Approximations of Markov Chains and Bayesian Inference

James E. Johndrow
Duke University, Durham, NC USA
james.johndrow@duke.edu

Jonathan C. Mattingly
Duke University, Durham, NC USA
jonm@math.duke.edu

Sayan Mukherjee
Duke University, Durham, NC USA
sayan@stat.duke.edu

David Dunson
Duke University, Durham, NC USA
dunson@duke.edu

Wednesday 27th January, 2016

Abstract

The Markov Chain Monte Carlo method is the dominant paradigm for posterior computation in Bayesian analysis. It has long been common to control computation time by making approximations to the Markov transition kernel. Comparatively little attention has been paid to convergence and estimation error in these approximating Markov Chains. We propose a framework for assessing when to use approximations in MCMC algorithms, and how much error in the transition kernel should be tolerated to obtain optimal estimation performance with respect to a specified discrepancy measure and computational budget. The results require only ergodicity of the exact kernel and control of the kernel approximation accuracy. The theoretical framework is applied to approximations based on random subsets of data, low-rank approximations of Gaussian processes, and a novel approximating Markov chain for discrete mixture models.

1 Introduction

The fundamental entity in Bayesian statistics is the posterior distribution

$$
\Pi(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int \theta p(x \mid \theta)p(\theta)},
$$

the conditional distribution of the model parameters $\theta$ given the data $x$. In the models most commonly used in applications, the integral in the denominator of (1) is not available in closed form. A common approach is to construct an ergodic Markov chain with invariant distribution $\Pi(\theta \mid x)$, and then collect samples $\theta_1, \ldots, \theta_t$ from the chain. Statistical inference then relies on properties of the ergodic measure $\frac{1}{t} \sum_{k=0}^{t-1} \delta_{\theta_k}$, associated ergodic averages $\frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k)$ for functions $f$, and other quantities. This is referred to as Markov Chain Monte Carlo (Robert and Casella 2004, Gamerman and Lopes 2006) or MCMC.

We consider Markov chains that result from approximating the transition kernel $P(\theta, \cdot)$ by another kernel $P_\epsilon(\theta, \cdot)$ satisfying $\|P(\theta, \cdot) - P_\epsilon(\theta, \cdot)\|_{TV} < \epsilon$. The use of approximate kernels — often without showing such an error bound — is common practice in Bayesian analysis, and is usually
computationally motivated, i.e. obtaining samples from $P_{\epsilon}(\theta, \cdot)$ requires less computation than sampling from $P(\theta, \cdot)$. Our main contributions are as follows. We provide results showing bounds in total variation and expected $L_2$ estimation error for finite-time ergodic averages, under simple assumptions on the original chain and the approximating kernel. We also provide a general result on the computational advantage and approximation error tradeoff, providing an explicit criterion for the level of error to tolerate in the approximate transition kernel. We include an illustration to three approximate MCMC (aMCMC) algorithms in which we verify the approximation error assumption, and show practical performance.

While being arguably the dominant algorithm for Bayesian inference, MCMC is computationally demanding in high-dimensional settings, e.g. where either $p$ (the dimension of $\theta$) or $n$ (the number of observations) is large. MCMC is inherently serial, and often thousands of iterations are required to reach apparent stationarity. Moreover, high autocorrelation is common in applications, so that many samples must be taken from the invariant measure. Since each step $\theta_t \to \theta_{t+1}$ is commonly similar in computational complexity to a single iteration of an optimization algorithm for estimating the parameters of a similar model on the same data, MCMC can be an order of magnitude slower than alternatives or worse.

Despite alternatives – including variational Bayes approximations [Attias 1999; Wainwright and Jordan 2008], integrated nested Laplace approximations [Rue et al. 2009], and approximate Bayesian computation [Diggle and Gratton 1984; Marin et al. 2012], among others – MCMC remains highly utilized among practitioners. To more easily apply MCMC in complex settings, it is common to approximate $P(\theta, \cdot)$ with a kernel that is simpler or faster to sample from. One example is inference for Gaussian process models, bypassing $O(n^3)$ matrix inversion through approximations or reparametrization (Banerjee et al. 2008; Banerjee et al. 2013; Hughes and Haran 2013). Another prevalent example is the use of Laplace or Gaussian approximations to obtain conditional conjugacy. Guhaniyogi et al. 2014 proposes an algorithm that replaces some sampling steps with point estimates. Korattikara et al. 2013 approximate Metropolis-Hastings acceptance decisions using subsets of the data. It is also common to approximate intractable full conditionals by simpler distributions, with Bhattacharya and Dunson 2010 using a beta approximation, O’brien and Dunson 2004 replacing the logistic with a $t$ distribution, and Ritter and Tanner 1992 discretizing.

While approximating $P(\theta, \cdot)$ by $P_{\epsilon}(\theta, \cdot)$ is common, literature addressing convergence and approximation error of these algorithms is recent. Pillai and Smith 2014 present perhaps the most complete treatment to date, utilizing the theoretical foundation in Joulin et al. 2010 to show error bounds in the Wasserstein topology under fairly general conditions. Their results are applied to the algorithm in Korattikara et al. 2013 and similar subsampling based algorithms. Rudolf and Schweizer 2015 show results under effectively the same conditions, but use Lyapunov functions to eliminate exit probability terms from the bounds in Pillai and Smith 2014 that grow with $t$; the most recent version of Pillai and Smith 2014 uses a similar approach. Alquier et al. 2014 provide results in a similar context, but focusing on bounding the error between the ergodic measures of the approximate and exact chains. Earlier references show error bounds for perturbations of uniformly ergodic chains (Mitrophanov 2005) and geometrically ergodic chains (Ferré et al. 2013, Roberts et al. 1998; the latter focuses on perturbation resulting from numerical imprecision). Among these references, Pillai and Smith 2014 has a substantial focus on implications of the theoretical results for parameter estimation.

A number of authors propose algorithms for large $n$, moderate $p$ settings based on subsampling data. These include Bardenet et al. 2014, Korattikara et al. 2013, Chen and Gahramani 2015, Chen et al. 2014, Friel et al. 2015, Maire et al. 2015, Quiroz et al. 2014, Vollmer and Zygalakis...
2015 and Chen et al. 2015. Although the details differ in important ways, the basic concept is the use of data subsamples at each step of the Markov chain, analogous to stochastic gradient methods for optimization. Green et al. 2015, Zhu et al. 2014 and Bardenet et al. 2015 provide comprehensive reviews of this growing literature.

Our work differs from the precedents in several ways. Our results focus on estimation error and on interpretation of the error bounds, and the entire theoretical framework is constructed from a statistical perspective, i.e. with the view that samples from the Markov chain will form the basis of estimation via the empirical measure. All of our bounds improve with number of MCMC samples at the expected rate in $t$. We provide explicit criteria for determining the optimal level of approximation error given a speedup function quantifying the computational advantage of the approximation and a discrepancy measure quantifying the statistical performance of the approximate algorithm. This is perhaps the most unique aspect of our work, as precedents have not directly addressed the question of when an approximate chain is superior to an exact chain from the point of view of estimation, which is of critical relevance in applications. We further verify the usefulness of the results by applying them to three approximate samplers constructed from common MCMC algorithms for standard Bayesian models: one that employs random subsets of data, another for Gaussian process models using a low-rank covariance approximation, and a novel algorithm for mixture models for high-dimensional contingency tables. Thus, we consider a broader variety of approximate MCMC algorithms than precedents, which have focused almost exclusively on subsets of data for large $n$.

2 Ergodicity and Approximation Error

This section provides error bounds for statistical estimators constructed from approximate MCMC chains. In particular, we provide bounds in total variation and expected $L_2$ loss for posterior functions using samples from approximate chains. These bounds are then compared to similar bounds for exact chains to illustrate the relative computational efficiency of the approximate chain as a function of computational clock time. Because the bounds obtained for the exact kernel are tight, the comparison of the bounds leads naturally to a novel notion of computational optimality that we refer to as compminimax. Under this optimality criterion, approximate chains are optimal for surprisingly long computation times, though the advantage relative to the exact chain diminishes with computational time.

2.1 Approximate MCMC

Consider a family of likelihoods $p(x \mid \theta)$ parametrized by $\theta \in \Theta$. We assume that $X \sim p(x \mid \theta)$ takes values in a Polish space $\mathcal{X}$. In general, the spaces $\Theta$ of interest will be equipped with a dominating measure $m^*(\cdot)$. We are concerned with Markov chain Monte Carlo algorithms, which obtain samples from the posterior distribution in (1) by constructing an ergodic Markov chain with invariant distribution $\Pi(\theta \mid x)$. To obtain useful bounds on the error from use of an approximate kernel, we require the original Markov chain to satisfy some minimal convergence and mixing properties. One such condition is given in Assumption 2.1.

**Assumption 2.1** (Doeblin condition for exact chain). There exists a constant $0 < \alpha < 1$ such that

$$\sup_{\theta, \theta^* \in \Theta \times \Theta} \| \mathcal{P}(\theta, \cdot) - \mathcal{P}(\theta^*, \cdot) \|_{TV} < 1 - \alpha$$

(2)
where \( \| P - Q \|_{TV} \) is the total variation distance between probability measures \( P \) and \( Q \). When a kernel \( \mathcal{P} \) satisfies this condition we write \( d(\mathcal{P}) = \alpha \).

Assumption 2.1 implies uniform ergodicity. An immediate corollary of Assumption 2.1 is that
\[
\| \Pi - \nu P^t \|_{TV} \leq (1 - \alpha)^t \| \nu - \Pi \|_{TV}.
\]

If the transition kernel to which an approximation is being made does not correspond to a Markov chain with good convergence properties, it is difficult to obtain useful bounds on the approximation error. Although related results can be obtained under a weaker geometric ergodicity condition, the resulting bounds are more complex (e.g. Pillai and Smith [2014], Rudolf and Schweizer [2015]). The Doeblin condition has the advantage of leading to a simple characterization of the approximation accuracy and computational time tradeoff. In practice, the condition can be shown in a variety of cases involving compact state spaces. Compactness is not an overly restrictive assumption in practice, as choosing priors with bounded support is justified in most applications.

Consider a family of alternative transition kernels \( \mathcal{P}_{\epsilon} \), whose members approximate \( \mathcal{P} \). We will require the condition on \( \mathcal{P}_{\epsilon} \) given in Assumption 2.2.

**Assumption 2.2 (Conditions on the approximating kernel).** There exists a constant \( 0 < \epsilon < \alpha / 2 \) such that
\[
\sup_{\theta \in \Theta} \| \mathcal{P}(\theta, \cdot) - \mathcal{P}_{\epsilon}(\theta, \cdot) \|_{TV} < \epsilon
\]
for every \( \mathcal{P}_{\epsilon}(\theta, \cdot) \in \mathcal{P}_{\epsilon}(\mathcal{P}) \).

Although we state most results in terms of \( \mathcal{P}_{\epsilon}(\theta, \cdot) \), a generic element of \( \mathcal{P}_{\epsilon}(\mathcal{P}) \), they should be understood to hold for every member of \( \mathcal{P}_{\epsilon}(\mathcal{P}) \), and apply to chains constructed using an arbitrary sequence of members of \( \mathcal{P}_{\epsilon}(\mathcal{P}) \); this simplification is made for brevity and notational convenience. Assumption 2.2 can be weakened; for example, requiring that the approximation error bound hold only on a subset of the parameter space and some structure, such as a Foster-Lyapunov function, which ensures return to that subset. However, we prefer to keep the assumptions and resulting bounds simple and transparent.

### 2.2 Main results

The main results of this section relate the convergence properties of the original chain and the approximation error of the kernel \( \mathcal{P}_{\epsilon}(\theta, \cdot) \) to the approximation error for \( \Pi(\theta \mid x) \). First, define the following quantities
\[
\hat{\Pi}_P^t f = \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k), \quad \hat{\Pi}_{\epsilon P}^t f = \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_{\epsilon k}), \quad \Pi f = \int_\Theta f(\theta) \Pi(d\theta \mid x).
\]

We often omit the subscripts \( \mathcal{P} \) and \( \mathcal{P}_{\epsilon} \) in the above notation when considering transition kernels with a particular invariant measure \( \Pi \).

The focus is on the computational efficiency of statistical estimators constructed from sample paths of approximating kernels. To this end, consider any \( \mathcal{P} \) corresponding to a MCMC algorithm and a discrepancy measure \( D \) that quantifies the statistical value of sample paths of length \( t \) from \( \mathcal{P}_{\epsilon} \in \mathcal{P}_{\epsilon}(\mathcal{P}) \). Two natural choices for \( D \) that we consider here are
\[
D_{TV}(\Pi, \mathcal{P}_{\epsilon}, t) = \| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}_{\epsilon}^k \|_{TV}
\]
$$D_L^2(\Pi, \mathcal{P}_\epsilon, t) = \sup_{f: |f| < 1} \mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k^\epsilon) \right)^2 \right], \quad (5)$$

where in (5), the expectation is taken with respect to the distribution of the first $t$ steps of the Markov chain.

The potential advantage of aMCMC is that longer sample paths can be obtained in equal computational (wall clock) time. For any transition kernel $\mathcal{P}$, let $\tau_P(t)$ be the computational time required to obtain a sample path of length $t$. Define the speedup function $s(\epsilon)$ of a class of approximations $\mathcal{P}_\epsilon$ by

$$s(\epsilon) = \frac{\tau_P(t)}{\inf_{P_\epsilon \in \mathcal{P}_\epsilon} \tau_{P_\epsilon}(t)}, \quad (6)$$

which we assume is constant as a function of $t$. Since we focus on cases where aMCMC provides a computational advantage, it makes sense to restrict attention to speedup functions that are monotone nondecreasing in $\epsilon$ on the interval $0 < \epsilon < \alpha/2$, and satisfy $s(0) = 1$. For simplicity, we assume that every member of $\mathcal{P}_\epsilon$ having approximation error $\epsilon_0$ has speedup $s(\epsilon_0)$, so that in the sequel the infimum in the denominator of (6) is redundant. Without loss of generality, we also take $\tau_P(t) = t$ so that speedup can be interpreted as the number of samples obtained from $\mathcal{P}_\epsilon$ in the time required to obtain one sample from $\mathcal{P}$.

When $s(\epsilon)$ is not constant, there exists the potential that for finite computational budgets, some member of $\mathcal{P}_\epsilon$ will be superior to $\mathcal{P}$ with respect to a discrepancy measure $D$, because the longer sample paths obtained from $\mathcal{P}_\epsilon$ might more than compensate for any bias and difference in convergence/mixing properties. To make this rigorous, we define a notion of statistical optimality that we refer to as “computational minimax” (compminimax) approximation error due to its conceptual similarity to minimax estimators.

**Definition: Compminimax** Fix a computational budget $\tau_{\max}$ and a discrepancy measure $D$. An approximation error $\epsilon_c(\tau_{\max})$ is compminimax if

$$\epsilon_c(\tau_{\max}) = \arg\inf_{\epsilon < \alpha/2} \sup_{P_\epsilon \in \mathcal{P}_\epsilon} D(\Pi, \mathcal{P}_\epsilon, \max_t \{ t : \tau_{P_\epsilon}(t) < \tau_{\max} \}) \quad (7)$$

With the assumption that $\tau(t) = t$, we have $\max_t \{ t : \tau_{P_\epsilon}(t) < \tau_{\max} \} = [s(\epsilon)\tau_{\max}]$.

The definition of compminimax effectively gives a decision rule that assures optimal performance in the worst case scenario when the available information is the value of $\alpha$, $s(\epsilon)$, and a computational budget. Using only assumptions [2.1] and [2.2] we obtain the simple estimation error results in Theorem 2.3 which allow evaluation of minimax computational efficiency of aMCMC with respect to the discrepancy measures in [5] and [4].

**Theorem 2.3** (Estimation error for aMCMC ergodic averages). Suppose $\mathcal{P}$ satisfies Assumption [2.1] and $\mathcal{P}_\epsilon$ satisfies [2.2] and $f$ is bounded. Let $\theta_0, \theta_0^\epsilon \sim \nu$ for any probability measure $\nu$ on $(\Theta, \mathcal{F}_0)$, where $\mathcal{F}_0$ is the $\sigma$-field generated by $\Theta$. Then

$$\left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu^{P_k} \right\|_{TV} \leq \frac{1 - (1 - \alpha)^t}{\alpha t} ||\Pi - \nu||_{TV}, \quad (8)$$
\[ \left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}_k \right\|_{TV} \leq \frac{\epsilon}{\alpha} + \frac{(1 - (1 - \alpha)^t) \|\Pi_e - \nu\|_{TV}}{t \alpha}, \tag{9} \]

and

\[ \mathbb{E} \left[ (\Pi f - \hat{\Pi}^t f)^2 \right] \leq \frac{4 \|f\|_*^2 (1 - (1 - \alpha)^t) \|\Pi_e - \nu\|_{TV}}{\alpha t} + \frac{\|f\|_*^2 S(t, \alpha)}{\alpha}, \tag{10} \]

\[ \mathbb{E} \left[ (\Pi f - \hat{\Pi}^t f)^2 \right] \leq \frac{4 \|f\|_*^2 (1 - (1 - \alpha_e)^t) \|\Pi_e - \nu\|_{TV}}{t \alpha_e} + \frac{\|f\|_*^2 S(t, \alpha_e)}{\alpha_e^2} + \frac{8 \|f\|_*^2 \epsilon (1 - (1 - \alpha_e)^t)}{\alpha e} + \frac{4 \epsilon^2 \|f\|_*^2}{\alpha_e^2},\tag{11} \]

where \( \alpha_e = \alpha - 2 \epsilon, \|f\|_* = \inf_{c \in \mathbb{R}} \|f - c\|_\infty \), and

\[ S(t, \alpha) = \left( \frac{2}{\alpha t} + \frac{2}{\alpha t^2} + \frac{2(1 - \alpha)^{t+1}}{\alpha^2 t^2} - \frac{1}{t} - \frac{2}{\alpha^2 t^2} \right). \]

Expectations are taken with respect to the measure of the first \( t \) steps of the Markov chain.

Remark 2.1 characterizes sharpness of the bounds in Theorem 2.3.

Remark 2.1 (Sharpness of bounds). The bound in (9) is sharp; that is, for every \( \alpha \), there exists a transition kernel \( \mathcal{P} \) satisfying the Doeblin condition with \( d(\mathcal{P}) = \alpha \) for which equality holds in (9) for every \( t \). In addition, for every \( \alpha \), there exists a perturbation \( \mathcal{P}_e \) of a Markov kernel \( \mathcal{P} \) with \( d(\mathcal{P}) = \alpha \) that satisfy Assumption 2.2 for which

\[ \|\Pi - \Pi_e\|_{TV} = \frac{\epsilon}{\alpha}, \]

and a distinct perturbation \( \hat{\mathcal{P}}_e \) that achieves

\[ \left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \hat{\mathcal{P}}_k \right\|_{TV} = \frac{(1 - (1 - \alpha_e)^t) \|\Pi_e - \nu\|_{TV}}{t \alpha_e} \]

with \( \alpha_e = \alpha - 2 \epsilon \). Finally, under the additional technical conditions that the operator

\[ Fg(\theta) := \int g(\theta') \mathcal{P}(\theta, \theta') d\theta' \tag{12} \]

is self-adjoint and compact, then for every \( \alpha \) there exists a Markov chain such that

\[ \mathbb{E} \left[ (\Pi f - \hat{\Pi}^t f)^2 \right] = \frac{C^* \|f\|_*^2 (1 - (1 - \alpha)^t) \|\Pi_e - \nu\|_{TV}}{\alpha t} + \frac{\|f\|_*^2 S(t, \alpha)}{\alpha}, \]

is achieved for every \( t \) with \( C^* \geq 2 \). These conditions would hold, for example, for a reversible Markov chain on a finite state space.

Thus, the total variation bound for the exact kernel is sharp, and the \( L_2 \) bound for the exact kernel is achieved up to a factor of two in the first term in some special cases. Although the bounds for the approximate chain may not be sharp, the two components of the total variation bound, which
arose from applications of the triangle inequality, are both achieved. Thus, the bounds provide reasonable estimates of the supremum in (7) when \( D = D_{TV} \) or \( D_{L^2} \), which provides support for the quantitative comparisons that follow.

Both bounds for approximate kernels in Theorem 2.3 contain an asymptotic bias term for estimates obtained from \( P_\epsilon \) that is zero for estimates based on \( P \). The relative performance is governed by the magnitude of this bias, differences in the worst-case convergence rate, and the speedup. For \( D_{L^2} \), the bias is \( \frac{4\epsilon^2||f||^2}{\alpha^2} \), while for \( D_{TV} \), the asymptotic bias is \( \epsilon/\alpha \). The \( D_{L^2} \) bound has additional terms involving \( \epsilon \) that disappear in the infinite-time limit. Also, the convergence rate and worst-case autocorrelations for the approximation can be worse than that of the exact chain, since \( \alpha_\epsilon < \alpha \); however, this will not always be the case, and in some cases approximate algorithms will have better mixing properties. As a result, the results that follow will sometimes understate the benefits of using \( P_\epsilon \). Finally, note that similar results could be obtained for \( f \) with \( ||f||_\infty = \infty \) using either concentration and tail assumptions or moment assumptions, but the convergence rate in \( t \) and the scale of the bias would not change, so we retain the boundedness assumption throughout this section.

An important interpretation is that for relatively short path lengths, the bounds in (9) and (11) are dominated by terms related to the mixing/convergence properties of the chain, assuming \( \epsilon \) is small relative to \( \alpha \). For \( \epsilon \ll \alpha \), this is the term \( \frac{4\epsilon^2||f||^2}{\alpha^2} \), which is similar in magnitude to the bound in (6) when \( \epsilon \ll \alpha \). This term decays with \( t \), so that eventually the bias term \( \epsilon/\alpha \) becomes dominant. Thus, for relatively short path lengths, there should exist a range of \( \epsilon \) values for which aMCMC offers better performance in the compminimax sense. For longer path lengths, the values of \( \epsilon \) for which an advantage persists will tend to be small relative to \( \alpha \).

A similar analysis applies to (11), where the leading term is \( ||f||_\infty^2 S(t, \alpha_\epsilon) \). This is effectively a variance term that is bounded by the covariances for worst case functions. So for shorter path lengths, the variance term will dominate the overall estimation error and aMCMC will offer better performance. For longer path lengths, the bias term \( \frac{4\epsilon^2||f||^2}{\alpha^2} \) is more important. One factor that is clear from (11) but is not revealed by (9) is that aMCMC can still have a significant advantage even when a burn-in period is used and the first \( t_b \) samples are discarded. Although this results in the term \( ||\Pi - \nu||_{TV} \) being small – since we would now replace \( \nu \) by \( \nu P^{t_b} \) – the leading term \( ||f||_\infty^2 S(t, \alpha) \) is unaffected. In other words, burn-in cannot cure the problem of high autocorrelations in a chain with small \( \alpha \), resulting in the variance term being dominant for relatively long path lengths even after discarding a burn-in.

### 2.3 Analysis of compminimax approximation error

We now apply Theorem 2.3 to analysis of the compminimax approximation accuracy for different computational budgets. In light of Remark 2.1, the bounds in (9) and (11) provide reasonable estimates of the supremum in (7) when \( D = D_{TV} \) or \( D = D_{L^2} \). In the sequel, we focus on these discrepancy measures, and approximate \( \epsilon_\epsilon(\tau_{\text{max}}) \) for different values of \( \tau_{\text{max}} \) and functional forms for \( s(\epsilon) \) by minimizing the upper bounds in (9) and (11). Since the bounds are tight for \( \epsilon = 0 \), when the analysis suggests that \( \epsilon_\epsilon(\tau_{\text{max}}) > 0 \), it will always be the case that some \( \epsilon > 0 \) is compminimax; however, the optimal value may actually be larger than that computed by minimizing the (possibly loose) upper bounds for \( P_\epsilon \) in Theorem 2.3.

Empirical analysis of \( \epsilon_\epsilon \) requires choices of \( \alpha \) values and speedup functions \( s(\epsilon) \). We consider values between \( \alpha = 0.1 \) and \( \alpha = 10^{-4} \). These values are chosen by considering the upper bound on
the $\delta$-mixing time $t_\delta$ of the chain

$$t_\delta = \inf \{ t : \| \nu P^t - \Pi \|_{TV} < \delta \}$$

(13)

A corollary of Assumption 2.1 is that (13) is upper bounded by $\log(\delta)/\log(1-\alpha)$ when $\| \nu - \Pi \|_{TV} = 1$ and $\mathcal{P}$ satisfies the Doeblin condition. The corresponding worst-case $\delta$-mixing times for a few values of $\delta$ and the four values of $\alpha$ considered are given in Table 1. This range of $\alpha$ values gives mixing times between about 45 and 92,000 for $\delta \in (10^{-2}, 10^{-4})$, which reflects the empirical performance of many MCMC algorithms. In particular, a very rapidly mixing MCMC algorithm may reach apparent stationarity in only a few iterations. On the other hand, it is not uncommon that MCMC algorithms for complex models may require a burn-in period of tens of thousands of iterations.1

| $\alpha$ | $\delta = 0.01$ | $\delta = 0.001$ | $\delta = 0.0001$ |
|----------|-----------------|-----------------|------------------|
| 0.1      | 44              | 66              | 87               |
| 0.01     | 458             | 687             | 916              |
| 0.001    | 4,603           | 6,904           | 9,206            |
| 0.0001   | 46,049          | 69,074          | 92,099           |

We consider four functional forms for $s(\epsilon)$: logarithmic, linear, quadratic, and exponential. Constants are chosen such that $s(0) = 1$ and $s(\alpha/2) = 100$. Plots of the four functions for $\alpha = 10^{-4}$ are shown in Figure 1.

Figure 1: Speedup functions used in analysis of compminimax.

For each choice of $s(\epsilon)$ and a grid of values of $\tau_{\max} \in [1, 10^5]$, we approximate $\epsilon_c(\tau_{\max})$ by minimizing the upper bound in (9) with $t = s_\tau \tau_{\max}$, corresponding to our standing assumption that $\tau_{\max} = \tau$. Results are summarized in Figure 2. The top two panels show results for $D_{TV}$. When $\tau = D_{TV}$, it is clear that over a range of values of $\tau_{\max}$ substantially larger than the mixing times, the optimal value of $\epsilon$ is nonzero, regardless of the form of $s(\epsilon)$. As $\tau_{\max}$ increases, the (approximate) optimal value of $\epsilon$ decreases.

1We acknowledge that the criteria used to select burn-in times can result in burn-in periods that do not correspond to a mixing time, particularly when the posterior is strongly multimodal and the transition kernel has small conductance. However, comparing mixing times and burn-in periods still provides a useful heuristic, and in most cases violation of the criteria used to select a burn-in period is sufficient to guarantee that the chain has not mixed.
To a first approximation, the results can be understood in terms of the second derivative of the speedup function and the fact that $d(P_{\epsilon})$ can be as small as $\alpha - 2\epsilon$. When the second derivative is positive, the computational benefit of increases in $\epsilon$ is largest for large values of $\epsilon$, so for relatively small values of $\tau_{\text{max}}$, the optimal $\epsilon$ is large. However, large values of $\epsilon$ incur a relatively high cost in terms of worst-case convergence rates and autocorrelations, since $d(P_{\epsilon})$ can be as small as $\alpha - 2\epsilon$. Thus, $\epsilon_{\tau_{\text{max}}}$ goes to zero more rapidly when the second derivative of $s_{\epsilon}$ is positive compared to the case where it is zero or negative. In the latter case, small values of $\epsilon$ offer relatively large computational benefits relative to the bias introduced by using the approximation and magnitude of the difference between $d(P)$ and $d(P_{\epsilon})$ in the worst case, so nonzero values of $\epsilon$ are optimal even for large values of $\tau_{\text{max}}$. Notably, for all forms of $s(\epsilon)$ except the exponential, the optimal value of $\epsilon$ is nonzero for values of $\tau_{\text{max}}$ greater than the $\delta$-mixing times for all three values of $\delta$ considered. Finally, the observation that the optimal value of $\epsilon$ is zero for very small values of $\tau_{\text{max}}$, then increases rapidly to its maximum value, is a result of the difference between the bounds on $d(P)$ and $d(P_{\epsilon})$. For small $t$, this has a significant effect on the upper bound in (9). It is important to note that this is a worst-case bound, and that many aMCMC algorithms may actually have better convergence...
and mixing properties than the exact algorithm, so these results should be viewed as less favorable toward aMCMC than what will often be observed in practice.

The bottom panel in Figure 2 shows results for $D = D_{L^2}$. In this case, we assume the chain starts close to its stationary distribution by putting $||\Pi - \nu||_{TV} = 10^{-4}$ in (11). This corresponds to the situation in which a substantial number of burn-in samples are discarded. The choice of $D_{L^2}$ instead of $D_{TV}$ results in approximate values of $\epsilon_c(\tau_{\text{max}})$ that are larger at every value of $\tau_{\text{max}}$. Additionally, values of $\epsilon$ significantly larger than zero remain optimal well beyond the maximum value of $t$ considered in each case ($5,000$ when $\alpha = 0.1$ and $10^5$ when $\alpha = 10^{-4}$). This reflects the fact that high autocorrelations for worst-case functions make variance of MCMC ergodic averages the dominating factor in the $L^2$ error bounds even for relatively long sample paths, and these autocorrelations are unaffected by discarding a burn-in. Even when autocorrelations are relatively low, as in the case where $\alpha = 0.1$, nonzero $\epsilon$ is optimal for relatively large computational budgets when the speedup function is nonconvex.

### 3 Algorithm case studies

We apply the theoretical results of Section 2.2 to three approximate MCMC algorithms: for mixture models for contingency tables using approximations to Gibbs sampler full conditionals, for logistic regression based on subsets of data, and for Gaussian processes using low-rank approximations. For the first example, we verify both Assumption 2.1 and 2.2. For the other two examples, the focus is on verification of Assumption 2.2 which we show holds with high probability. An important conclusion is that it is usually possible to construct kernels that satisfy Assumption 2.2 with high probability, but doing so requires adapting the approximation to the current state of the Markov chain.

#### 3.1 Distributional approximations to full conditionals

In this example, we consider distributional approximations to full conditionals in Gibbs samplers, where both the approximation and the exact distribution rely on the complete data. The motivation for this type of approximation is that sampling from the approximating distribution may be much faster, either because the sufficient statistics are cheaper to calculate or the sampling algorithm itself scales better in the number of observations.

The specific example we consider is a mixture model for contingency tables and nonnegative matrices. Suppose we have $p$ categorical variables $y = (y_1, \ldots, y_p)$, which for simplicity each take $d$ possible values. We consider a variation on the model of Dunson and Xing [2009], replacing a stick-breaking prior with a Dirichlet:

\begin{align*}
Pr(y_1 = c_1, \ldots, y_p = c_p) &= \pi_{c_1, \ldots, c_p} = \sum_{h=1}^K \nu_h \prod_{j=1}^p \lambda_h^{(j)}, \\
\lambda_h^{(j)} &\sim \text{Dirichlet}(a_h^{(j)}), \quad \nu \sim \text{Dirichlet}(\alpha, \ldots, \alpha).
\end{align*}

In MCMC algorithms for discrete mixture models, it is common to employ data augmentation. Specifically, re-write the likelihood conditional on a latent class variable $z$ as

\[ Pr(y_1 = c_1, \ldots, y_p = c_p | z_i = h) = \prod_{j=1}^p \lambda_h^{(j)}, \quad Pr(z_i = h) = \nu_h. \]
When multiple observations with identical values of \(y_1, \ldots, y_p\) exist, the data are more compactly represented as a \(d^p\) contingency table, where \(n(c) = \sum_{i=1}^n \prod_{j=1}^p I_{[y_{ij} = c_j]}\) and \(c = (c_1, \ldots, c_p)\) is a multi-index identifying the cell of the table. Let \(C^+ = \{c : n(c) > 0\}\), and for each \(c \in C^+\), let \(Z(c)\) be a \(1 \times K\) vector with entries \(Z(c)_h = \sum_{i=1}^n I_{[y_i = c]} I_{[z_i = h]}\).

A Gibbs sampling algorithm for this model is given by

1. For each \(c \in C^+\), sample
   \[
   Z(c) | \nu, \lambda, Y \sim \text{Multinomial}(n(c), \tilde{v}_h) = \frac{\nu_h \prod_{j=1}^p \lambda^{(j)}_{hc_j}}{\sum_{h=1}^K \nu_h \prod_{j=1}^p \lambda^{(j)}_{hc_j}}.
   \]
   \[\text{(15)}\]

2. Sample \(\lambda^{(j)}_h\) for \(h = 1, \ldots, K\) and \(j = 1, \ldots, p\) from
   \[
   \lambda^{(j)}_h \sim \text{Dirichlet}\left( a^{(j)}_{h1} + \sum_{c|c_j=1} Z(c)_h, \ldots, a^{(j)}_{h1} + \sum_{c|c_j=d_j} Z(c)_h \right).
   \]

3. Sample \(\nu\) from
   \[
   \nu \sim \text{Dirichlet}\left( \alpha + \sum_c Z(c)_1, \ldots, \alpha + \sum_c Z(c)_K \right).
   \]

The dominating step is sampling of \(Z(c)\), which has computational complexity increasing linearly in \(n(c)\), so that each Gibbs iteration consumes at least order \(N\) operations just to sample the \(Z(c)\).

An approximating sampler that facilitates scaling to large \(N\) replaces the Multinomial sampling step for \(Z(c)\) with the following procedure:

1. Let \(H = \{h : n(c)_h \tilde{v}_h > n_{\min}\}\), \(K_H = |H|\), with \(n_{\min}\) a pre-specified threshold. For any set \(A \subset \{1, \ldots, K\}\) and \(K\)-vector \(v\), define \(v_A = \{v_h, h \in A\}\).

2. For entries \(h \in H\), sample from the Gaussian approximation to the multinomial,
   \[
   W \sim \text{Normal}\left( n(c)\tilde{v}_H, n(c)[\text{diag}(\tilde{v}_H) - \tilde{v}_H\tilde{v}_H^\top] \right),
   \]
   and set \(Z(c)_H\) equal to \(W\) with the elements rounded to the nearest integers.

3. If \(K_H < K\), sample \(Z(c)_{H^c}\) from Multinomial \(n(c) - \sum_{h' \in H} Z(c)_{h'}\).

4. Repeat steps (1)-(3) at every MCMC scan.

The Gaussian approximation can be sampled with computational complexity \(O(|C^+K^3|)\), resulting in substantial speedup when \(N\) is large. The other sampling steps are unchanged. The possible values of \(n_{\min}\) define a collection of approximate transition kernels \(\mathcal{P}_\epsilon = \{P_{c_0} : ||P_{c_0}(\theta_0,d\theta) - P_{c_0}(\theta_0,d\theta)||_{TV} < \epsilon\}\). We allow rounding to negative integers in step (2) for convenience in proving Lemma 3.1 guaranteeing Assumption 2.2. In practice, negative integers very rarely occur, and in such cases we set them equal to zero.
Lemma 3.1 (Uniform error bounds for normal approximations). Consider any approximate MCMC algorithm that replaces some full conditionals in Gibbs steps with the discretized Gaussian approximation to the multinomial described in step (2). For every $\epsilon \in (0,1)$ there exists $n_{\text{min}}$ such that

$$
\sup_{\theta \in \Theta} ||P(\theta, \cdot) - P_\epsilon(\theta, \cdot)||_{TV} < \epsilon,
$$

where $P_\epsilon(\theta, \cdot)$ corresponds to the algorithm with threshold $n_{\text{min}}$. Moreover, there exist constants $C(K_H - 1)$ depending only on $K_H - 1$ for which

$$
n(\epsilon) > \frac{C(K_H(\epsilon) - 1)^2}{|C^+|^2} \left( \sum_{h \in H(\epsilon)} \frac{(1 - \hat{\nu}_h)(1 - 2\hat{\nu}_h + 2\hat{\nu}_h^2)(1 + P_h/\hat{\nu}_K)}{\sqrt{\hat{\nu}_h(1 - \hat{\nu}_h)}} \right)^2,
$$

for every $\epsilon \in C^+$ implies $||P(\theta, \cdot) - P_\epsilon(\theta, \cdot)||_{TV} < \epsilon$, where $P_h = \sum_{h':\hat{\nu}_{h'} > \hat{\nu}_h} \hat{\nu}_h$ and we have explicitly indicated the dependence of $H$ on $\epsilon$ through the notation $H(\epsilon)$.

Lemma 3.1 shows that by adapting the kernel to the current state, it is possible to construct an algorithm that satisfies Assumption 2.2. Adaptivity enters because the threshold $n_{\text{min}}$ is applied to $n(\epsilon)\hat{\nu}_h$, which depends on the current state. We include the condition in (16) to illustrate an interesting connection with condition numbers of covariance matrices of the parameters that will be revisited in later examples. Suppose that $H = \{1, \ldots, K\}$ and let $\hat{\nu}_{\text{max}}, \hat{\nu}_{\text{min}}$ be the largest and smallest entries of $\hat{\nu}$, respectively. The quantity on the right of (16) will be large when $\frac{\hat{\nu}_{\text{max}}}{\hat{\nu}_{\text{min}}}$ is large. The Multinomial($n(\epsilon), \hat{\nu}$) distribution has covariance $n(\epsilon)\text{(diag}(\hat{\nu}) - \hat{\nu}\hat{\nu}^\top)$. Applying inequalities from Golub [1973], the smallest eigenvalue of this matrix is bounded below by $n(\epsilon)\hat{\nu}_{\text{min}}$, and the largest is bounded above by $n(\epsilon)\left(\hat{\nu}_{\text{max}} + \sum_{h=1}^K \hat{\nu}_h^2\right)$. Thus the condition number is at least $\frac{\hat{\nu}_{\text{max}}}{\hat{\nu}_{\text{min}}}$, so the quantity on the right side of (16) will always be large when the condition number of the covariance is large, meaning that we require a larger sample in cell $\epsilon$ for an accurate approximation. In fact, one way to think about the adaptive approximation is that by excluding categories $h$ with small $\hat{\nu}_h$ – thereby resulting in larger $\hat{\nu}_{\text{min}}$ – the covariance matrix in the Gaussian approximation is better conditioned, ensuring a more accurate approximation. Analogous conclusions are reached for the other two example algorithms in the sequel.

Lemma 3.1 also allows for analysis of the order of the speedup function $s(\epsilon)$ for this algorithm. (16) shows that $\epsilon = \mathcal{O}(n^{-1/2})$, so we need to increase the threshold at the rate $\sqrt{n}$ for linear decreases in $\epsilon$. This requires substituting order $n$ computation for order $K^3$ computation. To a first approximation, this indicates that the speedup function is roughly $s(\epsilon) = \sqrt{\epsilon}$. For a concave speedup function, one expects relatively small values of $\epsilon$ to be optimal, a conclusion we will revisit in Section 4.

Remark 3.1 shows that the above exact Gibbs sampler for model (14a)-(14b) satisfies Assumption 2.1. In this example, the state space is compact, and the latent variable $Z$ is discrete, which makes verification of Assumption 2.1 fairly straightforward.

Remark 3.1 (Mixture model conditions). The Gibbs sampling algorithm described above for the model in (14a)-(14b) satisfies Assumption 2.1.

3.2 Approximations based on subsets of data

A variety of aMCMC algorithms that utilize subsets of the data have been proposed. An example of this class is provided in Korattikara et al. [2013], where $V \subset \{1, \ldots, N\}$ is a random subset of
indices adaptively chosen to obtain a pre-specified type I error in a Metropolis-Hastings acceptance decision. Specifically, they cast the MH sampling procedure as proposal of a candidate $\theta^*$, sampling $u \sim \text{Uniform}(0, 1)$, calculation of

$$\mu_0 = \frac{1}{N} \log \left[ \frac{p(\theta_t)q(\theta^* | \theta_t)}{p(\theta^*)q(\theta^* | \theta^*)} \right], \quad \mu = \frac{1}{N} \sum_{i=1}^{N} [l(x_i | \theta^*) - l(x_i | \theta_t)],$$

and acceptance of $\theta^*$ if $\mu > \mu_0$. Here $l(x | \theta) = \log p(x | \theta)$ is the log-likelihood. They replace this step by testing of the hypothesis $H_0 : \mu > \mu_0$ at level $\epsilon$ by utilizing a random sample $x_V$ of the data. This procedure substitutes an estimate of the mean of the log likelihood ratio $\bar{l}_V = \frac{1}{|V|} \sum_{i \in V} [l(x_i | \theta^*) - l(x_i | \theta_t)]$ for $\mu$. They give a uniform bound on $||\mathcal{P}_t(\theta, \cdot) - \mathcal{P}(\theta, \cdot)||_{TV}$ under the condition that the joint distribution of $\bar{l}_V : V \in \mathcal{V}$ for any collection of subsets $\mathcal{V}$ is multivariate normal. This algorithm has been referred to as the “austerity” framework.

The austerity algorithm can be viewed as a use of the likelihood approximation

$$L_\epsilon(x | \theta) = \left( \prod_{i \in V} L(x_i | \theta) \right)^{N/|V|},$$

adaptively choosing $|V|$ at each step to achieve a pre-specified approximation error. We consider an alternative use of subsampling within a Gibbs sampler for logistic regression; rather than using the likelihood approximation directly, we use a subsampling-based approximation of a covariance matrix in a Gibbs sampling step. Computing the covariance is the major computational bottleneck, so we achieve similar computational speedup to use of the full likelihood approximation, and are able to obtain theoretical guarantees on approximation error under weaker conditions on the data.

### 3.3 Model and computational algorithm

Consider a logistic regression model with a Gaussian prior on regression coefficients

$$y_i \sim \text{Bernoulli} \left( \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right), \quad \beta \sim \text{Normal}(\mu, \Sigma).$$

Polson et al. [2013] show that the likelihood in this model satisfies

$$L(y_i | \beta) = \left( \frac{e^{y_i x_i \beta}}{1 + e^{x_i \beta}} \right)^{\alpha} \exp(\kappa_i x_i \beta) \int_0^\infty \exp \left\{ -\omega_i (x_i \beta)^2 / 2 \right\} p(\omega_i | 1, 0),$$

where $\kappa_i = y_i - 1/2$ and $p(\omega_i | 1, 0)$ is the density of a Polya-Gamma random variable with parameters 1 and 0, which we represent as PG(1, 0). This results in the Gibbs sampler:

$$\omega_i | \beta \sim \text{PG}(1, x_i \beta) \quad \text{(19a)}$$

$$\beta | y, \omega \sim \text{Normal}(m_N, S_N) \quad \text{(19b)}$$

where $S_N = (X' \Omega X + B^{-1})^{-1}$, $m_N = S_N (X' \kappa + B^{-1} 1)$, and $\Omega = \text{diag}(\omega_1, \ldots, \omega_N)$. 


3.3.1 Approximation

When \( N \) is large and \( p \) – the dimension of \( \beta \) – is moderate, the main computational bottleneck is calculating \( X^\top \Omega X \). This step has computational complexity \( \mathcal{O}(N^2p^3) \). An approximating Markov chain that uses the likelihood approximation described above will reduce the computational complexity of each step to \( \mathcal{O}(|V|^2p^3) \), which can result in large computational speedup when \(|V| \ll N\). The downside is that, assuming a random subset \( V \) is chosen at each iteration, the estimated posterior variance of \( \beta \) will be inflated.

We analyze an approximating Gibbs sampler with the update rule

\[
V | \beta \sim \text{Subset}(\{V, \{1, \ldots, N\}\}),
\]

\[
\omega_i | \beta, V \sim \text{PG}(1, x_i \beta), \quad i \in V;
\]

\[
\beta | y, \omega, V \sim \text{Normal}(S_V X' \kappa, S_V),
\]

where \( S_V = \left( \frac{N}{|V|} X_V' \Omega V X_V + B^{-1} \right)^{-1} \) uses a subsample-based approximation to \( X^\top \Omega X \) and \(|V|\) may depend on \( \beta \). Choi and Hobert [2013] showed that the algorithm in (19a)-(19b) is uniformly ergodic. Theorem 3.1 shows that if \(|V|\) is chosen adaptively depending on \( \beta \), Assumption 2.2 is satisfied with high probability at any step of the chain.

**Theorem 3.1** (Error for random subset approximations). Suppose the rows of \( X \) are iid realizations with a log-concave density that is symmetric about the origin. Let \( b = 0, B = \eta I_p \) for \( \eta > 0 \). Let \( \mathcal{P}(\theta, \cdot) \) be the transition kernel of Gibbs sampler (19a) - (19b), and \( \mathcal{P}_\varepsilon(\theta, \cdot) \) be the transition kernel of sampler (20a)-(20c). Then, for every \( \epsilon > 0 \), there exists a kernel \( \mathcal{P}_\varepsilon(\theta, \cdot) \) that sets \(|V| \leq N\) as a function of \( \beta \), for which

\[
\sup_{\theta \in \Theta} ||\mathcal{P}_\varepsilon(\theta, \cdot) - \mathcal{P}(\theta, \cdot)||_{TV} \leq \epsilon
\]

with probability \((1 - q)^2\), where \( q \) decreases exponentially in \(|V|\).

The following remark provides insight into the achievable rates in \(|V|\) of the probabilities \( q \) and the approximation error \( \epsilon \).

**Remark 3.2** (Rates in Theorem 3.1). Let \( \omega \sim \text{PG}(1, x_\beta) \) and \( x = (x_1, \ldots, x_p) \) denote realizations for a random subject, and

\[
\Sigma = \text{cov}(\omega^{1/2} x | \beta), \quad \Sigma_N = \frac{1}{N} X^\top \Omega X, \quad \Sigma_V = \frac{1}{|V|} X_V' \Omega V X_V.
\]

Choosing \(|V| \geq pCM^4\delta^{-2}\log^2(2M^2\delta^{-2})\), \(||\Sigma_N - \Sigma_V|| < \delta ||\Sigma||\) with probability \((1 - e^{-cM\sqrt{p}})^2\), where \( C \) and \( c \) are absolute constants. Subsets of size \(|V| = \mathcal{O}(\sqrt{N})\) result in \( M \approx \frac{N^{1/8}}{\log^2(2N^{1/4})}\), achieving \( q \) slightly larger than \( e^{-N^{1/8}}\). The required value of \( \delta \) to achieve \(||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\varepsilon(\theta, \cdot)||_{TV} < \epsilon\) is

\[
\delta = \frac{2\sqrt{2} ep^{-1/2}(\lambda_{\min}(\beta)/2)^2}{(\lambda_{\max}(\beta) + \lambda_{\min}(\beta)/2)^{3/2}} \left\{ \frac{e^2(\lambda_{\min}(\beta)/2)}{p(\lambda_{\max}(\beta) + \lambda_{\min}(\beta)/2)} \right\}^{1/4} \wedge \frac{ep^{-1/2}\lambda_{\min}(\beta)}{\lambda_{\min}(\beta) + \lambda_{\max}(\beta)}.
\]

where \( \lambda_{\min}(\beta) \) and \( \lambda_{\max}(\beta) \) are the smallest and largest eigenvalues of \( \Sigma \), respectively.
Remark 3.2 shows that the required subset size to achieve approximation error $\epsilon$ with probability $(1-q)^2$ depends on the reciprocal condition number of $\Sigma = \text{cov}(\omega^{1/2}x | \beta)$. This is similar to the adaptivity result in Remark 3.1 and again indicates that algorithms achieving uniform bounds must be adaptive to the current state. Although the exact rates include unknown constants $C$ and $c$, the result is still useful in constructing a practical algorithm. The condition number of $\Sigma$ can be approximated by computing $X_i^\top \Omega X_v$, which is required by the subsetting algorithm anyway. An adaptive subsample size can be chosen by starting with a subsample of fixed size, computing the condition number of $X_i^\top \Omega X_v$, and increasing the subset size when this value is large relative to, say, its ergodic average.

Remark 3.2 also characterizes the speedup function for this algorithm. The computational cost of increasing $|V|$ is $O(|V|^3)$, while to a second-order approximation $\delta = O(|V|^{-1/2} \log V)$. Moreover, (21) shows that $C_\beta \epsilon^2 < \delta < C_\beta \epsilon$, so $\epsilon = O(|V|^{-1/2} + \log V)$ as well. Therefore, the speedup function is approximately $\epsilon^{q/2}$ up to log factors. This implies good performance for relatively large values of $\epsilon$ when the computational budget is small, with advantages possibly decaying rapidly as computational budget increases. This behavior may not be seen empirically for very small subsamples or for poorly conditioned $X$, since the constants can be very large in such cases.

### 3.4 Low-rank approximations to Gaussian processes

Exact MCMC algorithms involving Gaussian processes scale as $O(n^3)$, leading to numerous proposals for approximations. Prominent examples include the predictive process (Banerjee et al. 2008) and subset of regressors (Smola and Bartlett 2001), which both employ low-rank approximations to the Gaussian process covariance matrix.

#### 3.4.1 Model

Consider the nonparametric regression model

$$y_i = f(x_i) + \eta_i, \quad \eta_i \sim \text{Normal}(0, \sigma_i^2 I_n), \quad i = 1, \ldots, n,$$

(22)

where $y_i$ are responses, $x_i \in X$ are $p \times 1$ covariate vectors, and $f$ is an unknown function. A typical Bayesian approach assigns a Gaussian process prior to $f$, $f \sim \text{GP}(\mu(\beta), c(\gamma))$, with $\mu(\cdot; \beta)$ a mean function with parameter $\beta$ and $c(\cdot; \gamma)$ a covariance function parametrized by $\gamma$, so that for $x_1, x_2 \in X$, $\mathbb{E}[f(x_1)] = \mu(x_1; \beta)$ and $\text{cov}(f(x_1), f(x_2)) = c(x_1, x_2; \gamma)$. Here we will assume $\mu(x; \beta) = 0$ for all $x \in X$, so that the model parameters consist of $\theta = (\sigma^2, \gamma)$. Although we focus on model (22), our analysis applies to general settings involving Gaussian processes (e.g., for spatial data and latent processes).

The covariance kernel $c(x_1, x_2; \gamma)$ is positive definite, so that the $n \times n$ covariance matrix $S$ given by $S_{ij} = c(x_i, x_j; \gamma)$ is full rank. However, as noted by Banerjee et al. 2013, in many cases when $n$ is large, the matrix $S$ is poorly conditioned and nearly low-rank. This motivates low-rank approximations of $S$. As an example, consider the squared exponential kernel $c(x_1, x_2; \gamma) = \tau^2 \exp(-\phi \|x_1 - x_2\|^2)$, with $\gamma = (\tau^2, \phi)$ consisting of a decay parameter $\phi$ and scale $\tau^2$. In this case we write $S = \tau^2 \Sigma$, where $\Sigma_{ij} = \exp(-\phi \|x_i - x_j\|^2)$, and we have $\theta = (\sigma^2, \tau^2, \phi)$. We adopt the common prior structure

$$\phi \sim \text{DiscUnif}(\phi_1, \ldots, \phi_d), \quad \tau^{-2} \sim \text{Gamma}(a_\tau, b_\tau), \quad \sigma^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma).$$

(23)
We consider two MCMC algorithms for posterior computation: the data augmentation sampler in Banerjee et al. [2013] and the marginal sampler in e.g. Finley et al. [2009]. The data augmentation sampler iterates

1. Sample \( f \sim N(\mu_f, \Sigma_f) \) where \( \mu_f = \Sigma_f y \) and \( \Sigma_f = (\tau^2 \Sigma + \sigma^2 I)^{-1} \);
2. Sample \( \sigma^{-2} \sim \text{Gamma}(a_\sigma + n/2, b_\sigma + (y - f)'(y - f)/2) \);
3. Sample \( \tau^{-2} \sim \text{Gamma}(a_\tau + n/2, b_\tau + f'\Sigma^{-1}f'/2) \); and
4. Set \( p_l \propto \det (\tau^2 \Sigma^{(l)})^{-1/2} \exp (-f'(\tau^2 \Sigma^{(l)})^{-1} f) \), where \( \Sigma^{(l)}_{ij} = \exp(-\phi_l \|x_i - x_j\|^2) \), and sample \( \phi \sim \text{Disc} \{\{\phi_1, \ldots, \phi_d\}, (p_1, \ldots, p_d)\} \).

The marginal sampler retains the “griddy Gibbs” step for \( \phi \) but does not condition on the sampled value of the process

1. Sample \( \sigma^{-2}|y, \tau^2, \phi \) using a Metropolis-Hastings step with random walk on \( \log(\sigma^2) \) as a proposal.
2. Sample \( \tau^{-2}|y, \phi, \sigma^2 \) using a Metropolis-Hastings step with random walk on \( \log(\tau^2) \) as a proposal.
3. Set \( p_l \propto \det (\tau^2 \Sigma^{(l)} + \sigma^2 I_n)^{-1/2} \exp (-y'(\tau^2 \Sigma^{(l)} + \sigma^2 I_n)^{-1} y) \), where \( \Sigma^{(l)} \) is defined as in the data augmentation sampler, and sample \( \phi \sim \text{Disc} \{\{\phi_1, \ldots, \phi_d\}, (p_1, \ldots, p_d)\} \).

For the former sampler, the quadratic forms \( f'(\tau^2 \Sigma)^{-1} f \) and \( f' \Sigma^{-1} f \) both appear, while the latter marginalizes out \( f \) to instead use \( y'(\tau^2 \Sigma + \sigma^2 I)^{-1} y \). For the low-rank approximation to be accurate, all but \( r \) of the eigenvalues of \( \Sigma \) should be bounded above by \( \kappa \) for some small \( \kappa > 0 \) and \( r \ll n \). Under these conditions, the condition number of \( \Sigma \) will be huge, and calculation of \( f'(\tau^2 \Sigma)^{-1} f \) and \( f' \Sigma^{-1} f \) numerically unstable. In contrast, as long as \( \sigma^2 \) is not close to zero, \( (\tau^2 \Sigma + \sigma^2 I)^{-1} \) is well-conditioned. Hence, we focus on the marginal sampler.

### 3.4.2 Approximate MCMC for Gaussian processes

We replace \( \Sigma \) with a low-rank approximation \( \Sigma_e \approx \Sigma \) to construct a transition kernel \( \mathcal{P}_e (\theta, \cdot) \) for an approximate MCMC algorithm. We focus on approximations of the form

\[
\Sigma \approx U_r \Lambda_r U_r' = \Sigma_e, \tag{24}
\]

where \( U_r \) is orthonormal, and \( \Lambda_r \) is nonnegative and diagonal.

All of the steps of the marginal sampler contain the quadratic form \( y'(\tau^2 \Sigma + \sigma^2 I)^{-1} y \), and the process \( f \) is sampled from

\[
p(f \mid y, \theta) \sim N(\Psi y, \Psi), \quad \Psi = (\tau^2 \Sigma + \sigma^2 I)^{-1}
\]

for the purposes of obtaining interval estimates. The approximation replaces this with

\[
p_e(f \mid y, \theta) \sim N(\Psi_e y, \Psi_e), \quad \Psi_e = (\tau^2 \Sigma_e + \sigma^2 I)^{-1}.
\]

For algorithms in this class, we obtain the result in Theorem 3.2.
Theorem 3.2 (Gaussian process approximation error bounds). Suppose data are generated according to (22), with \( c(x_1, x_2; \gamma) = \tau^2 \exp(-\phi |x_1 - x_2|^2) \), and \( \Sigma_c = U_c \Lambda_c U_c' \) for \( U_c \) a \( n \times r \) matrix and \( \Lambda_c \) an \( r \times r \) matrix with \( r < n \). For every \( \epsilon > 0 \) there exists \( \delta > 0 \), which depends on \( \theta \), such that if \( \| \Sigma - \Sigma_c \|_F < \delta \),

\[
\| p(f \mid y, \theta) - p_c(f \mid y, \theta) \|_{TV} < \epsilon.
\]  

(25)

If additionally \( \Sigma_c \) is a partial rank-\( r \) eigendecomposition of \( \Sigma \) and a joint Metropolis-Hastings step is used for \((\sigma^2, \tau^2)\), then for every \( \epsilon > 0 \), there exists a \( P_c(\theta, \cdot) \) that replaces \( \Sigma \) with \( \Sigma_c \) achieving \( \| \Sigma - \Sigma_c \| < \delta \) with probability \( 1 - q \), where \( \delta \) depends on \( \theta \), such that

\[
\sup_{\theta \in \Theta} \| P_c(\theta, \cdot) - P(\theta, \cdot) \|_{TV} < \epsilon
\]  

(26)

with probability \( 1 - q \).

In practice, although we cannot calculate an exact partial eigendecomposition, Algorithm 2 of Banerjee et al. [2013] provides an accurate approximation, which is equivalent to the adaptive randomized range finder (Algorithm 4.2) combined with the eigenvalue decomposition via Nyström approximation (algorithm 5.5) in Halko et al. [2011]. Algorithm 2 attains approximation error \( \| \Sigma - \Sigma_c \|_F < \delta \) with probability \( 1 - 10^{-d} \) where both \( \delta \) and \( d \) can be specified. We provide empirical evidence that the partial eigendecomposition approximation is accurate in the Appendix. Not all low-rank approximation of \( \Sigma \) approximate a partial eigendecomposition, so Theorem 3.2 suggests an advantage of Algorithm 2 of Banerjee et al. [2013] over alternatives.

The following remark describes the achievable rates in \( \delta \) as a function of \( \epsilon \) and \( n \).

Remark 3.3 (Rates for aMCMC for Gaussian process). The value of \( \delta \) for (25) is

\[
\delta = \frac{\epsilon^2 \sigma^4}{\tau^2 \sqrt{n(\tau^2 \lambda_{\max}(\Sigma_c) + \sigma^2)}} \wedge \frac{\epsilon^2 \sigma^2}{\tau^2},
\]

where \( \lambda_{\max}(\Sigma_c) \) is the largest eigenvalue of \( \Sigma_c \). Controlling \( \delta \) to satisfy Assumption 2.2 requires that

\[
\left| \exp \left( \frac{n - r}{2} \left[ \frac{\sigma^2 - \sigma^2_\ast}{\sigma^2_\ast \sigma^2} - \log \frac{\sigma^2_\ast}{\sigma^2} \right] \right) - \exp \left( \frac{n - r}{2} \left[ \frac{\tau^2 \delta + \sigma^2 - \tau^2_\ast \delta - \sigma^2_\ast}{(\tau^2_\ast \delta + \sigma^2_\ast)(\tau^2 \delta + \sigma^2)} \right] - \log \frac{\tau^2 \delta + \sigma^2_\ast}{\tau^2_\ast \delta + \sigma^2} \right) \right|
\]

be small, where \( \sigma^2_\ast, \tau^2_\ast \) are the proposed values of \( \sigma^2, \tau^2 \) in the Metropolis-Hastings algorithm. To achieve constant approximation error, \( \delta \) must decrease with \( n \); if the spectrum of \( \Sigma_c \) decays faster than linearly, the decrease can be slow. In addition, a smaller value of \( \delta \) is required when \( \tau^2 \) is large relative to \( \sigma^2 \), suggesting that a higher signal to noise ratio requires better approximations.

Remark 3.3 implies that for the weaker (25) to hold, \( \epsilon = O \left( \sqrt{\delta} \right) \): no effective estimate is available from the proof of (26). The algorithm scales as \( O(n^2 \tau) \), where \( \tau \) is the rank of \( \Sigma_c \), so increasing \( \tau \) to achieve a better approximation has computational cost \( n^2 \). However, the relationship between \( r \) and \( \delta \) – and therefore between \( r \) and \( \epsilon \) – depends on the spectrum of \( \Sigma \). If, for example, \( \lambda_c \propto c^{-r} \), then the speedup will be exponential. At the other extreme, the speedup could easily be concave if the spectrum decays too slowly. This is ultimately an empirical question, and is revisited in Section 4 when a specific dataset is analyzed.
4 Computational examples

This section contains empirical studies of the algorithms presented in section 3. The focus is on illustrating (1) concordance with the theory in section 2.2, (2) providing an approach to empirically assess convergence rates and approximation error for aMCMC, and (3) empirically assessing the importance of the requirement in section 2 that we have $\epsilon < \alpha/2$.

4.1 Preliminaries

We focus on two types of empirical assessments. The first is performance of aMCMC in estimation. This is done in several ways. If the application is a simulation example so that the “true” parameter values are known, then we compare the performance of the exact and approximate algorithms in parameter estimation by comparing to the “true” parameter value. This applies to the simulation example for aMCMC for mixture models. In this case, we also assess the performance of aMCMC as an approximation to the exact posterior by measuring discrepancy relative to sample paths obtained using the exact algorithm. If the application is a real data example where the exact algorithm is computationally feasible, we obtain approximations to posterior quantities of interest based on sample paths for the exact MCMC, and compare performance of aMCMC by measuring discrepancy relative to the estimates obtained from the exact algorithm, neglecting any simulation error. This applies to the analysis of aMCMC for logistic regression. Finally, if the exact algorithm is computationally intractable and the example is a real data example, we assess performance for various aMCMC algorithms that are computationally tractable by comparing performance in prediction on a test set. This applies to the Gaussian process application.

We also empirically assess the mixing properties and approximation error of aMCMC, particularly for the logistic regression example and, to a lesser extent, for the mixture model example. Our approach is general and could be applied to other aMCMC algorithms. The main interest is in approximating the value of $\alpha$ in Assumption 2.1 and the value of $\epsilon$ in Assumption 2.2. Although many MCMC algorithms may not satisfy the uniform condition in Assumption 2.1, most will satisfy the weaker geometric ergodicity condition

**Geometric ergodicity** A Markov chain evolving according to a transition kernel $P(\theta; \cdot)$ on a state space $\Theta$ with invariant measure $\Pi$ is geometrically ergodic if there exist constants $\rho \in (0, 1)$ and $B < \infty$ and a function $V: \Theta \to [1, \infty)$ such that

$$||P^k(\theta_0; \cdot) - \Pi||_{TV} \leq BV(\theta_0)\rho^k.$$ 

The parameter $\rho$ is the geometric convergence rate of the chain, which is analogous to the Doeblin constant $1 - \alpha$ in Assumption 2.1. Unlike uniformly mixing chains, the properties of geometrically ergodic chains depend on the initial state in a manner determined by the function $V$. However, that similar bounds to those in Section 2.2 are obtained by Pillai and Smith [2014] and Rudolf and Schweizer [2015], which assume only geometric ergodicity, suggests that many of the intuitions for uniformly mixing chains apply to geometrically ergodic chains as well. In particular, when the chain is rapidly mixing, so that $\rho < 1$, we expect that aMCMC will be favorable for computation times corresponding to relatively shorter path lengths of the exact chain than when $\rho$ is near one. On the other hand, relatively larger values of the approximation error $\epsilon$ may be tolerable when $\rho$ is small.

To assess these predictions, it is necessary to empirically estimate the convergence rate $\rho$ for real algorithms and data. We use the following approach based on sample path autocorrelations. When
Assumption 2.2, then $K$ kernel. Specifically, let Minsker et al. [2014] suggest a sample estimate of the Wasserstein distance with respect to a metric samples is difficult in moderate to high dimensions, and the estimates tend to be very noisy. Instead, distance suggesting more accurate approximations.

the (unnormalized) isotropic Gaussian kernel for which there exists a simple sample estimator (see equation 2.12 in Minsker et al. [2014]). Choosing to this metric is quite general conditions, Minsker et al. [2014] shows the Wasserstein-1 distance defined with respect in estimate of a lower bound on $\epsilon$ over the different sample paths. However, since we are mainly interested in illustrating the approach, P is reversible, there exist $B$ and $V$ in Definition such that $\rho^k = ||F^k||$ for the forward operator $F$ defined in [12]. When $F$ is compact, we also have

$$\sup_{f \in L_2(\Pi)} \text{Corr}(f(\theta_0), f(\theta_k)) = ||F^k||$$

where $L_2(\Pi)$ is the space of $(\Pi)$ square-integrable functions (for a more detailed discussion of this equivalence, see section 3 of Yang and Dunson [2013]). Although not all MCMC algorithms are reversible – and in particular, Gibbs samplers often are not – in practice, a useful lower bound estimate of $\rho$ can usually be obtained from the autocorrelations. Specifically, if $\text{Corr}(f(\theta_0), f(\theta_k)) \approx \rho^k$, then we can estimate $\rho$ from

$$\hat{\varphi}_{\max} = \max_{j \leq p} \max_{k \leq k_{\max}} \hat{\varphi}_{j,k}^{1/k},$$

where $\hat{\varphi}_{j,k}$ is an estimate of the lag-$k$ autocorrelation for the $j$th component of $\theta$ and $k_{\max} \ll k$. This estimator can be unreliable when $\hat{\varphi}_{j,k}$ is near zero, particularly for large $k$. Thus, when using the maximum likelihood estimator of the sample autocorrelations to compute (27), we consider only the values of $j$ for which $\hat{\varphi}_{j,k}$ exceeds $\frac{\phi^{-1}(0.95/k_{\max})}{\sqrt{k_{\max}}}$, which corresponds to a union bound multiplicity correction at the 0.95 level based on the asymptotic distribution of the MLE. To improve the accuracy of this estimate, it is ideal to run several chains using different starting points, then average $\hat{\varphi}_{\max}$ over the different sample paths. However, since we are mainly interested in illustrating the approach, we do not run multiple chains in the analyses that follow.

Quantitative estimates of $\epsilon$ are also of interest. If $\mathcal{P}$ satisfies Assumption 2.1 and $\mathcal{P}_c$ satisfies Assumption 2.2 then $||\Pi - \Pi_c||_{TV} \leq \frac{\epsilon}{2}$, so point estimates of $\alpha$ and $||\Pi - \Pi_c||_{TV}$ give a plug-in estimate of a lower bound on $\epsilon$. If these assumptions are not satisfied, we would still expect $||\Pi - \Pi_c||_{TV}$ to provide an indication of the magnitude of $\epsilon$, with smaller values of the total variation distance suggesting more accurate approximations.

Unfortunately, estimating the total variation distance between two distributions on the basis of samples is difficult in moderate to high dimensions, and the estimates tend to be very noisy. Instead, Minsker et al. [2014] suggest a sample estimate of the Wasserstein distance with respect to a metric kernel. Specifically, let $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a reproducing kernel and $\mathcal{H}$ a Hilbert space, and define

$$d_K(\theta_1, \theta_2) := ||K(\cdot, \theta_1) - K(\cdot, \theta_2)||_{\mathcal{H}},$$

where $||f - g||_{\mathcal{H}} = \sqrt{\langle f, g \rangle}$ is the usual Hilbert space norm. For probability measures $P, Q$ satisfying quite general conditions, Minsker et al. [2014] shows the Wasserstein-1 distance defined with respect to this metric is

$$W_{1,d_K}(P, Q) := \left\| \int_{\Omega} k(\theta, \cdot) d(P - Q)(\theta) \right\|_{\mathcal{H}},$$

for which there exists a simple sample estimator (see equation 2.12 in Minsker et al. [2014]). Choosing the (unnormalized) isotropic Gaussian kernel

$$K(\theta_1, \theta_2) = \frac{1}{\sigma} \exp(-\phi(\theta_1 - \theta_2)^t(\theta_1 - \theta_2)),$$

results in $d_K \leq \frac{1}{2}$. As default choices, we put $\phi, \sigma = 1$, giving $0 \leq d_K \leq 1$. Moreover, we have $W_{1,d_K}(P, Q) = 0$ if and only if $P = Q$. Since $d_K \leq 1$, this provides a lower bound on the total variation distance via $2W_{1,d_K}(P, Q) \leq ||P - Q||_{TV}$. 

\newpage

J. E. Johndrow, J.C. Mattingly, S. Mukherjee, D. Dunson
4.2 Distributional approximations – Mixture model

We simulated a $1000 \times 1000 = d \times d$ contingency table according to model (14a)-(14b), with $a^{(j)} = 1/d$, $\alpha = 1$, and $k = 7$, and implemented either aMCMC or exact MCMC for 10,000 iterations after a burn-in of 10,000 iterations. During the burn-in, data were gradually added to prevent the chain from becoming trapped in a local mode (for details see the Appendix). As a result, the sample size during much of the burn-in was relatively small. Since a threshold on $n(c)\hat{\nu}_h$ was used to determine whether to use the approximation, the approximation is initially used for only a small subset of the cells, and is gradually used for a larger proportion of the cells as more data are added. We focus comparisons on the samples after burn-in.

Table 2 shows RMSE and MAE for estimation of $\pi$ for two different sample sizes, and either $n_{\text{min}} = 100$ or $n_{\text{min}} = \infty$ (the exact algorithm), calculated for either the 100 largest entries of $\pi$ or 100 random entries of $\pi$. The value of both losses is small and depends only weakly on $\epsilon$ (which is a function of $n_{\text{min}}$). The sample sizes of $10^6$ and $10^9$ correspond to a total count equal to the number of cells, and a total count 1000 times larger than the number of cells. As expected, with $N = 10^9$, the posterior is very concentrated around the true values of $\pi$.

Table 3 shows effective samples per second (ES/sec) for aMCMC (with $n_{\text{min}} = 100$) and MCMC for $N = 10^6$ and $N = 10^9$. In every case, the effective sample size was at least 90 percent of the number of MCMC samples taken. The Geweke scores do not raise serious concerns about stationarity. With $N = 10^6$, aMCMC – with the value of $\epsilon$ implied by the threshold $n_{\text{min}} = 100$ – is less efficient (as measured by ES/sec), since for most cells $n(c)$ is not large relative to $k^3$. However, for $N = 10^9$, aMCMC is $\sim 10$ times more efficient than the exact sampler. This is consistent with the finding that $s(\epsilon)\propto\sqrt{\epsilon}$ for this algorithm. For concave loss functions, relatively small values of $\epsilon$ are usually optimal. So the results suggest that the optimal value of $n_{\text{min}}$ is greater than 100 for this algorithm when $N = 10^9$. The different result with $N = 10^9$ can be viewed as changing the constant in the speedup function. Because the computational cost of the exact algorithm is roughly 1000 times greater per step when $N = 10^9$ compared to $N = 10^6$, the speedup provided by any fixed value of $\epsilon$ is larger when the sample size is larger, but the shape of the speedup function does not change.

We also analyzed the convergence properties and approximation error of this algorithm. Table 4 shows estimates of $\hat{\varphi}_{\text{max}}$ for the 100 largest entries of $\pi$. Results are shown for different values of $N$ for the exact and approximate algorithms. The results suggest relatively good convergence.
Table 3: Effective samples per second for samples of the 100 largest entries of $\pi$ (left half of table) and 100 random entries of $\pi$ (right half of table) at two different sample sizes, using either the exact or approximate sampler. Also shown is the percentage of the Geweke convergence diagnostic $z$-scores that exceed 1.96 in absolute value.

| $n_{\text{min}} \rightarrow$ | 100 largest | 100 random |
|-----------------------------|-------------|------------|
| ES/sec - $N = 10^6$        | 5.01 4.60   | 4.85 4.38  |
| ES/sec - $N = 10^9$        | 0.07 0.66   | 0.06 0.62  |
| Geweke - $N = 10^6$        | 0.01 0.04   | 0.07 0.06  |
| Geweke - $N = 10^9$        | 0.10 0.04   | 0.06 0.09  |

properties, with a maximum value of 0.98. By remark 3.1, the exact transition kernel satisfies the Doeblin condition, so this is a lower bound estimate of $1-\alpha$. The maximum value of 0.98 corresponds to a $\delta$-mixing time of 228 iterations, a fast converging chain in the context of MCMC. There is no evidence that the mixing and convergence properties of the exact chain are superior to those of the approximate chain.

Table 4: Estimates of $\hat{\varphi}_{\text{max}}$ for the exact and approximate algorithms

| $n_{\text{min}} = \infty$ | $n_{\text{min}} = 100$ | $n_{\text{min}} = \infty$ | $n_{\text{min}} = 100$ |
|-----------------------------|-------------------------|-----------------------------|-------------------------|
| $N = 10^6$                  | 0.92 0.93               | 0.95 0.97                   |                         |
| $N = 10^7$                  | 0.93 0.92               | 0.98 0.95                   |                         |
| $N = 10^8$                  | 0.91 0.92               | 0.94 0.94                   |                         |
| $N = 10^9$                  | 0.93 0.90               | 0.94 0.95                   |                         |

Figure 3 shows estimates $\hat{W}_{1,dK} \left( \frac{1}{t_0} \sum_{k=1}^{t_0} \delta_{\beta_k}^{(t)}, \frac{1}{T} \sum_{k=1}^{T} \delta_{\theta_k} \right)$ of $W_{1,dK} \left( \sum_{k=1}^{t_0} P_e^k(\theta_0, \cdot), \Pi \right)$ based on different length sample paths $t_0$ corresponding to the computation times shown on the horizontal axis. Results are shown for both the exact and approximate algorithms with $N = 10^9$, and are based on samples of the 100 largest entries of $\pi$. For computation times up to around 30 minutes (1800 seconds), the approximate algorithm with $n_{\text{min}} = 100$ performs better with respect to this discrepancy measure than the exact algorithm. In both cases, the estimated values of $W_{1,dK}$ are very small, indicating that the posterior is quite concentrated for these parameters. The relatively short computation time over which the approximation is optimal is consistent with the finding that the exact chain appears to mix rapidly.

4.3 Logistic regression using subsets

We performed a more comprehensive analysis of the sampler in (20a)-(20c), consisting of both an analysis of performance in estimation and empirical analysis of the convergence rate and approximation error. For the latter, we also consider a more complicated hierarchical prior structure on $\beta$ and a larger number of covariates.
Figure 3: \( \hat{W}_{1,4}\), between the empirical measure of the entire sample path from the exact algorithm and the empirical measure of sample paths from the exact and approximate transition kernels for different computation times \( \tau \) with \( N = 10^9 \).

4.3.1 Estimation

We used the sampler in \((20a)-(20c)\) for logistic regression using the SUSY dataset (Baldi et al. [2014]). The dataset consists of 5 million observations of a binary outcome with 18 continuous covariates. The data are divided into a training set consisting of 4.5 million observations and a test set of 0.5 million observations. Computation was performed for a range of seven subset sizes between \( |V| = 1,000 \) and \( |V| = 4,500,000 \). In each case, the following functionals were estimated based on the Markov chain for \( \theta_t \). We do not adapt the subset sizes based on the state of the chain for this example, so it is not expected to achieve the uniform error bound.

1. The mean of the regression coefficients \( \beta \), based on \( \frac{1}{t} \sum_{k=0}^{t-1} \beta_k \). Root mean square error (RMSE) was used as the loss.

2. The median of the regression coefficients, given by \( m = \arg\max_{m^*} \left( \frac{1}{t} \sum_{k=0}^{t-1} \mathbb{1}_{\{\beta_k < m^*\}} \right) < 0.5 \). Mean absolute error was chosen for the loss function.

3. The endpoints of 95 percent posterior credible intervals, \( m_q = \arg\max_{m^*} \left( \frac{1}{t} \sum_{k=0}^{t-1} \mathbb{1}_{\{\beta_k < m^*\}} \right) < q \) for \( q = 0.025, 0.975 \). The loss function is RMSE. In addition, each 95 percent credible interval was classified according to whether it included zero, and the \( L_0 \) loss for this classification was calculated.

4. Prediction of the outcome \( y \) on the test set. Predictive accuracy was measured with the area under curve metric.

5. The \( L_1 \) and \( L_2 \) norms of the regression coefficients, \( \sum_j |\beta_j| \) and \( \sum_j \beta_j^2 \), respectively.

We focus on the accuracy of estimates based on samples from the (putative) stationary distributions of each of the scaled-likelihood samplers with subsample sizes \( |V| = 1,000, 5,000, 10,000, \ldots \).
Table 5: Posterior loss for estimation of various functionals at different values of $|V|$ for logistic regression example on SUSY data.

|               | 1,000  | 5,000  | 10,000 | 50,000 | 250,000 | 500,000 | 4,500,000 |
|---------------|--------|--------|--------|--------|---------|---------|-----------|
| RMSE          | 0.12   | 0.04   | 0.02   | 0.01   | 0.00    | 0.00    | 0.00      |
| RMSE (c.i.)   | 1.36   | 0.58   | 0.44   | 0.17   | 0.06    | 0.03    | 0.00      |
| AUC           | 0.86   | 0.86   | 0.86   | 0.86   | 0.86    | 0.86    | 0.86      |
| L1 norm beta  | 19.71  | 18.59  | 18.33  | 17.86  | 17.76   | 17.76   | 17.76     |
| L2 norm beta  | 52.27  | 45.50  | 44.13  | 42.27  | 42.05   | 42.05   | 42.09     |
| MAE           | 0.08   | 0.03   | 0.01   | 0.00   | 0.00    | 0.00    | 0.00      |
| Mean L0 Error (c.i. cross zero) | 0.53 | 0.16 | 0.16 | 0.11 | 0.11 | 0.05 | 0.00 |
| Clock time (seconds) | 24.64 | 48.54 | 78.16 | 333.24 | 1585.94 | 3207.83 | 29876.50 |
| Effective sample size | 413.99 | 405.65 | 405.28 | 424.36 | 394.24 | 372.06 | 369.70 |

50,000, 100,000, 250,000, 500,000 and 4,500,000. The first 1,000 samples were discarded and the subsequent 1,000 samples used to compute ergodic averages. All loss functions used an estimate based on 1,000 samples from the exact Markov chain after a 1,000 sample burn-in as the “truth.” Because the exact sampler mixes rapidly and has low autocorrelation (as measured by effective sample size, see Table 5), the error in these “true” posterior estimates is expected to be small.

Table 5 shows posterior loss for the parameters described above. As expected, the loss invariably decreases as $|V|$ grows, which corresponds to smaller median values of $\epsilon$. However, there are substantial differences in the rate at which the loss converges to zero as $\epsilon$ decreases. For example, $|V|=1,000$ is sufficient to obtain the best possible out of sample predictive performance measured by AUC, while even with $|V|=500,000$, one of the 18 regression coefficients is improperly classified as having a posterior credible interval that includes zero. Similarly, RMSE for estimation of $\beta$ decreases more slowly with $\epsilon$ than MAE for estimation of $\beta$ (by the median of the sample path).

4.3.2 Mixing and convergence properties

We also analyzed the mixing and convergence properties of a sampler with stationary measure corresponding the posterior under a more complex hierarchical prior on $\beta$, specifically

$$\beta_j \sim N(0, \tau^2\lambda_j^2), \quad \tau \sim C_+(0, 1), \quad \lambda_j \sim C_+(0, 1),$$

where $C_+(0, 1)$ is the Cauchy distribution with location 0 and scale 1 restricted to the real positive half-line. This is referred to as the Horseshoe shrinkage prior (Carvalho et al. [2010]). Sampling $\tau$ and $\lambda_j$ adds two slice sampling steps to the update rule (see the supplement to Polson et al. [2014]). Thus, this provides an opportunity to demonstrate the applicability of our empirical approach to analyzing the properties of an approximate algorithm that has not been characterized theoretically. As before, our $P_\epsilon$ corresponds to using subsamples of minibatches of data to approximate $X'\Omega X + B^{-1}$, where here $B = \text{diag}(\tau^2\lambda^2)$. To increase model complexity, we also add two and three way interactions into the model, while removing half of the 18 covariates in the original model, resulting in $p = 92$. Half of the features in the original model were nonlinear functions of the other features, which were known to contain substantial information about the response. Thus this analysis represents a more typical approach to modeling the linear predictor as a nonlinear function of covariates.

Empirically, this algorithm was found to converge rapidly (see below), so we obtained sample paths of length 1,200, and, where burn-in was used, discarded the first 200 samples. The wall clock time to obtain a sample path of this length using all the data was in excess of 12 hours. Figure 4 shows some results. In the left panel are RMSE for estimation of $\beta$ by its ergodic average as a function of
computation time $\tau$. RMSE was calculated by treating the ergodic average from the full sample path with the exact algorithm, discarding burn-in, as the exact posterior mean. The RMSE calculations use burn-in, so for the larger sample sizes the graph originates away from $\tau = 0$. The threshold computational budget at which the exact algorithm performs best is approximately two hours (7200 seconds). Notably, sample sizes of $|V| = 100,000$ and $|V| = 50,000$ perform relatively well, with $|V| = 100,000$ optimal for computation times between about 100 and 900 seconds. Conversely, the estimates with $|V| = 10,000$ were much less accurate, and the scale of RMSE was so different for $|V| = 1,000$ that the result is not shown.

In the right panel are estimates $\hat{W}_{1,dK} \left( \frac{1}{t_0} \sum_{k=1}^{t_0} \delta_{\hat{\theta}_k^{(\epsilon)}}, \frac{1}{t} \sum_{k=1}^{t} \delta_{\theta_k} \right)$ of $W_{1,dK} \left( \sum_{k=1}^{t_0} \mathcal{P}_\epsilon (\theta_0, \cdot), \Pi \right)$ based on different length sample paths $t_0$ corresponding to the computation times on the horizontal axis. Results are shown for several minibatch sizes and for the full data, without discarding burn-in. Here, we again approximate $\Pi$ using the entire sample path for the complete data. The smallest minibatch size $|V| = 1,000$ has $\hat{W}_{1,dK} \approx 1$ for any computational budget, suggesting that the corresponding value of $\epsilon$ is effectively too large to be useful with this algorithm. The other sample sizes all result in meaningful approximation to the posterior by this metric, with $|V| = 500,000$ being optimal over most computation times up to about two hours. For budgets greater than two hours, $|V| = 500,000$ still provides a very accurate approximation. It is likely that values of $|V|$ satisfying $500,000 < |V| < 4,500,000$ would be optimal for larger computation times.

![Graph showing RMSE(\beta) and \hat{W}_{1,dK}](image)

**Figure 4:** Logistic regression RMSE for estimation of $\beta$ (left) and approximate $W_{1,dK}$ distance to the exact posterior (right) as a function of computation time $\tau$ in seconds.

We also assessed the convergence rate of the exact and approximate algorithms and approximate values of $\epsilon$. To assess the magnitude of $\epsilon$, we estimate $W_{1,dK} (\Pi_\epsilon, \Pi)$, by

$$\hat{W}_{1,dK} \left( \frac{1}{t} \sum_{k=1}^{t} \delta_{\hat{\theta}_k^{(\epsilon)}}, \frac{1}{t} \sum_{k=1}^{t} \delta_{\theta_k} \right),$$

which gives an estimated lower bound on $||\Pi_\epsilon - \Pi||_{TV}$. Results are shown in Table 6. A lower bound on $\rho$ is estimated from $\hat{\varphi}_{\text{max}}$. The values of $\hat{\varphi}_{\text{max}}$ in $\hat{\varphi}_{\text{max}}$ are all about 0.98, indicating a chain with fairly rapid mixing – if $\mathcal{P}$ satisfied the Doeblin condition, the estimated values of $\hat{\varphi}_{\text{max}}$ would correspond to $\delta$-mixing times of about 228 for $\delta = 0.01$, which resulted in the choice of 200 for the
burn-in time. The estimates of \( \hat{W}_{1,dK}(\Pi_\epsilon, \Pi) \) suggest that the approximation error for \(|V| \leq 50,000 \) may be relatively large. Based on the value of \( \hat{\varphi}_{\max} \), large values of \( \epsilon \) may be unacceptable, at least in terms of convergence. However, the results in Figure 4 suggest that the only value of \( \epsilon \) that is “too large” in the sense that the resulting sample paths are effectively useless as an approximation to the posterior correspond to \(|V| \geq 50,000 \). There is also no evidence that the convergence rate decreases as \( \epsilon \) increases. Thus, the results shown in section 2.2 suggest that more accurate approximations – corresponding to \(|V| \leq 500,000 \) – are optimal for computational times up to two hours.

| \( |V| \) | 1,000 | 10,000 | 50,000 | \( 10^4 \) | \( 5 \times 10^5 \) | \( 4.5 \times 10^6 \) |
|---|---|---|---|---|---|---|
| \( \hat{W}_{1,dK}(\Pi_\epsilon, \Pi) \) | 0.98 | 0.29 | 0.05 | 0.03 | 0.01 | 0.00 |
| \( \hat{\varphi}_{\max} \) | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |

### 4.4 Low-rank Gaussian process

Computation was performed for the low-rank Gaussian process approximations as described in section 3.4. Computation for the exact transition kernel is infeasible due to the need to invert a large matrix, so we focus solely on performance of the approximate algorithm in prediction for different levels of approximation error. Six values of \( \delta \) – corresponding to approximation error for \( \Sigma \) in the Frobenius norm of \( \delta = 0.001, 0.01, 0.02, 0.03, 0.04, \) and \( 0.05 \) – were chosen to assess the computation time-approximation accuracy tradeoff. We do not adapt \( \delta \) to the state of the chain in this example. The model in (22) with prior in (23) was estimated on Sarcos robot arm data (see Vijayakumar et al. [2005]). A grid of \( \phi \) values corresponding to correlations between 0.99 and 0.01 at the maximum pairwise distance in \( X \) was used for the prior on \( \phi \), and Gamma \( (1, 1) \) priors chosen on \( \tau^{-2} \) and \( \sigma^{-2} \).

The data consist of 48,933 observations on 21 continuous covariates and one continuous outcome. Of these, 4,449 observations are commonly designated the test set. We divided the dataset into ten subsets of approximately equal size and performed computation independently on each subset. The results provided here are combined over the ten independent datasets.

Table 7 shows loss for estimation of various functionals of \( y_{test} \), the vector of response values in the test set. As in previous examples, this table is based on estimates obtained from the chains at putative stationarity. In particular, \( t = 1,000 \) samples were gathered after discarding \( B = 1,000 \) samples as burn-in. In summary, the loss functions, parameters, and estimators are:

1. **RMSE (mean of \( y_{test} \))**: RMSE for out of sample prediction of \( y \) calculated using \( \frac{1}{t} \sum_{k=B+1}^{B+t} y_{test}^k \) as the point estimate.

2. **MAE (median of \( y_{test} \))**: MAE for out of sample prediction of \( y \) calculated using \( m = \arg \max_m \beta_m^* : \left( \frac{1}{t} \sum_{k=B+1}^{B+t} \mathbb{1}_{\{\beta_m^* < \beta_k \}} \right) < 0.5 \) as the point estimate.

3. **MAE [c.i. coverage - 0.95]**: MAE for coverage of credible intervals (difference between empirical
Approximate MCMC

Table 7: Summaries of results for the aMCMC algorithm in the Gaussian process regression example for varying levels of approximation accuracy of the covariance.

| $\delta_e$ | 0.001 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|------------|-------|------|------|------|------|------|
| RMSE (mean of $y_{test}$) | 3.57  | 3.60 | 3.66 | 3.72 | 3.80 | 3.85 |
| MAE (median of $y_{test}$) | 2.43  | 2.47 | 2.52 | 2.53 | 2.61 | 2.64 |
| MAE [c.i. coverage - 0.95] | 0.01  | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 |
| Effective Size per sample | 1.01  | 1.01 | 1.00 | 1.00 | 1.00 | 1.00 |
| Geweke test proportion | 0.06  | 0.06 | 0.06 | 0.07 | 0.06 | 0.06 |
| Seconds per sample | 0.17  | 0.12 | 0.11 | 0.11 | 0.11 | 0.11 |

coverage and 0.95) for out of sample predictive intervals for $y$:

$$m_q = \arg\max_{m*} : \left(\frac{1}{t} \sum_{k=0}^{t-1} \mathbb{1}_{\{\beta_k < m_*\}}\right) < q$$

for $q = 0.025, 0.975$ as the point estimate for the credible intervals.

Also shown are the effective sample size per iteration, the proportion of the Geweke convergence $z$-scores that are greater than 1.96 in magnitude, and the computational intensity.

That the seconds per sample in Table 7 increases by less than a factor of two over the entire range of $\delta$ values considered reflects the empirical finding that the spectrum of the covariance matrix $\Sigma$ decays relatively slowly. As such, increasing $\delta$ does not result in a large decrease in $r$. Our analysis of the approximate algorithm for the Gaussian process indicated that when the spectrum decays very rapidly, the speedup function will be convex, and could even be exponential in some cases. However, in this case the speedup function is more likely concave, which explains why smaller values of $\delta$ appear to give noticeable performance improvements for small computational cost.

Figure 5 shows loss as a function of computation time for the six different $\delta_e$ values. The exact chain was not run because of computational cost and numerical instability. The patterns are consistent with the discussion in Section 2.2, in that the less accurate approximations are superior with smaller computational budget, but have asymptotic bias, so that there exists a threshold computational budget for which it is better to obtain fewer samples from a more accurate approximate chain. In this case, the threshold time at which the most accurate approximation is preferred is relatively low. The most accurate approximate chain, with $\delta_e = 0.001$, is optimal with respect to RMSE among the six values tested for out of sample prediction with a computational budget of 10 seconds or greater, and is optimal with respect to MAE with a budget of 15 seconds or greater. This provides further empirical support for the hypothesis that the speedup function for this algorithm on this dataset is concave, a consequence of the slow-decaying spectrum of $\Sigma$.

5 Discussion

We have used strong conditions on the ergodicity of the exact chain and required uniform error bounds for the approximation in the total variation norm. Similar results can be obtained with weaker assumptions on the ergodicity of the original chain and in more general metrics, as shown in Pillai and Smith [2014]. However, our Assumption 2.2 is consistent with that used in Pillai and Smith [2014] and elsewhere, so the results in Section 5 are largely transferable to settings with weaker
conditions on $P$. In fact, our work suggests that Assumption 2.2 is often not overly restrictive, so long as the approximation is adapted to the state of the chain. Nonetheless, regardless of how strong the conditions assumed, the performance of the approximate chain depends critically on constants that cannot in most cases be readily computed. In the setting of section 2.2, these are $\epsilon$ and $\alpha$; their existence can be verified and empirical estimates of lower bounds computed using the methods we suggest, but their exact values remain elusive. Moreover, in most cases these calculations must be performed for every algorithm. For geometrically ergodic chains, the error bounds will depend on a function of the starting point, complicating interpretation.

An important and interesting finding of this work is that in many cases, exact MCMC may not be statistically optimal when an approximation that offers computational advantages is available. aMCMC offers longer sample paths at the same computational cost, which, when the approximation is fairly accurate, can easily outweigh the effects of bias and potentially slower convergence of aMCMC. This tradeoff is formalized through the characterization of speedup functions and the compminimax notion of optimal approximation error. It has long been recognized in optimization that noisy gradients are often far superior to exact gradients, but this rationale has only recently entered into the MCMC literature (Korattikara et al. 2013, Ahn et al. 2012), which has mainly pursued approximations when exact MCMC is considered computationally intractable. Another way to view this possibly surprising result is as a parallel to the well-characterized phenomenon that biased estimators have lower risk than unbiased estimators (see e.g. Stein 1956). The superiority of aMCMC with respect to $D_{L_2}$ for large computational budgets is conceptually similar – that is, biased Markov chains can often have superior statistical properties to those of unbiased ones – though in this context the optimal level of bias depends on the computation time and the fundamental reason for the improved performance is quite different. That approximate MCMC may offer optimal performance for sample path lengths that exceed those found in typical applications of MCMC in Bayesian statistics suggests there is much room for expanding the use of aMCMC.

The theory of approximate MCMC provides a guide to what can go wrong when approximate kernels are employed, and how to check whether difficulties are likely to occur. This is exemplified by our consideration of low-rank approximations for Gaussian processes and logistic regression with subsets. Optimal algorithms for a particular model, regardless of the convergence properties of
the original chain, can only be approximately determined through numerical approximation of the constants or obtaining theoretically upper and lower bounds. These issues are not conceptually different from the long-standing issue of empirically assessing MCMC convergence, and are important problems for which no definitive solution currently exists. In the interim, the results presented here should provide a level of comfort for practitioners that approximate MCMC algorithms can often result in better performance in statistical estimation with limited computational resources.
A Table of notation

Table 8: Notation used in the paper

| Symbol | definition |
|--------|------------|
| \( \mathcal{P}(\theta, \cdot) \), \( \mathcal{P} \) | a transition kernel for a Markov process |
| \( \mathcal{P}_\epsilon(\theta, \cdot), \mathcal{P}_\epsilon \) | a transition kernel that approximates \( \mathcal{P} \) for which \( ||\mathcal{P} - \mathcal{P}_\epsilon||_{TV} < \epsilon \) |
| \( \Theta \) | the parameter space; equivalently, the state space of a Markov chain |
| \( \theta_t \) | the state of a Markov chain at time \( t \); or the corresponding random variable |
| \( p(\xi | \zeta) \) | generic notation for full conditional in a Gibbs sampler |
| \( q(\theta | \theta_0) \) | A generic Metropolis Hastings proposal law |
| \( \Pi(\theta | x) \) | the posterior distribution of \( \theta \) given data \( x \) |
| \( f \) | A function \( f: \Theta \rightarrow \mathbb{R} \) |
| \( \Pi f \) | the expectation of \( f(\theta) \) with respect to the posterior \( \Pi(\theta | x) \) |
| \( \bar{\Pi} f \) | ergodic average of \( f(\theta) \) over \( \theta_0, \ldots, \theta_{t-1} \) |
| \( \bar{\Pi}_\epsilon f \) | ergodic average of \( f(\theta^*) \) over \( \theta_0, \ldots, \theta^*_{t-1} \) |
| \( ||x|| \) | where \( x \) is a vector; the Euclidean norm \( \sqrt{x^T x} \) |
| \( ||M|| \) | where \( M \) is a matrix; the spectral norm \( \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{||Mx||}{||x||} \) |

B Proof of Theorem 2.3

B.1 Preparatory results

The following is a standard result of the Doeblin condition in Assumption 2.1.

**Theorem B.1 (Convergence under Doeblin condition).** Under assumption 2.1, there exists a unique stationary measure \( \Pi \) for \( \mathcal{P} \). Furthermore for any initial probability measures \( \nu_1, \nu_2 \), one has

\[
||\nu_1 \mathcal{P}^t - \nu_2 \mathcal{P}^t||_{TV} \leq (1 - \alpha)^t ||\nu_1 - \nu_2||_{TV}.
\]

In particular, taking \( \nu_1 = \Pi \), we have

\[
||\Pi - \nu_2 \mathcal{P}^t||_{TV} \leq (1 - \alpha)^t ||\Pi - \nu_2||_{TV} \leq (1 - \alpha)^t.
\]

**Proposition B.2.** Under Assumptions 2.1 and 2.2, any stationary measure \( \Pi_\epsilon \) of \( \mathcal{P}_\epsilon \) satisfies

\[
||\Pi - \Pi_\epsilon||_{TV} \leq \frac{\epsilon}{\alpha}.\]

**Proof.**

\[
||\Pi - \Pi_\epsilon||_{TV} \leq ||\Pi \mathcal{P} - \Pi_\epsilon \mathcal{P}_\epsilon||_{TV} + ||\Pi_\epsilon \mathcal{P} - \Pi_\epsilon \mathcal{P}_\epsilon||_{TV} \leq (1 - \alpha) ||\Pi - \Pi_\epsilon||_{TV} + \epsilon
\]

The first inequality follows from the triangle inequality the second used Assumption 2.1 for the first term and Assumption 2.2 for the second term. Rearranging the resulting inequality produces the result.

**Proposition B.3.** Let Assumption 2.1 and 2.2 hold. For any \( \epsilon \in (0, \alpha/2) \) Assumption 2.1 holds for the Markov operator \( \mathcal{P}_\epsilon \) with the constant \( \alpha' \) equal to \( \alpha - 2\epsilon \), which is less than 1 by construction. Hence for such \( \epsilon \) the chain has a unique stationary distribution \( \Pi_\epsilon \) to which it converges exponentially.
Proof. We have

$$\|\mathcal{P}_c(\theta, \cdot) - \mathcal{P}_c(\theta', \cdot)\|_{TV} = \|\mathcal{P}_c(\theta, \cdot) - \mathcal{P}(\theta, \cdot) + \mathcal{P}(\theta, \cdot) - \mathcal{P}_c(\theta', \cdot)\|_{TV}$$

$$\leq \|\mathcal{P}_c(\theta, \cdot) - \mathcal{P}(\theta, \cdot)\|_{TV} + \|\mathcal{P}(\theta, \cdot) - \mathcal{P}_c(\theta', \cdot)\|_{TV}$$

$$\leq \epsilon + \|\mathcal{P}(\theta, \cdot) - \mathcal{P}(\theta', \cdot)\|_{TV} + \|\mathcal{P}(\theta', \cdot) - \mathcal{P}_c(\theta', \cdot)\|_{TV}$$

$$\leq \epsilon + (1 - \alpha) + \epsilon = 1 - (\alpha - 2\epsilon)$$

Corollary B.4. If \( \mathcal{P}_c \) satisfies assumption \[2.2\] with \( \epsilon < \alpha/2 \), and \( \mathcal{P} \) satisfies Assumption \[2.1\] then for any initial state measure \( \nu \)

$$\|\nu \mathbb{P}_c^t - \Pi\|_{TV} \leq (1 - (\alpha - 2\epsilon))^t \|\nu - \Pi\|_{TV} + \frac{\epsilon}{\alpha}.$$  

Proof. This follows by applying the triangle inequality and the results of Propositions \[B.2\] and \[B.3\].

Corollary B.5 (Upper bounds on covariances). Suppose \( \mathcal{P} \) satisfies assumption \[2.1\]. Let \( f \) and \( g \) be bounded functions. Then

$$\text{cov}(f(\theta), g(\theta)) \leq (1 - \alpha)^{t-s}\|f\|_*\|g\|_*,$$

where \( \|f\|_* = \inf_{c \in \mathbb{R}} \|f - c\|_X \).

Proof. Our strategy follows some of the discussion in Yang and Dunson \[2013\]. Suppose \( f \) satisfies \( \Pi f = 0 \), and \( f \in L_2(\Pi) \). Define the forward operator

$$Ff(\theta) := f(\theta')\mathbb{P}(\theta, \theta')d\theta' = \mathbb{E}\left[f(\theta) \mid \theta_0 = \theta'\right].$$

From Lemma 12.6.4 in Liu \[2008\],

$$\sup_{f,g \in L_2(\Pi)} \text{corr}(f(\theta_0), g(\theta_0)) = \sup_{\|f\|_1 = 1, \|g\|_1 = 1} \langle F^t f, g \rangle = \|F^t\|,$$  

where \( \|F^t\| \) is the operator norm of \( F^t \). Since \( F^t f(\theta') = \mathbb{E}\left[f(\theta) \mid \theta_0 = \theta'\right] \), we have that

$$F^t f - \Pi f = \mathbb{E}\left[f(\theta) \mid \theta_0 = \theta'\right] - \Pi f \leq \|f\|_X (1 - \alpha)^t$$

by Theorem \[B.1\] so \( \langle F^t f, g \rangle < \|f\|_X \|g\|_X (1 - \alpha)^t \), giving \( \|F^t\| \leq (1 - \alpha)^t \). Now, since

$$\text{corr}(f(\theta_0), g(\theta_0)) = \text{corr}(f(\theta_0) - c, g(\theta_0) - c')$$

for any \( c, c' \in \mathbb{R} \), the bound in (28) also holds for functions with nonzero expectation with respect to \( \Pi \). Therefore

$$\sup_{f,g \in L_2(\Pi)} \text{cov}(f(\theta_0), g(\theta_0)) \leq \|f\|_*\|g\|_*(1 - \alpha)^t.$$  

Finally, since the above holds for any starting measure \( \theta_0 \sim \nu \), and \( \text{cov}(\theta_2, \theta_0) = \text{cov}(\theta_0, \theta_2) \), we obtain

$$\text{cov}(f(\theta), g(\theta)) \leq (1 - \alpha)^{t-s}\|f\|_*\|g\|_*.$$  

\[\square\]
B.2 Error bounds for exact chain

We want to show upper bounds on
\[ \mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) \right)^2 \right] \quad \text{and} \quad \left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k \right\|_{\text{TV}}. \]

A simple way to obtain a bound on the expected square is to proceed analogously to a bias-variance decomposition
\[
\begin{align*}
\mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) \right)^2 \right] &= \mathbb{E} \left[ \left( \Pi f + \frac{1}{t} \sum_{k=1}^{t-1} (\nu \mathcal{P}^k f - \nu \mathcal{P}^k f) + \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) \right)^2 \right] \\
&= \mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) - \nu \mathcal{P}^k f \right)^2 \right] \\
&= \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k f \right)^2 + \mathbb{E} \left[ \left( \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) - \nu \mathcal{P}^k f \right)^2 \right] \\
&= \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k f \right)^2 + \frac{1}{t^2} \sum_{k=0}^{t-1} \sum_{j=0}^{t-1} \text{cov}(f(\theta_k), f(\theta_j)).
\end{align*}
\]

Now applying Corollary B.5 and Theorem B.1
\[
\begin{align*}
\mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) \right)^2 \right] &\leq 4\|f\|^2 \left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k \right\|_{\text{TV}} + \frac{\|f\|^2}{t^2} \sum_{k=0}^{t-1} \sum_{j=0}^{t-1} (1 - \alpha)^{|j-k|} \\
&\leq 4\|f\|^2 (1 - (1 - \alpha)^t) \||\Pi - \nu||_{\text{TV}} + \frac{\|f\|^2}{t^2} \sum_{k=0}^{t-1} \sum_{j=0}^{t-1} (1 - \alpha)^{|j-k|}.
\end{align*}
\]

Concentrating on the second term, we have
\[
\frac{\|f\|^2}{t^2} \sum_{k=0}^{t-1} \sum_{j=0}^{t-1} (1 - \alpha)^{|j-k|} = \frac{\|f\|^2}{\alpha t^2} \sum_{k=0}^{t-1} \left( \sum_{j=0}^{k} (1 - \alpha)^{k-j} + \sum_{j=k+1}^{t-1} (1 - \alpha)^{j-k} \right)
\]
\[
= \frac{\|f\|^2}{\alpha t^2} \sum_{k=0}^{t-1} \left( (1 - \alpha)^{k+1} - (1 - \alpha)^{t-k} \right)
\]
\[
= \frac{\|f\|^2}{\alpha t} \left( \frac{2t + 2}{\alpha} + \frac{2(1 - \alpha)^{t+1}}{\alpha^2} - t - \frac{2}{\alpha^2} \right)
\]
\[
= \frac{\|f\|^2}{\alpha t} \left( \frac{2}{\alpha} + \frac{2}{\alpha^2} + \frac{2(1 - \alpha)^{t+1}}{\alpha^2} - \frac{1}{t} - \frac{2}{\alpha^2} \right),
\]

which gives the result.
To get a total variation bound, just apply Theorem B.1
\[
\left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k \right\|_{TV} = \frac{1}{t} \sum_{k=0}^{t-1} \|\Pi - \nu \mathcal{P}^k\|_{TV} \\
\leq \frac{1}{t} \sum_{k=0}^{t-1} (1 - \alpha)^k \|\Pi - \nu\|_{TV} = \frac{(1 - (1 - \alpha)^t)}{\alpha t} \|\Pi - \nu\|_{TV}.
\]

B.3 Basic closeness properties of \( \mathcal{P}_\epsilon \)

Here we follow a similar approach as with the exact chain, except an additional asymptotic bias term will appear. Start with the \( L_2 \) bound
\[
\mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k) \right)^2 \right] = \mathbb{E} \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} (\Pi_\epsilon - \nu \mathcal{P}_\epsilon^k) f + \frac{1}{t} \sum_{k=0}^{t-1} (\Pi_\epsilon - \nu \mathcal{P}_\epsilon^k) f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k') \right)^2 \right] \\
= \mathbb{E} \left[ \left( (\Pi - \Pi_\epsilon) f + \frac{1}{t} \sum_{k=0}^{t-1} (\Pi_\epsilon - \nu \mathcal{P}_\epsilon^k) f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k') \right)^2 \right] \\
= ((\Pi - \Pi_\epsilon) f)^2 + \left( \frac{1}{t} \sum_{k=0}^{t-1} (\Pi_\epsilon - \nu \mathcal{P}_\epsilon^k) f \right)^2 + \mathbb{E} \left[ \left( \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k') - \nu \mathcal{P}_\epsilon^k f \right)^2 \right] \\
+ \frac{2}{t} ((\Pi - \Pi_\epsilon) f) \left( \sum_{k=0}^{t-1} (\Pi_\epsilon - \nu \mathcal{P}_\epsilon^k) f \right) \\
\leq \frac{4\epsilon^2 \|f\|_*^2}{\alpha^2} \sum_{k=0}^{t-1} \sum_{j=0}^{t-1} \text{cov} \left( f(\theta_k'), f(\theta_j') \right) \\
+ \frac{2 \|f\|_*^2 \epsilon}{t} \sum_{k=0}^{t-1} (1 - (1 - \alpha - 2\epsilon)^t) \|\Pi_\epsilon - \nu\|_{TV} \\
+ \frac{2}{t} \frac{\alpha}{\alpha - 2\epsilon} \frac{\|f\|_*^2 (1 - (1 - \alpha - 2\epsilon)^t)}{t(\alpha - 2\epsilon)} \|\Pi_\epsilon - \nu\|_{TV} \\
+ \|f\|_*^2 \left( \frac{2}{\alpha t} + \frac{2}{\alpha t^2} + \frac{2(1 - \alpha)\epsilon}{\alpha^2 t^2} - \frac{1}{t} \frac{2}{\alpha_t^2 t^2} \right) \\
+ \frac{8 \|f\|_* \epsilon (1 - (1 - \alpha\epsilon)^t) \|\Pi_\epsilon - \nu\|_{TV}}{t \alpha \epsilon},
\]
where \( \alpha_\epsilon = \alpha - 2\epsilon \). Now for the total variation result
\[
\mathbb{E} \left[ \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k') \right] = \mathbb{E} \left[ (\Pi - \Pi_\epsilon) f + \frac{1}{t} \sum_{k=0}^{t-1} (\Pi_\epsilon - \nu \mathcal{P}_\epsilon^k) f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k') \right] \\
\leq \frac{2 \|f\|_* \epsilon}{\alpha} + \frac{2 \|f\|_* (1 - (1 - \alpha\epsilon)^t) \|\Pi_\epsilon - \nu\|_{TV}}{t \alpha_\epsilon},
\]
\[ \left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu P_k^\epsilon \right\|_{TV} = \sup_{f: \|f\|_\infty \leq 1} \frac{1}{2} E \left[ \|f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k^\epsilon) \|_{TV} \right] \leq \frac{\epsilon}{\alpha} + \frac{(1 - (1 - \alpha_t)^t) \|\Pi - \nu\|_{TV}}{t \alpha_\epsilon}, \]

since \( \|f\|_\infty \leq 1 \) implies \( \|f\|_* \leq 1. \)

C Proof of Remark 2.1

Now we show that the total variation bound for the exact chain is tight by exhibiting a Markov chain satisfying the assumptions that achieves the bound. Let

\[ P = \begin{pmatrix} 1 - a & a \\ a & 1 - a \end{pmatrix} \]

for \( a \leq 1/2 \). It is easy to verify by direct calculation that the invariant measure is \( \Pi = (1/2, 1/2) \) and \( P \) satisfies the Doeblin condition with \( \alpha = 2a \). \( P \) has eigenvectors

\[ \phi_1 = (1/2, 1/2), \quad \phi_2 = (-1/2, 1/2) \]

with eigenvalues 1 and \( 1 - 2a \), respectively. Any possible starting measure \( \nu \) can be expressed as \( \nu_\gamma = (\gamma, 1 - \gamma) \) for some \( \gamma \leq 1/2 \) (if \( \gamma > 1/2 \), just switch the definitions of the two states). Then \( \|\nu_\gamma - \Pi\|_{TV} = \frac{1}{2} \left( \frac{1}{2} \right) (\gamma - 1/2) = |\gamma - 1/2| \) when \( \gamma < 1/2 \). This can be expressed in terms of the eigenvectors as

\[ (\gamma, 1 - \gamma) = (1/2, 1/2) + (1 - 2\gamma)(-1/2, 1/2). \]

So then

\[ \nu_\gamma P^k = (1/2, 1/2) + (1 - 2\gamma)(1 - 2a)^k(-1/2, 1/2) = (1/2, 1/2) + \|\nu_\gamma - \Pi\|_{TV} (1 - \alpha)^k(-1/2, 1/2) \]

and so

\[ \frac{1}{t} \sum_{k=0}^{t-1} \nu P^k = (1/2, 1/2) + (1 - 2\gamma) \frac{1 - (1 - \alpha)^t}{2at} (-1/2, 1/2) \]

So

\[ \left\| \Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu P^k \right\|_{TV} = \frac{1}{2} \left( \left\| \frac{1}{2} \left( \frac{1}{2} - \frac{1}{2} \|\nu_\gamma - \Pi\|_{TV} \frac{1 - (1 - \alpha)^t}{at} \right) \right\| \right) \]

\[ + \frac{1}{2} \left( \left\| \frac{1}{2} + \frac{1}{2} \|\nu_\gamma - \Pi\|_{TV} \frac{1 - (1 - \alpha)^t}{at} \right\| \right) \]

\[ = \frac{(1 - (1 - \alpha)^t) \|\nu_\gamma - \Pi\|_{TV}}{at}, \]
as required.

Now, we note that the perturbation
\[
\mathcal{P}_\epsilon = \begin{pmatrix}
1 - (a - \epsilon) & a - \epsilon \\
\alpha + \epsilon & 1 - (a + \epsilon)
\end{pmatrix},
\]
satisfies \(\sup_{\theta \in \Theta} ||\mathcal{P}_\epsilon(\theta, \cdot) - \mathcal{P}(\theta, \cdot)||_{TV} = \epsilon\). For \(\epsilon < \alpha/2\), \(\mathcal{P}_\epsilon\) satisfies the Doeblin condition with \(\alpha_\epsilon = 2a = \alpha\), and has invariant measure \(\Pi_\epsilon = \left(\frac{a + \epsilon}{2a}, \frac{a - \epsilon}{2a}\right)\). Therefore, we have
\[
||\Pi - \Pi_\epsilon||_{TV} = \frac{1}{2} \left(\frac{1}{2} \frac{a + \epsilon}{2a} + \frac{1}{2} \frac{a - \epsilon}{2a}\right) = \frac{1}{2} \left(\frac{2a - 2(a + \epsilon)}{2(2a)} + \frac{2a - 2(a - \epsilon)}{2(2a)}\right) = \frac{1}{2} \left(\frac{\epsilon}{\alpha} + \frac{\epsilon}{\alpha}\right) = \frac{\epsilon}{\alpha}
\]
for this chain, showing that for every \(\alpha < 1/2\) and \(\epsilon < \alpha/2\), there exists a Markov chain satisfying both the Doeblin condition and uniform approximation error conditions for which \(||\Pi - \Pi_\epsilon||_{TV} = \frac{\epsilon}{\alpha}\).

A similar perturbation
\[
\mathcal{P}_\epsilon = \begin{pmatrix}
1 - (a - \epsilon) & a - \epsilon \\
a - \epsilon & 1 - (a - \epsilon)
\end{pmatrix},
\]
can be represented as
\[
\mathcal{P}_\epsilon = \begin{pmatrix}
1 - a_\epsilon & a_\epsilon \\
a_\epsilon & 1 - a_\epsilon
\end{pmatrix},
\]
for \(a_\epsilon = a - \epsilon\), and has \(\alpha_\epsilon = 2a - 2\epsilon = \alpha - 2\epsilon\), and invariant measure \((1/2, 1/2)\). So applying the result proved for \(\mathcal{P}, \mathcal{P}_\epsilon\) achieves
\[
= \frac{1}{\alpha\epsilon} \left(1 - (1 - \alpha_\epsilon)^t \left||\nu_\gamma - \Pi\right||_{TV}\right) = \frac{1}{\alpha\epsilon} \left(1 - (1 - (\alpha - 2\epsilon))^t \left||\nu_\gamma - \Pi\right||_{TV}\right).
\]

So there exist perturbations that achieve both of the components of the bound for \(||\Pi - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k\||_{TV}\), but the perturbations exhibited differ.

Now, recall that
\[
E \left[\left(\Pi f - \frac{1}{t} \sum_{k=0}^{t-1} f(\theta_k)\right)^2\right] = \left(\Pi f - \frac{1}{t} \sum_{k=0}^{t-1} \nu \mathcal{P}^k f\right)^2 + \frac{1}{t^2} \sum_{k=0}^{t-1} \sum_{j=0}^{t-1} \text{cov}(f(\theta_k), f(\theta_j))\).
\]

For a discussion of tightness of the covariance bound \(\text{cov}(f(\theta_0), f(\theta_t)) \leq ||F||\) when the forward operator is compact and self-adjoint, see Yang and Dunson [2013]. Now, note that
\[
\mathcal{P} = \begin{pmatrix}
1 - a & a \\
a & 1 - a
\end{pmatrix}
\]
is the transition matrix of a reversible Markov chain on a finite state space, so $F$ is compact and self-adjoint. We showed that

$$
\frac{1}{t} \sum_{k=0}^{t-1} \nu P^k = (1/2, 1/2) + (1 - 2\gamma) \frac{1 - (1 - 2\alpha)^t}{2at} (-1/2, 1/2).
$$

The only non-trivial functions on this state space have different values in the two states. To make $|f| \leq 1$, put $f(0) = -1$ and $f(1) = 1$. Then

$$
\Pi f - \frac{1}{t} \sum_{k=0}^{t-1} \nu P^k = \begin{cases} 1 & \text{w.p. } \frac{||\Pi - \nu||_{TV}(1 - (1 - \alpha)^t)}{\alpha t}, \\ -1 & \text{w.p. } \frac{||\Pi - \nu||_{TV}(1 - (1 - \alpha)^t)}{\alpha t} \end{cases}
$$

so

$$
E \left[ \left( \Pi f - \frac{1}{t} \sum_{k=0}^{t-1} \nu P^k \right)^2 \right] = \frac{2||\Pi - \nu||_{TV}(1 - (1 - \alpha)^t)}{\alpha t} = \frac{2||f||_p^2 ||\Pi - \nu||_{TV}(1 - (1 - \alpha)^t)}{\alpha t}.
$$

\section{D Alternative to Assumption 2.1}

We give an alternative set of conditions that are together equivalent to Assumption 2.1 but are easier to verify. A classic idea in Markov chain theory is that a minorization condition on the state space, $\inf_{\theta \in \Theta} P(\theta, \cdot) \geq \gamma m(\cdot)$ where $m(\cdot)$ is a probability measure, implies the Doeblin condition stated in Assumption 2.1 \cite{Nummelin1978, AthreyaNey1978, MeynTweedie2009, Rosenthal1994}. Here we use a slight variation on the standard minorization condition. Specifically, we divide the state space into a good set $\Theta_0$ and a bad set $\Theta_c$, with standard minorization conditions holding on the good set and a lower bound on the probability of transitioning from the bad set to the good set; these conditions are stated in Assumption D.1. One can show that two steps of any Markov chain satisfying Assumption D.1 will satisfy the standard minorization condition. This implies that satisfying Assumption D.1 is equivalent to satisfying Assumption 2.1.

\begin{assumption}
(Minorization and return condition for exact chain) \label{assumption:D.1}
Let $P$ be a Markov transition kernel with state space $\Theta$. There exist $\Theta_0 \subset \Theta$, constants $\gamma, \beta > 0$, and a probability measure $m(\cdot)$ supported on $\Theta$ such that

\begin{align}
\inf_{\theta \in \Theta_0} P(\theta, \cdot) & \geq \gamma m(\cdot), \\
\inf_{\theta \in \Theta_c} P(\theta, \Theta_0) & \geq \beta,
\end{align}

and $m(\Theta_0) > 0$.
\end{assumption}

\section{E Proof of Remark 3.1}

Here we show that there exists a probability measure $m(\cdot)$, a set $\Theta_0 \subset \Theta$, and constants $\gamma, \beta > 0$ such that the Gibbs sampling algorithm in Section 3.1 for the mixture model in (14a)-(14b) satisfies Assumption D.1 and hence Assumption 2.1.
The state space for this Gibbs sampler is given by \( \Theta = \Lambda \times \mathcal{N} \times \mathcal{Z} \), where
\[
\Lambda = \prod_{j=1}^{p} \prod_{h=1}^{K} \Delta^{(d_j-1)}, \quad \mathcal{N} = \Delta^{(K-1)}, \quad \mathcal{Z} = \times_{c \in \mathcal{C}^+} \mathcal{Z}_c
\]
\[
\mathcal{Z}_c = \left\{ Z(c) \in \mathbb{N}^K : \sum_{h=1}^{K} Z(c)_h = n(c) \right\},
\]
\( \mathbb{N} \) are the nonnegative integers, \( n(c) \) is the observed count in cell \( c \) and \( c \in \times_{j=1}^{p} \{1, \ldots, d_j\} \). Fix \( 0 < \delta < 1 \) and define \( \Theta_0 = \Lambda_0 \times \mathcal{N}_0 \times \mathcal{Z} \), where
\[
\Lambda_0 = \prod_{j=1}^{p} \prod_{h=1}^{K} \Delta_0^{(d_j-1)}, \quad \Delta_0^{(d_j-1)} = \{ \lambda \in \Delta^{(d_j-1)} : \delta < \lambda_c < 1 - \delta \ \forall \ c \}, \quad \mathcal{N}_0 = \Delta_0^{(K-1)}.
\]
\( \Delta^K \) is the \( K \)-dimensional unit simplex, and \( \times \) represents a Cartesian product.

**E.1 Minorization condition**

First we construct a measure \( m(\cdot) \) such that
\[
\inf_{\theta \in \Theta_0} \mathbb{P}(\theta, \cdot) \geq \gamma m(\cdot).
\]
For a function of two variables \( f(x, y) \), let \( f_{\inf}(y)(x) = \inf_y f(x, y) \) be the function defined by the pointwise infimum over \( y \). Let \( p(\nu \mid Z) \) and \( p(\lambda \mid Z) \) be the conditional densities of \( \nu, \lambda \) given \( Z \) in the Gibbs sampling algorithm.

It is enough to show that (1) every configuration of \( Z \) has positive probability for \( \nu, \lambda \in \mathcal{N}_0 \times \Lambda_0 \) and (2) the functions \( p_{\inf}(Z)(\lambda), p_{\inf}(Z)(\nu) \) satisfy \( \int_{\mathcal{N} \times \Lambda} p_{\inf}(Z)(\nu) d\nu > 0, \int_{\mathcal{N} \times \Lambda} p_{\inf}(Z)(\lambda) > 0 \).

The conditional distribution for \( Z \) given \( \lambda, \nu \) is
\[
Z(c) \mid \nu, \lambda, Y \sim \text{Multinomial}(n(c), \tilde{\nu}), \quad \tilde{\nu}_h = \frac{\nu_h \prod_{j=1}^{p} \lambda_{hc_j}^{(j)}}{\sum_{l=1}^{K} \nu_l \prod_{j=1}^{p} \lambda_{lc_j}^{(j)}}
\]
so that for any \( \theta \in \Theta_0, \tilde{\nu}_h > \frac{1}{K} \left( \frac{\delta}{1-\delta} \right)^{p+1} \) for every \( h \in \{1, \ldots, K\} \). This immediately implies that
\[
\inf_{\lambda \in \Lambda_0, \nu \in \mathcal{N}_0} p(Z \mid \lambda, \nu) > 0.
\]

To show (2), note that \( p(\nu \mid Z) \) and \( p(\lambda^{(j)}_h \mid Z) \) are both Dirichlet densities (since \( \lambda^{(j)}_h \) are conditionally independent given \( Z \), it is enough to show (2) for an arbitrary \( \lambda^{(j)}_h \)). The parameter of \( p(\nu \mid Z) \) is \( \alpha(Z) = \alpha + \sum_{c \in \mathcal{C}^+} Z(c) \), with density
\[
p(\nu \mid Z) = \frac{1}{B(\alpha(Z))} \prod_{h=1}^{K} \nu^{\alpha(Z)_h - 1},
\]
where \( B(\alpha(Z)) = \frac{\Gamma(\alpha(Z)_h)}{\Gamma(\alpha(Z))} \). Consider any compact subset of \( \mathcal{N} \) with nonzero Lebesgue measure that has empty intersection with the boundaries of the simplex. For simplicity, we can take \( \mathcal{N}_0 \).

Because \( \mathcal{Z} \) is a finite set, and for any \( Z \in \mathcal{Z}, \alpha(Z)_h > 1 \) for all \( h \) so long as \( \alpha > 0 \), \( \inf_{\nu \in \mathcal{N}_0} p_{\inf}(Z)(\nu) = \gamma^* > 0 \). This is enough to give \( \int_{\mathcal{N}_0} p_{\inf}(Z)(\nu) d\nu > \gamma^* \text{Vol}(\mathcal{N}_0) > 0 \), where \( \text{Vol}(\mathcal{N}_0) \) is the Lebesgue measure of the set \( \mathcal{N}_0 \). A result for \( \lambda^{(j)}_h \) follows by a similar argument.
E.2 Return condition

Now we show that for $\theta \in \Theta_0$, $\inf_{\theta \in \Theta_0} P(\theta | \Theta) > 0$. Since $\lambda \perp \nu \mid Z$ and $\nu \perp \lambda \mid Z$, the return probability does not depend on $(\lambda, \nu)$ but only on $Z$. Conditional on any value of $Z$, $P(\Theta_0 | Z)$ is strictly positive so long as the prior hyperparameters $a^{(j)}$ and $\alpha$ have strictly positive entries. Since $Z$ is finite, minimize over all elements of $Z$ to obtain $\beta = \bigwedge_{Z \in Z} P(\Theta_0 | Z) > 0$, a lower bound for $P(\Theta_0 | \theta)$ that holds for any $\theta \in \Theta$, so in particular it holds for any $\theta \in \Theta_0$.

E.3 Nonnegativity condition

Now we just want $m(\Theta_0) > 0$, but this is easy since we showed that $\inf_{\nu} P(\Theta_0 | \nu)$ are bounded below on $\Theta_0$.

F Proof of Lemma 3.1

We rely on the Berry-Esseen result in Weiss [1978]. The result is given for a Multinomial $(n, \nu)$ distribution with number of classes $K$ that may be increasing in $n$, but in our setting $n$ is fixed so we state the result in this special case.

F.1 Result from Weiss [1978]

Suppose there exists $\delta > 0$ such that $\min_{1 \leq h \leq K} (1 - \nu_h) > \delta$. Let $W(n)$ be a random variable having distribution given by the usual normal approximation to the Multinomial, so that $W \sim \text{Normal}(n\nu, n[\text{diag}(\nu) - \nu\nu^T])$, and for $h = 1, \ldots, K - 1$, define the random variable $\tilde{W}_h$ as the closest value to $W_h$ (in the $L_1$ sense) which makes $n\nu_h + \sqrt{n}\nu_h \tilde{W}_h$ an integer. $\tilde{W}_K(n)$ is given by the identity

$$\sum_{h=1}^{K} \sqrt{\nu_h} \tilde{W}_h = 0.$$  

Note that this is equivalent to rounding the entries $1, \ldots, K - 1$ to the nearest integer and defining the final entry to ensure that the full vector $\tilde{W}$ sums to $n$.

Let $\mu_{\tilde{W}}(\cdot)$ be the measure on $\mathbb{Z}^K$ induced by the definition of $\tilde{W}$ and let $\mu_Y(\cdot)$ be the Multinomial $(n, \nu)$ measure. The result in Weiss [1978] is

$$||\mu_{\tilde{W}} - \mu_Y||_{TV} \leq C(K - 1) \frac{\sum_{h=1}^{K-1} (1 - \nu_h)(1 + P_h/\nu_K)(1 - 2\nu_h + 2\nu_h^2)}{\sqrt{n}\nu_h(1 - \nu_h)} ,$$

where $P_h = \sum_{h' \leq h} \nu_{h'}$ and $C(K - 1)$ is a constant depending on $K - 1$.

By constructing a result of this sort from first principles, it should be possible to obtain a bound on the magnitude of $C(K)$, as has been shown for Berry-Esseen results in other settings. However, as our goal is only to show that any approximation error can be obtained with sufficiently large $n$, we do not pursue this here.

F.2 Construction of $P_\epsilon$

To construct $P_\epsilon$ satisfying assumption 2.2, let $p_c(Z | c, \nu, \lambda)$ be the pmf of a random variable corresponding to the measure $\mu_{\tilde{W}}(\cdot)$.
Use the independence of $Z(c)$ conditional on $\lambda, \nu$ to obtain

$$||p_r(Z | \lambda, \nu) - p(Z | \lambda, \nu)||_{TV} \leq \sum_{c \in C^+} ||p_r(Z(c) | \lambda, \nu) - p(Z(c) | \lambda, \nu)||_{TV}.$$  

This implies that $||p_r(Z(c) | \lambda, \nu) - p(Z(c) | \lambda, \nu)||_{TV} < \epsilon/N_z$, where $N_z = |C^+|$, is sufficient for $||P(\theta, \cdot) - P_r(\theta, \cdot)||_{TV} < \epsilon$. For any $c$ and any set $H \in \{1, \ldots, K\}$, recall that $Z(c)_H = Z(c)_h$, $h \in H$. Define $n^*_H(c) = n(c) - Z(c)_H$, and $\bar{\nu}_H = 1 - \sum_{h \in H} \bar{\nu}_h$. Let $Z^*(c)_H = (Z(c)_H, n^*_H(c))$, which is distributed Multinomial$(n(c), (\bar{\nu}_H, \bar{\nu}^*_H))$. Put

$$\mathcal{H}_e = \{H \in \{1, \ldots, K\} : ||p_r(Z^*(c)_H | \lambda, \nu) - p(Z^*(c)_H | \lambda, \nu)||_{TV} < \epsilon/N_z\},$$

and for each $c$ define the subset $H$ by

$$H = \left\{ H \in \mathcal{H}_e : \sum_{h \in H} \bar{\nu}_h = \bigvee_{H \in \mathcal{H}_e} \sum_{h \in H} \bar{\nu}_h \right\},$$

where $\bigvee$ is the max function. Define $P_r$ by the update rule:

1. For every $c \in C^+$, sample $Z^*(c)_H$ from the normal approximation $\bar{W}$ defined above.
2. Conditional on $n^*_H(c)$, sample $Z(c)_H \sim$ Multinomial$(n^*(c)_H, \bar{\nu}_H)$ from its exact multinomial distribution.
3. Sample $\nu, \lambda$ from their exact full conditionals.

This chain satisfies assumption 2.2

**G. Proof of Theorem 3.1**

First we show a lemma that is used in the proof of the main result.

**Lemma G.1.** The PG$\{1, \alpha\}$ distribution is a log-concave probability law.

*Proof.* If $\omega \sim$ PG$\{1, \alpha\}$, then it is equal in distribution to the infinite sum of Exponentials

$$\omega \sim \sum_{k=0}^{\infty} \varphi_k, \quad \varphi_k = \frac{g_k}{\pi^2(k-1/2)^2 + \alpha^2/2},$$

where $g_k \sim \text{Exp}(1)$, $\varphi_k \sim \text{Exp}(\pi^2(k-1/2)^2 + \alpha^2/2)$, and $\varphi_k$ has a log-concave probability distribution since $\text{Exp}(\lambda)$ is log-concave for all finite $\lambda$ (see e.g. Bagnoli and Bergstrom [2005]). Consider the sequence of random variables

$$\omega_n \sim \sum_{k=0}^{n} \frac{g_k}{\pi^2(k-1/2)^2 + \alpha^2/2} = \sum_{k=0}^{\infty} \varphi_k$$

for $n = 0, \ldots, \infty$. For any finite $n$, $\omega_n$ has a log-concave distribution since the sum of independent random variables having log-concave distributions is log-concave (see Proposition 3.5 in Saumard et al. [2014]). As $\omega_n \overset{D}{\to} \omega$ (indicating convergence in distribution), $\omega$ is log concave from Proposition 3.6 in Saumard et al. [2014].
G.1 Proof of main result

We want to show sup_{θ ∈ Θ} ||P (θ, ·) − P_ϵ (θ, ·)||_{TV} < ϵ with high probability. Here, P (θ, ·) is the transition kernel based on the full sample of N observations for the Gibbs sampler in (19a)-(19c), and P_ϵ (θ, ·) uses subsets of data of size |V| ≤ N to approximate X′ΩX by \( \frac{N}{|V|} X'_V \Omega V X_V \), in accordance with the update rule in (20a)-(20c).

We begin by showing how to construct a transition kernel P_ϵ (θ, ·) that achieves ||P (θ, ·) − P_ϵ (θ, ·)||_{TV} < ϵ conditional on the current state θ = (β, ω), then show that we can control the supremum. First, notice that the Gibbs sampling update rule in (20a)-(20c) depends on θ only through β, so we need only condition on β. Define

\[
\Sigma(β) = \text{cov}(ω_{1/2} x | β), \quad Σ_N(β) = \frac{1}{N} X'ΩX, \quad Σ_V(β) = \frac{1}{|V|} X'_V \Omega V X_V
\]

\[
S_N(β) = \frac{1}{N} (Σ_N(β) + B^{-1}/N)^{-1}, \quad S_V(β) = \frac{1}{N} (Σ_V(β) + B^{-1}/N)^{-1};
\]

we will sometimes suppress dependence on β for notational convenience. Recall that the distribution of β_{t+1} given ω_{t+1} is Normal (Σ_N X'κ, Σ_N), with κ = y−1/2. Let N(·; m, M) be the measure induced by a normal random variable with mean m and covariance M.

We first show that for every δ and every 0 < q < 1 there exists a |V| for which

\[
||Σ_N − Σ_V|| ≤ δ||Σ||
\]

with probability 1 − q whenever N > |V|. In practice, the achievable q with |V| < N will depend on N and δ. We then apply this to bound the Kullback-Leibler divergence

\[
\text{KL}(N(·; S_V X'κ, S_V) \| N(·; S_N X'κ, S_N)) = \frac{1}{2} \left( \text{tr} \left( S_N^{-1} S_V \right) - p + \log \left( \frac{|S_N|}{|S_V|} \right) + Q \right),
\]

with Q = (Σ_N X'κ − S_V X'κ)S_N^{-1}(Σ_N X'κ − S_V X'κ). We then use Pinsker’s inequality to obtain a total variation bound. We will choose δ as a function of ϵ and quantities depending on β to obtain ||P (θ, ·) − P_ϵ (θ, ·)||_{TV} < ϵ; thus, the supremum is controlled by adaptive choice of δ. When this requires |V| > N, put V = \{1, ..., N\} and obtain the exact kernel.

We proceed in four steps:

1. Showing we can control ||Σ_V − Σ_N|| with high probability;
2. Obtaining bounds on the eigenvalues of Σ_V and Σ_N when ||Σ_V − Σ_N|| < δ||Σ||;
3. Using (a) and (b) to control the KL;
4. Showing how to choose δ as a function of β to achieve uniform control of ||P (θ, ·) − P_ϵ (θ, ·)||_{TV}.

Part (a): Control of ||Σ_V − Σ_N||.

The sample covariance matrix of z_i = ω_{i/2} x_i is X′ΩX. The z_i are iid given β, since

\[
f_z(z_1, ..., z_N | \bar{β}) = \int \left( \prod_{i=1}^N f_z(z_i | x_i, \bar{β}) g_{x_i}(x_i | \bar{β}) \right) dx_1 ... dx_N,
\]

39
we show that \( \text{E} [\omega^{1/2}x] = 0 \). Since \( x \sim f_x(x; \alpha) \),
\[
\text{E} [\omega^{1/2}x] = \int_{x \in \mathbb{R}^p} \int_{\omega \in \mathbb{R}^+} \omega^{1/2} xe^{-(x, \beta)^2/\omega} f_\omega(\omega) \cosh \left( \frac{x, \beta}{2} \right) f_x(x; \alpha) d\omega dx,
\]
where \( f_\omega(\omega) \) is the PG(1, 0) density and \( f_x(x; \alpha) \) is symmetric about 0, so the expectation is zero. Since \( f_\sqrt{\omega}(y) = 2y f_\omega(y^2) \) is the density of \( \sqrt{\omega} \), and \( f_\omega(y) \) is log-concave by Lemma G.1, \( f_\sqrt{\omega}(y) \) is log-concave. Since the product of log-concave functions is log-concave, and \( f_x(x; \alpha) \) is log-concave by assumption, the distribution of \( \omega^{1/2}x \) is log-concave. This allows us to apply the following Theorem from Adamczak et al. [2010].

**Theorem G.1** (Adamczak 2010, Theorem 4.1). Let \( Z_1, \ldots, Z_N \) be i.i.d. random vectors distributed according to an isotropic, log-concave probability measure on \( \mathbb{R}^p \). For every \( \delta \in (0, 1) \) and \( M > 1 \) there exists \( C(\delta, M) > 0 \) such that if \( C(\delta, M) p \leq N \), then with probability at least \( 1 - e^{-cM\sqrt{p}} \),
\[
||\Sigma_N - I_p|| \leq \delta,
\]
where \( c > 0 \) is an absolute constant and \( \Sigma_N \) is the sample covariance matrix based on \( N \) samples. Moreover, one can take \( C(\delta, M) = CM^4\delta^{-2} \log^2(2M\delta^{-2}) \), where \( C \) is an absolute constant.

Here, \( ||\Sigma|| \) is the spectral norm of \( \Sigma \), i.e.
\[
\sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{||\Sigma x||_2}{||x||_2}
\]

Adamczak et al. [2010] notes that when the distribution of \( Z \) is not isotropic but does have zero mean, we instead have
\[
||\Sigma_N - \Sigma|| < \delta||\Sigma||
\]
with the same probability.

Note that at best we can achieve probabilities on the order of \( 1 - e^{-N^{1/4}} \), since \( C(\delta, M) \) grows like \( M^4 \) up to a log factor. Now, fix a \( q \) and suppose that for \( |V| < N \) we can achieve \( ||\Sigma_N - \Sigma|| < \frac{\delta}{2}||\Sigma|| \) with probability at least \( 1 - q \), for some value of \( \delta \) to be determined subsequently. Then with probability \( (1 - q)^2 \) we have \( ||\Sigma_N - \Sigma|| < \frac{\delta}{2}||\Sigma|| \), and by the triangle inequality, with the same probability we have \( ||\Sigma_N - \Sigma_V|| < \delta ||\Sigma|| \). We now show that for sufficiently small \( \delta \), this allows us to bound the eigenvalues of \( \Sigma_N \) and \( \Sigma_V \).

**Part (b) : Control of eigenvalues of \( \Sigma_N \) and \( \Sigma_V \)**
If \( ||\Sigma_N - \Sigma|| < \frac{\delta}{2}||\Sigma|| \) then
\[
||\Sigma_N + B^{-1}/N - (\Sigma + B^{-1}/N)|| = ||\Sigma_N - \Sigma|| \leq \frac{\delta}{2}||\Sigma||.
\]
Now, use $||\Sigma_N - \Sigma||_F^2 \leq p \, ||\Sigma_N - \Sigma||^2$, and that $\Sigma, \Sigma_N$ are Hermitian, and apply the Hoffman-Weilandt inequality \cite{bhatia2013matrix, hoffman1953eigenvalues, tao2015topics}, which ensures existence of a permutation $\rho$ of the eigenvalues of $\Sigma_N$ such that

$$
\sum_{j=1}^{p}(\lambda_{\rho(j)}(\Sigma_N) - \lambda_j(\Sigma))^2 < ||\Sigma_N - \Sigma||_F^2 \leq (\sqrt{p}||\Sigma_N - \Sigma||)^2 \leq p \delta^2 ||\Sigma||^2,
$$

where $\lambda_j(\Sigma)$ is the $j$th eigenvalue of the matrix $\Sigma$. So there exists a $j$ such that

$$
(\lambda_{\max}(\Sigma_N) - \lambda_j(\Sigma))^2 < p \delta^2 ||\Sigma||^2,
$$

(32)

where $\lambda_{\max}(\Sigma_N)$ is the largest eigenvalue of $\Sigma_N$. This implies that

$$
\lambda_{\max}(\Sigma_N) < \lambda_{\max}(\Sigma) + \sqrt{p} \delta ||\Sigma||.
$$

(33)

This is immediate if $j = 1$ in (32). If $j > 1$ in (32), then we must have (33), since otherwise

$$
(\lambda_{\max}(\Sigma_N) - \lambda_j(\Sigma))^2 \geq (\lambda_{\max}(\Sigma) + \sqrt{p} \delta ||\Sigma|| - \lambda_j(\Sigma))^2 \geq p \delta^2 ||\Sigma||^2.
$$

Furthermore, there exists a $j'$ for which

$$
(\lambda_{\min}(\Sigma_N) - \lambda_{j'}(\Sigma))^2 < p \delta^2 ||\Sigma||^2,
$$

with $\lambda_{\min}(\Sigma_N)$ the smallest eigenvalue of $\Sigma_N$, implying

$$
\lambda_{\min}(\Sigma_N) > \lambda_{\min}(\Sigma) - \sqrt{p} \delta ||\Sigma||
$$

by analogous argument. So if

$$
\delta < p^{-1/2} \frac{\lambda_{\min}(\Sigma)}{(\lambda_{\max}(\Sigma) + \lambda_{\min}(\Sigma))},
$$

we have $\lambda_{\min}(\Sigma_N) > \lambda_{\min}(\Sigma)/2$, ensuring the smallest eigenvalue of $\Sigma_N$ is bounded away from zero, and $\lambda_{\max}(\Sigma_N) < \lambda_{\max}(\Sigma) + \lambda_{\min}(\Sigma)/2$. Now, put $\ell_{\max}(\beta) = \lambda_{\max}(\Sigma(\beta)) + \lambda_{\min}(\Sigma(\beta))/2$ and $\ell_{\min}(\beta) = \lambda_{\min}(\Sigma(\beta))/2$. With $B = \eta I_p$,

$$
\lambda_{\min}((\Sigma_N + B^{-1}/N)^{-1}) \geq \frac{1}{\ell_{\max}(\beta) + (N\eta)^{-1}}, \quad \lambda_{\max}((\Sigma_N + B^{-1}/N)^{-1}) \leq \frac{1}{\ell_{\min}(\beta) + (N\eta)^{-1}}
$$

$$
\lambda_{\min}((\Sigma_V + B^{-1}/N)^{-1}) \geq \frac{1}{\ell_{\max}(\beta) + (N\eta)^{-1}}, \quad \lambda_{\max}((\Sigma_V + B^{-1}/N)^{-1}) \leq \frac{1}{\ell_{\min}(\beta) + (N\eta)^{-1}},
$$

where the result for $\Sigma_V$ follows because we also have $||\Sigma_V - \Sigma|| < \frac{\delta}{2} ||\Sigma||$. 

41
Part (c): Control of KL Divergence

Now we show control of $Q$, assuming that $||\Sigma - \Sigma_N|| \leq \delta||\Sigma||$.

\[
Q = \left( \left( \frac{N}{|V|} X_V^\top \Omega_V X_V + B^{-1} \right)^{-1} X' \kappa - (X' \Omega X + B^{-1})^{-1} X' \kappa \right)' (X' \Omega X + B^{-1})
\]
\[
= \left( \left( \frac{N}{|V|} X_V^\top \Omega_V X_V + B^{-1} \right)^{-1} X' \kappa - (X' \Omega X + B^{-1})^{-1} X' \kappa \right)' \left( \frac{1}{N} X' \Omega X + B^{-1} \right)
\]
\[
= \frac{1}{N} \left( \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right)^{-1} X' \kappa - \left( \frac{1}{N} X' \Omega X + B^{-1} \right)^{-1} X' \kappa \right)'
\]
\[
\leq \frac{\ell_{\max} + (N \eta)^{-1}}{N} \left\| \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right)^{-1} X' \kappa - \left( \frac{1}{N} X' \Omega X + B^{-1} \right)^{-1} X' \kappa \right\|^2
\]
\[
\leq \frac{\ell_{\max} + (N \eta)^{-1}}{N \ell_{\min} + (N \eta)^{-1}} \left\| \left( I - \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right) \right) \left( \frac{1}{N} X' \Omega X + B^{-1} \right)^{-1} \right\|^2 \left\| X' \kappa \right\|^2
\]
\[
\leq \frac{\ell_{\max} + (N \eta)^{-1}}{N \ell_{\min} + (N \eta)^{-1}} \left\| I - \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right) \right\|^2 \left\| \frac{\delta^2 ||X' \kappa||^2 \ell_{\max}^2}{\ell_{\min} + (N \eta)^{-1}} \right\|^2 \left\| X' \kappa \right\|^2
\]
\[
\leq \frac{\ell_{\max} + (N \eta)^{-1}}{N \ell_{\min} + (N \eta)^{-1}} \left( \frac{\delta^2 \ell_{\max}^2}{\ell_{\min} + (N \eta)^{-1}} \right) \left( \frac{\ell_{\max} + (N \eta)^{-1}}{\ell_{\min} + (N \eta)^{-1}} \right)^2 \left\| I - \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right) \right\|^2 \left\| \frac{\delta^2 \ell_{\max}^2}{\ell_{\min} + (N \eta)^{-1}} \right\|^2 \left\| X' \kappa \right\|^2
\]

where various steps used Cauchy-Schwartz, assume $X$ is standardized to unit variance, $\kappa_i \in \{-1/2, 1/2\}$, $||\Sigma_V - \Sigma_N|| < \delta \ell_{\max} + (N \eta)^{-1}$, and

\[
\left\| I - \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right) \left( \frac{1}{N} X' \Omega X + B^{-1} \right)^{-1} \right\| \leq \left\| \frac{1}{N} X' \Omega X + B^{-1} \right\| \left\| \left( \frac{1}{N} X_V^\top \Omega_V X_V + B^{-1} \right)^{-1} \right\|.
\]
To bound the other terms in the KL, first note that
\[
\tr \left( (X'\Omega X + B^{-1}) \left( \frac{N}{|V|} X'_V \Omega_V X_V + B^{-1} \right)^{-1} \right) - p = \\
= \tr \left( N \left( \Sigma_N + \frac{B^{-1}}{N} \right) \frac{1}{N} \left( \Sigma_V + \frac{B^{-1}}{N} \right)^{-1} - I \right) \\
= \tr \left( (\Sigma_N - \Sigma_V) \left( \Sigma_V + \frac{B^{-1}}{N} \right)^{-1} \right) \\
\leq \lambda_{\text{max}}(\Sigma_N - \Sigma_V) \tr ((\Sigma_V + B^{-1}/N)^{-1}) \\
\leq \frac{p \delta \ell_{\text{max}}(\beta)}{\ell_{\min}(\beta) + (N\eta)^{-1}}.
\]
Further, from Lemma B.2 in [Pati et al. 2014], since $S_V$ and $S_N$ are both positive definite for $|V| > p,$
\[
\log |S_N S_V^{-1}| < \tr\left(S_N^{-1} S_V - I\right).
\]
So putting all of the bounds together,
\[
\KL\left( \mathcal{N}(\cdot; S_V X', S_V) \mid \mid \mathcal{N}(\cdot; S_N X', S_N) \right) \leq \frac{1}{2} \left( \ell_{\text{max}}(\beta) + (N\eta)^{-1} \right) \left( \frac{\delta^2 p \ell_{\text{max}}(\beta)^2}{4(\ell_{\min}(\beta) + (N\eta)^{-1})^2} \right) \\
+ \frac{p \delta \ell_{\text{max}}(\beta)}{\ell_{\min}(\beta) + (N\eta)^{-1}} \\
\leq \frac{\delta^2 p \ell_{\text{max}}(\beta)^3}{8 \ell_{\min}(\beta)^4} + \frac{p \delta \ell_{\text{max}}(\beta)}{\ell_{\min}(\beta)}.
\]

**Part (d): Uniform control of $||P(\theta, \cdot) - P_\epsilon(\theta, \cdot)||_{TV}$**

Notice that
\[
\frac{\lambda_{\text{min}}(\Sigma)}{\lambda_{\text{max}}(\Sigma) + \lambda_{\text{min}}(\Sigma)} = \frac{2 \ell_{\min}(\beta)}{\ell_{\min}(\beta) + \ell_{\max}(\beta)} > \frac{\ell_{\min}(\beta)}{\ell_{\max}(\beta)}.
\]
So, put
\[
\delta = \frac{2 \sqrt{2} \epsilon p^{-1/2} \ell_{\min}(\beta)^2}{\ell_{\max}(\beta)^{3/2}} \land \frac{\epsilon^2 \ell_{\min}(\beta)}{p \ell_{\max}(\beta)} \land \frac{\epsilon p^{-1/2} \lambda_{\min}(\Sigma)}{\lambda_{\min}(\beta) + \lambda_{\max}(\beta)}.
\]
with $0 < \epsilon < 1,$ where now we explicitly indicate the dependence of $\lambda_{\text{min}}(\Sigma), \lambda_{\text{max}}(\Sigma)$ on $\beta$ through the notation $\lambda_{\text{min}}(\beta), \lambda_{\text{max}}(\beta),$ and the final term ensures we satisfy the earlier condition on $\delta.$ Thus we obtain
\[
\KL\left( \mathcal{N}(\cdot; S_V X', S_V) \mid \mid \mathcal{N}(\cdot; S_N X', S_N) \right) \leq 2 \epsilon^2.
\]
Thus by adaptively choosing $\delta$ as a function of $\beta,$ we obtain approximation error that does not depend on $\beta.$ Now, apply Pinsker’s inequality, so that
\[
|| \mathcal{N}(\cdot; S_V X', S_V) - \mathcal{N}(\cdot; S_N X', S_N) ||_{TV} \leq \epsilon
\]
with probability at least
\[(1 - e^{-cM^2\epsilon})^2\]
whenever \(|V| \geq pCM^4\delta^{-2} \log^2(2M^2\delta^{-2})

H Proof of Theorem 3.2

The results in this section concern the model in (22) with priors in (23). The transition kernel \(P\) is induced by the marginal sampler defined in section 3.4 and the approximating kernel \(P_e\) substitutes \(\Sigma_e = U^e\Lambda^eU^e\) for \(\Sigma\) where \(U^e\) is \(n \times r\), \(\Lambda^e\) is \(r \times r\), and \(r \leq n\).

H.1 Result for predictive \(p(f \mid \theta)\)

First we show that for every \(\epsilon \in (0, 1)\) there exists a \(\delta\) depending on the state \(\theta = (\sigma^2, \tau^2, \phi)\) such that \(||\Sigma - \Sigma_e|| < \delta\) implies
\[\|p(f \mid \theta) - p_e(f \mid \theta)\|_{TV} < \epsilon,\]
where \(f\) is the latent Gaussian process in (22), \(p(f \mid \theta)\) is its full conditional in the exact MCMC algorithm (we repress the dependence on \(y\) for notational brevity), and \(p_e(f \mid \theta)\) is its full conditional in the approximate sampler. The strategy is to show a bound on
\[\text{KL}(p(f \mid \theta) \| p_e(f \mid \theta)) = \frac{1}{2} \left( \text{tr}((\Psi_e)^{-1}\Psi) - n + \log(||\Psi_e^{-1}\Psi||) + y'(\Psi_e - \Psi)^{-1}(\Psi_e - \Psi)y \right)\]
where \(\Psi = (\tau^2\Sigma + \sigma^2I)^{-1}\) and \(\Psi_e = (\tau^2\Sigma_e + \sigma^2I)^{-1}\), then use Pinsker’s inequality. We now bound each term separately following the proof of Theorem 3.1.

The eigenvalues of \(\Psi\) and \(\Psi_e\) satisfy
\[
\lambda_{\min}(\Psi) > \frac{1}{\tau^2\lambda_{\max}(\Sigma) + \sigma^2}, \quad \lambda_{\max}(\Psi) < \frac{1}{\sigma^2}, \\
\lambda_{\min}(\Psi_e) > \frac{1}{\tau^2\lambda_{\max}(\Sigma_e) + \sigma^2}, \quad \lambda_{\max}(\Psi_e) < \frac{1}{\sigma^2}.
\]

We assume that the approximation achieves \(||\Sigma_e - \Sigma||_F < \delta\) with probability \(1 - 10^{-d}\). So then using the strategy in the proof of Theorem 3.1,

\[
Q = (\Psi_e y - \Psi y)^{-1}(\Psi_e y - \Psi y) \\
\leqslant (\tau^2\lambda_{\max}(\Sigma_e) + \sigma^2)||\Psi_e y - \Psi y||^2 \\
\leqslant (\tau^2\lambda_{\max}(\Sigma_e) + \sigma^2)||\Psi||^2||\Psi^{-1}\Psi_e - I||^2 ||y||^2 \\
\leqslant \frac{\tau^2\lambda_{\max}(\Sigma_e) + \sigma^2}{\sigma^4}||\Psi^{-1}\Psi_e - I||^2 ||y||^2 \\
\leqslant \frac{\tau^2\lambda_{\max}(\Sigma_e) + \sigma^2}{\sigma^4}||\Sigma - \Sigma_e||^2 ||\Psi_e||^2 ||y||^2 \\
\leqslant \frac{\tau^4\delta^2||y||^2(\tau^2\lambda_{\max}(\Sigma_e) + \sigma^2)}{\sigma^8}
\]
\[
\leq \frac{\tau^4 \delta^2 n (\tau^2 \lambda_{\text{max}}(\Sigma_c) + \sigma^2)}{\sigma^4},
\]
where we used that \( y \) is standardized to unit variance. Now since
\[
\text{tr} (\Psi^{-1} \Psi) - n = \text{tr} ((\Psi^{-1}_c - \Psi^{-1}) \Psi) \leq \|\Psi\| \text{tr} ((\Psi^{-1}_c - \Psi^{-1}))
\leq \frac{n \tau^2 \delta}{\sigma^2},
\]
applying Lemma B.2 in [Pati et al. 2014], we obtain the KL bound
\[
\text{KL} (p(f \mid \theta) \mid p_c(f \mid \theta)) \leq \frac{\tau^4 \delta^2 n (\tau^2 \lambda_{\text{max}}(\Sigma_c) + \sigma^2)}{2 \sigma^4} + \frac{n \tau^2 \delta}{\sigma^2}.
\]
Apply Pinsker’s inequality and get
\[
\|p(f \mid \theta) - p_c(f \mid \theta)\|_{\text{TV}} \leq \sqrt{n \left( \frac{\tau^4 \delta^2 (\tau^2 \lambda_{\text{max}}(\Sigma_c) + \sigma^2)}{4 \sigma^4} + \frac{\delta \tau^2}{2 \sigma^2} \right)}.
\]
So choose
\[
\delta = \frac{\epsilon^2 \sigma^4}{\tau^2 \sqrt{n (\tau^2 \lambda_{\text{max}}(\Sigma_c) + \sigma^2)}} \wedge \frac{\epsilon^2 \sigma^2}{n \tau^2}
\]
for \( 0 < \epsilon < 1 \) to achieve TV error of \( \epsilon \). By adapting the required accuracy \( \delta \) to the state, and noting that one can always achieve \( \delta = 0 \) by utilizing the exact \( \Sigma \) so that no value of \( \delta \) is unachievable, this is sufficient to show that the total variation error can be controlled uniformly. Note we did not need the assumption that \( \Sigma_c \) is a partial eigendecomposition; this will be used below.

**H.2 Result for \( \mathcal{P}_\epsilon \)**

We first prove a lemma that will be used to obtain the main result. We will in general use \( \theta_* \) to represent the proposal value in Metropolis-Hastings algorithms and \( \theta \) to represent the current state.

**Lemma H.1.** Consider transition kernels \( \mathcal{P} (\theta, \cdot), \mathcal{P}_\epsilon (\theta, \cdot) \) constructed by Metropolis-Hastings algorithms with identical proposal distributions and acceptance probabilities \( p(\theta \to \theta_*), p_c(\theta \to \theta_*) \) for any \( \theta, \theta_* \in \Theta \). If
\[
\sup_{\theta_* \in \Theta} \sup_{\theta \in \Theta} |p(\theta \to \theta_*) - p_c(\theta \to \theta_*)| < \frac{\epsilon}{2},
\]
then
\[
\sup_{\theta} \|\mathcal{P} (\theta, \cdot) - \mathcal{P}_\epsilon (\theta, \cdot)\|_{\text{TV}} < \epsilon.
\]

**Proof.** Let \( Q(\theta; d\theta_*) \) denote the proposal distribution, which may depend on the current state \( \theta \). Then
\[
\mathcal{P}(\theta, \theta_*) = \int p(\theta \to \theta_*) Q(\theta; d\theta_*) + \delta_\theta(\theta_*) \int (1 - p(\theta \to \theta_*)) Q(\theta; d\theta_*)
\]
\[ P_c(\theta, \theta_*) = \int p_c(\theta \to \theta_*) Q(\theta; d\theta_*) + \delta_0(\theta_*) \int (1 - p_c(\theta \to \theta_*) ) Q(\theta; d\theta_*) \]

Hence, we have

\[
\sup_{\theta} ||P (\theta, \cdot) - P_c (\theta, \cdot)||_{TV} \\
= \sup_{\theta \in \Theta} \sup_{A \subset \Theta} \left| \int_A p(\theta \to \theta_*) Q(\theta; d\theta_*) + 1_{\{\theta \in A\}} \int_A (1 - p(\theta \to \theta_*) ) Q(\theta; d\theta_*) \right| \\
- \int_\Theta p_c(\theta \to \theta_*) Q(\theta; d\theta_*) - 1_{\{\theta \in A\}} \int_A (1 - p_c(\theta \to \theta_*) ) Q(\theta; d\theta_*) \\
\leq \sup_{\theta \in \Theta} \left| \int_\Theta p(\theta \to \theta_*) Q(\theta; d\theta_*) + \int_\Theta (1 - p(\theta \to \theta_*) ) Q(\theta; d\theta_*) \right| \\
- \int_\Theta p_c(\theta \to \theta_*) Q(\theta; d\theta_*) - \int_\Theta (1 - p_c(\theta \to \theta_*) ) Q(\theta; d\theta_*) \\
\leq \sup_{\theta \in \Theta} \left| \int_\Theta [p(\theta \to \theta_*) - p_c(\theta \to \theta_*)] Q(\theta; d\theta_*) \right| + \left| \int_\Theta [p_c(\theta \to \theta_*) - p(\theta \to \theta_*)] Q(\theta; d\theta_*) \right| \\
\leq \sup_{\theta \in \Theta} \left| \int_\Theta [p(\theta \to \theta_*) - p_c(\theta \to \theta_*)] Q(\theta; d\theta_*) \right| + \left| \int_\Theta [p_c(\theta \to \theta_*) - p(\theta \to \theta_*)] Q(\theta; d\theta_*) \right| \\
\leq \epsilon \int_\Theta Q(\theta; d\theta_*) \leq \epsilon \\
\]

\[ \square \]

H.2.1 Main result: approximation error for GP MH steps

We now show that for every \( \epsilon > 0 \), the kernel \( P_c \) that replaces \( \Sigma \) with \( \Sigma_\epsilon \), achieving \( ||\Sigma - \Sigma_\epsilon|| < \delta \) with probability \( 1 - q \), satisfies Assumption 2.2 (also with probability \( 1 - q \)). This result uses the additional assumption that \( \Sigma_\epsilon \) is a partial eigendecomposition. To simplify the exposition and reduce length, the result is obtained for a joint Metropolis-Hastings step for \( (\sigma^2, \tau^2) \). A similar result could be obtained for the sequential Metropolis-Hastings steps in the marginal sampler described in Section 3.4 by appropriately re-defining the acceptance probability. Applying lemma H.1, we need only control \( p_{\text{sup}} \). The absolute difference in MH acceptance probabilities for the marginal sampler is

\[ D_c(\theta, \theta_*) = |p_c(\theta \to \theta_*) - p(\theta \to \theta_*)| \]

\[ = \left| \frac{L_c(y \mid \theta_*) p(\theta_*) q(\theta \mid \theta_*)}{L_c(y \mid \theta) p(\theta) q(\theta_*) \mid \theta} \right| - \left| \frac{L(y \mid \theta_*) p(\theta_*) q(\theta \mid \theta_*)}{L(y \mid \theta) p(\theta) q(\theta_*) \mid \theta} \right| \]

\[ = |(r_c(\theta \to \theta_*) \wedge 1) - (r(\theta \to \theta_*) \wedge 1)|. \]

Initially focus on the case where both \( r_c(\theta \to \theta_*) \) and \( r(\theta \to \theta_*) \) are less than one, and set \( M(\theta, \theta_*) = \frac{p(\theta_*) q(\theta \mid \theta_*)}{p(\theta) q(\theta_*) \mid \theta} \).

Then

\[ D_c(\theta, \theta_*) = M(\theta, \theta_*) \left| \frac{L_c(y \mid \theta_*)}{L_c(y \mid \theta)} - \frac{L(y \mid \theta_*)}{L(y \mid \theta)} \right| \]

46
\[
M(\theta, \theta_a) = \frac{\tau^2 \Sigma + \sigma^2 I}{|2\pi(\tau^2 \Sigma + \sigma^2 I)|^{-1/2}} \exp(-y'(\tau^2 \Sigma + \sigma^2 I)^{-1}y/2) \\
- \frac{2\pi(\tau^2 \Sigma + \sigma^2 I)}{|2\pi(\tau^2 \Sigma + \sigma^2 I)|^{-1/2}} \exp(-(y'(\tau^2 \Sigma + \sigma^2 I)^{-1}y/2) \\
M(\theta, \theta_a) \left( \frac{\prod_{i=1}^{n} \tau_2^2 \lambda_i + \sigma^2}{|2\pi(\tau^2 \Sigma + \sigma^2 I)|^{-1/2}} \exp(-y'(\tau^2 \Sigma + \sigma^2 I)^{-1}y/2) \right) \\
- \frac{2\pi(\tau^2 \Sigma + \sigma^2 I)}{|2\pi(\tau^2 \Sigma + \sigma^2 I)|^{-1/2}} \exp(-(y'(\tau^2 \Sigma + \sigma^2 I)^{-1}y/2) \\
M(\theta, \theta_a) \left( \frac{\prod_{i=1}^{n} \tau_2^2 \lambda_i + \sigma^2}{|2\pi(\tau^2 \Sigma + \sigma^2 I)|^{-1/2}} \exp(-y'(\tau^2 \Sigma + \sigma^2 I)^{-1}y/2) \right)
\]

Now use that \( \Sigma \) is a rank \( r \) partial eigendecomposition of \( \Sigma \) satisfying \( ||\Sigma - \Sigma||_F < \delta \), implying the following

\[
\tau^2 \Sigma + \sigma^2 I = U(\tau^2 \Lambda + \sigma^2 I)U', \quad \tau^2 \Sigma + \sigma^2 I = U(\tau^2 \Lambda + \sigma^2 I)U'
\]

where \( \Lambda_i = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) \), and \( \lambda_i \) is the \( i \)th eigenvalue of \( \Sigma \). Now put \( y_U = yU \), with \( i \)th entry \( y_{U,i} \), and obtain \( D_\epsilon(\theta, \theta_a) = \)

\[
M(\theta, \theta_a) \left( \frac{1}{2} \sum_{i=1}^{n} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \right) \\
- \frac{1}{2} \sum_{i=r+1}^{n} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \\
M(\theta, \theta_a) \left( \frac{1}{2} \sum_{i=1}^{r} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \right)
\]

Put

\[
M_1(\theta, \theta_a) = M(\theta, \theta_a) \exp \left( -\frac{1}{2} \sum_{i=1}^{r} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \right)
\]

and

\[
D_\epsilon(\theta, \theta_a) = M_1(\theta, \theta_a) \left( \frac{1}{2} \sum_{i=1}^{n} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \right)
\]

\[
- \exp \left( -\frac{1}{2} \sum_{i=r+1}^{n} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \right) \\
- \exp \left( -\frac{1}{2} \sum_{i=1}^{r} \log \frac{\tau_2^2 \lambda_i + \sigma^2}{\tau_2^2 \lambda_i + \sigma^2} \right)
\]

\[
47
\]
Call the term inside the absolute value \( \Delta(\delta, \theta, \theta_\ast) \), and simplify to obtain

\[
\Delta(\delta, \theta, \theta_\ast) = \left| \exp \left( \frac{1}{2} \sum_{i=r+1}^{n} \frac{1}{\sigma_{\ast i}^2} \left[ 1 - \frac{1}{\sigma_i^2} \right] \right) - 1 \right| \sum_{i=r+1}^{n} \log \frac{\sigma_{\ast i}^2}{\sigma_i^2}
\]

\[
- \exp \left( \frac{1}{2} \sum_{i=r+1}^{n} \frac{1}{\sigma_{\ast i}^2} \left[ 1 - \frac{1}{\sigma_i^2} \right] \right) \left[ \frac{1}{\tau^2 \lambda_i + \sigma_{\ast i}^2} - 1 \right] \sum_{i=r+1}^{n} \log \frac{\tau^2 \lambda_i + \sigma_{\ast i}^2}{\tau^2 \lambda_i + \sigma_i^2}
\]

\[
\leq \exp \left( \frac{n-r}{2} \left( \frac{\sigma^2 - \sigma_{\ast}^2}{\sigma_{\ast}^2 \sigma^2} - \log \frac{\sigma_{\ast}^2}{\sigma^2} \right) \right)
\]

\[
- \exp \left( \frac{n-r}{2} \left[ \frac{\tau^2 \delta + \sigma^2 - \tau^2 \delta - \sigma_{\ast}^2}{(\tau^2 \delta + \sigma_{\ast}^2)(\tau^2 \delta + \sigma^2)} - \log \frac{\tau^2 \delta + \sigma_{\ast}^2}{\tau^2 \delta + \sigma^2} \right] \right).
\]

Taking \( \delta \to 0 \), \( \Delta(\delta, \theta, \theta_\ast) \) can be made arbitrarily small.

Finally, because the prior on \( \phi \) is finitely supported and depends only on likelihood ratios of the same form as those considered above, control of the approximation error for sampling of \( \phi \) follows easily. Thus, the following algorithm achieves sup\( P_\theta(\theta_\ast) - P_\ast(\theta_\ast) \)TV < \( \epsilon \): (a) Take a draw from \( q(\theta_\ast \mid \theta) \); (b) Choose \( \delta \) such that \( \Delta(\delta, \theta, \theta_\ast) < \frac{\epsilon}{2\max\{r(\theta \rightarrow \theta_\ast)\}} \); (c) Compute \( r_\ast(\theta \rightarrow \theta_\ast) \); (d) Use this quantity in the MH acceptance decision; and (e) Sample \( \phi \) from its discrete full conditional distribution.

Although the only case considered above was that where \( r_\ast(\theta \rightarrow \theta_\ast) < 1 \), note that if \( |r_\ast(\theta \rightarrow \theta_\ast) - r(\theta \rightarrow \theta_\ast)| < \frac{\epsilon}{2} \), then

\[
|(1 \land r_\ast(\theta \rightarrow \theta_\ast)) - (1 \land r(\theta \rightarrow \theta_\ast))| < \frac{\epsilon}{2}.
\]

Noting that \( \delta = 0 \) is always achievable by taking \( \Sigma_\epsilon = \Sigma \), this is sufficient to control the approximation error everywhere in the state space.

# I Simulation study: accuracy of approximate eigendecompositions

In each simulation, 1000 points are generated in \( \mathbb{R}^k \) and pairwise distances computed. A grid of 100 values of \( \phi \) is constructed, corresponding to evenly spaced values such that the minimum value of \( \phi \) corresponds to a correlation of 0.99 at the maximum observed distance, and the maximum value of \( \phi \) corresponds to a correlation of 0.01. In every simulation, \( \tau^2 \) is set to one. Four approaches to generating pairwise distances were considered: (1) the points \( x \) were evenly distributed on the interval \([0.001, 1]\) (Grid case); (2) the points \( x \) were sampled uniformly on the unit interval; (3) the points \( x \) were sampled from Gamma \((1, 1)\); and, (4) the points are vectors in \( \mathbb{R}^5 \) with independent standard normal entries. Naturally, the first three cases correspond to approximately low-rank \( \Sigma \), while (4) corresponds to a \( \Sigma \) with a much more slowly decaying spectrum.

To assess the accuracy of the approximate partial eigendecomposition, both a complete eigendecomposition and an approximate partial eigendecomposition were computed, producing \( \Sigma = U \Lambda U' \) and \( \Sigma_\ast = U_\ast \Lambda_\ast U_\ast' \), where \( U \) is \( n \times m \) and \( \Lambda \) \( m \times m \) with \( m < n \). The approximate partial eigendecomposition was computed using Algorithms 4.2 and 5.5 of [Halko et al. 2011] with \( \delta_\ast = 0.001 \). Let \( \Lambda^\ast \) and \( U^\ast \) be the diagonal matrix consisting of the largest \( m \) eigenvalues of \( \Sigma \) and the corresponding
Table 9: Results of simulation study for approximation error using approximate eigendecomposition. The median, maximum, and minimum values of $C$, $R$, and $F$ are shown across the 100 values of $\phi$ specified in the text.

|        | $R(\Lambda^*, \Lambda_\subset)$ | $F(\mathbb{U}^*, U_\subset)$ | $C(\mathbb{U}^*, U_\subset)$ |
|--------|---------------------------------|---------------------------------|---------------------------------|
|        | median  | max    | min    | median  | max    | min    | median  | max    | min    |
| Grid   | 2.55e-12| 3.562e-10| 8.21e-16| 1.54e-07| 7.42e-15| 0.00e+00| 1.00    | 1.00    | 1.00    |
| Uniform| 1.70e-12| 4.41e-10| 1.04e-17| 1.36e-07| 1.19e-15| 0.00e+00| 1.00    | 1.00    | 1.00    |
| Gamma  | 2.56e-12| 1.41e-09| 1.14e-16| 2.60e-07| 2.54e-15| 0.00e+00| 1.00    | 1.00    | 1.00    |
| Normal | 9.36e-09| 1.24e-08| 1.14e-09| 3.87e-01| 3.65e-02| 0.00e+00| 0.99    | 1.00    | 0.90    |

$m$ eigenvectors, respectively. We then compute

\[ R(\Lambda^*, \Lambda_\subset) = \sqrt{\sum_{i=1}^{m}(\lambda^*_i - \lambda_{\subset,i})^2}, \quad F(\mathbb{U}^*, U_\subset) = ||I - U'_\subset \mathbb{U}^*||_F/\sqrt{n}, \quad \text{and} \]

\[ C(\mathbb{U}^*, U_\subset) = \text{Corr} (y, U_\subset (U'_\subset U_\subset)^{-1}U'_\subset y), \]

where $y = \mathbb{U}^* \beta$ for $\beta_j \sim N(0, 1)$ a random $m \times 1$ vector with independent standard normal entries. Essentially, $R$ measures the quality of the approximation to the eigenvalues and $F$ and $C$ measure the quality of approximation to the column space of $\mathbb{U}^*$. Table 9 shows results. For the Grid, Uniform, and Gamma cases, the approximation is extremely accurate; the approximate eigendecomposition is almost identical to the partial eigendecomposition. For the Normal case, the approximation to the eigenvalues is still very accurate, but there is noticeable error in the column space approximation. It should be noted that for the first three cases, typical values of $m$ ranged from 10 to 50, whereas in the Normal case, $m$ is nearly 500 for most values of $\phi$. In general, we expect the approximate eigendecomposition to be less accurate in cases where the spectrum decays very slowly, so the results in Table 9 are not surprising.

References

Radoslaw Adamczak, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010.

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

P Alquier, N Friel, R Everitt, and A Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 25(1):1–19, 2014.

K.B Athreya and P. Ney. A new approach to the limit theory of recurrent markov chains. *Trans. Amer. Math. Soc.*, 245:493–501, 1978.

Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 21–30. Morgan Kaufmann Publishers Inc., 1999.
Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic theory*, 26(2):445–469, 2005.

P Baldi, P Sadowski, and D Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:1–9, 2014.

Anjishnu Banerjee, David B Dunson, and Surya T Tokdar. Efficient gaussian process regression for large datasets. *Biometrika*, 100:75–89, 2013.

Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 405–413, 2014.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*, 2015.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Abhishek Bhattacharya and David B Dunson. Nonparametric bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 102(2):851–865, 2010.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, page asq017, 2010.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. Preprint, June 2015.

Yutian Chen and Zoubin Ghahramani. Subsampling-based approximate Monte Carlo for discrete distributions. *arXiv preprint arXiv:1506.09039*, 2015.

Yutian Chen, Vikash Mansinghka, and Zoubin Ghahramani. Sublinear approximate inference for probabilistic programs. *arXiv preprint arXiv:1411.1690*, 2014.

Hee Min Choi and James P Hobert. The poly-a-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

Peter J Diggle and Richard J Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227, 1984.

David B Dunson and Chuanhua Xing. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487), 2009.

Déborah Ferré, Loïc Hervé, James LEDoux, et al. Regular perturbation of V-geometrically ergodic Markov chains. *Journal of applied probability*, 50(1):184–194, 2013.

Andrew O Finley, Huiyan Sang, Sudipto Banerjee, and Alan E Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884, 2009.
Nial Friel, Antonietta Mira, Chris J Oates, et al. Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. *Bayesian Analysis*, 2015.

Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2 edition, 2006.

Gene H Golub. Some modified matrix eigenvalue problems. *Siam Review*, 15(2):318–334, 1973.

Peter J Green, Krzysztof Latuszyński, Marcelo Pereyra, and Christian P Robert. Bayesian computation: a perspective on the current state, and sampling backwards and forwards. *arXiv preprint arXiv:1502.01148*, 2015.

Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian conditional density filtering. *arXiv preprint arXiv:1401.3632*, 2014.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Alan J Hoffman, Helmut W Wielandt, et al. The variation of the spectrum of a normal matrix. *Duke Math. J.*, 20(1):37–39, 1953.

John Hughes and Murali Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159, 2013.

Aldéric Joulin, Yann Ollivier, et al. Curvature, concentration and error estimates for markov chain monte carlo. *The Annals of Probability*, 38(6):2418–2442, 2010.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in mcmc land: Cutting the metropolis-hastings budget. *arXiv preprint arXiv:1304.5299*, 2013.

Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Florian Maire, Nial Friel, and Pierre Alquier. Light and widely applicable MCMC: Approximate Bayesian inference for large datasets. *arXiv preprint arXiv:1503.04178*, 2015.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*, 2014.

A Yu Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, pages 1003–1014, 2005.

Esa Nummelin. A splitting technique for Harris recurrent markov chains. *Z. Wahrsch. Verw. Gebiete.*, 43:309–318, 1978.
Sean M O’brien and David B Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004.

Debdeep Pati, Anirban Bhattacharya, Natesh S Pillai, David Dunson, et al. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102–1130, 2014.

Natesh S Pillai and Aaron Smith. Ergodicity of approximate mcmc chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*, 2014.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Nicholas G Polson, James G Scott, and Jesse Windle. The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):713–733, 2014.

Matias Quiroz, Mattias Villani, and Robert Kohn. Speeding up MCMC by efficient data subsampling. *arXiv preprint arXiv:1404.4178*, 2014.

Christian Ritter and Martin A Tanner. Facilitating the Gibbs sampler: The Gibbs stopper and griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87:861–868, 1992.

Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer, 2 edition, 2004.

Gareth O Roberts, Jeffrey S Rosenthal, and Peter O Schwartz. Convergence properties of perturbed markov chains. *Journal of Applied Probability*, pages 1–11, 1998.

Jeffrey S. Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *Journal of the American Statistical Association*, 90:558–566, 1994.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *arXiv preprint arXiv:1503.04123*, 2015.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Adrien Saumard, Jon A Wellner, et al. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45–114, 2014.

Alex J Smola and Peter Bartlett. Sparse greedy gaussian process regression. In *Advances in Neural Information Processing Systems 13*, 2001.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

Terence Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices. [https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices/#more-3341](https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices/#more-3341) 2015. Accessed: 2015-11-02.
Sethu Vijayakumar, Aaron D’souza, and Stefan Schaal. Incremental online learning in high dimensions. *Neural computation*, 17(12):2602–2634, 2005.

S. J. Vollmer and K. C. Zygalakis. (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *ArXiv e-prints*, January 2015.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Lionel Weiss. The error in the normal approximation to the multinomial with an increasing number of classes. *Naval Research Logistics Quarterly*, 25(2):257–261, 1978.

Yun Yang and David B Dunson. Sequential markov chain monte carlo. *arXiv preprint arXiv:1308.3861*, 2013.

Jun Zhu, Jianfei Chen, and Wenbo Hu. Big learning with Bayesian methods. *arXiv preprint arXiv:1411.6370*, 2014.