simpler total variation norm.

Example 2. For $n \in \mathbb{N}$ and $p \in (0, 0.5)$, define the transition kernel $Q = Q_{n,p}$ on $[n]$ by

$$\begin{aligned} Q(i, i+1) &= \tfrac{1}{2}p, & i < n, \\ Q(i, i-1) &= \tfrac{1}{2}(1-p), & i > 1, \\ Q(i, i) &= \tfrac{1}{2}, & 1 < i < n, \end{aligned}$$

and $Q(1, 1) = \frac{1}{2}(2-p)$, $Q(n, n) = \frac{1}{2}(1+p)$. Next, for $0 < \varepsilon < 1$, define $K = K_{n,p,\varepsilon}$ by

$$K(i, j) = (1-\varepsilon)Q(i, j) + \varepsilon \mathbf{1}_{\{j=n\}}.$$

Informally, K takes a step from Q with probability $1-\varepsilon$ and teleports to n with probability ε.

We consider the regime where $p \in (0, 0.5)$ is fixed and $\varepsilon = \varepsilon_n = C_n/n$ for some sequence $C_n \to \infty$. The proofs of the following four facts can be found in Appendix A:

(i) The distances between kernels $d_{\text{TV}}(Q, K) \approx C_n/n$.

(ii) The mixing times satisfy $\tau_{\text{mix}}(Q) \approx n$, $\tau_{\text{mix}}(K) \gtrsim n/C_n$.

(iii) The stationary distribution μ of Q assigns probability $\mu([0, n/3]) \to 1$ as $n \to \infty$.

(iv) The stationary distribution ν of K assigns probability $\nu([2n/3, n]) \to 1$ as $n \to \infty$.

By items (iii) and (iv), we can see that $d_{\text{TV}}(\mu, \nu) \to 1$, so the stationary measures are very far apart. On the other hand, by items (i) and (ii), the error $\min\{\tau_{\text{mix}}(Q), \tau_{\text{mix}}(K)\} d_{\text{TV}}(Q, K) \lesssim C_n$ grows arbitrarily slowly.

However, as we will see, (6) is far from sharp for many special cases.

Example 3. Consider $0.01 < p < \tilde{p} < 0.99$. Let μ be the distribution of random vector $X = (X_1, X_2, \ldots, X_{n^2})$ on $\{0, 1\}^{n^2}$, where all the random variables X_i, $i = 1, 2, \ldots, n^2$, are independently and identically distributed and take values over a Bernoulli distribution with parameter p. Denote by ν the analogous distribution with parameter $\tilde{p}$. Denote by Q, K the usual Gibbs samplers with targets μ, ν. It is well known that their mixing times are $\tau_{\text{mix}}(Q)$, $\tau_{\text{mix}}(K) \approx n^2 \log(n)$, uniformly in the choice of $0.01 < p < \tilde{p} < 0.99$. It is straightforward to check that $d_{\text{TV}}(Q(x, \cdot), K(x, \cdot)) \approx |\tilde{p} - p|$. For $0 < p < \tilde{p} < 1$, [1, (2.15)] says

$$d_{\text{TV}}(\mu, \nu) \leq \frac{\sqrt{\text{e}}}{2} \cdot \frac{C(\tilde{p} - p)}{(1 - C(\tilde{p} - p))^2}, \qquad \text{with } C(x) := x\sqrt{\frac{n^2 + 2}{2p(1-p)}}.$$

This shows that we can allow perturbations of size up to $d_{\text{TV}}(Q, K) \lesssim n^{-1}$, even though the mixing times are $\tau_{\text{mix}}(Q)$, $\tau_{\text{mix}}(K) \approx n^2 \log(n)$. This motivates us to generalize this behavior and find a more relaxed condition on the perturbation to ensure robustness.

We think of the main result of this paper as an extension of the phenomenon in Example 3. To give a hint as to how we might prove this, we give an informal calculation for a simple graphical model. Take the Ising model at a high enough temperature on a two-dimensional lattice, for example. The decay-of-correlation property yields that, for any sequence $\omega_n \gtrsim \log(n)$,

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

mixing occurs at a point x before any influence from outside the surrounding box of side-length $O(\omega_n)$ can propagate to that point; see [25] for rigorous mathematical definitions. This means that, when looking at perturbation bounds, we can effectively focus on a box growing at rate $\log(n)^2$ rather than the usual n^2. Thus, naive estimates with perturbation size up to $O(\log(n)^{-2})$ are enough to tell us that the marginal distributions at the center of the box are close. If we can leverage the independence property, we expect to be able to get bounds on the whole chain that are qualitatively similar to those in Example 3.

We will use the following subadditivity property of the Hellinger distance, proved in [8], to leverage this independence property.

Lemma 1. *If μ and ν are two Markov random fields on V with common factorization structure as (4), then we have*

$$d_{\mathrm{H}}^2(\mu, \nu) \le \sum_{j=1}^{\Gamma} d_{\mathrm{H}}^2(\mu|_{S_j \cup \Pi_j}, \nu|_{S_j \cup \Pi_j}). \tag{7}$$

4. General results

Throughout this section, we assume that μ is the Gibbs measure of a factor graph $G = (V, \Phi)$ on state space $\Omega = [q]^V$ with factorization structure (4), that Q is the associated Gibbs sampler, and that K is a perturbed Gibbs sampler with the same factor graph in the sense of Definition 2. We set notation for sampling from the first component of K: if $(\eta, b) \sim K((\sigma, a), \cdot)$, we say that $\eta \sim K((\sigma, a), \cdot)|_\Omega$. We also denote by $\widehat{\nu}$ the stationary measure of K and by ν its marginal distribution on Ω.

Fix a metric d on probability measures. Our results rely on the following four quantitative assumptions.

Assumption 1. (Decay of correlations.) *There exist positive constants m, $C_1 < \infty$ such that, for all configuration s σ, $\eta \in \Omega$, all $j \in [\Gamma]$, and any $r > 0$,*

$$d_{\mathrm{H}}\Big(\mu\big|_{S_j \cup \Pi_j}^{\sigma|_{B_r^c(S_j \cup \Pi_j)}}, \mu\big|_{S_j \cup \Pi_j}^{\eta|_{B_r^c(S_j \cup \Pi_j)}}\Big) \le C_1 \mathrm{e}^{-mr}.$$

Assumption 2. (Relationship to Hellinger.) *There exists a positive constant $C_2 < \infty$ such that, for any probability measures μ, ν on Ω, $d_{\mathrm{H}}(\mu, \nu) \le C_2 d(\mu, \nu)$.*

Assumption 3. (Propagation of perturbations.) *There exists a positive constant $C_3 < \infty$ and an increasing convex function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for all configurations $\sigma \in \Omega$ and all $j \in [\Gamma]$ and any $r > 0$,*

$$d\Big(\mu\big|_{S_j \cup \Pi_j}^{\sigma|_{B_r^c(S_j \cup \Pi_j)}}, \nu\big|_{S_j \cup \Pi_j}^{\sigma|_{B_r^c(S_j \cup \Pi_j)}}\Big) \le C_3 f(r) \sup_{(\sigma,a) \in \Omega \times \mathbb{A}} d(Q(\sigma, \cdot), K((\sigma, a), \cdot)|_\Omega).$$

Assumption 4. (Small perturbations) *There exists a positive constant $C_4 < \infty$ such that*

$$\sup_{(\sigma,a) \in \Omega \times \mathbb{A}} d(Q(\sigma, \cdot), K((\sigma, a), \cdot)|_\Omega) \le \frac{C_4}{\sqrt{\Gamma} f(\log \Gamma)}.$$

Remark 2. Assumption 1, also known as strong spatial mixing, plays a key role in the study of exponential ergodicity of Glauber dynamics for discrete lattice spin systems (see [25, Section 2.3] for details). It implies that a local modification of the boundary condition has an

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

influence on the corresponding Gibbs measure that decays exponentially fast with the distance from the site of the perturbation. There is a large literature devoted to studying the regimes where strong spatial mixing holds for particular models. In particular, the Ising and Potts models [7, 10, 13, 14, 26, 27, 40], the hard-core model [41], and q-colorings [12] have received special attention.

Assumption 2 places restrictions on the relationship between the distance measure d and the Hellinger distance d_{H}. Many popular norms satisfy this condition. For example, by [23, Lemma 20.10] the L^2 norm satisfies the condition with $C_2 = 1$.

Assumption 3 is very similar to the usual perturbation bound for discrete-time Markov chains, but specialized to subsets of the state space. See, e.g., [16, 30, 32, 38] for proofs of results of this form. There are many ways to check Assumption 3 with bounds of this form; see Remark 4 for a longer explanation of a prototypical argument of this sort.

Assumption 4 is the main assumption of our result, and must be checked for individual approximate chains. We think of this assumption as a nearly-sharp criterion for obtaining small error in the stationary measure, and we expect that users should *design* their MCMC algorithms to satisfy this criterion.

We are now in a position to state our main result.

Theorem 2. *Let the sequences of measures $\mu = \mu^{(\Gamma)}$ and $\nu = \nu^{(\Gamma)}$ have common factorization structure (4), and their associated Gibbs samplers $Q = Q^{(\Gamma)}$, $K = K^{(\Gamma)}$ satisfy Assumptions 1–4 with constants C_1, C_2, C_3, and C_4 that do not depend on Γ. Fix any $\delta > 0$. Let*

$$D = \sup\left\{M > 0 : \frac{f(M \log 2)}{f(\log 2)} \leq \frac{\delta}{2C_2C_3C_4}\right\}.$$

Then, for all $\Gamma \geq \max\{2, (2C_1/\delta)^{1/(mD-0.5)}\}$, we have $d_{\mathrm{H}}\big(\mu^{(\Gamma)}, \nu^{(\Gamma)}\big) \leq \delta$.

Proof. Under these assumptions, we have the following basic calculation for any Γ (which we omit in the notation for simplicity), any $j \in [\Gamma]$, and any $r > 0$:

$$\begin{aligned}
&d_{\mathrm{H}}\Big(\mu\big|_{S_j\cup\Pi_j}^{\sigma|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}, \nu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}\Big) \\
&\qquad \leq d_{\mathrm{H}}\Big(\mu\big|_{S_j\cup\Pi_j}^{\sigma|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}, \mu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}\Big) + d_{\mathrm{H}}\Big(\mu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}, \nu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}\Big) \\
&\qquad \leq C_1\mathrm{e}^{-mr} + C_2 d\Big(\mu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}, \nu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}\Big) \\
&\qquad \leq C_1\mathrm{e}^{-mr} + C_2C_3 f(r) \sup_{(\sigma,a)\in\Omega\times\mathbb{A}} d(Q(\sigma,\cdot), K((\sigma,a),\cdot)|_{\Omega}).
\end{aligned}$$

We now consider what happens as the number Γ of partitions gets large. Set $r = D\log\Gamma$. Using Assumption 4, we have

$$d_{\mathrm{H}}\Big(\mu\big|_{S_j\cup\Pi_j}^{\sigma|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}, \nu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\mathrm{c}}(S_j\cup\Pi_j)}}\Big) \leq C_1\Gamma^{-mD} + \frac{C_2C_3C_4 f(D\log\Gamma)}{\sqrt{\Gamma} f(\log\Gamma)}. \tag{8}$$

In particular, it is easy to check that for $\Gamma \geq (2C_1/\delta)^{1/(mD-0.5)}$ the first term of the right-hand-side in (8) is less than $\delta/2\sqrt{\Gamma}$. Furthermore, since f is an increasing convex function, for $\Gamma \geq 2$ we have

$$\frac{f(D\log\Gamma)}{f(\log\Gamma)} \leq \frac{f(D\log 2)}{f(\log 2)} \leq \frac{\delta}{2C_2C_3C_4}.$$

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Putting these together yields that, for $\Gamma \geq \max\{2, (2C_1/\delta)^{1/(mD-0.5)}\}$,

$$d_{\mathrm{H}}\Big(\mu\big|_{S_j\cup\Pi_j}^{\sigma|_{B_r^c(S_j\cup\Pi_j)}}, \nu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^c(S_j\cup\Pi_j)}}\Big) \leq \frac{\delta}{\sqrt{\Gamma}}.$$

Plugging this into (7) gives $d_{\mathrm{H}}^2(\mu, \nu) \leq \sum_{j=1}^{\Gamma} d_{\mathrm{H}}^2\big(\mu|_{S_j\cup\Pi_j}, \nu|_{S_j\cup\Pi_j}\big) \leq \delta^2$, which allows us to complete the proof. □

Remark 3. In the special case where $\Gamma = |V|$, $d = d_{L^2}$, and f is polynomial, this shows that we have 'small' perturbation error in the Hellinger distance even with 'large' kernel perturbations up to order $(\sqrt{|V|}\,\mathrm{polylog}\,(|V|))^{-1}$, even though the sample space is $|V|$-dimensional and the mixing time is about $|V|$. In other words, compared with generic bounds such as [30, 38] as summarized in our simple heuristic bound (6), our Assumption 4 allows for much larger perturbations.

Remark 4. (*Prototypical verification of Assumption 3.*) We recall that most papers in the perturbation theory literature (e.g. [16, 30, 32, 38]) assume (i) upper bounds on the difference $d(Q, K)$ between the kernels of interest, and (ii) lower bounds on some notion of convergence rate $\lambda(Q|_S)$ (such as the spectral gap) for one of the kernels. When using these bounds to verify Assumption 3, it is *not* enough to have bounds of this sort for the Markov chain as a whole. However, in the typical applications of these results, we expect these bounds to come from slightly stronger bounds, and these will be enough to verify Assumption 3.

First, the bound on $d(Q, K)$ is typically achieved as a consequence of a bound on $\sup_S d(Q|_S, K|_S)$. Second, for 'rapidly mixing' spin systems, we expect that $\lambda(Q|_S) \approx 1/|S|$ scales nicely with the size of S, and we assume the bounds of this sort are achievable. Then, in order to satisfy Assumption 3, it suffices to estimate

$$d(\mu|_S, \nu|_S) \lesssim \frac{d(Q, K)}{\max\{\lambda(Q|_S), \lambda(K|_S)\}} \lesssim |S| \sup_S d(Q|_S, K|_S),$$

where the first inequality can be directly derived by the usual perturbation theory of [16, 30, 32, 38]. Consequently, in practical applications, we require that $\sup_S d(Q|_S, K|_S)$ meets the condition in Assumption 4, namely $\sup_S d(Q|_S, K|_S) \leq C_4/(\sqrt{\Gamma} f(\log \Gamma))$.

Of course, these assumptions are not easy to verify in most situations. Our point is that the assumptions in all perturbation papers can be hard to verify, and that verifying our Assumption 3 is not much harder than verifying the bounds common to the rest of the usual perturbation-theory literature. The only real change is that our bounds are assumed to scale in a certain 'typical' way with the size of the system.

Remark 5. There exists an implicit trade-off between Assumptions 3 and 4 that is captured by the number Γ of partitions. The local perturbation bound, on the one hand, can be more easily estimated within a relatively small restricted region, which corresponds to fewer conditional dependencies and a larger Γ. On the other hand, a smaller number of partitions allows for a greater permissible perturbation between transition kernels. Our theorem states that, given that the maximal degree of dependencies between vertices is fixed, as the number of vertices (and hence the number of partitions) in the graphical model increases, the perturbed Gibbs sampler will converge to the original Gibbs sampler.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

5. Application to pseudo-marginal algorithms

The main goal of this section is to illustrate one way to design an algorithm for which our perturbation result applies. As the example shows, this is not terribly difficult, but does require some effort. Our main tool for designing an algorithm of the correct form is the pseudo-marginal algorithm of [3].

Note that this worked example is meant only to illustrate how to design the overall MCMC algorithm, *given* plug-in estimators for the likelihood. It is *not* meant to illustrate how to design good plug-in estimators. The latter task is well-studied in the context of perturbed MCMC chains, but the best estimators are often problem-specific and giving advice is beyond the scope of the current paper. See, e.g., [35] for practical mainstream advice, largely based around choosing good control variates.

The worked example in this section is based on the problem of inferring values sampled from a latent hidden Markov random field given observations at each of its vertices. We begin with the underlying hidden Markov random field. Throughout this section, denote by $G = (V, \Phi_1)$ a factor graph associated with state space $\Omega = [q]^V$ and Gibbs measure f with factorization structure (4). Next, we define the distributions for the observations at each vertex. Denote by $\{\mathrm{e}^{\ell(i,\cdot)}\}_{i\in[q]}$ a family of distributions on $[q]$, indexed by the same set $[q]$. For each vertex $v \in V$, fix $n_v \in \mathbb{N}$; this represents the 'number of observations' associated with vertex v.

This data lets us define the joint distribution of a sample from the hidden Markov random field and the associated observations. First, the latent sample from the hidden Markov random field: $\sigma \sim f$. Conditional on σ, sample independently the observations $Z_{ij} \overset{\text{ind}}{\sim} \mathrm{e}^{\ell(\sigma(i),\cdot)}$, $j \in [n_i]$. This completes the description of the data-generating process. The usual statistical problem, such as the image denoising problem [24], is to infer σ given the observed Z.

Given the independent observations $Z = \{z_{ij}\}_{i\in V, j\in[n_i]}$ with $Z_i = \{z_{ij}\}_{j\in[n_i]}$, we denote the log-likelihood of vertex i by $\phi_i(\sigma, Z_i) = \sum_{j=1}^{n_i} \ell(\sigma(i), z_{ij})$ and define $\Phi_2 = \{\phi_i\}$. Note that our model on the random variables (σ, Z) is then equivalent to a Markov random field on the larger vertex set $V \cup \{(i, j)\}_{i\in V, j\in[n_i]}$ and with augmented collection of factors $\Phi = \Phi_1 \cup \Phi_2$. The associated posterior distribution on the hidden σ given the observed Z is

$$\mu(\sigma \mid Z) \propto f(\sigma) \cdot \exp\left(\sum_{\phi_i\in\Phi_2} \phi_i(\sigma, Z_i)\right).$$

When many values of n_i are large, an exact evaluation of the likelihood is typically expensive. A natural approach is to rely on a moderately sized random subsample of the data in each step of the algorithm. Let $a := (a_i)_{i\in V} \in \mathbb{A}$ be a vector of auxiliary variables with each element a_i a subset of $[n_i]$; we think of this as corresponding to the subset of observations to include when estimating $\phi_i(\sigma, Z_i)$. Fix a probability distribution g on $\mathbb{A}$ of the form $g(a) = \prod_{i\in V} g_i(a_i)$. We next define an estimator $\hat{L}_\phi(\sigma, a; Z)$ for each element $\phi \in \Phi$:

$$\hat{L}_\phi(\sigma, a; Z) = \begin{cases} \phi(\sigma), & \phi \in \Phi_1, \\ (n_i/|a_i|) \sum_{j\in a_i} \ell(\sigma(i), z_{ij}), & \phi = \phi_i \in \Phi_2. \end{cases}$$

The associated approximate posterior distribution on the augmented space $\Omega \times \mathbb{A}$ is then

$$\hat{\nu}(\sigma, a \mid Z) \propto g(a) \exp\left(\sum_{\phi\in\Phi} \hat{L}_\phi(\sigma, a; Z)\right). \tag{9}$$

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Algorithm 2 Step of alternating Gibbs sampler.

Require: Initial state $(\sigma, a) \in \Omega \times \mathbb{A}$, distribution g on $\mathbb{A}$, and observations Z.

1: Sample variable index $v \sim \mathrm{Unif}(V)$.

2: For $i \in [q]$, calculate $s_i^a = \sum_{\phi \in A[v]} \hat{L}_\phi(\sigma_{v \mapsto i}, a; Z)$, and construct distribution p_1 over $[q]$ according to $p_1(i) \propto \exp(s_i^a)$.

3: Sample $k \sim p_1$, and define η as

$$\eta(w) = \begin{cases} k, & w = v, \\ \sigma(w), & w \neq v. \end{cases}$$

4: Sample variable index $v' \sim \mathrm{Unif}(V)$, and choose two observations $z_1 \sim \mathrm{Unif}(a_{v'})$ and $z_2 \sim \mathrm{Unif}(n_{v'} \setminus a_{v'})$.

5: For $i \in \{1, 2\}$, calculate $l_i^\eta = \hat{L}_{\phi_{v'}}\left(\eta, a_{v'}^{(i)}; Z\right)$, where $a_{v'}^{(1)} = a_{v'}$ and $a_{v'}^{(2)} = a_{v'} \cup \{z_2\} \setminus \{z_1\}$, and define distribution p_2 over $\{1, 2\}$ by $p_2(i) \propto g_{v'}\left(a_{v'}^{(i)}\right) \exp(l_i^\eta)$.

6: Draw $k \sim p_2$, and let $b = (b_w)$ where

$$b_w = \begin{cases} a_{v'}^{(k)}, & w = v', \\ a_w, & w \neq v'. \end{cases}$$

7: Return (η, b).

Algorithm 2 gives a step of an alternating Gibbs sampler that targets (9) in algorithmic form. The process can be succinctly described as follows. It first performs Gibbs sampler $K^{(a)}$ on σ targeting $\hat{\nu}(\sigma \mid a, Z)$, and then conducts Gibbs sampler $K^{(\eta)}$ on a by updating one observation in the subset of some vertex given updated configuration η from the previous step. This leads to the alternating Gibbs sampler, denoted by K, which can be represented as $K = K^{(a)} \cdot K^{(\eta)}$.

Notably, Algorithm 2 gives a perturbed Gibbs sampler in the sense of Definition 2 (see Remark 1(ii) for an explanation). Since g is a product measure on $\prod_{i \in V} \mathbb{A}_i$,

$$\begin{aligned} \nu(\sigma \mid Z) &\propto \int_{\mathbb{A}} g(a) \prod_{\phi \in \Phi} \exp(\hat{L}_\phi(\sigma, a; Z))\, \mathrm{d}a \\ &= f(\sigma) \cdot \int_{\mathbb{A}} g(a) \prod_{\phi_i \in \Phi_2} \exp(\hat{L}_{\phi_i}(\sigma, a; Z))\, \mathrm{d}a \\ &= f(\sigma) \cdot \prod_{\phi_i \in \Phi_2} \int_{\mathbb{A}_i} (\exp(\hat{L}_{\phi_i}(\sigma, a; Z)) + \log(g_i(a_i)))\, \mathrm{d}a_i, \end{aligned}$$

which is of the form (3). Moreover, it is obvious that f, $\mu(\sigma \mid Z)$, and $\nu(\sigma \mid Z)$ share the same dependence structure since they only differ in the product terms related to $\phi_i \in \Phi_2$, which do

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

not depend on other vertices except some single vertex. This ensures that if the prior f exhibits a factorization structure as (4), then the same holds true for $\mu(\sigma|Z)$ and $\nu(\sigma|Z)$.

Remark 6. (*The reason for such a small move on such a large augmented state space*.) In Algorithm 2, steps 4 to 6 involve swapping out a *single* element z_1 from a *single* set $a_{v'}$. It is natural to ask: why not make a larger Gibbs move, perhaps regenerating *all* elements from $a_{v'}$ or perhaps even *all* elements from *all* such subsets?

In principle, it is straightforward to write down an algorithm that would do this. However, even regenerating a single set $a_{v'}$ would require an enormous amount of work – the analogue of the distribution p_2 in step 5 would have support on a set of size $\binom{n_{v'}}{|a_{v'}|}$ rather than 2. Unless $|a_{v'}| = 1$, this grows much more quickly than the cost of simply keeping the full set of observations at all times.

Conversely, updating a single entry as we do in steps 4–6 can be very quick. In the extreme case that p_2 is always uniform, standard results from [23] on the mixing time of a Gibbs sampler on the hypercube indicate that $a_{v'}$ can be resampled in $O(n_{v'} \log(n_{v'}))$ steps, regardless of $a_{v'}$.

While this paper is not focused on providing 'near-optimal' perturbed algorithms, this choice allows us to provide an algorithm that has reasonable performance in the regime of interest.

5.1. Worked example: A tree-based Ising model

In the context of the above Gibbs samplers, we show that it is possible to apply Theorem 2 to a particular family of graphs and distributions f, ℓ. This amounts to verifying Assumption 1, since (as noted immediately after they are stated) Assumption 2 holds for $d = d_{L^2}$ and it is straightforward to force Assumptions 3 and 4 by choosing a sufficiently good approximate likelihood $\hat{L}$ and sufficiently large subsamples.

Unless otherwise noted, we keep the notation used throughout Section 5.

Example 4. (*Tree-based Potts model*.) Fix $n \in \mathbb{N}$ and $q \in \mathbb{N}_+$. Let $T_n = T = (V, E)$ be the usual depth-n binary tree with $\sum_{j=0}^{n} 2^j$ vertices, labelled as in Figure 1.

Each vertex takes a value in $[q]$. We take the prior distribution f on $\Omega = [q]^V$ as the Gibbs measure of the Potts model associated with fixed inverse temperature $\beta \in (0, \infty)$: each configuration $\sigma \in \Omega$ is assigned a prior probability

$$f(\sigma) \propto \exp\left(\beta \sum_{(v,w)\in E} \psi(|\sigma(v) - \sigma(w)|)\right),$$

where ψ is a monotonic decreasing function. Given a configuration $\sigma \in \Omega$, we have the set of observations $Z = \{Z_{ij} : i \in V, j \in [n_i]\}$, where each observation Z_{ij} takes a value in $[q]$ and

$$\mathbb{P}\{Z_{ij} = a \mid \sigma(i) = b\} \propto \exp(-g(|a - b|)),$$

where g is a monotonic increasing function.

The remainder of this section is an analysis of Example 4.

Remark 7. (*The reason for the subsample*.) We note that the fraction $\overline{ik} = (1/n_i) \sum_{j=1}^{n_i} \mathbf{1}_{Z_{ij}=k}$ of observations at node i with label $k \in [q]$ are sufficient statistics for this model. In this setting,

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

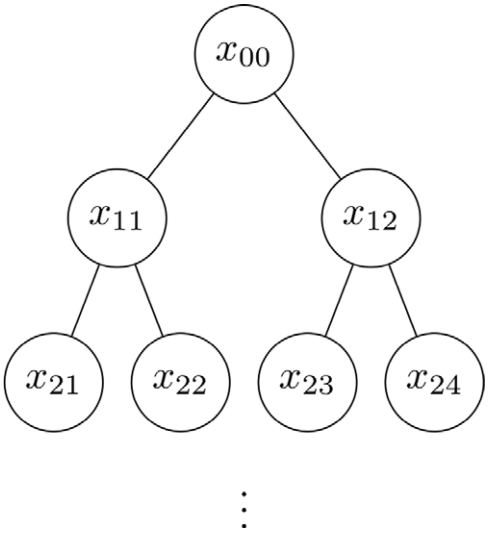

FIGURE 1. Tree structure of the Potts model.

it is natural to simply precompute the sufficient statistics, so it might be surprising to see an algorithm based on subsampling.

We give two replies. The minor answer is that this is primarily a pedagogical example – the proofs for simple exponential models are shorter and simpler than the proofs for more complicated ones.

The main answer is that this is one (of many) situations in which sufficient statistics may not be helpful. In the regime that the number n_i of observations per vertex is small compared to the number of allowed values q, the dimension of the sufficient statistics is larger than that of the original data vector. Consequently, storing and computing with these sufficient statistics does not simplify the process compared to using the original data. However, in this regime, our subsampling-based algorithm can still provide useful speedups. (In practice, we would use a sparse data structure, and this could give a very small speedup. This has essentially no impact on the conclusions of the remark.)

Recall that the log-likelihood of observations on node i is denoted by

$$\phi_i(\sigma, Z_i) = \sum_{j\in[n_i]} \ell(\sigma(i), z_{ij}) \propto -\sum_{j\in[n_i]} g(|\sigma(i) - z_{ij}|).$$

The posterior distribution on Ω is

$$\begin{aligned}\mu(\sigma \mid Z) &\propto f(\sigma)\cdot \exp\left(\sum_{i\in V}\phi_i(\sigma, Z_i)\right)\\ &\propto \exp\left(\beta\sum_{(v,w)\in E}\psi(|\sigma(v)-\sigma(w)|) - \sum_{i\in V}\sum_{j\in[n_i]} g(|\sigma(i)-z_{ij}|)\right),\end{aligned}$$

which is essentially the Gibbs measure of the generalized Potts model with some external field.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Due to the tree structure of the graph, it is easy to check that $\mu(\sigma|Z)$ factorizes as

$$\mu(\sigma \mid Z) = \mu|_{x_{00}}(\sigma) \cdot \prod_{i=1}^{n} \prod_{j=1}^{2^{i-1}} \prod_{k=0}^{1} \mu\big|_{x_{i,2j-k}}^{\sigma|_{x_{i-1,j}}}(\sigma), \tag{10}$$

where

$$\mu\big|_{x_{i,2j-k}}^{\sigma|_{x_{i-1,j}}}(\sigma_{x_{i,2j-k}\mapsto a}) \propto \exp\left(\beta\psi(|a-\sigma(x_{i-1,j})| - \sum_{j\in[n_{x_{i,2j-k}}]} g(|a-z_{x_{i,2j-k},j}|))\right).$$

Hence, in this example we take $\{S_i\}$ as an enumeration of the vertices of T and take $S_i \cup \Pi_i$ as the vertex S_i and its parent. Correspondingly, $\Gamma = \sum_{j=0}^{n} 2^j$.

Say that a sequence of vertices $\gamma_1, \ldots, \gamma_k \in V$ is a *path* if $(\gamma_i, \gamma_{i+1}) \in E$ for each $i \in [k-1]$. In view of the factorization structure of (10), we make the following observation.

Proposition 1. *Fix a path $\gamma_1, \gamma_2, \ldots, \gamma_k$ in the tree T. Sample σ, Z as in Example 4. Then, conditional on Z, the sequence $\sigma(\gamma_1), \sigma(\gamma_2), \ldots, \sigma(\gamma_k)$ is a (time-inhomogeneous) Markov chain.*

This lets us check the following.

Lemma 2. *Fix a path $\gamma_1, \gamma_2, \ldots, \gamma_k$ in the tree T. Sample Z as in Example 4. Then sample $\sigma^{(+)}$ (respectively $\sigma^{(-)}$) from the distribution in Example 4, conditional on both (i) the value Z and (ii) the value $\sigma^{(+)}(\gamma_1) = 1$ (respectively $\sigma^{(-)}(\gamma_1) = q$). If $0 \le \psi \le \alpha_1$ and $0 \le g \le \alpha_2$, then*

$$d_{\mathrm{TV}}(\mathcal{L}(\sigma^{(+)}(\gamma_j), \ldots, \sigma^{(+)}(\gamma_k)), \mathcal{L}(\sigma^{(-)}(\gamma_j), \ldots, \sigma^{(-)}(\gamma_k))) \le (1-\varepsilon_\beta)^{j-1},$$

where $\varepsilon_\beta = \mathrm{e}^{-(3\alpha_1\beta+\alpha_2 n)}$.

Proof. To simplify the notation, let $X_t = \sigma^{(+)}(\gamma_{t+1})$ and $Y_t = \sigma^{(-)}(\gamma_{t+1})$. We couple these according to the greedy coupling from Definition 1 and let $\tau = \min\{t : X_t = Y_t\}$ be the usual coupling time.

For $j \ge 1$, we can calculate

$$\begin{aligned}
\mathbb{P}(\tau = j \mid \tau > j-1) &= \mathbb{P}(X_j = Y_j \mid X_{j-1} \ne Y_{j-1}) \\
&= 1 - \mathbb{P}(X_j \ne Y_j \mid X_{j-1} \ne Y_{j-1}) \\
&\ge \min_{a,b\in[q]} \left(1 - d_{\mathrm{TV}}\left(\mu\big|_{\gamma_{j+1}}^{\sigma^{(+)}|_{\gamma_j}}(\sigma^{(+)}_{\gamma_j\to a} \mid Z), \mu\big|_{\gamma_{j+1}}^{\sigma^{(-)}|_{\gamma_j}}(\sigma^{(-)}_{\gamma_j\to b} \mid Z)\right)\right) \\
&\ge \min_{\sigma\in\Omega} \min_{a,b\in[q]} (1 - d_{\mathrm{TV}}(\mu(\sigma_{\gamma_j\mapsto a} \mid Z), \mu(\sigma_{\gamma_j\mapsto b} \mid Z))),
\end{aligned}$$

where the penultimate line follows by the optimal coupling between the distributions of times after j, and the last inequality comes from looking at the worst-case scenario regarding the values of the neighbors of γ_{j+1} excluding γ_j. Noting that each vertex in the tree has degree at most 3, $0 \le \psi \le \alpha_1$, and $0 \le g \le \alpha_2$, we have

$$\mathrm{e}^{-(3\alpha_1\beta+\alpha_2 n)} \le \frac{\mu(\sigma_{\gamma_j\mapsto a} \mid Z)}{\mu(\sigma_{\gamma_j\mapsto b} \mid Z)} \le \mathrm{e}^{3\alpha_1\beta+\alpha_2 n}.$$

Applying Lemma 3 from Appendix B,

$$\mathbb{P}(\tau = j \mid \tau > j-1) \ge 1 - \frac{1-\mathrm{e}^{-(3\alpha_1\beta+\alpha_2 n)}}{1+\mathrm{e}^{-(3\alpha_1\beta+\alpha_2 n)}} = \frac{2\mathrm{e}^{-(3\alpha_1\beta+\alpha_2 n)}}{1+\mathrm{e}^{-(3\alpha_1\beta+\alpha_2 n)}} \ge \varepsilon_\beta.$$

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

Continuing by induction on j, this yields $\mathbb{P}(\tau > t) \le (1-\varepsilon_\beta)^t$ for $t > 0$. In view of the property that $X_s = Y_s$ for all $s > j$ if $X_j = Y_j$, the claim then follows from the standard coupling inequality (see [23, Theorem 5.4]). □

Note that the boundary of $B_r(S_j \cup \Pi_j)$ in the tree T has at most 2^{r+1} vertices, which implies that there are at most 2^{r+1} paths starting from $S_j \cup \Pi_j$ to the boundary. Denote by τ_{global} the global coupling time of the boundary. In view of the union bound and Lemma 2,

$$\mathbb{P}\{\tau_{\text{global}} > r\} \le 2^{r+1} \cdot \mathbb{P}\{\tau > r\} \le 2[2(1-\varepsilon_\beta)]^r.$$

Thus, as long as $1-\varepsilon_\beta < \frac{1}{2}$, considering the optimal coupling again yields that the influence of the perturbation decays exponentially fast with the distance from the support of the perturbation. That is,

$$d_{\text{TV}}\left(\mu\big|_{S_j\cup\Pi_j}^{\sigma|_{B_r^{\text{c}}(S_j\cup\Pi_j)}}, \nu\big|_{S_j\cup\Pi_j}^{\eta|_{B_r^{\text{c}}(S_j\cup\Pi_j)}}\right) \le 2\text{e}^{-\log[2(1-\varepsilon_\beta)]^{-1}r}.$$

By virtue of the relationship (2), the decay-of-correlation Assumption 1 holds for $C_1 = 2$ and $m = \frac{1}{2}\log[2(1-\varepsilon_\beta)]^{-1}$.

Suppose that we apply an alternating Gibbs sampler as outlined in Algorithm 2 to sample from the posterior distribution. Let Q and K be the transition matrices of the classical Gibbs sampler and its approximate pseudo-marginal version, respectively. For simplicity, we denote by $K((\sigma, a), \cdot)|_{B_r(S_j\cup\Pi_j)}$ the transition kernel of σ restricted to the region $B_r(S_j \cup \Pi_j)$, and omit the superscript for the configuration values conditioned on $B_r^{\text{c}}(S_j \cup \Pi_j)$ in the perturbation bounds below.

Since the state space in our setting is finite, $Q|_{B_r(S_j\cup\Pi_j)}$ is always strongly ergodic with some positive constant ρ_j. Let $\rho = \max_{j\in[\Gamma]} \rho_j$. Then an immediate consequence of Theorem 1 is

$$\begin{aligned} &d_{L^2(\mu)}(\mu|_{S_j\cup\Pi_j}, \nu|_{S_j\cup\Pi_j}) \\ &\quad\le d_{L^2(\mu)}(\mu|_{B_r(S_j\cup\Pi_j)}, \nu|_{B_r(S_j\cup\Pi_j)}) \\ &\quad\lesssim \frac{1}{1-\rho} \sup_{(\sigma,a)\in\Omega\times\mathbb{A}} d(Q|_{B_r(S_j\cup\Pi_j)}(\sigma,\cdot), K((\sigma,a),\cdot)|_{B_r(S_j\cup\Pi_j)}) \\ &\quad\le \frac{1}{1-\rho} \sup_{j\in[\Gamma]} \sup_{(\sigma,a)\in\Omega\times\mathbb{A}} d(Q|_{B_r(S_j\cup\Pi_j)}(\sigma,\cdot), K((\sigma,a),\cdot)|_{B_r(S_j\cup\Pi_j)}). \end{aligned}$$

Note that we can replace $\rho \in (0, 1)$ with $\rho = 1 - 1/C_3 r^2$ where C_3 is a finite constant, which gives

$$\begin{aligned} &d_{L^2(\mu)}(\mu|_{S_j\cup\Pi_j}, \nu|_{S_j\cup\Pi_j}) \\ &\qquad\le Mr^2 \sup_{j\in[\Gamma]} \sup_{(\sigma,a)\in\Omega\times\mathbb{A}} d(Q|_{B_r(S_j\cup\Pi_j)}(\sigma,\cdot), K((\sigma,a),\cdot)|_{B_r(S_j\cup\Pi_j)}). \end{aligned}$$

Suppose that the observation subset $a_i \subset [n_i]$ of each vertex i is close enough to the entire set $[n_i]$ such that, for a positive constant $C_4 < \infty$,

$$\sup_{j\in[\Gamma]} \sup_{(\sigma,a)\in\Omega\times\mathbb{A}} d(Q|_{B_r(S_j\cup\Pi_j)}(\sigma,\cdot), K((\sigma,a),\cdot)|_{B_r(S_j\cup\Pi_j)}) \le \frac{C_4}{\sqrt{\Gamma}\log\Gamma}.$$

For any fixed $\delta > 0$, let $D = \sqrt{\delta/(2C_3C_4)}$. Then our Theorem 2 ensures that, for all $\Gamma \ge \max\{2, (4/\delta)^{1/(mD-0.5)}\}$, we have $d_{\text{H}}(\mu^{(\Gamma)}, \nu^{(\Gamma)}) \le \delta$.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

6. Conclusions and open questions

Under very generic conditions, it is well known that the stationary measures of two Markov chains P, Q are 'close' as long as the distance between P, Q is 'small' compared to some notion of the convergence rate of P or Q. The main conclusion of this paper is that much larger perturbations, of size roughly the *square root* of the convergence rate, are possible under certain conditions (which we think of as being related to a sort of approximate spatial independence). Furthermore, by looking at simple examples such as random walk on the hypercube, we can see that this 'square-rooting' is essentially optimal.

Our work leaves a number of open questions, and we describe the two that seem most important to us. The first is related to the following assumption: P, Q must have stationary measures that are both exactly Gibbs measures, with common factorization (4). This assumption was used in order to apply the exact summation formula (7), and feels (to us) rather unsatisfying. Our intuition, previous work on related problems [11], and some pen-and-paper calculations for specific examples all suggest that a weaker form of spatial mixing or approximate independence should be sufficient. One approach to tackling this problem would be to prove an approximate version of (7). Obtaining *some* generalization appears to be straightforward, but we don't have any conjectures about the *best* generalization.

Although we find this assumption unsatisfying, we should note that it is not terribly difficult to achieve in most applications. Perturbation bounds are most often used for either Markov chains that come from statistical algorithms or for those coming from physical models. In the former case, the algorithm designer can create Markov chains that satisfy this assumption. In both cases, the most natural perturbed Markov chains seem to satisfy this condition already.

Our other main question is: how can and should we force Assumption 4 to hold for Markov chains designed for statistical computing? Our paper does include an initial answer to this question – as we pointed out, it is typically easy to force Assumption 4 to hold when designing an algorithm. For example, in Example 4 we can force the assumption to hold by using a sufficiently high subsampling rate (we can prove this works by a standard application of a concentration inequality). Some early papers on perturbation bounds and subsampling in MCMC suggested that this sort of answer was 'good enough' – getting a perturbation bound for *some* subsampling rate was enough. However, as was pointed out in [5] and other sources, many naive subsampling algorithms have very bad performance if the subsampling rate is chosen on the basis of perturbation estimates. It is now widely known that naive subsampling algorithms typically do not give performance improvements if the subsampling rate is large enough for standard perturbation-style bounds to apply [5, 18, 31, 39]. While our results do allow for much larger perturbation errors, it is straightforward to check that a qualitatively similar conclusion often holds in the context of our paper as well. This is the main reason that we have not included a careful analysis of the quantitative tradeoff between the per-step cost of an algorithm and the minibatch size required by Assumption 4: we know that, for naive minibatch algorithms, the tradeoff would be quite poor.

Fortunately, all is not lost. While perturbation bounds often give qualitatively poor estimates for naive subsampling algorithms, the recent survey [39] describes many examples for which more sophisticated algorithms can be satisfactorily analyzed with perturbation methods. A natural next step would be to choose an algorithm of this sort and compare the subsampling rate required by standard perturbation estimates and those required by our work. Back-of-the-envelope calculations suggest that, when existing perturbation bounds provide useful estimates

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

for a reasonable subsampling rate, our bounds will provide useful estimates for a much lower (and thus computationally cheaper) subsampling rate.

Appendix A. Proof sketches for Example 2

We give short proof sketches for the four claims in Example 2, in order.

Claim 1. *The distances between kernels is* $d_{\mathrm{TV}}(Q, K) \approx C_n/n$.

Proof. Since $Q(i, \cdot), K(i, \cdot)$ are supported on at most three points, this is a straightforward calculation. For $i \neq n$, $d_{\mathrm{TV}}(Q(i, \cdot), K(i, \cdot)) = 2\varepsilon \approx C_n/n$. For $i = n$ a similar calculation holds. □

Claim 2. *The mixing times are* $\tau_{\mathrm{mix}}(Q) \approx n$, $\tau_{\mathrm{mix}}(K) \gtrsim n/C_n$.

Proof. Note that there is a monotone coupling of two walks $X_t, Y_t \sim Q$ with X_t started at $X_0 = n$ and Y_t started at any point Y_0. That is, it is possible to couple these chains so that $X_t \geq Y_t$ for all t.

Given such a monotone coupling, a standard coupling argument says that

$$\tau_{\mathrm{mix}} \leq 4\mathbb{E}[\min\{t : X_t = n\}].$$

Since Q is the standard simple random walk with drift, it is well known that $\mathbb{E}[\min\{t : X_t = n\}] \approx n$. This gives the upper bound $\tau_{\mathrm{mix}}(Q) \lesssim n$. To get the matching lower bound, first note that any walk $X_t \sim Q$ must satisfy $|X_t - X_{t+1}| \leq 1$ for all t. Thus, for $t < n/2$,

$$d_{\mathrm{TV}}(Q^t(n, \cdot), \mu) \geq \mu([0, n/3]).$$

Applying Claim 3 completes the proof that $\tau_{\mathrm{mix}}(Q) \gtrsim n$.

Next, we analyze the mixing time of K. The lower bound is very similar to the lower bound on Q. Let $Y_t \sim K$ be a chain started at $Y_0 = 0$ and let $\tau(n) = \min\{t : Y_t = n\}$. For $t < n/2$, we have

$$d_{\mathrm{TV}}(K^t(0, \cdot), \nu) \geq \mathbb{P}[\tau(n) > t] \cdot \nu([2n/3, n]) = (1 - \varepsilon)^t \cdot \nu([2n/3, n]).$$

Considering $t < (n \log 4)/C_n$ and applying Claim 4 to the second term completes the proof that $\tau_{\mathrm{mix}}(K) \gtrsim n/C_n$. □

Claim 3. *The stationary distribution* μ *of* Q *assigns probability* $\mu([0, n/3]) \to 1$ *as* $n \to \infty$.

Proof. We recognize that Q is obtained by taking a simple birth-and-death chain and inserting a holding probability of $\frac{1}{2}$ at each step. Thus, we can use the standard exact formula for the stationary measure of a birth-and-death chain to check this. □

Claim 4. *The stationary distribution* ν *of* K *assigns probability* $\nu([2n/3, n]) \to 1$ *as* $n \to \infty$.

Proof. Let X_t be a Markov chain drawn from K, and let $\tau_{k+1} = \min\{t > \tau_k : X_t = n\}$ represent the successive times at which the chain hits n. We assume $X_0 = n$ and set $\tau_0 = 0$.

Define $\delta_j = \tau_j - \tau_{j-1}$ as the times between successive visits to n. We note that $\delta_j, j \geq 1$, are independent and identically distributed. Let $S_k = |\{\tau_k \leq t < \tau_{k+1} : X_t < 2n/3\}|$ be the amount of time between τ_k and τ_{k+1} that the chain spends below $2n/3$. By the strong law of large numbers,

$$\nu([0, 2n/3)) = \lim_{K \to \infty} \frac{\sum_{k=1}^{K} S_k}{\sum_{k=1}^{K} \delta_k} = \frac{\lim_{K \to \infty} (1/K) \sum_{k=1}^{K} S_k}{\lim_{K \to \infty} (1/K) \sum_{k=1}^{K} \delta_k} = \frac{\mathbb{E}[S_1]}{\mathbb{E}[\delta_1]}$$

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

holds almost surely. On the one hand, since there is a probability ε of teleportation at each time step, we have $\mathbb{E}[S_1] \le 1/\varepsilon$. On the other hand, when hitting the state below $2n/3$, the chain must have spent at least $n/3$ time steps above $2n/3$. This leads to $\mathbb{E}[\delta_1] \ge n/3$. Putting this together, we have

$$\nu([0, 2n/3)) = \frac{\mathbb{E}[S_1]}{\mathbb{E}[\delta_1]} \le \frac{1/\varepsilon}{n/3} = \frac{3}{C_n},$$

which goes to 0 as long as $C_n \to \infty$. □

Appendix B. Short calculation on TV distances

Lemma 3. *Fix $0 < r < 1$ and let μ, ν be two distributions on $\{-1, +1\}$ satisfying*

$$r \le \frac{\mu(x)}{\nu(x)} \le r^{-1}$$

for $x \in \{-1, 1\}$. Then

$$d_{\mathrm{TV}}(\mu, \nu) \le \frac{1-r}{1+r}.$$

Proof. By symmetry, the worst-case pair μ, ν must satisfy $\mu(+1) = \nu(-1) = \delta$ and $\mu(-1) = \nu(+1) = 1 - \delta$ for some $0 < \delta < \frac{1}{2}$ satisfying $\delta/(1-\delta) = r$. Solving this for δ, we have $\delta = r/(1+r)$. Then the total variation distance is

$$d_{\mathrm{TV}}(\mu, \nu) = \nu(+1) - \mu(+1) = 1 - 2\delta = \frac{1-r}{1+r}.$$ □

Funding information

The first author was supported in part by the Innovation Program of Hunan Province (Grant No. CX20220259) and the China Scholarship Council Program (Project ID: 202206370087). The second author was supported by the National Natural Science Foundation of China (Grant No. 12471148) and the Natural Science Foundation of Hunan (Grant No. 2020JJ4674). The third author was supported by the NSERC Discovery Grant Program.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Adell, J. A. and Jodra, P. (2006). Exact Kolmogorov and total variation distances between some familiar discrete distributions. *J. Inequal. Appl.* **2006**, 1–8.
[2] Alquier, P., Friel, N., Everitt, R. and Boland, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statist. Comput.* **26**, 29–47.
[3] Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.
[4] Baker, J., Fearnhead, P., Fox, E. B. and Nemeth, C. (2019). Control variates for stochastic gradient MCMC. *Statist. Comput.* **29**, 599–615.
[5] Bardenet, R., Doucet, A. and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.* **18**, 1515–1557.
[6] Breyer, L., Roberts, G. O. and Rosenthal, J. S. (2001). A note on geometric ergodicity and floating-point roundoff error. *Statist. Prob. Lett.* **53**, 123–127.
[7] Bubley, R., Dyer, M., Greenhill, C. and Jerrum, M. (1999). On approximately counting colorings of small degree graphs. *SIAM J. Comput.* **29**, 387–400.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

[8] DASKALAKIS, C. AND PAN, Q. (2017). Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proc. Mach. Learn. Res.* **65**, 697–703.
[9] DASKALAKIS, C. AND PAN, Q. (2021). Sample-optimal and efficient learning of tree Ising models. In *Proc. 53rd Ann. ACM SIGACT Symp. Theory of Computing*, pp. 133–146.
[10] DING, J., SONG, J. AND SUN, R. (2023). A new correlation inequality for Ising models with external fields. *Prob. Theory Relat. Fields* **186**, 477–492.
[11] DURMUS, A. O. AND EBERLE, A. (2023). Asymptotic bias of inexact Markov chain Monte Carlo methods in high dimension. Preprint, arXiv:2108.00682.
[12] GAMARNIK, D. AND KATZ, D. (2012). Correlation decay and deterministic FPTAS for counting colorings of a graph. *J. Discrete Alg.* **12**, 29–47.
[13] GOLDBERG, L. A., JALSENIUS, M., MARTIN, R. AND PATERSON, M. (2006). Improved mixing bounds for the anti-ferromagnetic Potts model on $\mathbb{Z}^2$. *LMS J. Comput. Math.* **9**, 1–20.
[14] GOLDBERG, L. A., MARTIN, R. AND PATERSON, M. (2005). Strong spatial mixing with fewer colors for lattice graphs. *SIAM J. Comput.* **35**, 486–517.
[15] GUYON, X. AND HARDOUIN, C. (2002). Markov chain Markov field dynamics: Models and statistics. *Statistics* **36**, 339–363.
[16] JOHNDROW, J. AND MATTINGLY, J. (2017). Error bounds for approximations of Markov chains used in Bayesian sampling. Preprint, arXiv:1711.05382.
[17] JOHNDROW, J., ORENSTEIN, P. AND BHATTACHARYA, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *J. Mach. Learn. Res.* **21**, 1–61.
[18] JOHNDROW, J., PILLAI, N. AND SMITH, A. (2020). No free lunch for approximate MCMC. Preprint, arXiv:2010.12514.
[19] KATO, T. (1995). *A Perturbation Theory for Linear Operators*. Springer, Berlin.
[20] KOLLER, D. AND FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.
[21] KORATTIKARA, A., CHEN, Y. AND WELLING, M. (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *Proc. 31st Int. Conf. Machine Learning*, Vol. 32, pp. 181–189.
[22] KUNSCH, H. (1984). Time reversal and stationary Gibbs measures. *Stoch. Process. Appl.* **17**, 159–166.
[23] LEVIN, D. A., PERES, Y. AND WILMER, E. L. (2009). *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI.
[24] LI, S. Z. (2009). *Markov Random Field Modeling in Image Analysis*. Springer, New York.
[25] MARTINELLI, F. (1999). Lectures on Glauber dynamics for discrete spin models. In *Lectures on Probability Theory and Statistics* (Lect. Notes Math. 1717), eds J. Bertoin, F. Martinelli and Y. Peres. Springer, New York, pp. 93–191.
[26] MARTINELLI, F. AND OLIVIERI, E. (1994). Approach to equilibrium of Glauber dynamics in the one phase region: I. The attractive case. *Commun. Math. Phys.* **161**, 447–486.
[27] MARTINELLI, F., OLIVIERI, E. AND SCHONMANN, R. H. (1994). For 2-D lattice spin systems weak mixing implies strong mixing. *Commun. Math. Phys.* **165**, 33–47.
[28] MEDINA-AGUAYO, F. J., LEE, A. AND ROBERTS, G. O. (2016). Stability of noisy Metropolis–Hastings. *Statist. Comput.* **26**, 1187–1211.
[29] MEDINA-AGUAYO, F., RUDOLF, D. AND SCHWEIZER, N. (2020). Perturbation bounds for Monte Carlo within Metropolis via restricted approximations. *Stoch. Process. Appl.* **130**, 2200–2227.
[30] MITROPHANOV, A. Yu. (2005). Sensitivity and convergence of uniformly ergodic Markov chains. *J. Appl. Prob.* **42**, 1003–1014.
[31] NAGAPETYAN, T., DUNCAN, A. B., HASENCLEVER, L., VOLLMER, S. J., SZPRUCH, L. AND ZYGALAKIS, K. (2017). The true cost of stochastic gradient Langevin dynamics. Preprint, arXiv:1706.02692.
[32] NEGREA, J. AND ROSENTHAL, J. S. (2021). Approximations of geometrically ergodic reversible Markov chains. *Adv. Appl. Prob.* **53**, 981–1022.
[33] PARISI, G. (1981). Correlation functions and computer simulations. *Nucl. Phys. B* **180**, 378–384.
[34] QUIROZ, M., KOHN, R. Villani, M. and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *J. Amer. Statist. Assoc.* **114**, 831–843.
[35] QUIROZ, M., KOHN, R., VILLANI, M., TRAN, M.-N. AND DANG, K.-D. (2018). Subsampling MCMC – An introduction for the survey statistician. *Sankhya A* **80**, 33–69.
[36] RASTELLI, R., MAIRE, F. AND FRIEL, N. (2018). Computationally efficient inference for latent position network models. Preprint, arXiv:1804.02274.
[37] ROBERTS, G. AND ROSENTHAL, J. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Prob.* **2**, 13–25.
[38] RUDOLF, D. AND SCHWEIZER, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **24**, 2610–2639.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press

[39] RUDOLF, D., SMITH, A. AND QUIROZ, M. (2024). Perturbations of Markov chains. Preprint, arXiv:2404.10251.
[40] SALAS, J. AND SOKAL, A. D. (1997). Absence of phase transition for antiferromagnetic Potts models via the Dobrushin uniqueness theorem. *J. Statist. Phys.* **86**, 551–579.
[41] WEITZ, D. (2006). Counting independent sets up to the tree threshold. In *Proc. 38th Ann. ACM Symp. Theory of Computing*, pp. 140–149.
[42] WELLING, M. AND TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. 28th Int. Conf. Machine Learning*, ICML-11, pp. 681–688.

https://doi.org/10.1017/jpr.2024.102 Published online by Cambridge University Press