Error bounds for Approximations of Markov chains

James E. Johndrow
Stanford University

Jonathan C. Mattingly
Duke University

November 16, 2017

The first part of this article gives error bounds for approximations of Markov kernels under Foster-Lyapunov conditions. The basic idea is that when both the approximating kernel and the original kernel satisfy a Foster-Lyapunov condition, the long-time dynamics of the two chains – as well as the invariant measures, when they exist – will be close in a weighted total variation norm, provided that the approximation is sufficiently accurate. The required accuracy depends in part on the Lyapunov function, with more stable chains being more tolerant of approximation error. This is similar to the situation for uniformly ergodic chains, studied by the authors in [12], and others in [19, 1].

Our development is inspired by [11], and certainly has commonalities with [23], which addresses the same topic. Our proofs are rather different, and we give error bounds for approximation of expectations using Poisson equation arguments similar to [7, 15]. (Of course, such methods are intimately related to classical Martingale and potential methods [21] as well as classical ideas from dynamical systems.) We also show conditions under which the approximating chain will satisfy Harris' theorem. An ancillary implication of our results is that one need not explicitly construct a Markov kernel that targets the posterior measure to obtain an algorithm that is useful for Bayesian computation. While this point is certainly made elsewhere in the literature, the connection with approximating kernels has perhaps not been made in this way before.

We are motivated by the recent growth in proposals for scaling Markov chain Monte Carlo algorithms to large datasets by defining an approximating kernel that is faster to sample from [14, 25, 5, 2, 3]. Many of these proposals use only a small subset of the data points to construct the transition kernel, and we consider an application to this class of approximating kernel. We also consider applications to distribution approximations in Gibbs sampling which are again made for computational efficiency. Unlike the examples discussed in [12], here we will consider unbounded domains which necessitate the introduction of Lyapunov functions and weighted total variation norms rather than the classical total variation norm used in [12].

Another application in which approximating kernels are commonly used is in Metropolis algorithms for Gaussian process models common in spatial statistics and nonparametric regression. In this setting, there are typically two sources of approximation error: discretization error and approximation of Metropolis acceptance ratios. Because the approximating kernel is obtained by discretizing the state space, it is singular with respect to the exact kernel. To analyze this application, we give additional results in Wasserstein metrics in contrast to the proceeding examples which quantified the level of approximation in a total variation norm. Informally, the results require a uniform Wasserstein contraction condition for the exact kernel similar to [10] and a uniform approximation error condition. These conditions are appropriate for our application, which typically operates on a compact state space.

1 Bounds in weighted total variation

We begin by defining the weighted total variation metrics which will be used to quantify convergence, largely following [11]. We then introduce a approximating change and bound the shift in the invariant measure and time averages.

1.1 Basic mixing results

We assume conditions on the Markov kernel similar to those in [11, 16]. Let $\mathcal{P}(x,\,\cdot\,)$ be a Markov kernel on a Polish state space $\mathbf{X}$, which in many applications is $\mathbb{R}^p$, p-dimensional Euclidean space. We use $\mathcal{P}$ for operators defined on the set of measurable functions and the set of finite measures

$$(\mathcal{P}\phi)(x) = \int_{\mathbf{X}} \phi(y)\mathcal{P}(x,dy), \quad (\mathcal{P}\mu)(A) = \int_{\mathbf{X}} \mathcal{P}(x,A)\mu(dx).$$

We assume that $\mathcal{P}$ satisfies a Foster-Lyapunov drift condition

Assumption 1.1. *There exists a function $V : \mathbf{X} \to [0,\infty)$ such that for some $\gamma \in (0,1)$ and $K > 0$*

$$\mathcal{P}V(x) \le \gamma V(x) + K \tag{1}$$

for all $x \in \mathbf{X}$.

We also assume that sublevel sets of V are “small” in that they satisfy a uniform minorization condition.

Assumption 1.2. *For every $R > 0$, there exists $\alpha \in (0,1)$ (depending on R) such that*

$$\sup_{x,y\in\mathcal{C}(R)} \|\mathcal{P}(x,\,\cdot\,) - \mathcal{P}(y,\,\cdot\,)\|_{TV} \le 2(1-\alpha) \tag{2}$$

for $\mathcal{C}(R) = \{x : V(x) \le R\}$.

To quantify the rate of convergence to equilibrium, we procede in the spirit of [16] and define a family of weighted supremum norms indexed by a scale parameter $\beta > 0$ by

$$\|\phi\|_\beta = \sup_x \frac{|\phi(x)|}{1+\beta V(x)}$$

and the dual metric ρ_β on probability measures

$$\rho_\beta(\mu_1,\mu_2) = \sup_{\phi:\|\phi\|_\beta\le 1} \int_{\mathbf{X}} \phi(x)(\mu_1-\mu_2)(dx) = \int_{\mathbf{X}} (1+\beta V(x))|\mu_1-\mu_2|(dx),$$

a weighted total variation distance. Hairer and Mattingly [11] show that for β sufficiently small the Markov semigroup $\mathcal{P}$ is a contraction in the metric ρ_β under Assumptions 1.1 and 1.2. In [11], they also showed that these metrics are equivalent to the metric d_β on measures induced by

$$d_\beta(x,y) = \begin{cases} 0 & x=y \\ 2+\beta V(x)+\beta V(y) & x\neq y \end{cases}$$

To define d_β, one first defines a Lipschitz seminorm on measurable functions by

$$|||\phi|||_\beta = \sup_{x\neq y} \frac{|\phi(x)-\phi(y)|}{d_\beta(x,y)} .$$

This in turn induces the metric d_β on probability measures through

$$d_\beta(\mu_1,\mu_2) = \sup_{\phi:|||\phi|||_\beta\le 1} \int_{\mathbf{X}} \phi(x)(\mu_1-\mu_2)(dx),$$

for which it turns out that $|||\phi|||_\beta = \inf_{c\in\mathbb{R}} \|\phi+c\|_\beta$ and therefore $d_\beta = \rho_\beta$. In the sequel we freely interchange these metrics. We now give the convergence theorem from Hairer and Mattingly [11] which uses these metrics.

Theorem 1.3 (Theorem 1.3 of [11])**.** *Under Assumptionss 1.1 and 1.2, there exist an $\alpha\in(0,1)$ and $\beta>0$ so that*

$$d_\beta(\nu_1\mathcal{P},\nu_2\mathcal{P}) \le \alpha d_\beta(\nu_1,\nu_2)$$

for all probability measure ν_1 and ν_2.

1.2 Basic approximation results

Now consider a second transition kernel $\mathcal{P}_\epsilon$ that is "nearby" $\mathcal{P}$ in the following sense.

Assumption 1.4. *For some $\delta\ge 0$ and all x,*

$$d_1(\mathcal{P}(x,\,\cdot\,),\mathcal{P}_\epsilon(x,\,\cdot\,)) \le \epsilon(1+\delta V(x))$$
$$\Updownarrow$$
$$(\mathcal{P}-\mathcal{P}_\epsilon)\phi(x) \le \epsilon(1+\delta V(x)) \textit{ for all } |\phi|\le 1+V$$

The following basic pertubation bound is one of our main results.

Theorem 1.5. *Suppose assumptions 1.1, 1.2, and 1.4 hold. Then there exists a $\beta\in(0,1]$ and $\alpha_0\in(0,1)$ so that*

$$\mathcal{P}_\epsilon\phi(x) - \mathcal{P}\phi(y) \le \epsilon(1+\delta V(x)) + \alpha_0\, d_\beta(x,y)$$

for all $|\phi|\le 1+\beta V$.

Proof. We have

$$\begin{aligned} d_\beta(\delta_y\mathcal{P},\delta_x\mathcal{P}_\epsilon) &\le d_\beta(\delta_y\mathcal{P},\delta_x\mathcal{P}) + d_\beta(\delta_x\mathcal{P},\delta_x\mathcal{P}_\epsilon) \\ &\le \alpha_0 d_\beta(x,y) + \epsilon(1+\delta V(x)), \end{aligned}$$

where the first term followed from Assumptions 1.1 and 1.2 and [11, Theorem 3.1] and the second term from Assumption 1.4. □

This immediately gives a bound on the distance between the one-step transition kernels for any pair of starting measures.

Corollary 1.6. *Let μ and ν be two probability measures. Then*

$$d_\beta(\mu\mathcal{P}_\epsilon,\nu\mathcal{P}) \le \epsilon(1+\delta\mu V) + \alpha_0 d_\beta(\mu,\nu). \tag{3}$$

We can use this result to bound the distance between the invariant measure(s). If m and m_ϵ are invariant measures of $\mathcal{P}$ and $\mathcal{P}_\epsilon$ respectively then

$$d_\beta(m_\epsilon,m) \le \frac{\epsilon}{1-\alpha_0}(1+\delta\, m_\epsilon V),$$

and iterating the estimate in (3) gives

$$d_\beta(\mu\mathcal{P}_\epsilon^n,\nu\mathcal{P}^n) \le \epsilon\sum_{k=1}^n \alpha_0^{n-k}(1+\delta\mu\mathcal{P}_\epsilon^{k-1}V) + \alpha_0^n d_\beta(\mu,\nu), \tag{4}$$

a finite-time error bound. If we now assume

Assumption 1.7. *For some $\gamma_\epsilon\in(0,1)$ and $K_\epsilon>0$*

$$\mathcal{P}_\epsilon V(x) \le \gamma_\epsilon V(x) + K_\epsilon \tag{5}$$

for all x.

so that V is also a Lyapunov function of $\mathcal{P}_\epsilon$, then

$$\mu\mathcal{P}_\epsilon^j V \le \gamma_\epsilon^j\mu V + \frac{K_\epsilon}{1-\gamma_\epsilon},$$

and in place of (4) we can use the bound

$$\begin{aligned} d_\beta(\mu P_\epsilon^n,\nu P^n) &\le \frac{\epsilon}{1-\alpha_0}(1+\delta\frac{K_\epsilon}{1-\gamma_\epsilon}) + \epsilon\delta(\mu V)\sum_{k=1}^n \alpha_0^{n-k}\gamma_\epsilon^k + \alpha_0^n d_\beta(\mu,\nu) \\ &\le \frac{\epsilon}{1-\alpha_0}(1+\delta\frac{K_\epsilon}{1-\gamma_\epsilon}) + \epsilon\delta(\mu V)(\alpha_0\vee\gamma_\epsilon)^{n-1}n + \alpha_0^n d_\beta(\mu,\nu). \end{aligned}$$

We note that Assumption 1.7 is implied by Assumption 1.4 when $\epsilon\delta < 1-\gamma$; a simple argument is given in the proof of Remark 1.9. Also under Assumption 1.7, if m and m_ϵ are the invariant measures of $\mathcal{P}$ and $\mathcal{P}_\epsilon$ respectively then we have the bound

$$d_\beta(m,m_\epsilon) \le \frac{\epsilon}{1-\alpha_0}(1+\delta\frac{K_\epsilon}{1-\gamma_\epsilon}).$$

We now show that under the following additional condition, one can prove Harris' theorem for $\mathcal{P}_\epsilon$.

Assumption 1.8. *For every* $0 < R < \frac{2(K+\epsilon)}{1-(\gamma+\epsilon\delta)}$ *there exists* $\zeta < \alpha$ *(depending on* R*) such that*

$$\sup_{x \in \mathcal{C}} \|\mathcal{P}_\epsilon(x,\,\cdot\,) - \mathcal{P}(x,\,\cdot\,)\|_{TV} \le \zeta$$

for $\mathcal{C} = \{x : V(x) \le R\}$.

This result is included mainly for completeness. It is common in the MCMC literature to prove Harris' theorem, and many practitioners mistakenly interpret it as a guarantee of good finite-time performance. It is clear from Theorem 1.10 that this is not necessary to obtain the kind of variation bounds that are desired in MCMC applications, but the following result may nonetheless be of interest.

Remark 1.9. *Suppose Assumptions 1.2, 1.4, 1.1, and 1.8 hold and* $\delta\epsilon < 1 - \gamma$. *Then there exists* $\bar{\alpha}_\epsilon < 1$ *and* $\beta > 0$ *such that*

$$\rho_\beta(\mathcal{P}_\epsilon \mu, \mathcal{P}_\epsilon \nu) \le \bar{\alpha}_\epsilon \rho_\beta(\mu, \nu)$$

for any probability measures μ, ν *on* $\mathbf{X}$.

Proof. We have

$$\begin{aligned}\mathcal{P}_\epsilon V &= (\mathcal{P} + \mathcal{P}_\epsilon - \mathcal{P})V \\ &\le \gamma V + K + \epsilon(1 + \delta V),\end{aligned}$$

and for any $x, y \in \mathcal{C}$ with $\mathcal{C} = \{x : V(x) \le R\}$

$$\begin{aligned}\|\mathcal{P}_\epsilon(x,\,\cdot\,) - \mathcal{P}_\epsilon(y,\,\cdot\,)\|_{TV} &\le \|\mathcal{P}_\epsilon(x,\,\cdot\,) - \mathcal{P}(x,\,\cdot\,)\|_{TV} \\ &\quad + \|\mathcal{P}(x,\,\cdot\,) - \mathcal{P}(y,\,\cdot\,)\|_{TV} + \|\mathcal{P}(y,\,\cdot\,) - \mathcal{P}_\epsilon(y,\,\cdot\,)\|_{TV} \\ &\le \zeta + 2(1-\alpha) + \zeta = 2(1 - (\alpha - \zeta)).\end{aligned}$$

for every $R \le \frac{2(K+\epsilon)}{1-(\gamma+\epsilon\delta)}$. □

Our main error bound is given by the following result.

Theorem 1.10. *Assume that Assumptions 1.1, 1.2, 1.4 and 1.7 hold. Then there exists* $C < \infty$ *so that*

$$\mathbf{E}\left(\frac{1}{n}\sum_{k=0}^{n-1} \phi(X_k^\epsilon) - \mu\phi\right)^2 \le 6C^2\epsilon\delta\left(\frac{K_\epsilon}{1-\gamma_\epsilon} + \frac{1-\delta}{\delta}\right) + \frac{6C^2}{n}\left(\frac{K_\epsilon + 1}{1-\gamma_\epsilon}\right)(1+V) + \mathcal{O}\left(\frac{1}{n^2}\right).$$

The proof is deferred to the Appendix. The bound consists of an error term that goes to zero when $\epsilon \to 0$, terms proportional to V that goes to zero at the rate n^{-1}, and terms going to zero like n^{-2}.

2 Applications

We begin this section with a result which is helpful in verifying the Assumption 1.4. We then apply the previous sections to a simple pedagogical example, a basic Gibbs sampling application, and Minibatching Metropolis-Hastings.

2.1 Achieving the error condition

For the bounds above to be practical, we need a way to construct approximations $\mathcal{P}_\epsilon$ that achieve Assumption 1.4 that is broadly applicable. Often, it is easier to construct approximations satisfying a condition like

$$\|\mathcal{P}(x,\,\cdot\,) - \mathcal{P}_\epsilon(x,\,\cdot\,)\|_{TV} < \epsilon$$

than to directly construct an approximating kernel with error depending on the Lyapunov function. However, uniform total variation error control is not enough to show Assumption 1.4, so we seek an adaptive total variation error condition that gives Assumption 1.4.

If V is a Lyapunov function of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$, observe that

$$\sup_{|\phi|<V} \int \phi(y)\mathcal{P}(x,dy) < \gamma V(x) + K.$$

Because this family is integrable for each x, for every $\epsilon > 0$ there exists $M_\epsilon(x) < \infty$ such that

$$\begin{aligned}\sup_{|\phi|<V} \int \phi(y)\mathbf{1}\{|\phi(y)| > M_\epsilon(x)\}\mathcal{P}(x,dy) &< \int V(y)\mathbf{1}\{V(y) > M_\epsilon(x)\}\mathcal{P}(x,dy) \\ &< \frac{\epsilon}{4}. \end{aligned} \tag{6}$$

A similar condition holds for $\mathcal{P}_\epsilon(x,dy)$ for each x; redefine $M_\epsilon(x)$ so that the condition in (6) holds for both $\mathcal{P}$ and $\mathcal{P}_\epsilon$. Now suppose

$$\|\mathcal{P}(x,dy) - \mathcal{P}_\epsilon(x,dy)\|_{TV} < \frac{\epsilon\gamma V(x)}{2M_\epsilon(x)} + \frac{\epsilon}{4M_\epsilon(x)}. \tag{7}$$

Then setting $\mathcal{A}_\epsilon = \{|\phi(y)| > M_\epsilon(x)\}$

$$\begin{aligned}&\sup_{|\phi|<V} \int \phi(y)(\mathcal{P}(x,dy) - \mathcal{P}_\epsilon(x,dy)) \\ &\qquad = \sup_{|\phi|<V} \left(\int \phi(y)\mathbf{1}_{\mathcal{A}_\epsilon}(\mathcal{P}(x,dy) - \mathcal{P}_\epsilon(x,dy)) \right. \\ &\qquad\qquad \left. + \int \phi(y)\mathbf{1}_{\mathcal{A}_\epsilon^c}(\mathcal{P}(x,dy) - \mathcal{P}_\epsilon(x,dy)) \right) \\ &\qquad \le \epsilon + \gamma\epsilon V(x) = \epsilon(1 + \gamma V(x)).\end{aligned}$$

In other words, total variation control is good enough, assuming we tune the approximation error in total variation to the current state of the chain. This is consistent with several of our other results. In some sense, this is obvious, since we have a weighted total variation norm, so it should be enough to have a total variation approximation error that adapts to the state. The main argument here is to use the integrability condition. Notice that if we can compute $V(x)/M_\epsilon(x)$, then an algorithm that allows us to choose the total variation approximation error at each step is good enough to achieve our approximation error condition in the V-weighted norm. Note that this condition only requires that there exist *some* Lyapunov function V_ϵ of $\mathcal{P}_\epsilon$, since we can then always take $V \wedge V_\epsilon$ as a Lyapunov function of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$.

2.2 A (simple) example

Consider the Gaussian autoregressive model of order 1 with unit variance

$$\mathcal{P}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-(\rho x - y)^2/2}$$

for $\rho \in (-1,1)$. The function $V(x) = x^2$ is a Lyapunov function since

$$\int_{-\infty}^{\infty} y^2 \mathcal{P}(x,dy) = \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-(\rho x-y)^2/2} dy = x^2\rho^2 + 1$$

and $\rho^2 < |\rho| < 1$. For any $\xi > 0$ we have

$$\int_{-\infty}^{\rho x - \xi} y^2 \mathcal{P}(x,dy) + \int_{\rho x + \xi}^{\infty} y^2 \mathcal{P}(x,dy) = \sqrt{\frac{2}{\pi}} \xi e^{-\frac{\xi^2}{2}} + 2(1 + x^2\rho^2)\Phi(-\xi),$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function. Using an inequality we later use in the proof of Proposition 2.1

$$\sqrt{\frac{2}{\pi}} \xi e^{-\frac{\xi^2}{2}} + 2(1+x^2\rho^2)(1-\Phi(\xi)) = \sqrt{\frac{2}{\pi}} \xi e^{-\frac{\xi^2}{2}} + \frac{2(1+x^2\rho^2)}{\sqrt{2\pi}} \frac{e^{-\frac{\xi^2}{2}}}{\xi + \sqrt{\xi^2 + h(\xi)}},$$

with $\frac{8}{\pi} < h(\xi) < 4$. Taking $\xi_\epsilon(x) = \sqrt{8 \log(\sqrt{\frac{4}{\epsilon}} + \rho x)}$ we have the upper bound

$$\begin{aligned} &\sqrt{\frac{2}{\pi}} \frac{\sqrt{8\log(\sqrt{4/\epsilon} + \rho x)}}{(\sqrt{4/\epsilon} + \rho x)^4} + \sqrt{\frac{2}{\pi}} \frac{(1+x^2\rho^2)}{(\sqrt{4/\epsilon}+\rho x)^4} \frac{1}{2\sqrt{8\log(\sqrt{4/\epsilon}+\rho x)}} \\ &\quad \leq \frac{1}{(\sqrt{4/\epsilon}+\rho x)^2} < \frac{\epsilon}{4}, \end{aligned}$$

with the corresponding value

$$M_\epsilon(x) = \left(|\rho| x + \sqrt{8\log(\sqrt{\tfrac{4}{\epsilon}} + \rho x)}\right)^2.$$

Now consider the perturbation

$$\mathcal{P}_\epsilon(x, dy) = \frac{1}{\sqrt{2\pi}} e^{-\frac{((\rho+\delta)x-y)^2}{2}} dy$$

for $-1 + \rho < \delta < 1 - \rho$. This is just another first order autoregressive model, with $\rho_\epsilon = \rho + \delta$, so x^2 is still a Lyapunov function and definining

$$M_\epsilon(x) = \left(|\rho + (\delta \vee 0)| x + \sqrt{8 \log \left\{ \sqrt{\tfrac{4}{\epsilon}} + (\rho + (\delta \vee 0))x \right\}} \right)^2 .$$

we have

$$\sup_{|\phi|<V} \int \phi(y) \mathbf{1}\{|\phi(y)| > M_\epsilon(x)\} (\mathcal{P}_\epsilon(x, dy) - \mathcal{P}(x, dy)) \leq \frac{\epsilon}{2}$$

and if $\delta > 0$ then

$$\|\mathcal{P}(x, dy) - \mathcal{P}_\epsilon(x, dy)\|_{TV} \leq \frac{\epsilon \gamma x^2 + \epsilon/4}{\left(\rho x + \sqrt{8 \log \left\{ \sqrt{\tfrac{4}{\epsilon}} + \rho x \right\}} \right)^2} . \tag{8}$$

Figure 1 shows a particular example of the function on the right side of (8) for $\epsilon = 0.1, \gamma = 0.9, \delta = 0.05$, and $\rho = 0.8$. Outside of the interval $[-1, 1]$, the total variation error can be strictly larger than 0.005 and still achieve the error condition in Assumption 1.4. The error requirement becomes non-binding when $x < -2$, but for positive values of x, total variation error of less than 0.05 is required over the entire range of states that the chain is likely to visit. Certainly this will lead to some computational improvement over an algorithm achieving uniform error of, say, 0.01 everywhere.

2.3 Application to Gibbs sampling

In this section we consider approximating a Gibbs sampler for Probit regression with Gaussian priors. The model is

$$y_i \sim \text{Binomial}(n_i, \Phi(w_i \beta)),\ i = 1, \dots, N, \quad \beta \sim N(0, B)$$

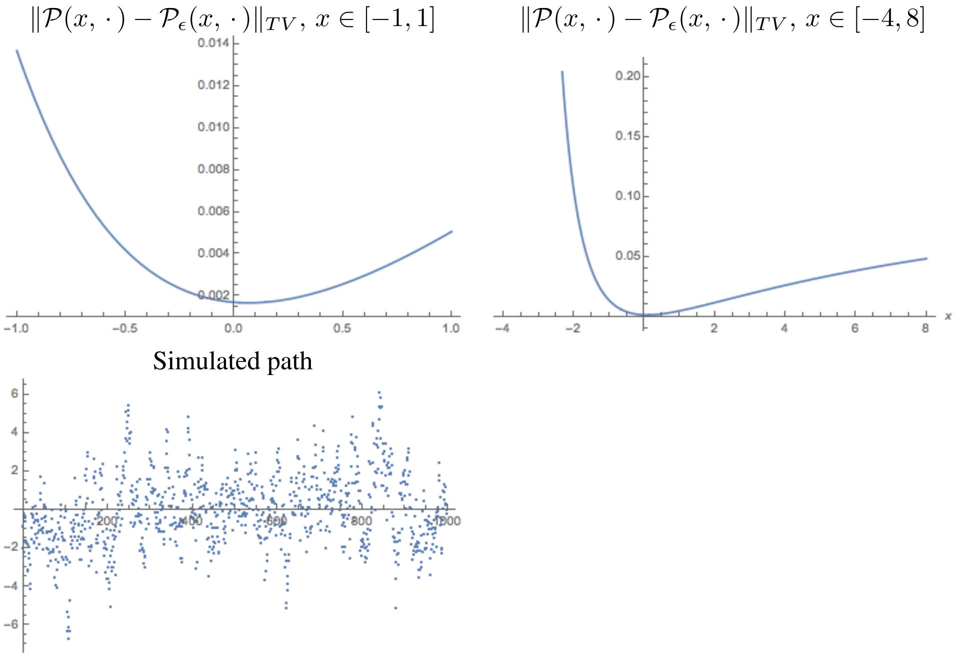
where each w_i is a $1 \times p$ real vector. In general we will assume that the (y_i, w_i) are independent.

A transition kernel $\mathcal{P}$ with invariant measure the posterior in this model is defined by the update rule

$$\omega_i = \sum_{j=1}^{y_i} Z_{ij}^+ + \sum_{j=1}^{n_i - y_i} Z_{ij}^-$$

$$p_{Z_{ij}^+}(z) \propto \frac{1}{\sqrt{2\pi}} e^{-(z - w_i\beta)/2} \mathbf{1}\{z > 0\}, \quad p_{Z_{ij}^-}(z) \propto \frac{1}{\sqrt{2\pi}} e^{-(z - w_i\beta)/2} \mathbf{1}\{z \leq 0\}$$

$$\beta \sim \text{Normal}((B^{-1} + W'DW)^{-1} W'\Omega, (B^{-1} + W'DW)^{-1})$$



$\|\mathcal{P}(x,\,\cdot\,)-\mathcal{P}_\epsilon(x,\,\cdot\,)\|_{TV}$, $x\in[-1,1]$

$\|\mathcal{P}(x,\,\cdot\,)-\mathcal{P}_\epsilon(x,\,\cdot\,)\|_{TV}$, $x\in[-4,8]$

Simulated path

Figure 1: Top two panels: Plots of the state x on the horizontal axis vs the right side of (8) for value of parameters as given in text. Bottom panel: simulated path of length 1000 for AR-1 process with autocorrelation 0.85 (the mean of ρ and $\rho+\delta$).

where D is a diagonal matrix with diagonal entries $n_1,\ldots,n_N$, W is a $N\times p$ matrix with rows consisting of the w_i's, and $\Omega=(\omega_1,\ldots,\omega_N)$ is a $N\times 1$ vector of the ω_i's. It is standard in the literature to analyze $\mathcal{P}$ in the special case of a flat prior on β, in which case the update is simplified slightly so that the last step becomes

$$\beta\sim\text{Normal}((W'DW)^{-1}W'\Omega,(W'DW)^{-1}).\tag{9}$$

A perturbation $\mathcal{P}_\epsilon$ of $\mathcal{P}$ can be generated by replacing $Z^+=\sum_{j=1}^{y_i}Z_{ij}^+$ and $Z^-=\sum_{j=1}^{n_i-y_i}Z_{ij}^-$ by Gaussian approximations U^+ and U^- given by

$$\begin{aligned}U_i^+&\sim\text{Normal}(y\mathbf{E}(Z_{ij}^+),y\mathbf{V}(Z_{ij}^+))\\U_i^-&\sim\text{Normal}((n-y)\mathbf{E}(Z_{ij}^-),(n-y)\mathbf{V}(Z_{ij}^-)).\end{aligned}$$

Suppose we use a flat prior on β, so that the second step of the update is given by (9). We show a Lyapunov function of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$ in this case when $0<\min_i y_i/n_i<1$.

Proposition 2.1. *$V:\mathbb{R}^p\to\mathbb{R}_+$ given by $V(\beta)=\beta'W'DW\beta$ is a Lyapunov function of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$. In particular*

$$\mathcal{P}V\le\gamma V+K,\quad\mathcal{P}_\epsilon V\le\gamma V+K$$

for some $0 < \gamma < 1$ *and* $K > 0$.

The proof of Proposition 2.1 is given in the appendix.

Now we give a bound on $\|\mathcal{P} - \mathcal{P}_\epsilon\|_{TV}$. Denote by $\mathrm{KL}(\mu\|\nu)$ the Kullback-Leibler divergence between probability measures μ, ν that are absolutely continuous with respect to a dominating measure λ, which in this example we can take to be Lebesgue measure. Denote by $\mu(\beta \mid \Omega)$ and $\mu_\epsilon(\beta \mid \Omega)$ the conditional measure of β given Ω induced by the kernels $\mathcal{P}$ and $\mathcal{P}_\epsilon$, respectively. Observe

$$\begin{aligned}\mathrm{KL}(\mu\|\mu_\epsilon) &= \frac{1}{2}((W'DW)^{-1}W'(\Omega - \Omega_\epsilon))'W'DW((W'DW)^{-1}W'(\Omega - \Omega_\epsilon)) \\ &= \frac{1}{2}(\Omega - \Omega_\epsilon)'W(W'DW)^{-1}W'(\Omega - \Omega_\epsilon).\end{aligned}$$

Putting $\Psi = W(W'DW)^{-1}W'$, we have by Pinsker's inequality

$$\begin{aligned}\|\mu - \mu_\epsilon\|_{TV} &\le \sqrt{\frac{1}{4}(\Omega - \Omega_\epsilon)'\Psi(\Omega - \Omega_\epsilon)} \\ &\le \frac{1}{2}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}(\omega_i - \omega_i^\epsilon)(\omega_j - \omega_j^\epsilon)\psi_{ij}} \\ \mathbf{E}[\|\mu - \mu_\epsilon\|_{TV} \mid \beta_{k-1}] &\le \frac{1}{4}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{E}(\omega_i - \omega_i^\epsilon)(\omega_j - \omega_j^\epsilon)\psi_{ij}} \\ &= \frac{1}{4}\sqrt{\sum_{i=1}^{N}\mathbf{E}(\omega_i - \omega_i^\epsilon)^2\psi_{ii}} \\ &= \frac{\sqrt{2}}{4}\sqrt{\sum_{i=1}^{N}\mathrm{var}(\omega_i)\psi_{ii}}.\end{aligned}$$

This quantity will be roughly $n^{-1/2}$ when all $n_i = n$, so the error converges to zero in the total variation norm at the expected rate.

2.4 Application to Minibatching Metropolis-Hastings

Consider a generic Metropolis-Hastings algorithm with target measure $\mu(dx)$, proposal kernel $Q(x, dy) = q(x, y)\mu(dy)$ and transition kernel $\mathcal{P}$. Suppose V is a Lyapunov function of $\mathcal{P}$ satisfying

$$\mathcal{P}V \le \gamma V + K$$

for $0 < K < \infty, \gamma \in (0, 1)$. Let

$$\begin{aligned}\beta(x, y) &= \frac{q(y, x)}{q(x, y)} \\ \alpha(x, y) &= 1 \wedge \beta(x, y)\end{aligned}$$

Then we can write $\mathcal{P}V$ as

$$\begin{aligned}\mathcal{P}V(x) &= \int V(y)\alpha(x,y)Q(x,dy) + V(x)\left(1 - \int \alpha(x,y)Q(x,dy)\right) \\ &\leq \gamma V(x) + K\end{aligned} \tag{10}$$

Let $\mathcal{P}_\epsilon$ be the transition kernel of another Metropolis algorithm with the same proposal distribution, but which replaces $\alpha(x,y)$ with $\alpha_\epsilon(x,y)$. Then

$$\begin{aligned}\|\mathcal{P}(x,\,\cdot\,) - \mathcal{P}_\epsilon(x,\,\cdot\,)\|_{TV} &= \sup_{|\phi|<1}\left[\int \phi(y)\{\alpha(x,y) - \alpha_\epsilon(x,y)\}Q(x,dy)\right. \\ &\quad \left. + \phi(x)\int \{\alpha_\epsilon(x,y) - \alpha(x,y)\}Q(x,dy)\right] \\ &= \sup_{|\phi|<1}\int (\phi(y) - \phi(x))\{\alpha(x,y) - \alpha_\epsilon(x,y)\}Q(x,dy) \\ &\leq 2\int |\alpha(x,y) - \alpha_\epsilon(x,y)|Q(x,dy).\end{aligned}$$

Suppose V is a Lyapunov function of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$. Then to achieve (7) it is enough to have

$$2\int |\alpha(x,y) - \alpha_\epsilon(x,y)|Q(x,dy) \leq \frac{\epsilon\gamma V(x)}{2M_\epsilon(x)} + \frac{\epsilon}{4M_\epsilon(x)}. \tag{11}$$

To apply this to an algorithm we need a Lyapunov function and an estimate of $M_\epsilon(x)$. Consider the simple case of intercept-only logistic regression with normal prior

$$z \sim \text{Binomial}\left(N, \frac{e^x}{1+e^x}\right), \quad x \sim \text{Normal}(0, B)$$

and let

$$Q(x,dy) = \frac{1}{2c}\mathbf{1}_{\{x-c<y<x+c\}}dy, \tag{12}$$

so the proposal is uniform on the current state plus or minus c. Let

$$p(\theta) = \binom{N}{z}\left(\frac{e^\theta}{1+e^\theta}\right)^z\left(\frac{1}{1+e^\theta}\right)^{N-z}\frac{1}{\sqrt{2\pi B}}e^{-\frac{\theta^2}{2B}}$$

be the posterior density, and define $\hat{\theta} = \text{argmax}_\theta\, p(\theta)$, the posterior mode. It is easy to see that $\hat{\theta}$ satisfies

$$\frac{\hat{\theta}}{B} + N\frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} = z.$$

We show a Foster-Lyapunov condition and multistep minorization on small sets for this algorithm. The proofs are a generalization of the proof of [13, Theorem 3.1] and use a similar argument, so we defer them to the appendix.

Theorem 2.2. *Assume $N > e^2$ and $1 < z < N-1$, and put $B = \log N$, $c = 2\log N$. Then $V(x) = e^{|x|}$ satisfies $(\mathcal{P}V)(x) \leq \gamma V(x) + K$ where $K = N^3$ and*

$$\gamma = \frac{1}{2} + \frac{5}{8\log N} - \frac{1}{4\log N}\left(\frac{2}{N^2} + \frac{1}{2N^4}\right)$$

and

$$\mathcal{P}^5(x, \{z : |z - \hat{\theta}| < 1\}) > \frac{1}{128}\left(1 + \frac{1}{e^2}\right)$$

for any $x \in \mathcal{C}_0$, where $\hat{\theta}$ is the unique mode of the target and $\mathcal{C}_0 = \left\{x : V(x) \leq \frac{2K}{1-\gamma}\right\}$.

We now describe the minibatching approximation. Observe that the likelihood can also be represented as

$$z_i \sim \text{Bernoulli}\left(\frac{e^x}{1+e^x}\right), \quad i = 1, \ldots, N$$

with $z = \sum_{i=1}^N z_i$. A minibatching appoximation is constructed by taking a random sample of size N_0 of the z_i. Let A be the set of sampled indices and $z_0 = \sum_{i \in A} z_i$. The minibatching approximation replaces z with $z_0 N/N_0$ to calculate $\alpha_\epsilon(x,y)$. In the intercept-only case, there is little motivation for such an approximation since the per-iteration computational cost does not increase with N. In a general logistic regression, calculation of $\alpha(x,y)$ grows linearly with N, and this minibatching strategy reduces per iteration cost from $\mathcal{O}(pN)$ to $\mathcal{O}(pN_0)$, where p is the number of covariates. We study the intercept-only case because sharp and informative bounds are achievable, with explicit constants. We then study the general case numerically.

Observe that because the bounds in Theorem 2.2 do not depend on z, they also apply to the minibatch approximation so long as $0 < z_0 < N_0$. Recall the definition

$$\int V(y)\mathbf{1}\{V(y) > M_\epsilon(x)\}\mathcal{P}(x, dy) < \frac{\epsilon}{4}.$$

With the uniform random walk kernel in (12), $y \in [x-c, x+c]$ with probability one, so the worst we could do is $M_\epsilon(x) = e^{|x|+c} = e^c V(x)$, and with $c = 2\log N$ as in the proof of Theorem 2.2, we obtain

$$\begin{aligned} \frac{1}{c}|\alpha(x,y) - \alpha_\epsilon(x,y)| &\leq \frac{\epsilon\gamma V(x)}{2N^2 V(x)} + \frac{\epsilon}{4N^2} \\ |\alpha(x,y) - \alpha_\epsilon(x,y)| &\leq \frac{\epsilon\gamma\log N}{N^2} + \frac{\epsilon\log N}{2N^2}, \end{aligned} \tag{13}$$

which is stronger than the condition

$$\Delta(x,y,\epsilon) \equiv |\alpha(x,y) - \alpha_\epsilon(x,y)| \leq \epsilon. \tag{14}$$

We now consider how difficult it is to achieve (14) using minibatching. Let N_0 be the minibatch size, and z_0 the number of successes in the minibatch sample. We have that

$$\alpha(x,y) = \left(\frac{e^x}{e^y}\right)^z \left(\frac{1+e^y}{1+e^x}\right)^N e^{-(x^2-y^2)/2B}$$

$$\alpha_\epsilon(x,y) = \left(\frac{e^x}{e^y}\right)^{\frac{z_0 N}{N_0}} \left(\frac{1+e^y}{1+e^x}\right)^N e^{-(x^2-y^2)/2B}$$

with $z_0 \sim \text{HyperGeometric}(N, z, N_0)$. We note that since Theorem 2.2 holds only for $1 < z < N-1$, the condition in (13) may not be enough to guarantee the error condition when $z_0 < 2$. So the subsequent analysis is in some sense conservative, as $z_0 < 2$ will occur with non-negligible probability in some of the cases under consideration.

We consider the probability of achieving (13) at various points on the interval $x \in [-\log N, \log N]$ for the uniform proposal considered above. Specifically, we estimate by simulation the probability that $|\alpha(x,y) - \alpha_\epsilon(x,y)| < \epsilon$ for a grid of values in the interval $[-\log N, \log N]$ for different values of z and N by simulating from the hypergeometric distribution of z_0. Figure 2 shows results. The probability of achieving the error condition is well-controlled throughout the region $[-\log N, \log N]$ when $z/N = 0.5$, but when z/N is small, there is a significant portion of the region over which the probability of achieving the error condition is less than 0.9. This problem is more acute when N is smaller. A heuristic for thinking about this problem is to consider how much the random variable $R_0 = Z_0/N_0$ differs from its expectation z/N. From [8] we have the following Bernstein-like inequality for the hypergeometric distribution

$$\mathbf{P}\left[\left|\frac{Z_0}{N_0} - \frac{z}{N}\right| > \frac{\lambda}{\sqrt{N_0}}\right] \le \exp\left(-\frac{\lambda^2}{2\sigma_N^2(1-f_{N_0}) + \frac{2\lambda}{3\sqrt{N_0}}}\right) \tag{15}$$

where $f_{N_0} = \frac{N_0-1}{N-1}$ and $\sigma_N^2 = \frac{z(N-z)}{N^2}$. Clearly, the distribution becomes more concentrated around the "exact" value z/N as N_0 increases. The effect of z is also evident in (15), when $z \approx N/2$, $\sigma_N^2 \approx 1/4$, while when $z \ll N$, $\sigma_N^2 \approx N^{-1}$, thus making the tail probabilities for fixed λ and N_0 much larger.

We now give a more complicated but related numerical example for logistic regression. Consider the general model

$$\mathbf{P}(z_i = 1 \mid w_i, x) = \frac{e^{w_i' x}}{1+e^{w_i' x}}, \quad x \sim N(0, B)$$

for $x \in \mathbb{R}^p$ and a random-walk Metropolis algorithm with proposal kernel

$$Q(x, dy) = |2\pi\Sigma|^{-1/2} e^{-(y-x)'\Sigma^{-1}(y-x)/2} dy.$$

We use the adaptive Metropolis algorithm of [9] with the scaling factor suggested in [22] to construct Σ. We compare this to a minibatch algorithm defined by taking a sample A of size N_0 of the indices $i = 1, \ldots, N$ at each iteration and approximating the acceptance ratio $\alpha(x,y)$ by $\alpha_\epsilon(x,y)$ defined by

$$\log \alpha_\epsilon(x,y) = \frac{N}{N_0}\left(\sum_{i \in A} w_i'(x-y) - \log\frac{1+e^{w_i' x}}{1+e^{w_i' y}}\right) + \frac{1}{2}(y-x)'B^{-1}(y-x),$$

which saves computation by approximating the acceptance ratio using only a subset of the data of size N_0. We assess the accuracy of $\alpha_\epsilon(x,y)$ as an approximation to

α by computing $\Delta(x, y, \epsilon)$ as defined in (14) at different points in $\mathbf{X} \times \mathbf{X}$. We use $p = 2$, generate w_i iid from a normal distribution with identity covariance, and consider different values of N_0 with $N = 100,000$. Since it is not feasible to compute Δ everywhere in $\mathbf{X} \times \mathbf{X}$ – and the value matters only in regions where either the exact or approximating chain is likely to reside – we compute it by running the exact algorithm and computing $\alpha_\epsilon(x, y)$ in addition to $\alpha(x, y)$ at each step.

Figure 3 shows results. We plot Δ as a function of the Mahalanobis distance

$$D_{\hat{\Sigma}}(x, \hat{x}) \equiv (x - \hat{x})\hat{\Sigma}^{-1}(x - \hat{x})$$

where $\hat{x}$ and $\hat{\Sigma}$ are estimates of the posterior mean and covariance based on samples of the exact algorithm after discarding a burn-in. We also estimate

$$\mathbf{P}[\Delta(x, y, \epsilon) < \epsilon]$$

as a function of $D_{\hat{\Sigma}}(x, \hat{x})$ using local regression (LOESS) for $\epsilon = 0.1$. Results are shown for the case of independent normal w_i with identity covariance. When the current state is near the "center" of the state space – that is, close to $\hat{x}$ with respect to the metric D – Δ has larger mean and the distribution is almost symmetric around 0.5. Similarly, the probability of achieving $\Delta(x, y, \epsilon) < \epsilon$ decreases as the state moves closer to the posterior mean. Naturally, the larger the value of N_0, the higher the probability of achieving $\Delta < \epsilon$, though it is notable that more than half the data are necessary to make this probability greater than 0.5 in a $D_{\hat{\Sigma}}$ neighborhood of the mean of radius greater than one. This suggests the minibatching strategy will give small computational advantage if the goal is to achieve a condition such as Assumption 1.4. These results are generally consistent with those of Bardenet et al. [4].

3 An application where $\mathcal{P}_\epsilon$ and $\mathcal{P}$ are mutually singular

The applications considered in the previous section were amenable to bounds in the weighted total variation norm that was the focus of Section 1 since $\mathcal{P}(x, \cdot)$ and $\mathcal{P}_\epsilon(x, \cdot)$ were jointly absolutely continuous for every $x \in \mathbf{X}$. In this section, we consider an application to approximations commonly used in MCMC for Gaussian process models in which $\mathcal{P}(x, \cdot)$ and $\mathcal{P}_\epsilon(x, \cdot)$ are mutually singular for every $x \in \mathbf{X}$. This motivates bounds in Wasserstein metrics.

Consider a Gaussian process model with squared exponential (or "radial basis") kernel

$$\begin{aligned} z(w) &= x_3 f(w) + \epsilon, \quad \epsilon \sim N(0, x_3^2) \\ \operatorname{cov}(f(w_i), f(w_j)) &= x_2 \exp(-x_1 \|w_i - w_j\|_2^2). \end{aligned} \tag{16}$$

The parameters of the model are $x = (x_1, x_2, x_3) \in \mathbb{R}_+^3 = \mathbf{X}$, the positive orthant in $\mathbb{R}^3$. The points $W = w_1, \ldots, w_N$ at which the process is sampled are treated as fixed and known, and the observations of the process at these points are denoted $z = (z(w_1), \ldots, z(w_N))$. Bayesian inference on x requires choice of a prior distribution. A common choice is an inverse Gamma prior on x_3, and independent uniform priors on x_1, x_2 restricted to compact intervals I_1, I_2. In general, the left endpoints of $I_1 = [a_1, b_1]$, $I_2 = [a_2, b_2]$ are bounded away from zero. The complete prior can be written

$$\begin{aligned} p(x_3) &= \frac{b^a}{\Gamma(a)}(x_3^2)^{-\frac{a}{2}-1}e^{-\frac{b}{2x_3^2}} \\ p(x_1) &= \frac{1}{|I_1|}\mathbf{1}\{x_1 \in I_1\}, \quad p(x_2) = \frac{1}{|I_2|}\mathbf{1}\{x_2 \in I_2\}, \end{aligned}$$

with $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$. Integration over x_3^2 is available in closed form, leading to the likelihood for z marginal of x_3

$$L(z \mid x_1, x_2, W) \propto \frac{1}{|I + x_2\Sigma(x_1, W)|^{\frac{1}{2}}}\Big(b + z'(I + x_2\Sigma(x_1, W))^{-1}z\Big)^{-\frac{a+N}{2}} \tag{17}$$

where $\Sigma(x_1, W)$ is an $N \times N$ symmetric, positive definite matrix with entries $\Sigma(x_1, W)_{ij} = \exp(-x_1\|w_i - w_j\|_2^2)$.

We consider the case where $x_2 = 1$ is known and we target the posterior for x_1, so the state space for our Markov chain is the interval I_1. For simplicity, we drop subscripts from now on since the state space is one dimensional. We define $\mathcal{P}$ by a Metropolis algorithm with a wrapped Gaussian random walk on the interval I centered at x with variance v. Without loss of generality, take $|I| = 2\pi$ with midpoint m so that

$$Q(x, dy) := \frac{1}{\sqrt{2\pi v}}\sum_{k=-\infty}^{\infty} e^{-\frac{(y-x+2\pi k)^2}{2v}}\mathbf{1}\{m - \pi < y < m + \pi\}\, dy. \tag{18}$$

This algorithm is computationally expensive because it requires that we compute the determinant of $I + \Sigma(x_1, W)$ and a quadratic form in its inverse at every step. We consider an approximating kernel $\mathcal{P}_\epsilon$ that saves computation by discretizing the state space for x_1. Observe that

$$(I + \Sigma)^{-1} = (I + U\Lambda U')^{-1} = U(I + \Lambda)^{-1}U',$$

so that if we have the spectral decomposition of Σ available, we can easily compute the inverse appearing in (17) and its determinant. Therefore, we discretize $\mathbf{X}_1$ to a ϵ-grid of points. Denote these points as $\{\theta_k\}_{k\in\mathbb{N}}$. In practice, one would pre-compute the spectral decomposition at some small set of support points that are likely to be visited frequently by the chain, and then expand this set as necessary while the algorithm runs. When N is very large, computing the likelihood at even one point may be prohibitive; we consider an algorithm designed for this setting in the next section.

Define $\mathcal{P}_\epsilon$ by sampling y^* from (18), then proposing

$$y = \text{argmin}_{\theta_k}|y^* - \theta_k|,$$

the closest support point to y^*. Since $\mathcal{P}_\epsilon$ and $\mathcal{P}$ are mutually singular, the weighted total variation bounds we have used until now to study approximating kernels are not useful for this application. We now derive some bounds in Wasserstein metrics to study this algorithm.

3.1 Uniform Wasserstein contraction

We construct a contraction in the Wasserstein metric that is natural for this application. Let d be a lower semicontinuous metric on $\mathbf{X}$. For $f : \mathbf{X} \to \mathbb{R}$ define

$$|f|_{\mathrm{Lip}(d)} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$$

and for probability measures μ, ν define the dual metric on probability measures

$$d(\mu,\nu) = \sup_{|\phi|_{\mathrm{Lip}(d)} < 1} \int \phi(x)(\mu - \nu)(dx).$$

Henceforth we will consider the distance

$$d(x,y) = 1 \wedge \frac{|x-y|}{\delta}, \tag{19}$$

which generates the same topology as the standard distance but is localized on a scale δ and capped at one. Notice that because d is capped at one, if $\phi \in \mathrm{Lip}(d)$ then ϕ is necessarily bounded.

Our first condition on $\mathcal{P}$ stats that $\mathcal{P}$ is locally Lipschitz in the initial condition:

Assumption 3.1. *There exists $C < \infty$ such that for $|x-y| < \delta$*

$$d(\delta_x \mathcal{P}, \delta_y \mathcal{P}) < C|x-y|.$$

Letting $\mathcal{C}(\delta_x \mathcal{P}, \delta_y \mathcal{P})$ be the space of all couplings of $\delta_x \mathcal{P}$ and $\delta_y \mathcal{P}$, our second condition is a form of uniform topological irreducibility.

Assumption 3.2. *For all $\gamma > 0$ and $(x,y) \in \mathbf{X} \times \mathbf{X}$ there exists $\Gamma_{x,y} \in \mathcal{C}(\delta_x \mathcal{P}, \delta_y \mathcal{P})$ and $\alpha_\gamma > 0$ such that $\Gamma_{x,y}((a,b) : |a-b| < \gamma) > \alpha_\gamma$.*

Under these assumptions, we have the following contractility result which implies exponential convergence in the Wasserstein metric.

Theorem 3.3. *Suppose Assumptions 3.1 and 3.2 hold. Then there exists $\bar{\alpha} < 1$ such that*

$$d(\delta_x \mathcal{P}, \delta_y \mathcal{P}) \leq \bar{\alpha} d(x,y).$$

Proof. This proof largely follows Section 2.1 from Hairer and Mattingly [10]. First suppose $|x-y| < \delta$ and $\gamma < \delta < \frac{1}{C}$. Then

$$d(\delta_x \mathcal{P}, \delta_y \mathcal{P}) \leq C|x-y| \leq C\delta\big(1 \wedge \tfrac{|x-y|}{\delta}\big) \leq C\delta d(x,y).$$

On the other hand if $|x-y| > \delta$ then defining $\Delta_\gamma = \{(a,b) \in \mathbf{X} \times \mathbf{X} : |a-b| < \gamma\}$

$$\begin{aligned} d(\delta_x \mathcal{P}, \delta_y \mathcal{P}) &\leq \int_{\Delta_\gamma} d(a,b)\Gamma_{x,y}(da,db) + \int_{\Delta_\gamma^c} d(a,b)\Gamma_{x,y}(da,db) \\ &\leq \frac{\gamma}{\delta}\alpha_\gamma + (1-\alpha_\gamma) = 1 - \left(1 - \frac{\gamma}{\delta}\right)\alpha_\gamma \leq \bar{\alpha}_\gamma = \bar{\alpha}_\gamma d(x,y). \end{aligned}$$

Putting $\bar{\alpha} = \bar{\alpha}_\gamma \wedge C\delta$ completes the proof. □

3.2 Wasserstein contraction for Metropolis-Hastings

We now give some sufficient conditions for establishing Assumptions 3.2 and 3.1, and use these conditions to establish Theorem 3.3 for our application. In this section, we will assume that the target μ and the proposal kernel $Q(x,\,\cdot\,)$ are absolutely continuous with respect to Lebesgue measure. The following condition implies Assumption 3.2 and is easy to show for our application since the state space is compact.

Remark 3.4. *Write $\mu(dx) = m(x)dx$ and let $\mathcal{B}_\delta(z)$ be a ball of diameter δ with center z. Suppose that for some z^* and $\delta > 0$ one has*

$$\inf_x \inf_{z \in \mathcal{B}_\delta(z^*)} q(z,x) = c_0 > 0, \quad \sup_x \sup_{z \in \mathcal{B}_\delta(z^*)} q(x,z) = c_1 < \infty \tag{20}$$

with $Q(x,dz) = q(x,z)\mu(dz)$ and if $\mu(dx) = m(x)dx$ the target density m satisfies

$$\inf_{z \in \mathcal{B}_\delta(z^*)} m(z) = C_0 > 0 \tag{21}$$

Then Assumption 3.2 holds for the independence coupling $\Gamma_{x,y}(du,dv) = \mathcal{P}(x,du)\mathcal{P}(y,dv)$.

Proof. Clearly

$$\frac{c_0}{c_1} < \alpha(x,z) \le 1$$

uniformly over $(x,z) \in \mathbf{X} \times \mathcal{B}_\delta(z^*)$. Let $I_\gamma \subset \mathcal{B}_\delta(z^*)$ be ball of diameter γ. Then

$$\inf_x \mathcal{P}(x, I_\gamma) \ge |I_\gamma| \inf_x \inf_{z \in \mathcal{B}_\delta(z^*)} q(x,z)m(z)\alpha(x,z) \ge \gamma C_0 \frac{c_0^2}{c_1}$$

Consider the coupling $\Gamma_{x,y}(du,dv) = \mathcal{P}(x,du)\mathcal{P}(y,dv)$. We have

$$\begin{aligned}\inf_{(x,y)\in\mathbf{X}\times\mathbf{X}} \Gamma_{x,y}((a,b) : |a-b| < \gamma) &\ge \inf_{(x,y)\in\mathbf{X}\times\mathbf{X}} \mathcal{P}(x, \mathcal{B}_{\frac{\gamma}{2}}(z^*))\mathcal{P}(y, \mathcal{B}_{\frac{\gamma}{2}}(z^*)) \\ &\ge \frac{\gamma^2}{4} C_0^2 \frac{c_0^4}{c_1^2}\end{aligned}$$

establishing the result. □

We show these conditions for our application. Define

$$M(x,W) \equiv I + \Sigma(x,W)$$

so the eigenvalues of M satisfy

$$\begin{aligned}\lambda_{\min}(M(x,W)) &\ge 1 \\ \lambda_{\max}(M(x,W)) &\le 1 + N \\ \mathrm{trace}(M(x,W)) &\le 2N,\end{aligned} \tag{22}$$

and since M is positive definite

$$|M|^{1/N} \le N^{-1}\,\mathrm{trace}(M),$$

so $|M| \leq 2^N$. In the application to Gaussian process models, $\mu(dx)$ is given by (17), so

$$\mu(dx) \geq \frac{1}{(1+N)^{N/2}}(b + \|z\|_2^2)^{-(N+a)/2} = C_0,$$
$$\mu(dx) \leq \left(b + \frac{\|z\|_2^2}{1+N}\right)^{-(N+a)/2} = C_1,$$

and

$$c_0 = \frac{1}{C_1}\frac{1}{\sqrt{2\pi v}}e^{-\frac{\pi^2}{2v}} \leq q(x,y) \leq \frac{1}{C_0}\frac{1}{\sqrt{2\pi v}} = c_1,$$

which shows 3.4 for the Gaussian process application. An easily verifiable condition for our example that implies Assumption 3.1 is the following

Remark 3.5. *Consider a Metropolis-Hastings algorithm with proposal kernel $Q(x, dy)$ and acceptance probability $\alpha(x, y)$. Recalling the definition of the metric d from (19), suppose that*

$$\alpha(x,y) \in \mathrm{Lip}(d)$$

for every $y \in \mathbf{X}$, and that

$$d(Q(x,\,\cdot\,), Q(y,\,\cdot\,)) \leq C_0 d(x,y)\,. \tag{23}$$

Then Assumption 3.1 holds.

Proof. First observe that because $\phi \in \mathrm{Lip}(d)$ implies that ϕ is bounded we have that

$$\sup_{x\in\mathbf{X}} \int \phi(y) Q(x,dy) = C_1 < \infty\,.$$

Next observe that we have

$$\begin{aligned}\int \phi(z)(\mathcal{P}(x,dz) - \mathcal{P}(y,dz)) &= \int \phi(z)\alpha(x,z)Q(x,dz) + \int \phi(x)(1-\alpha(x,z))Q(x,dz)\\ &\quad - \int \phi(z)\alpha(y,z)Q(y,dz) - \int \phi(y)(1-\alpha(y,z))Q(y,dz)\\ &= \int \phi(z)[\alpha(x,z)Q(x,dz) - \alpha(y,z)Q(y,dz)]\\ &\quad - \phi(x)\int \alpha(x,z)Q(x,dz) + \phi(y)\int \alpha(y,z)Q(y,dz)\end{aligned}$$

Focusing now on the first term

$$\begin{aligned}\int \phi(z)[\alpha(x,z)Q(x,dz) - \alpha(y,z)Q(y,dz)] &= \int \phi(z)(\alpha(x,z) - \alpha(y,z))Q(x,dz)\\ &\quad + \int \phi(z)\alpha(y,z)(Q(x,dz) - Q(y,dz))\end{aligned}$$

$$
\begin{aligned}
&\leq \int \phi(z)|\alpha|_{\mathrm{Lip}(d)} d(x,y) Q(x,dz) \\
&+ \int \phi(z)(Q(x,dz) - Q(y,dz)) \\
&\leq C_1 |\alpha|_{\mathrm{Lip}(d)} d(x,y) + C_0 d(x,y)
\end{aligned}
$$

Recognizing that if $\alpha(x,z)$ is Lipschitz in its first argument, then so is $\varphi(x,z) = \phi(x)\alpha(x,z)$, and applying a similar argument to the above gives a similar bound for the term

$$\phi(y) \int \alpha(y,z) Q(y,dz) - \phi(x) \int \alpha(x,z) Q(x,dz).$$

□

We now show these conditions for the Gaussian process application. The next remark implies that we can work with the ratio of the target densities to verify that α is Lipschitz.

Remark 3.6. *Define* $\beta(x,y)$ *by*

$$\alpha(x,y) = \beta(x,y) \wedge 1,$$

so that

$$\beta(x,y) = \frac{q(y,x)}{q(x,y)}$$

is the target times the Hastings ratio. Then $\beta \in \mathrm{Lip}(d)$ *implies* $\alpha \in \mathrm{Lip}(d)$.

Proof. We have

$$\alpha(x,y) = \beta(x,y) \wedge 1$$

so for $d(x,y) \neq 0$

$$\frac{|\alpha(x,z) - \alpha(y,z)|}{d(x,y)} \leq \frac{|\beta(x,z) - \beta(y,z)|}{d(x,y)}.$$

□

This implies that, for example, it is enough to check that $\beta(x,y)$ has bounded derivative. We show that $\sup_x \frac{\partial}{\partial y}\beta(x,y) < \infty$ in Section 3.3. The proof of $\sup_y \frac{\partial}{\partial x}\beta(x,y) < \infty$ is similar and omitted. Finally, we show (23) for the Gaussian process application. Without loss of generality, take $x < y$. Then we have

$$
\begin{aligned}
d(Q(x,\cdot), Q(y,\cdot)) &= \sup_{|\phi|_{\mathrm{Lip}(d)} < 1} \int \phi(z)(Q(x,dz) - Q(y,dz)) \\
&= \sup_{|\phi|_{\mathrm{Lip}(d)} < 1} \int_{-\pi}^{\pi} \tfrac{1}{\sqrt{2\pi v}} \phi(z) \sum_{k=-\infty}^{\infty} (e^{-\frac{(z-x-2\pi k)^2}{2v}} - e^{-\frac{(z-y-2\pi k)^2}{2v}}) dz
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2\pi} \sup_{|\phi|_{\mathrm{Lip}(d)}<1} \int_{-\pi-x}^{\pi-x} \phi(\xi+x)\vartheta_3(-\tfrac{\xi}{2}, e^{-v/2})d\xi - \int_{-\pi-y}^{\pi-y} \phi(\xi+y)\vartheta_3(-\tfrac{\xi}{2}, e^{-v/2})d\xi \\
&\le \frac{|x-y|}{\delta} + \int_{\pi-y}^{\pi-x} \frac{1}{2\pi}\vartheta_3(-\frac{\xi}{2}, e^{-v/2})d\xi - \int_{-\pi-y}^{-\pi-x} \frac{1}{2\pi}\vartheta_3(-\frac{\xi}{2}, e^{-v/2})d\xi \\
&\le \frac{|x-y|}{\delta} + 2C|x-y|,
\end{aligned}
$$

where ϑ_3 is the third Jacobi theta function, and the last step followed because $\vartheta_3(-\frac{\xi}{2}, e^{-v/2})$ is clearly bounded since $e^{-v/2} \in (0,1)$.

3.3 Approximating Kernels for Gaussian Process Models

So far we have said nothing of approximations. Now we derive an approximation error condition that gives bounds in Wasserstein metrics that can be verified for our application. As in the proof of Theorem 1.10, we work initially with the generator and Poisson equation. Define

$$
L = \mathcal{P} - I, \quad U(x) = \sum_{k=0}^{\infty} \widetilde{\phi}(x),
$$

with $\widetilde{\phi} = \phi - \mu\phi$ and $\phi \in \mathrm{Lip}_1(d)$, so that $LU = -\widetilde{\phi}$. Then, with the same notation as in (29),

$$
\frac{1}{n}\sum_{k=0}^{n-1} \widetilde{\phi}(X_k^\epsilon) = \frac{U(X_0^\epsilon) - U(X_n^\epsilon)}{n} + \frac{1}{n}M_n^\epsilon + \frac{1}{n}\sum_{k=0}^{n-1}(\mathcal{P}_\epsilon - \mathcal{P})U(X_k^\epsilon).
$$

Observe that

$$
U(x) - U(y) = \sum_{k=0}^{\infty} \mathcal{P}^k\phi(x) - \mathcal{P}^k\phi(y) \le \sum_{k=0}^{\infty} \bar{\alpha}^k d(x,y) = \frac{1}{1-\bar{\alpha}} d(x,y),
$$

so that if $\phi \in \mathrm{Lip}_1(d)$, then $U \in \mathrm{Lip}_{\frac{1}{1-\bar{\alpha}}}(d)$. So since

$$
\mathbf{E}\frac{1}{n}\sum_{k=0}^{n-1} \widetilde{\phi}(X_k^\epsilon) = \frac{\mathbf{E}[U(X_0^\epsilon) - U(X_n^\epsilon)]}{n} + \frac{1}{n}\sum_{k=0}^{n-1}\mathbf{E}(\mathcal{P}_\epsilon - \mathcal{P})U(X_k^\epsilon),
$$

we need only bound $\mathcal{P}_\epsilon - \mathcal{P}$ in the Wasserstein metric. This motivates

Assumption 3.7. *Suppose $\mathcal{P}$ satisfies Assumptions 3.2 and 3.1, and $\bar{\alpha}$ is defined as in Theorem 3.3. Then*

$$
\sup_{x\in\mathbf{X}} d(P(x,\cdot), \mathcal{P}_\epsilon(x,\cdot)) < \epsilon(1-\bar{\alpha}).
$$

Consider the Gaussian process example from above. Define

$$
I_j = \{y : \mathrm{argmin}_k\, |y - \theta_k| = j\}
$$

and observe that

$$Q_\epsilon = \sum_{k=1}^{\infty} \delta_{\theta_k} Q(x, I_k)$$

For any ϕ we have

$$\begin{aligned}(\mathcal{P}\phi - \mathcal{P}_\epsilon\phi)(x) &= \int \phi(y)\alpha(x,y)Q(x,dy) + \int \phi(x)(1-\alpha(x,y))Q(x,dy) \\ &\quad - \int \phi(y)\alpha(x,y)Q_\epsilon(x,dy) - \int \phi(x)(1-\alpha(x,y))Q_\epsilon(x,dy) \\ &= \int \phi(y)\alpha(x,y)[Q-Q_\epsilon](x,dy) + \int \phi(x)(1-\alpha(x,y))[Q-Q_\epsilon](x,dy)\end{aligned}$$

So we would like to bound on $Q - Q_\epsilon$ in the Wasserstein-d metric. We have for any $f \in \text{Lip}_1(d_\beta)$

$$\begin{aligned}\int f(y)[Q-Q_\epsilon](x,dy) &= \sum_k \int_{I_k} f(y)[Q-Q_\epsilon](x,dy) \\ &\leq \left| \sum_k f(\theta_k)Q(x,I_k) - \int_{I_k} [f(\theta_k)+\epsilon]Q(x,dy) \right| \\ &\leq \int \epsilon Q(x,dy) = \epsilon.\end{aligned}$$

It is worth pointing out that so far this argument holds for any Q_ϵ obtained by an ϵ-discretization of the support of Q.

Now we need only show that $\alpha(x,y) \in \text{Lip}(d)$. By Remark 3.6, it is enough to show that $\beta(x,y)$ has uniformly bounded first derivative. The identity

$$\frac{\partial}{\partial y}\Sigma(y)^{-1} = -\Sigma^{-1}\frac{\partial \Sigma}{\partial y}\Sigma^{-1}$$

where

$$\left(\frac{\partial \Sigma}{\partial y}\right)_{ij} = \frac{\partial}{\partial y}\Sigma(y)_{ij}$$

will be used. We now compute the derivative

$$\begin{aligned}\frac{\partial}{\partial y}\beta(x,y) &= \frac{\partial}{\partial y}\frac{|I+\Sigma(y,W)|^{-1/2}\{b+z'(I+\Sigma(y,W))^{-1}z\}^{-\frac{a+N}{2}}}{L(z \mid x,W)} \\ &= \frac{1}{L(z \mid x,W)}\frac{\partial}{\partial y}|M(y,W)|^{-1/2}\{b+z'M(y,W)^{-1}z\}^{-\frac{a+N}{2}}.\end{aligned}$$

Defining $D(y)$ as the $N \times N$ matrix with entries

$$\{D(y)\}_{ij} = -\|w_i - w_j\|^2 e^{-y\|w_i-w_j\|^2} = -\|w_i-w_j\|^2\{\Sigma(y)\}_{ij}$$

we have

$$\frac{\partial}{\partial y_1}|M|^{-1/2} = \frac{1}{2}|M|^{1/2}\,\text{trace}\{MD\}$$

$$\frac{\partial}{\partial y_1}\{b + z'M^{-1}z\}^{-\frac{a+N}{2}} = \frac{a+N}{2}\{b + z'M^{-1}z\}^{-\frac{a+N+2}{2}} z'\{M^{-1}DM^{-1}\}z.$$

Observe that

$$\left|\frac{\partial}{\partial y}\beta(x,y)\right| \leq |M(x,W)|^{1/2}\{b + z'M(x,W)^{-1}z\}^{\frac{a+N}{2}}$$

$$\times\Bigg(|M(y,W)|^{-1/2}\frac{a+N}{2}\{b{+}z'M(y,W)^{-1}z\}^{-\frac{a+N+2}{2}}|z'\{M(y,W)^{-1}D(y)M(y,W)^{-1}\}z|$$

$$+ \{b + z'M(y,W)^{-1}z\}^{-\frac{a+N}{2}}\frac{1}{2}|M(y,W)|^{1/2}|\,\text{trace}\{M(y,W)D(y)\}|\Bigg).$$

We would like to bound this uniformly away from ∞. In addition to the bounds in (22), we will also need a bound on the norm of D. Observe that

$$\|D(y_1)\|_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\|w_i - w_j\|^2 e^{-y_1\|w_i - w_j\|^2} \leq \sum_{i=1}^{N}\sum_{j=1}^{N}\|w_i - w_j\|^2 \equiv \bar{D}^2.$$

It follows that

$$\lambda_{\max}(D(y_1)) \leq \bar{D}$$

and therefore applying standard inequalities for products of Hermitian matrices and quadratic forms we have

$$\left|\frac{\partial}{\partial y}\beta(x,y)\right| \leq 2^{\frac{N}{2}}(b + \|z\|_2^2)^{\frac{a+N}{2}} b^{-\frac{a+N+2}{2}}\left(\frac{a+N}{2}\bar{D}\|z\|_2^2 + 2^{\frac{N}{2}}N\bar{D}\right) \equiv C_1.$$

so the derivative is uniformly bounded. It follows that if Q_ϵ is obtained from an $\epsilon(1-\bar{\alpha})C_1^{-1}$-discretization of the support of Q, then Assumption 3.7 holds. Note that none of this required that y reside on a compact set; the compactness of the prior support will be used in the next example.

3.4 Use of low-rank approximations

In the previous example, the only source of approximation error was the use of an approximate proposal Q_ϵ, and it was enough to uniformly bound the derivative of α to control the approximation error. In this section, we consider a variation on the previous algorithm where both an approximate proposal Q_ϵ and an approximate acceptance probability α_ϵ are used.

When the number of points N at which the process is sampled is large, it is computationally and numerically difficult to compute a spectral decomposition of $\Sigma(x, W)$

at even a single point. Therefore in addition to discretizing the state space for x, it is common to approximate $\Sigma(x, W)$ by its partial spectral decomposition

$$\Sigma = U\Lambda U' \approx U\Lambda_\epsilon U'$$

where Λ_ϵ is a diagonal matrix that is equal to Λ in its first r diagonal entries and is zero in its remaining diagonal entries. The resulting algorithm therefore has both an approximate proposal Q_ϵ, where the approximation error arises from discretization, and an approximated acceptance probability α_ϵ, where the approximation error arises from using a partial spectral decomposition.

The approximate acceptance ratio α_ϵ can be expressed as

$$\begin{aligned}\alpha_\epsilon(x,y) &= 1 \wedge \frac{|U(I+\Lambda_\epsilon(x,W))U'|^{-\frac{1}{2}}\{b+z'U(I+\Lambda_\epsilon(x,W))^{-1}U'z\}^{-\frac{a+N}{2}}}{|U(I+x_2\Lambda_\epsilon(y,W))U'|^{-\frac{1}{2}}\{b+z'U(I+\Lambda_\epsilon(y,W))^{-1}U'z\}^{-\frac{a+N}{2}}} \\ &= 1 \wedge \frac{(b+\sum_{i=1}^r \frac{1}{1+\lambda_i(y,W)}+N-r)^{\frac{a+N}{2}}\prod_{i=1}^r(1+\lambda_i(y,W))^{1/2}}{(b+\sum_{i=1}^r \frac{1}{1+\lambda_i(x,W)}+N-r)^{\frac{a+N}{2}}\prod_{i=1}^r(1+\lambda_i(x,W))^{1/2}},\end{aligned}$$

for $\lambda_i(x, W)$ the ith largest eigenvalue of $\Sigma(x, W)$.

Now for any ϕ we have

$$\begin{aligned}(\mathcal{P}\phi - \mathcal{P}_\epsilon\phi)(x) = &\int \phi(y)\alpha(x,y)Q(x,dy) + \int \phi(x)(1-\alpha(x,y))Q(x,dy) \\ &- \int \phi(y)\alpha_\epsilon(x,y)Q_\epsilon(x,dy) - \int \phi(x)(1-\alpha_\epsilon(x,y))Q_\epsilon(x,dy),\end{aligned}$$

and adding and subtracting, we get

$$\alpha(x,y)Q(x,dy) - \alpha_\epsilon(x,y)Q_\epsilon(x,dy) = \alpha(x,y)(Q-Q_\epsilon)(x,dy) + (\alpha-\alpha_\epsilon)(x,y)Q_\epsilon(x,dy).$$

We already know how to deal with the first term, so it remains to handle the second term. For $f \in \mathrm{Lip}_1(d_\beta)$, we need

$$\int f(y)(\alpha-\alpha_\epsilon)(x,y)Q_\epsilon(x,dy) \leq \epsilon,$$

which depends on how well α approximates α_ϵ, rather than how well Q approximates Q_ϵ. Observe that if $f \in \mathrm{Lip}_1(d_\beta)$, then $|f|_\infty < 1$, so

$$\begin{aligned}\int f(y)(\alpha-\alpha_\epsilon)(x,y)Q_\epsilon(x,dy) &\leq |f|_\infty \int (\alpha-\alpha_\epsilon)(x,y)Q_\epsilon(x,dy), \\ &\leq \int (\alpha-\alpha_\epsilon)(x,y)Q_\epsilon(x,dy),\end{aligned}$$

so we need only make the integral on the right side small. We have

$$\int (\alpha-\alpha_\epsilon)(x,y)Q_\epsilon(x,dy) = \frac{1}{|\{k : \theta_k \in I(x)\}|}\sum_{\theta_k \in I(x)} (\alpha-\alpha_\epsilon)(x,\theta_k),$$

with $I(x) = \{y : Q(x, y) > 0\}$, so the desired bound will follow if

$$\sum_{\theta_k \in I(x)} (\alpha - \alpha_\epsilon)(x, \theta_k) < |\{k : \theta_k \in I(x)\}|\epsilon,$$

and a sufficient condition is

$$\sup_{k:\theta_k \in I(x)} (\alpha - \alpha_\epsilon)(x, \theta_k) < \epsilon. \tag{24}$$

It is always possible to make $\epsilon = 0$ by putting $r = N$, though naturally this would eliminate any computational advantage. Regardless, it is clear that for every ϵ and every x, y there exists $r(\epsilon, x, y) \leq N$ such that

$$(\alpha - \alpha_\epsilon)(x, y) < \epsilon,$$

so by choosing the rank of the partial spectral decomposition in an adaptive way depending on the state, the proposal, and the desired approximation error, we can achieve (24). Numerical experiments showing that this approximation can be very accurate in some cases using $r \ll N$ can be found in [12].

4 Discussion

There has been considerable interest in utilizing approximations of transition kernels to reduce the computational complexity of MCMC. The results we give in §1 indicate that when the approximating kernel is sufficiently accurate in a weighted total variation norm, pathwise quantities generated from the approximating kernel will provide useful approximations of posterior expectations for large classes of test functions. The applications we considered suggest that it is possible to verify the assumptions under which the generic bounds are given for real algorithms, albeit with some effort. We also outline a way to achieve our main error condition in Assumption 1.4 using state-adaptive control of total variation, which is likely to be easier to achieve in algorithm development.

Much of the existing literature focuses on minibatching algorithms, and we consider a simple example of such an algorithm in §2.4. The bounds we give are quite sharp, and the numerical experiments verify that even in simple examples, a large portion of the data is necessary to achieve reasonable bounds on the approximation error. A similar conclusion was reached by [4], which also considered minibatching Metropolis-Hastings. It may at this point be reasonable to conclude that approximating likelihood ratios based on subsamples of the full data is not a very efficient path to generating approximating kernels. There have been some recent proposals of algorithms that avoid the Metropolis-Hastings step entirely [5], though approximation error guarantees are not given. It seems likely that minibatching will be most valuable in combination with other strategies, such as divide-and-conquer algorithms [17, 18, 24]. Alternative strategies, such as the use of Gaussian approximations like the one we consider in §2.3, may hold more promise as standalone algorithms.

The spatial statistics and nonparametrics literature is replete with approximate algorithms because of the high computational cost of exact computation. Our analysis of Gaussian process models in §3 suggests that the common practice of discretization and low-rank approximation of the covariance will give useful approximation to the exact kernel when the number of eigenvectors used in the approximation is chosen adaptively. The relative computational advantage of the approximate algorithm will depend on the rate of decay of the eigenvalues of the covariance matrix. It would perhaps be more natural to study this problem in the infinite-dimensional limit, where the notion of decay rate can be made precise.

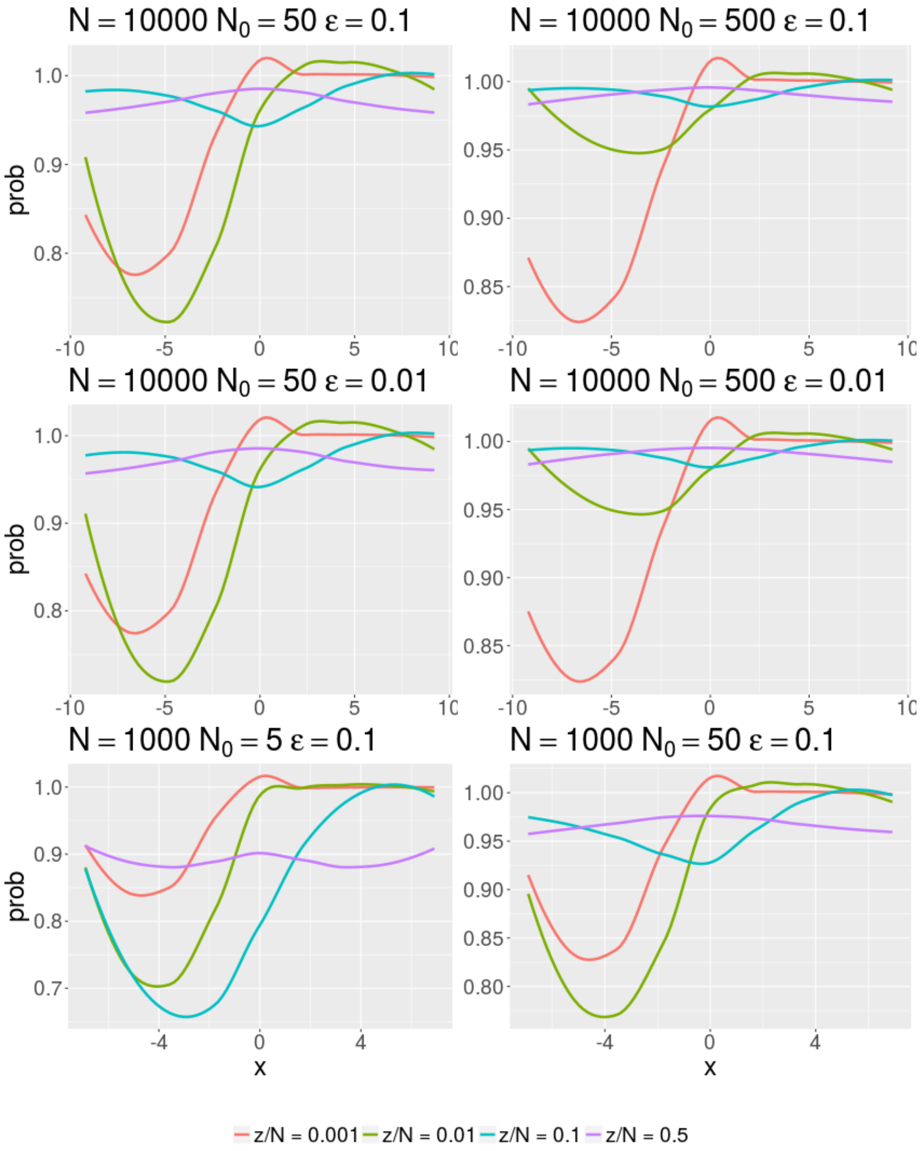


Figure 2: Probability of achieving (13) as a function of x for different values of N, N_0, ϵ and z.

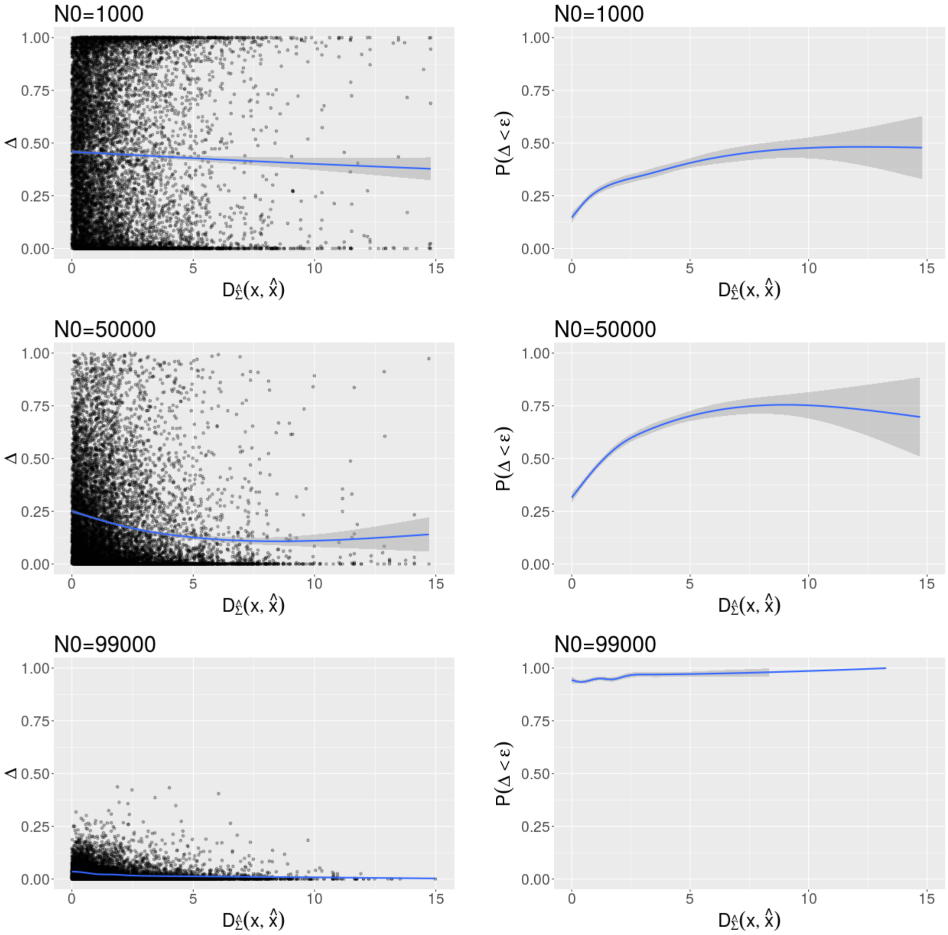


Figure 3: Samples of $\Delta(x, y, \epsilon)$ as a function of $D_{\hat{\Sigma}}(x, \hat{x})$ (left column) and estimated $\mathbf{P}[\Delta(x, y, \epsilon) < \epsilon]$ as a function of $D_{\hat{\Sigma}}(x, \hat{x})$ (right column) for different values of N_0; the function estimation uses LOESS local linear smoothing.

A Additional proofs

A.1 Proof of Theorem 1.10

For any ϕ with $\phi \leq V^{\frac{1}{2}}$ define $\widetilde{\phi} = \phi - \mu\phi$

$$U(x) = \sum_{k=0}^{\infty} \mathcal{P}^k \widetilde{\phi} \tag{25}$$

Now for $p \in (0, 1]$ we have $\gamma_p \in (0, 1)$ and $K_p > 0$ so that

$$\mathcal{P}V^p(x) \leq \gamma_p V^p(x) + K_p \tag{26}$$

and

$$|\mathcal{P}\widetilde{\phi}|_{(p)} \leq \alpha_p |\widetilde{\phi}|_{(p)}$$

is the weighted TV norm built on V^p with an appropriate β_p. Now observe that

$$|U|_{(p)} \leq \sum_{k=0}^{\infty} \alpha_p^k |\widetilde{\phi}|_{(p)} = \frac{|\widetilde{\phi}|_{(p)}}{1 - \alpha_p}$$

with $p = \frac{1}{2}$. Observe that since $\phi < V^{\frac{1}{2}}$, we have $\mu\phi < \mu V^{\frac{1}{2}} < \infty$, so $|\widetilde{\phi}| < \mu V^{\frac{1}{2}} + V^{\frac{1}{2}}$, and

$$\begin{aligned} |U(x)|_{1/2} &\leq (\mu V^{\frac{1}{2}} + V^{\frac{1}{2}}(x)) \frac{1}{1 - \alpha_{(1/2)}} \\ &\leq C(1 + V^{\frac{1}{2}}(x)) \end{aligned}$$

for

$$C = \frac{1}{1 - \alpha_{(1/2)}} \vee \frac{1}{1 - \alpha_{(1/2)}} \vee \mu V^{\frac{1}{2}}$$

This implies that

$$|U(x)| \leq C(1 + V^{\frac{1}{2}}(x)). \tag{27}$$

Note that

$$(\mathcal{P} - I)U(x) = -\widetilde{\phi}(x) \tag{28}$$

so

$$\begin{aligned} U(X_n^\epsilon) - U(X_0^\epsilon) &= \sum_{k=0}^{n-1} U(X_{k+1}^\epsilon) - U(X_k^\epsilon) = \sum_{k=0}^{n-1} [U(X_{k+1}^\epsilon) - \mathcal{P}_\epsilon U(X_k^\epsilon)] + \sum_{k=0}^{n-1} (\mathcal{P}_\epsilon - I)U(X_k^\epsilon) \\ &= \sum_{k=0}^{n-1} [U(X_{k+1}^\epsilon) - \mathcal{P}_\epsilon U(X_k^\epsilon)] + \sum_{k=0}^{n-1} (\mathcal{P} - I)U(X_k^\epsilon) + \sum_{k=0}^{n-1} (\mathcal{P}_\epsilon - \mathcal{P})U(X_k^\epsilon) \end{aligned}$$

Using (28) and defining $m^\epsilon_{k+1} = U(X^\epsilon_{k+1}) - \mathcal{P}_\epsilon U(X^\epsilon_k)$ and $M^\epsilon_n = \sum_{k=1}^n m^\epsilon_k$, we have

$$\frac{1}{n}\sum_{k=0}^{n-1}\phi(X^\epsilon_k) - \mu\phi = \frac{U(X^\epsilon_0) - U(X^\epsilon_n)}{n} + \frac{1}{n}M^\epsilon_n + \frac{1}{n}\sum_{k=0}^{n-1}(\mathcal{P}_\epsilon - \mathcal{P})U(X^\epsilon_k) \quad (29)$$

Now

$$\mathbf{E}[(m^\epsilon_{k+1})^2|\mathcal{F}_k] \leq \mathcal{P}_\epsilon(U^2)(X^\epsilon_k) - [\mathcal{P}_\epsilon(U)(X^\epsilon_k)]^2$$

and

$$\mathbf{E}(\frac{1}{n}M^\epsilon_n)^2 \leq \frac{1}{n^2}\sum_{k=1}^n \mathbf{E}(m^\epsilon_k)^2$$

It follows from (27) that

$$U^2(x) \leq 2C^2(1 + V(x)).$$

So then with $X^\epsilon_0 = x_0$

$$\begin{aligned}\mathbf{E}[\mathcal{P}_\epsilon(U^2)(X^\epsilon_k)] &\leq \mathcal{P}_\epsilon 2C(1 + \mathcal{P}^k_\epsilon V(x_0)) \\ &\leq 2C + 2C\mathcal{P}^{k+1}_\epsilon V(x_0).\end{aligned}$$

We proceed by bounding the square of each term on the right side of (29). We have

$$\begin{aligned}\mathcal{P}^{k+1}_\epsilon(1+V) &\leq \left(\gamma^{k+1}_\epsilon + \frac{1-\gamma^{k+1}_\epsilon}{1-\gamma_\epsilon}K_\epsilon\right)(1+V) \\ &\leq \left(\gamma^{k+1}_\epsilon + \frac{K_\epsilon}{1-\gamma_\epsilon}\right)(1+V)\end{aligned}$$

so

$$\begin{aligned}\sum_{k=0}^{n-1}\sum_{k=0}^{n-1}\mathbf{E}[(m^\epsilon_{k+1})^2] &\leq 2C^2\left(\frac{nK_\epsilon}{1-\gamma_\epsilon} + \frac{1-\gamma^n_\epsilon}{1-\gamma_\epsilon}\right)(1+V) \\ \frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{k=0}^{n-1}\mathbf{E}[(m^\epsilon_{k+1})^2] &\leq 2C^2\left(\frac{K_\epsilon}{n\{1-\gamma_\epsilon\}} - \frac{1-\gamma^n_\epsilon}{n^2\{1-\gamma_\epsilon\}}\right)(1+V).\end{aligned}$$

Now for the term

$$\frac{1}{n}\sum_{k=0}^{n-1}(\mathcal{P}_\epsilon - \mathcal{P})U(X^\epsilon_k).$$

Since $C^{-1}|U| < 1 + V^{\frac{1}{2}}$ and for $|\phi| < 1 + V$

$$(\mathcal{P}_\epsilon - \mathcal{P})(\phi) \leq \epsilon(1 + \delta V)$$

by Jensen's inequality

$$(\mathcal{P}_\epsilon - \mathcal{P})(\phi^{\frac{1}{2}}) \le \sqrt{\epsilon(1+\delta V)} \le \sqrt{\epsilon} + \sqrt{\epsilon \delta V} \le \sqrt{\epsilon}(1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}})$$

so we have

$$C^{-1} |(\mathcal{P}_\epsilon - \mathcal{P}) U(x)| \le \epsilon^{\frac{1}{2}} (1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(x)).$$

Using these inequalities, and taking $k \ge j$ without loss of generality, we get

$$\begin{aligned}
(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_k)(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_j) &\le C^2 \epsilon (1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(X^\epsilon_k))(1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(X^\epsilon_j)) \\
\mathbf{E}\left[(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_k)(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_j)\right] &\le C^2 \epsilon \mathbf{E}\left[\mathbf{E}\left[(1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(X^\epsilon_k)) \mid \mathcal{F}_j\right] (1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(X^\epsilon_j))\right] \\
&\le C^2 \epsilon \mathbf{E}\left[\mathcal{P}^{k-j}_\epsilon (1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(X^\epsilon_j))(1 + \delta^{\frac{1}{2}} V^{\frac{1}{2}}(X^\epsilon_j))\right] \\
&\le 2C^2 \epsilon \mathcal{P}^k_\epsilon (1 + \delta V) \qquad (30)
\end{aligned}$$

Now since (30) is bounded by

$$2C^2 \epsilon \mathcal{P}^k_\epsilon (1 + \delta V) \le 2C^2 \epsilon \delta \left(\gamma^k_\epsilon V + \frac{K_\epsilon}{1 - \gamma_\epsilon} + \frac{1 - \delta}{\delta} \right)$$

we get

$$\begin{aligned}
\sum_{k=0}^{n-1} \sum_{j=0}^{n-1} \mathbf{E}\left[(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_k)(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_j)\right] &\le n 2C^2 \epsilon \delta \left(\frac{1 - \gamma^n_\epsilon}{1 - \gamma_\epsilon} V + \frac{n K_\epsilon}{1 - \gamma_\epsilon} + \frac{n(1 - \delta)}{\delta} \right) \\
\frac{1}{n^2} \sum_{k=0}^{n-1} \sum_{j=0}^{n-1} \mathbf{E}\left[(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_k)(\mathcal{P}_\epsilon - \mathcal{P}) U(X^\epsilon_j)\right] &\le 2C^2 \epsilon \delta \left(\frac{1}{n} \frac{V}{1 - \gamma_\epsilon} + \frac{K_\epsilon}{1 - \gamma_\epsilon} + \frac{1 - \delta}{\delta} \right).
\end{aligned}$$

Finally we have

$$\begin{aligned}
\frac{(U(X^\epsilon_0) - U(X^\epsilon_n))^2}{n^2} &\le \frac{2U^2(X^\epsilon_0) + 2U^2(X^\epsilon_n)}{n^2} \\
\mathbf{E} \frac{(U(X^\epsilon_0) - U(X^\epsilon_n))^2}{n^2} &\le \frac{4C^2}{n^2} \left(\mathbf{E}[1 + V(X^\epsilon_0)] + \mathbf{E}[1 + V(X^\epsilon_n)] \right) \\
&\le \frac{4C^2}{n^2} \left(1 + V(x_0) + \gamma^n_\epsilon V(x_0) + \frac{1 - \gamma^n_\epsilon}{1 - \gamma_\epsilon} K_\epsilon \right) \\
&\le \frac{4C^2}{n^2} \left(1 + (1 + \gamma^n_\epsilon) V + \frac{K_\epsilon}{1 - \gamma_\epsilon} \right)
\end{aligned}$$

Giving us

$$\begin{aligned}
\mathbf{E} \left(\frac{1}{n} \sum_{k=0}^{n-1} \phi(X^\epsilon_k) - \mu\phi \right)^2 \le 6C^2 &\left(\frac{K_\epsilon}{n\{1 - \gamma_\epsilon\}} - \frac{1 - \gamma^n_\epsilon}{n^2\{1 - \gamma_\epsilon\}} \right) (1 + V) \\
&+ 6C^2 \epsilon \delta \left(\frac{1}{n} \frac{V}{1 - \gamma_\epsilon} + \frac{K_\epsilon}{1 - \gamma_\epsilon} + \frac{1 - \delta}{\delta} \right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{12C^2}{n^2}\left(1 + (1+\gamma^n)V(x_0) + \frac{K_\epsilon}{1-\gamma_\epsilon}\right) \\
&\le 6C^2\epsilon\delta\left(\frac{K_\epsilon}{1-\gamma_\epsilon} + \frac{1-\delta}{\delta}\right) + \frac{6C^2}{n}\left(\frac{K_\epsilon+1}{1-\gamma_\epsilon}\right)(1+V) + \mathcal{O}\left(\frac{1}{n^2}\right)
\end{aligned}
$$

concluding the proof of Theorem 1.10.

A.2 Proof of Proposition 2.1

Proof of proposition. We have

$$
\begin{aligned}
\mathbf{E}[V(\beta) \mid (\beta^*, \Omega^*)] &= \mathbf{E}[\mathbf{E}[V(\beta) \mid \Omega] \mid \beta^*] \\
\mathbf{E}[V(\beta) \mid \Omega] &= \operatorname{trace}(W'DW(W'DW)^{-1}) + \mathbf{E}[\beta \mid \Omega]'(W'DW)\mathbf{E}[\beta \mid \Omega] \\
&= p + ((W'DW)^{-1}W'\Omega)'(W'DW)((W'DW)^{-1}W'\Omega) \\
&= p + \Omega'W(W'DW)^{-1}W'\Omega
\end{aligned}
$$

so then

$$
\begin{aligned}
\mathbf{E}[\mathbf{E}[V(\beta) \mid \Omega] \mid \beta^*] &= \mathbf{E}[p + \Omega'W(W'DW)^{-1}W'\Omega \mid \beta^*] \\
&= p + \operatorname{trace}(W(W'DW)^{-1}W') + \mathbf{E}[\Omega \mid \beta^*]'W(W'DW)^{-1}W'\mathbf{E}[\Omega \mid \beta^*] \\
&= K + \mathbf{E}[\Omega \mid \beta^*]'W(W'DW)^{-1}W'\mathbf{E}[\Omega \mid \beta^*].
\end{aligned}
$$

Now we need to compute

$$
\begin{aligned}
E[\omega_i \mid \beta^*] &= (n_i - y_i)\left(w_i\beta^* - \frac{\phi(-w_i\beta^*)}{\Phi(-w_i\beta^*)}\right) + y_i\left(w_i\beta^* + \frac{\phi(-w_i\beta^*)}{1-\Phi(-w_i\beta^*)}\right) \\
&= n_i w_i \beta^* + y_i \frac{\phi(-w_i\beta^*)}{1-\Phi(-w_i\beta^*)} - (n_i - y_i)\frac{\phi(w_i\beta^*)}{1-\Phi(w_i\beta^*)} \\
&\le \gamma n_i w_i
\end{aligned}
$$

whenever $0 < y_i < n_i$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard Gaussian distribution function and density function, respectively.

From [20, Equations 7.8.1, 7.8.2], if $x \ge 0$,

$$
\frac{2}{x + (x^2+4)^{1/2}} < \frac{1-\Phi(x)}{\phi(x)} \le \frac{2}{x + (x^2 + 8/\pi)^{1/2}}
$$

so

$$
\frac{\phi(x)}{1-\Phi(x)} = \frac{1}{2}(x + \sqrt{x^2 + h(x)})
$$

for some function $h(x)$ satisfying $8/\pi < h(x) < 4$ for $x \ge 0$. It follows that for $x \le 0$ we have

$$
\frac{\phi(-x)}{1-\Phi(-x)} = \frac{1}{2}(-x + \sqrt{(-x)^2 + h(-x)}).
$$

On the other hand, for $x < 0$,

$$\phi(x) < \frac{\phi(x)}{1-\Phi(x)} < 2\phi(x)$$

and for $x > 0$

$$\phi(x) < \frac{\phi(-x)}{1-\Phi(-x)} < 2\phi(x).$$

Let $\xi_i = w_i\beta^*$. When $\xi_i < 0$, we have

$$\begin{aligned} n_i\xi_i + y_i\frac{\phi(-\xi_i)}{1-\Phi(-\xi_i)} - (n_i - y_i)\frac{\phi(\xi_i)}{1-\Phi(\xi_i)} &\leq n_i\xi_i + y_i\frac{1}{2}(-\xi_i + \sqrt{\xi_i^2 + h(-\xi_i)}) - (n_i - y_i)\phi(\xi_i) \\ &\leq (n_i - y_i)\xi_i \end{aligned}$$

while if $\xi_i > 0$

$$\begin{aligned} n_i\xi_i + y_i\frac{\phi(-\xi_i)}{1-\Phi(-\xi_i)} - (n_i - y_i)\frac{\phi(\xi_i)}{1-\Phi(\xi_i)} &\leq n_i\xi_i + 2y_i\phi(\xi_i) - (n_i - y_i)\frac{1}{2}(\xi_i + \sqrt{\xi_i^2 + h(\xi_i)}) \\ &\leq (n_i - (n_i - y_i))\xi_i + 2y_i\phi(\xi_i). \end{aligned}$$

For ξ_i sufficiently large, this is strictly less than $n_i\xi_i$; in particular, we just need

$$2y_i\phi(\xi_i) < (n_i - y_i)\xi_i$$

for which it is enough that

$$\xi_i > 2\sqrt{\log\left(\sqrt{\frac{2}{\pi}}\frac{y_i}{n_i - y_i}\right)}.$$

Define the set

$$\mathcal{B} = \left\{\beta : w_i\beta < 2\sqrt{\log\left(\sqrt{\frac{2}{\pi}}\frac{y_i}{n_i - y_i}\right)} \text{ for some } i\right\}.$$

This set is compact, and on this set,

$$\mathbf{E}[\Omega \mid \beta]'W(W'DW)^{-1}W'\mathbf{E}[\Omega \mid \beta] < C$$

for $C < \infty$. Outside this set, we have just showed that we have $\mathbf{E}[\omega_i \mid \beta] < \gamma n_i w_i\beta$ with $0 < \gamma < 1$. Then

$$\begin{aligned} \mathbf{E}[\Omega \mid \beta]'W(W'DW)^{-1}W'\mathbf{E}[\Omega \mid \beta] &< \beta'W'\gamma DW(W'DW)^{-1}W'\gamma DW\beta + C \\ &< \gamma^2\beta'(W'DW)\beta + C \end{aligned}$$

so

$$V(\beta) \leq \gamma^2 V(\beta) + K + C,$$

proving the result for $\mathcal{P}$. To get the result for $\mathcal{P}_\epsilon$, just observe that $\mathbf{E}[\Omega \mid \beta^*]$ is the same for $\mathcal{P}_\epsilon$ and $\mathcal{P}$, and the conditional update for β is unchanged in $\mathcal{P}_\epsilon$ compared to $\mathcal{P}$. Thus the constants γ, K, C are identical.

It is worth noting that if we make $\mathcal{B}$ "large enough" so that $2y_i\phi(\xi_i) \ll (n_i - y_i)\xi_i$, then the constant γ is approximately

$$\left(1 - \frac{y_i}{n_i}\right) \vee \frac{y_i}{n_i};$$

this is broadly consistent with the results in [13], which used conductance bounds. Of course, this Lyapunov function will only give us control over functions that grow like the Euclidean norm of β at infinity, whereas the conductance result gives a bound for $L^2(\mu)$ functions. □

A.3 Proof of Theorem 2.2

This proof is overall quite similar to the proof of [13, Theorem 3.1], except that here we do not assume that $z = 1$. As such, certain details are omitted.

We show that $p(\theta)$ has only one local maximum

Lemma A.1. *We have $p'(\theta) > 0$ for $\theta < \hat{\theta}$ and $p'(\theta) < 0$, for $\theta > \hat{\theta}$.*

Proof. Define $f(\theta) = \frac{\partial}{\partial\theta} \log p(\theta)$. By direct calculation

$$\begin{aligned} p'(\theta) &= p(\theta)\frac{\partial}{\partial\theta}\log p(\theta) \\ p'(\theta) &= p(\theta)\left(z - N\frac{e^\theta}{1+e^\theta} - \frac{\theta}{B}\right) = p(\theta)f(\theta) \end{aligned} \tag{31}$$

Observe that

$$f'(\theta) = -\frac{1}{B} - N\frac{e^\theta}{1+e^\theta}\frac{1}{1+e^\theta} \le -\frac{1}{B} < 0. \tag{32}$$

Since $f'(\theta) < 0$ for all $\theta \in \mathbb{R}$, it follows that $\{\theta : f(\theta) = 0\}$ has at most one point. Since $p(\theta) > 0$ for all $\theta \in \mathbb{R}$, we have by (31) that $\{\theta : p'(\theta) = 0\} = \{\theta : f(\theta) = 0\}$, so $\{\theta : p'(\theta) = 0\}$ has at most one point. Since $p'(\hat{\theta}) = 0$, the lemma follows. □

We estimate the mode for certain choices of B under weak conditions on the data.

Lemma A.2. *Suppose $1 < z < N - 1$ and put $B = \log N$. Then*

$$\log\frac{z-1}{N-(z-1)} < \hat{\theta} < \log\frac{z+1}{N-(z+1)}$$

Proof. The mode is the zero of the function

$$g(\theta) = \frac{\theta}{B} + N\frac{e^\theta}{1+e^\theta} - z$$

We have

$$g\left(\log \frac{z-1}{N-(z-1)}\right) = \frac{1}{B}\log \frac{z-1}{N+1-z} + N\frac{z-1}{N} - z$$
$$= \frac{1}{\log N}\log \frac{z-1}{N+1-z} - 1$$

so

$$\frac{1}{\log N}\log \frac{1}{N-1} - 1 < g\left(\log \frac{z-1}{N-(z-1)}\right) < \frac{1}{\log N}\log \frac{N-3}{3} - 1$$
$$-2 < g\left(\log \frac{z-1}{N-(z-1)}\right) < 0$$

while

$$g\left(\log \frac{z+1}{N-(z+1)}\right) = \frac{1}{B}\log \frac{z+1}{N-z-1} + N\frac{z+1}{N} - z$$
$$= \frac{1}{\log N}\log \frac{z+1}{N-z-1} + 1$$

so

$$\frac{1}{\log N}\log \frac{3}{N-4} + 1 < g\left(\log \frac{z+1}{N-(z+1)}\right) < \frac{1}{\log N}\log \frac{N-1}{1} + 1$$
$$0 < g\left(\log \frac{z+1}{N-(z+1)}\right) < 2,$$

giving the result. □

We now bound $\alpha(x, y)$ far from $\hat{\theta}$.

Lemma A.3. *For $\hat{\theta} \leq x \leq y$*

$$\alpha(x, y) \leq e^{-\frac{(y-x)(x+y-2\hat{\theta})}{2B}} \leq e^{-\frac{(x-\hat{\theta})(y-x)}{B}} \tag{33}$$

while for $y \leq x \leq \hat{\theta}$

$$\alpha(x, y) \leq e^{-\frac{(y-x)(x+y-2\hat{\theta})}{2B}} \leq e^{-\frac{(x-\hat{\theta})(y-x)}{B}}. \tag{34}$$

Proof. It follows from (32), the fact that $f(\hat{\theta}) = 0$, and the mean value theorem, that $f(w + \hat{\theta}) \leq \frac{-w}{B}$ for all $w \geq 0$. Combining with (31)

$$p'(w + \hat{\theta}) \leq -\frac{w}{B}p(w + \hat{\theta})$$

for all $w \geq 0$. So we can replace $w = \zeta + x - \hat{\theta}$ for any $\zeta \geq 0$ to obtain

$$p'(\zeta + x - \hat{\theta} + \hat{\theta}) \leq -\frac{(\zeta + x - \hat{\theta})}{B}p(\zeta + x - \hat{\theta} + \hat{\theta})$$

$$p'(\zeta + x) \le -\frac{(\zeta + x - \hat{\theta})}{B} p(\zeta + x).$$

Now we apply Grönwall's inequality and change variables inside the integral to $s = \zeta + x$ with Jacobian equal to one to obtain

$$\begin{aligned} p(\zeta + x) &\le p(x) \exp\left(-\int_x^{\zeta+x} \frac{s - \hat{\theta}}{B} ds \right) \\ &= p(x) \exp\left(-\frac{(s^2 - 2\hat{\theta}s)}{2B} \Big|_x^{\zeta+x} \right) = p(x) \exp\left(-\frac{\{(\zeta+x)^2 - x^2 - 2\hat{\theta}\zeta\}}{2B} \right) \\ &= p(x) \exp\left(-\frac{\zeta(\zeta + 2x - 2\hat{\theta})}{2B} \right). \end{aligned}$$

Since $y \ge x$, we can write $y = \zeta + x$ for $\zeta > 0$, so that $\zeta = y - x$, and

$$p(y) \le p(x) \exp\left(-\frac{(y-x)(y+x-2\hat{\theta})}{2B} \right),$$

which completes the proof of inequality (33). The proof of (34) is similar. □

We now construct a Lyapunov function for $\mathcal{P}$. This will show that X_n tends to drift toward the origin. Actually, it drifts toward $\hat{\theta}$, but this is "close" to the origin by Lemma A.2. Because we will later want to consider $\mathcal{P}_\epsilon$, which is generated by subsampling and therefore has a target with a different mode, it will be helpful to have a drift function that does not depend on the mode.

Lemma A.4. *Assume* $N > 13$ *and* $1 < z < N - 1$. *Put* $B = \log N$ *and* $c = 2 \log N$. *Then*

$$V(\theta) = e^{|\theta|}$$

is a Lyapunov function of $\mathcal{P}$.

Proof. We proceed in three cases.

Case 1 : $x > |\hat{\theta}| + c$. We have

$$\begin{aligned} 2c\mathbf{E}V(X) &= \int_{x-c}^{x} V(y)\alpha(x,y)dy + \int_x^{x+c} V(y)\alpha(x,y)dy + V(x)\int_{x-c}^{x+c} (1-\alpha(x,y))dy \\ &= \int_{x-c}^{x} (V(y) - V(x))dy + cV(x) + \int_x^{x+c} V(y)\alpha(x,y)dy + V(x)\int_x^{x+c} (1-\alpha(x,y))dy \\ &= \int_{x-c}^{x} V(y)dy + \int_x^{x+c} V(y)\alpha(x,y)dy + V(x)\int_x^{x+c} (1-\alpha(x,y))dy \end{aligned}$$

We bound the three terms on the right separately

$$\int_{x-c}^{x} V(y)dy = \int_{x-c}^{x} e^{|y|}dy = \int_{x-c}^{x} e^{y}dy = e^x\left(1 - e^{-c}\right) = V(x)\left(1 - e^{-c}\right).$$

Now since $x - \hat{\theta} \geq x - |\hat{\theta}| > c$

$$\begin{aligned}\int_x^{x+c} V(y)\alpha(x,y)dy &\leq \int_x^{x+c} e^y e^{-\frac{(x-\hat{\theta})(y-x)}{B}} dy \leq \int_x^{x+c} e^y e^{-\frac{c(y-x)}{B}} dy \\ &= e^{\frac{cx}{B}} \int_x^{x+c} e^{y(1-\frac{c}{B})} dy = e^{\frac{cx}{B}} \frac{B}{B-c} \left[e^{(x+c)\frac{B-c}{B}} - e^{x\frac{B-c}{B}} \right] \\ &= \frac{B}{B-c} e^{\frac{cx}{B}} e^{x\frac{B-c}{B}} \left[e^{c\frac{B-c}{B}} - 1 \right] = \frac{B}{B-c} V(x) \left[e^{c\frac{B-c}{B}} - 1 \right].\end{aligned}$$

Finally

$$\begin{aligned}V(x) \int_x^{x+c} (1-\alpha(x,y))dy &\leq V(x) \int_x^{x+c} (1 - e^{-\frac{c}{B}(y-x)})dy \\ &= V(x) \left(c - e^{\frac{cx}{B}} \int_x^{x+c} e^{-\frac{cy}{B}} dy \right) \\ &= V(x) \left(c + \frac{B}{c} \left[1 - e^{-\frac{c^2}{B}} \right] \right).\end{aligned}$$

We just need to show that for $c = 2\log N$ we have

$$1 - e^{-c} + \frac{B}{B-c} \left[e^{c\frac{B-c}{B}} - 1 \right] + c + \frac{B}{c} \left[1 - e^{-\frac{c^2}{B}} \right] < 2c;$$

we leave this to the end.

Case 2: $x < -|\hat{\theta}| - c$

$$\begin{aligned}2c\mathbf{E}V(X) &= \int_{x-c}^{x} V(y)\alpha(x,y)dy + \int_x^{x+c} V(y)\alpha(x,y)dy + V(x) \int_{x-c}^{x+c} (1-\alpha(x,y))dy \\ &= \int_x^{x+c} (V(y) - V(x))dy + cV(x) + \int_{x-c}^{x} V(y)\alpha(x,y)dy + V(x) \int_{x-c}^{x} (1-\alpha(x,y))dy \\ &= \int_x^{x+c} V(y)dy + \int_{x-c}^{x} V(y)\alpha(x,y)dy + V(x) \int_{x-c}^{x} (1-\alpha(x,y))dy\end{aligned}$$

so since $x + c < -|\hat{\theta}|$

$$\int_x^{x+c} V(y)dy = \int_x^{x+c} e^{-y} dy = e^{-x} - e^{-(x+c)} = V(x)(1 - e^{-c}).$$

Now since $-x + \hat{\theta} \geq -x - |\hat{\theta}| > c$

$$\begin{aligned}\int_{x-c}^{x} V(y)\alpha(x,y)dy &\leq \int_{x-c}^{x} e^{-y} e^{-\frac{(x-\hat{\theta})(y-x)}{B}} dy \leq \int_{x-c}^{x} e^{-y} e^{\frac{c(y-x)}{B}} dy \\ &= e^{\frac{-cx}{B}} \int_{x-c}^{x} e^{-y\frac{B-c}{B}} dy = \frac{B}{B-c} e^{\frac{-cx}{B}} e^{\frac{cx}{B}} e^{-x} \left(e^{c\frac{B-c}{B}} - 1 \right) \\ &= \frac{B}{B-c} V(x) \left(e^{c\frac{B-c}{B}} - 1 \right),\end{aligned}$$

and finally

$$\begin{aligned}
V(x)\int_{x-c}^{x}(1-\alpha(x,y))dy &\leq V(x)\int_{x-c}^{x}(1-e^{\frac{c}{B}(y-x)})dy \\
&= V(x)\left(c-e^{\frac{-cx}{B}}\int_{x-c}^{x}e^{\frac{cy}{B}}dy\right) \\
&= V(x)\left(c+e^{\frac{-cx}{B}}\frac{B}{c}\left[e^{\frac{cx}{B}}-e^{\frac{c(x-c)}{B}}\right]\right) \\
&= V(x)\left(c+\frac{B}{c}\left[1-e^{-\frac{c^2}{B}}\right]\right).
\end{aligned}$$

Notice this gives exactly the same result as Case 1, so there are no additional conditions to check on c.

Case 3: $-|\hat{\theta}|-c < x < |\hat{\theta}|+c$. In this case we have by Lemma A.2

$$V(x) \leq e^{|\hat{\theta}|+c} \leq e^{|\hat{\theta}|}e^{\log N}$$

□

It follows that if we put $B = \log N$ and $c = 2\log N$ then

$$\mathcal{P}V \leq \gamma V + K$$

and we can take

$$\begin{aligned}
4\gamma\log N &= 1-e^{-2\log N}+\frac{\log N}{-\log N}\left[e^{2\log N\frac{-\log N}{\log N}}-1\right]+2\log N+\frac{\log N}{2\log N}\left[1-e^{-\frac{(2\log N)^2}{\log N}}\right] \\
&= 1-\frac{1}{N^2}+\left[1-\frac{1}{N^2}\right]+2\log N+\frac{1}{2}\left[1-\frac{1}{N^4}\right] \\
&= \frac{5}{2}-\frac{2}{N^2}-\frac{1}{2N^4}+2\log N
\end{aligned}$$

and so

$$\gamma = \frac{1}{2}+\frac{5}{8\log N}-\frac{1}{4\log N}\left(\frac{2}{N^2}+\frac{1}{2N^4}\right),$$

which is < 1 when $N > e^2$. From case 3 we have $K \leq e^{2\log N}e^{\log N} = N^3$. Therefore

$$\frac{2K}{1-\gamma} = \frac{2N^3}{\frac{1}{2}-\frac{5}{8\log N}+\frac{1}{4\log N}\left(\frac{2}{N^2}+\frac{1}{2N^4}\right)}$$

and so $V(x) = \frac{2K}{1-\gamma}$ when

$$\begin{aligned}
|x| &= \log 2+3\log N-\log\left(\frac{1}{2}-\frac{5}{8\log N}+\frac{1}{4\log N}\left(\frac{2}{N^2}+\frac{1}{2N^4}\right)\right) \qquad (35)\\
&< 3\log N+\log 12
\end{aligned}$$

when $N \geq e^2$.

Now we show five step minorization on this set. We begin by lower bounding the measure assigned to an interval around the mode.

Lemma A.5. *For $\delta = 1$ and $c = 2\log N$ if $|\hat{\theta} - x| < c - \delta$, then*

$$\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \mathcal{P}(x, dy) > \frac{1}{2} + \frac{1}{2e^2}.$$

Proof. We again have several cases. Put $\zeta = \hat{\theta} - x$ so that $|\zeta| < c - \delta$, and set $A = [x - c, \hat{\theta} - \delta] \cup [\hat{\theta} + \delta, x + c]$.

Case 1: $x > \hat{\theta} + \delta$ so that $\zeta < -\delta$

$$\begin{aligned}
\int_{x-c}^{\hat{\theta}-\delta} \alpha(x,y)dy + \int_{\hat{\theta}+\delta}^{x+c} \alpha(x,y)dy &\leq \int_{x-c}^{\hat{\theta}-\delta} e^{-\frac{1}{B}(x-\hat{\theta})(y-x)}dy + \int_{\hat{\theta}+\delta}^{x+c} e^{-\frac{1}{B}(x-\hat{\theta})(y-x)}dy \\
&\leq \int_{x-c}^{\hat{\theta}-\delta} e^{\frac{\delta}{B}(y-x)}dy + \int_{\hat{\theta}+\delta}^{x} e^{\frac{\delta}{B}(y-x)}dy + \int_{x}^{x+c} e^{-\frac{\delta}{B}(y-x)}dy \\
&= \frac{B}{\delta}\left[2 - 2e^{-\frac{c\delta}{B}} + e^{\frac{\delta}{B}(\zeta-\delta)} - e^{\frac{\delta}{B}(\delta+\zeta)}\right] \\
&= \frac{B}{\delta}\left[2 - 2e^{-\frac{c\delta}{B}} + e^{\frac{\delta\zeta}{B}}(e^{-\frac{\delta^2}{B}} - e^{\frac{\delta^2}{B}})\right] \\
&\leq \frac{\log N}{\delta}\left[2 - 2e^{-\frac{c\delta}{B}} + e^{-\frac{\delta^2}{B}}(e^{-\frac{\delta^2}{B}} - e^{\frac{\delta^2}{B}})\right] \\
&= \frac{\log N}{\delta}\left[1 - 2e^{-2\delta} + e^{-\frac{2\delta^2}{\log N}}\right]
\end{aligned}$$

It follows that with $\delta = 1, c = 2\log N$

$$\begin{aligned}
\mathcal{P}(x, A) &\leq \frac{1}{4\log N}\frac{\log N}{\delta}\left[1 - 2e^{-2\delta} + e^{-\frac{2\delta^2}{\log N}}\right] \\
&\leq \frac{1}{4}\left[2 - \frac{2}{e^2}\right] = \frac{1}{2} - \frac{1}{2e^2} < 1 \\
\mathcal{P}(x, A^c) &\geq \frac{1}{2} + \frac{1}{2e^2}
\end{aligned}$$

Case 2: $x < \hat{\theta} - \delta$ so that $\zeta > \delta$

$$\begin{aligned}
\int_{x-c}^{\hat{\theta}-\delta} \alpha(x,y)dy + \int_{\hat{\theta}+\delta}^{x+c} \alpha(x,y)dy &\leq \int_{x-c}^{\hat{\theta}-\delta} e^{-(x-\hat{\theta})(y-x)}dy + \int_{\hat{\theta}+\delta}^{x+c} e^{-(x-\hat{\theta})(y-x)}dy \\
&\leq \int_{x-c}^{x} e^{\delta(y-x)}dy + \int_{x}^{\hat{\theta}-\delta} e^{-\delta(y-x)}dy + \int_{\hat{\theta}+\delta}^{x+c} e^{-\delta(y-x)}dy \\
&= \frac{B}{\delta}\left[2 - 2e^{-\frac{c\delta}{B}} + e^{\frac{\delta}{B}(-\delta-\zeta)} - e^{\frac{\delta}{B}(\delta-\zeta)}\right] \\
&= \frac{B}{\delta}\left[2 - 2e^{-\frac{c\delta}{B}} + e^{-\frac{\delta\zeta}{B}}(e^{-\frac{\delta^2}{B}} - e^{\frac{\delta^2}{B}})\right] \\
&\leq \frac{\log N}{\delta}\left[2 - 2e^{-2\delta} + e^{-\frac{\delta^2}{B}}(e^{-\frac{\delta^2}{B}} - e^{\frac{\delta^2}{B}})\right] \\
&= \frac{\log N}{\delta}\left[1 - \frac{2}{e^{2\delta}} + e^{-\frac{2\delta^2}{\log N}}\right],
\end{aligned}$$

the same as in case 1.

Case 3: $\hat{\theta} - \delta < x < \hat{\theta} + \delta$, so $-\delta < \hat{\theta} - x < \delta$.

$$\begin{aligned}
\int_{x-c}^{\hat{\theta}-\delta} \alpha(x,y)dy + \int_{\hat{\theta}+\delta}^{x+c} \alpha(x,y)dy &\leq \int_{x-c}^{\hat{\theta}-\delta} e^{-\frac{1}{B}(x-\hat{\theta})(y-x)}dy + \int_{\hat{\theta}+\delta}^{x+c} e^{-\frac{1}{B}(x-\hat{\theta})(y-x)}dy \\
&\leq \int_{x-c}^{\hat{\theta}-\delta} e^{\frac{\delta}{B}(y-x)}dy + \int_{\hat{\theta}+\delta}^{x+c} e^{-\frac{\delta}{B}(y-x)}dy \\
&= \frac{B}{\delta}\left[e^{\frac{\delta}{B}(-\delta-\zeta)} + e^{\frac{\delta}{B}(-\delta+\zeta)} - 2e^{-\frac{c\delta}{B}}\right] \\
&= \frac{B}{\delta}\left[e^{-\frac{\delta^2}{B}}(e^{\frac{\delta\zeta}{B}} + e^{-\frac{\delta\zeta}{B}}) - 2e^{-\frac{c\delta}{B}}\right] \\
&\leq \frac{B}{\delta}\left[e^{-\frac{\delta^2}{B}}(e^{\frac{\delta^2}{B}} + e^{-\frac{\delta^2}{B}}) - 2e^{-2\delta}\right] \\
&\leq \frac{\log N}{\delta}\left[1 - 2e^{-2\delta} + e^{-\frac{2\delta^2}{\log N}}\right]
\end{aligned}$$

again the same expression as before, giving the result. □

This showed minorization on the set

$$\mathcal{C}_0 = \{x : |\hat{\theta} - x| < c - 1\}, \tag{36}$$

an interval of width $4\log N - 2$ centered at the mode. This is not quite the entire sublevel set $\mathcal{C} = \{x : V(x) \leq \frac{2K}{1-\gamma}\}$, which is given by

$$\mathcal{C} = \left\{x : |x| \leq \log 2 + 3\log N - \log\left(\frac{1}{2} - \frac{5}{8\log N} + \frac{1}{4\log N}\left(\frac{2}{N^2} + \frac{1}{2N^4}\right)\right)\right\} \tag{37}$$

a consequence of (35). When $N \geq e^2$ and $1 < z < N - 1$,

$$\mathcal{C} \subseteq \{x : |x| \leq 3\log N + \log 12\}. \tag{38}$$

By Lemma A.2, we have $-\log N < \hat{\theta} < \log N$. The next Lemma lower bounds the probability of transitioning into $\mathcal{C}_0$ from any point in the set on the right side of (38) in four steps, which, combined with the previous result, shows five step minorization on $\mathcal{C}$.

Lemma A.6. *Let $x \in \mathcal{C} \setminus \mathcal{C}_0$ were C_0 and C are defined in* (36) *and* (37) *respectively. Then*

$$\mathcal{P}^4(x, \mathcal{C}_0) > \frac{1}{64}.$$

Proof. Take any point $x \in \mathcal{C} \setminus C_0$. Without loss of generality, assume that $\hat{\theta} < x$. Since $x \notin \mathcal{C}_0$, $\hat{\theta} + c - 1 < x < \hat{\theta} + 2c + \log 12$. Define $d(x) = x - \hat{\theta}$ with

$$c - 1 < d < 2c + \log 12 < 3c.$$

Suppose $d(x) < \frac{3c}{2}$. Then

$$\left[x - c, x - \frac{c}{2}\right] \subset \mathcal{C}_0.$$

Observe that when $d(x) < \frac{3c}{2}$

$$\mathcal{P}(x, \mathcal{C}_0) > \mathcal{P}\left(x, \left[x - c, x - \frac{c}{2}\right]\right) > \frac{1}{4}.$$

Alternatively, suppose $\frac{3c}{2} < d(x) < 2c$. Then

$$\mathcal{P}(x, \{y : c < d(y) < \frac{3c}{2}\}) > \frac{1}{4}$$

so that

$$\mathcal{P}^2(x, \mathcal{C}_0) > \frac{1}{16}.$$

Using similar arguments for the case where $2c < d(x) < \frac{5c}{2}$ and the case $\frac{5c}{2} < d(x) < 3c$ gives the result. □

References

[1] Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 25(1):1–19, 2014.

[2] 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.

[3] 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.

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

[5] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

[6] DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.16 of 2017-09-18. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[7] Peter W. Glynn and Sean P. Meyn. A Liapounov bound for solutions of the Poisson equation. *The Annals of Probability*, 24(2):916–931, 1996.

[8] Evan Greene, Jon A Wellner, et al. Exponential bounds for the hypergeometric distribution. *Bernoulli*, 23(3):1911–1950, 2017.

[9] Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[10] Martin Hairer and Jonathan C. Mattingly. Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. *Ann. of Math. (2)*, 164(3):993–1032, 2006. ISSN 0003-486X. doi: 10.4007/annals.2006.164.993. URL http://dx.doi.org/10.4007/annals.2006.164.993.

[11] Martin Hairer and Jonathan C Mattingly. Yet another look at harris' ergodic theorem for markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer, 2011.

[12] James E Johndrow and Jonathan C Mattingly. Coupling and decoupling to bound an approximating markov chain. *arXiv preprint arXiv:1706.02040*, 2017.

[13] James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. Inefficiency of data augmentation for large sample imbalanced data. *arXiv preprint arXiv:1605.05798*, 2016.

[14] Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *arXiv preprint arXiv:1304.5299*, 2013.

[15] Jonathan C. Mattingly, Andrew M. Stuart, and M. V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

[16] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[17] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664, 2014.

[18] 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.

[19] Alexander Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, pages 1003–1014, 2005.

[20] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010. Print companion to [6].

[21] D. Revuz. *Markov chains*. North-Holland Publishing Co., Amsterdam-Oxford; American Elsevier Publishing Co., Inc., New York, 1975. North-Holland Mathematical Library, Vol. 11.

[22] Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

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

[24] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

[25] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.