ERGODICITY OF APPROXIMATE MCMC CHAINS WITH APPLICATIONS TO LARGE DATA SETS

NATESH S. PILLAI† AND AARON SMITH‡

Abstract. In many modern applications, difficulty in evaluating the posterior density makes taking performing a single MCMC step slow; this difficulty can be caused by intractable likelihood functions, but also appears for routine problems with large data sets. Many researchers have responded by running approximate versions of MCMC algorithms. In this note, we develop very general quantitative bounds for showing the ergodicity of these approximate samplers. In particular, our bounds can be used to perform a bias-variance trade-off argument to give practical guidance on the quality of approximation that should be used for a given total computational budget. We present a few applications of our results in recently proposed algorithms, including the “austerity” framework, Stochastic Gradient Langevin Dynamics, exponential random graphs and ABC-MCMC algorithms.

1. Introduction

Markov chain Monte Carlo (MCMC) sampling is an indispensable tool for Bayesian computation. Most of the popular Metropolis-Hastings samplers require full evaluation of the posterior at two points at every step, while other MCMC samplers require even more information, such as gradients of the likelihood function. In many modern applications involving intractable likelihoods, the computational cost of this full evaluation of the likelihood function can be prohibitively expensive. For instance, a full likelihood evaluation might involve processing a massive amount data, or computing the solution of a partial differential equation representing an underlying physical phenomenon. The result is that the inferential performance of naive implementations of MCMC algorithms can deteriorate as the amount of data grows, unless available computational resources grow even more quickly (see [Jor13] for a broader discussion of this problem outside of the context of MCMC).

An increasingly common approach to circumvent this difficulty is to run only approximations of the desired MCMC dynamics. These approximations often rely on estimating, rather than evaluating, the posterior distribution of interest - for example, doing so based on a subsample of the available data [Bea03, OBB+00, WT11, KCW13, BDH14]. While many approximate MCMC methods seem to be very successful in practice, they do not have the same convergence guarantees as standard MCMC samplers. In this paper, we present some general convergence results for such ‘approximate’ MCMC algorithms and discuss their applications to some recently proposed algorithms. Our results give quantitative bounds on the convergence in distribution of the Markov chain as well as convergence of finite samples drawn from the Markov chain. These bounds allow us to provide advice on how to choose

†pillai@fas.harvard.edu, Department of Statistics Harvard University, 1 Oxford Street, Cambridge MA 02138, USA.
‡smith.aaron.matthew@gmail.com, Department of Mathematics and Statistics University of Ottawa, 585 King Edward Drive, Ottawa ON K1N 7N5, Canada.
parameters for various approximation schemes, and in particular to give modest conditions under which approximation schemes are more efficient than their underlying MCMC algorithms. While our examples focus on the problems posed by large data sets, our bounds are also relevant to MCMC samplers targeting intractable likelihoods; see [PM+13] for applications of related ideas in that context.

Throughout the paper, we are especially interested in how approximations perform outside of the simplest setting of uniformly good approximations of uniformly ergodic Markov chains (see e.g. [Mit05] for bounds in that setting; heuristic arguments presented in [KCW13] and [WT11] implicitly make such assumptions). It is well known that some approximation schemes that have been proposed in the past can fail badly even in innocuous settings when the approximations are not uniformly good (see Example 17 in [PM+13] for a sampler that fails to converge to a two-valued density on [0,1] and Theorem 1 of [LL12] for one that has difficulty sampling from a Gaussian). Similarly, even uniformly good approximations of chains without uniform ergodicity can fail to inherit good convergence properties. To see this, fix $\epsilon > 0$ and consider simple random walk on $\mathbb{N}$ as a uniformly good approximation of simple random walk on $\mathbb{N}$ with drift $\min\left(0, \epsilon \log(n)\right)$ towards 0 at point $n$. The latter chain has reasonable convergence properties; the former does not even have a stationary measure.

Our paper has three main contributions. The first is providing practical advice on how to choose the best approximate sampler for a given computational budget. The second is providing robust ergodicity results. Finally, we provide a quantitative discussion of convergence for many approximation algorithms that cannot be easily analyzed as globally-small perturbations. This includes non-reversible samplers, algorithms that are geometrically but not uniformly ergodic, and approximations that can be arbitrarily bad outside of “small” sets.

After presenting the main results, we discuss applications. Our first application is a formalization of the austerity framework proposed in [KCW13] for discussing the problem of running approximate MCMC algorithms under computational constraints imposed by data volume. In [KCW13], the authors create a family of approximate samplers $K_\epsilon$ parameterized by an error $\epsilon \geq 0$, with $K_0 = K$ being the original MCMC sampler. Heuristically, a larger value of $\epsilon$ corresponds to an algorithm that can run more steps for any given amount of computer time, but has a larger asymptotic bias. For a fixed amount of computer time, the goal is to choose the value of $\epsilon$ that minimizes the expected $L^2$ error of the resulting sample.

Although this bias-variance tradeoff is proposed in their paper, they do not provide bounds relating the expected finite-time error to $\epsilon$ or the total computing time. We use our general results to relate these quantities, and provide guidance as to the optimal value of $\epsilon$ for a given amount of computing time when the approximation $K_\epsilon$ of $K_0$ is uniformly good on the parameter space. We then extend these results to several large classes of base chains $K$ that are not uniformly ergodic, obtaining similar rates. We also extend our bounds to the exponential random graph model, using this as a test case for non-i.i.d. data and providing analogous bounds. One of the main goals of our study is to justify the fact that, for a small computational budget and large amount of data, MCMC dynamics $K_\epsilon$ with $\epsilon > 0$ can be more efficient than the usual dynamics $K_0$. Our approximation assumptions are stated in terms of families of metrics on kernels that include those used in [KCW13, Afeb14, Mit05, BDH14], and we show that our converge bounds are sharp in this generality. In examples, we also
point out that subsampling algorithms can have additional structure under which much faster convergence is possible.

In the following section, we apply our bounds to the Stochastic Gradient Langevin Dynamics (SGLD) of [WT11]. When these dynamics are uniformly ergodic and the SGLD approximation to the usual Langevin Dynamics are uniformly good, the bounds are similar to those for the austerity framework. Unfortunately, as pointed out in [WT11, AFEB14], these assumptions very rarely hold. We make progress by observing the fact that, for many natural examples, the shape of the tails of the likelihood function depends very little on the details of a data set. When this is the case, a subsample of data may result in very poor kernel approximations of the tails, but even a subsample consisting of a single data point is often enough to obtain global properties such as drift conditions. We find quantitative bounds based on this idea, and apply them to examples for which the heuristic convergence argument in [WT11] does not apply.

Our final section briefly discusses global properties that one might consider in trading bias for variance in MCMC algorithms, with applications to SGLD and to ABC. Briefly, the work in [KCW13] and related papers is built on the idea that one can take more steps of an MCMC algorithm if one can very cheaply approximate a transition kernel; the increased number of samples will decrease the variance of the resulting estimate, at some cost in bias. This is a local improvement, as we only improve the speed of individual steps without thought to the global properties of the algorithm. Our final section discusses a very different tradeoff, where the transition kernel \( K \) is approximated by a transition kernel \( \tilde{K} \) that may be just as expensive to sample from, but which has better convergence properties than \( K \). In one of our examples, \( \tilde{K} \) is geometrically ergodic while \( K \) is not. In other examples, we may increase the spectral gap. In all cases, these improved convergence properties should result in MCMC samplers that have a smaller variance for the same number of steps. While the approach and algorithms discussed here are very different from those in [KCW13], the idea is quite similar: for a quantifiable increase in bias, we can decrease the variance of our MCMC sample.

While we were finishing this paper, closely related work was released in the preprints [BDH14, AFEB14]. We feel that the preprint [BDH14] is complementary to the current paper: their paper focuses on providing guarantees that a given approximate algorithm is close to a base chain under certain conditions, and largely ignores what this implies for the convergence of the approximate algorithm. We focus on the convergence rates of approximate algorithms given guarantees of the sort found in [BDH14]; besides lemmas covering very simple approximation algorithms in Sections 5 and 6, we spend little time on developing such guarantees. The paper [AFEB14] is much closer to our work. Like us, the authors of [AFEB14] were concerned with the convergence of various approximate versions of standard MCMC algorithms, and are particularly interested in the problems associated with running MCMC algorithms when it is expensive to exactly evaluate the target distribution, even up to normalizing constant. Their main theoretical results are applications of the perturbation-theoretic theorems found in [Mit05]. In particular, their Corollary 2.2 gives similar bounds to our Lemma 3.2; like us, they give a few variations on this theme.

Since our main applications are very similar, we point out some important differences between our papers. Their theoretical results focus on the case of uniformly ergodic approximating chains that are uniformly close to their target chains, and these results can
be slightly sharper than ours in that generality. In contrast, we obtain theoretical results that apply to a much broader collection of chains, including approximating chains for which the quality of the approximation is not uniformly bounded over the state space, as well as dropping the requirement for uniform ergodicity. This allows us to provide useful bounds on several algorithms, and allows us to give a partial answer to a question on the convergence of stochastic gradient Langevin dynamics posed in Section 2.3 of [AFEB14]. In another direction, we also examine convergence in stronger metrics, such as Wasserstein distance of the empirical distribution of the chain to its target, and obtain sharper inequalities in these situations. Finally, in addition to proving convergence results related to Corollary 2.2 of [AFEB14], we prove a tradeoff bound that describes the optimal approximation quality one should use with a given computational budget, and describes the asymptotic variance of an MCMC sample in terms of this computational budget. Their paper covers a wider variety of examples than ours, and has several very useful simulated studies of different algorithms. Overall, their paper focuses more on serious empirical studies than ours, while ours focuses on more broadly-applicable theoretical convergence bounds and their consequences for sample properties.

The organization of our paper is as follows. We begin by setting up some notation in Section 2, and prove our main theoretical results in Section 3. In Section 4, we then study the austerity framework recently proposed in [KCW13]. In Section 5, we discuss how the same austerity framework extends to non-i.i.d. data by looking at exponential random graph models. We discuss a related idea, the stochastic gradient Langevin dynamics, in Section 6. These dynamics were proposed in [WT11], and they are generally very far from the parent algorithm outside of a small set; overcoming this difficulty is the critical step in making rigorous the heuristic convergence argument presented in [WT11]. Finally, in Section 7, we introduce and analyze some approximate MCMC samplers that are based on adding bias to improve global mixing properties rather than local step speeds.

2. Preliminaries

For a random variable $X$ and measure $\mu$, $X \sim \mu$ denotes that $X$ is distributed according to $\mu$. $\text{Unif}(A)$ denotes the uniform or Haar distribution on the set $A$ as appropriate. We will write $f = O(g)$ to mean that there exists a constant $C > 0$ so that $f(x) \leq Cg(x)$. We also write $f = o(g)$ if $\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$. For a random variable $X$, we will write $\mathcal{L}(X)$ to denote the law of $X$. Throughout the paper, the letter $C$ will denote a generic positive constant.

By $\| \cdot \|_{\text{TV}}$ we mean the total variation norm. We use the framework of curvature for operators used heavily in [JO10] and introduced in [Oll09]. Throughout, we will consider several kernels $K$ on several Polish spaces $(\Omega, d)$. Associated with each kernel and Polish space is a notion of curvature. Fix two measures $\mu, \nu$ on $\Omega$, and let $\Pi(\mu, \nu)$ be the set of all couplings of $\mu$ and $\nu$. The Wasserstein distance between $\mu$ and $\nu$ is defined as

$$W_d(\mu, \nu) = \inf_{\zeta \in \Pi(\mu, \nu)} \int_{x, y \in \Omega} d(x, y)\zeta(dx, dy).$$

In this paper, we frequently pass between this definition of the Wasserstein distance and the following version provided by the Kantorovitch-Rubinstein duality theorem (see Remark
6.5 of [Vil08]):

\[ W_d(\mu, \nu) = \sup_{\|f\|_{\text{Lip}} = 1} |\mu(f) - \nu(f)|. \]

The **Ricci curvature** of the kernel \( K \) at the pair of points \( x, y \) is defined to be:

\[ \kappa(x, y) = 1 - \frac{W_d(K(x, \cdot), K(y, \cdot))}{d(x, y)} \]

and the curvature of the entire chain is defined to be

\[ \kappa = \inf_{x,y \in \Omega} \kappa(x, y). \]

It is worth noting that in many cases of interest, it is sufficient to calculate \( \kappa(x, y) \) for \( d(x, y) \) ‘small’; see, e.g., Prop 19 of [Oll09].

In addition to the curvature, which describes the tendency of nearby points to coalesce, we also need several measures of variation from [JO10]. The **eccentricity** of a point \( x \in \Omega \) is given by:

\[ E(x) = \int_{\Omega} d(x, y)\pi(dy). \]

The **coarse diffusion** is defined as

\[ \sigma(x)^2 = \frac{1}{2} \int_{y,z \in \Omega} d(y, z)^2 K(x, dy)K(x, dz) \]

and the **local dimension** is given by

\[ m(x) = \inf_{\|f\|_{\text{Lip}} = 1} \frac{\int_{y,z \in \Omega} d(y, z)^2 K(x, dy)K(x, dz)}{\int_{y,z \in \Omega} |f(y) - f(z)|^2 K(x, dy)K(x, dz)}. \]

Finally, define the **granularity** to be

\[ \sigma_\infty = \frac{1}{2} \sup_{x \in \Omega} \text{diam Supp}(K(x, \cdot)). \]

For a running time \( T \geq 1 \), define

\[ \pi_T(f) = \frac{1}{T} \sum_{t=1}^{T} f(X_t) \quad (2.1) \]

for any \( f : \Omega \to \mathbb{R} \). Fix a Markov chain \( X_t \) with associated operator \( K \) and invariant measure \( \pi \) on the Polish space \((\Omega, d)\), and define the quantity

\[ V^2 = \frac{1}{\kappa T} \frac{\sigma^2(x)}{m(x)\kappa}. \]

Theorem 4 of [JO10] gives the following concentration inequality for any Lipschitz function \( f \) with \( \pi(f) = 0 \):

\[ \mathbb{P}\left( \frac{\|\pi_T(f) - \mathbb{E}[\pi(f)]\|}{\|f\|_{\text{Lip}}} > r \right) \leq 2e^{-r^2/(16V^2)} \quad (2.2) \]

\(^1\text{Without loss of generality, we assume the burn-in time of all of the MCMC algorithms studied in this paper to be 0; it is quite straightforward to modify our results including the burn-in time.}\)
3. Technical Results

3.1. Convergence of Approximate Chains under Contraction Assumptions. Next we consider a general non-stationary, time inhomogeneous Markov chain $X_t$ evolving according to some sequence of kernels $K_t$ that approximates a kernel of interest $K$ with stationary distribution $\pi$ on state space $\Omega$. The following calculation gives quantitative bounds on the mixing properties of the approximate chain $X_t$:

**Lemma 3.1 (Coupling Inequality).** Assume that $K$ satisfies

$$\|K^t(x, \cdot) - \pi(\cdot)\|_{TV} \leq C_x(1 - \alpha)^t$$

for some $\alpha > 0$ and all $t > 0$, and also assume that

$$\sup_{1 \leq t \leq T, x \in X} |K_t(x, \cdot) - K(x, \cdot)|_{TV} < \delta$$

for some set $X$. Then, if $X_1 = x \in X$, we have:

$$\|\mathcal{L}(X_T) - \pi\|_{TV} \leq C_x(1 - \alpha)^{T-1} + (T - 1)\delta + \sum_{t=1}^{T} K^t(x, X^c).$$

(3.1)

**Proof of Lemma 3.1.** We will couple the process \( \{X_t\}_{t=1}^T \) evolving according to the sequence of kernels \( \{K_t\}_{t=1}^{T-1} \) with two Markov chains \( \{Y_t\}_{t=1}^T, \{Y'_t\}_{t=1}^T \) both evolving according to the kernel $K$. The first chain, $Y_t$, starts at stationarity: $Y_1 \sim \pi$. The second chain, $Y'_t$, starts with $X_t$: $Y'_1 = x = X_1$.

We couple the Markov chains $\{Y_t\}$ and $\{Y'_t\}$ so that

$$\mathbb{P}[Y_T \neq Y'_T] = \|\mathcal{L}(Y_T) - \mathcal{L}(Y'_T)\|_{TV} \leq C_x(1 - \alpha)^{T-1}.\tag{3.2}$$

Next we construct a coupling between $Y'_t$ from (3.2) and $X_t$ as follows. At every step $t$, if $X_t \neq Y'_t$ or $Y'_t \not\in X$, choose $X_{t+1} \sim K_{t+1}(X_t, \cdot)$ independently of $Y'_{t+1}$. If $X_t = Y'_t$, choose $X_{t+1}$ from a distribution which satisfies

$$\mathbb{P}[X_{t+1} \neq Y'_{t+1}] = \|K_{t+1}(X_t, \cdot) - K(Y'_t, \cdot)\|_{TV}.$$ 

Define $\tau_1 = \inf\{t \geq 1 : X_t \neq Y'_t\}$, $\tau_2 = \inf\{t \geq 1 : Y'_t \not\in X\}$ and $\tau = \min(\tau_1, \tau_2)$. Then:

$$\|\mathcal{L}(X_T) - \pi\|_{TV} = \|\mathcal{L}(X_T) - \mathcal{L}(Y_T)\|_{TV}$$

$$\leq \mathbb{P}[X_T \neq Y_T]$$

$$\leq \mathbb{P}[X_T \neq Y'_T] + \mathbb{P}[Y'_T \neq Y_T]$$

$$\leq \sum_{t=1}^{T} \mathbb{P}[\tau = t] + C_x(1 - \alpha)^{T-1}$$

$$\leq \sum_{t=1}^{T} (\mathbb{P}[\tau_2 = t] + \mathbb{P}[\tau_2 > \tau_1 = t]) + C_x(1 - \alpha)^{T-1}.$$
\[ \leq \sum_{t=1}^{T} K^{t-1}(x, \mathcal{X}^c) + (T-1)\delta + C_x(1-\alpha)^{T-1} \]

and the proof is finished. \qed

Next we give a similar argument in the Wasserstein topology.

**Lemma 3.2.** Fix a kernel of interest \( K \) with stationary distribution \( \pi \), and let \((\Omega, d)\) be a Polish space that has eccentricity \( E(x) < \infty \) with respect to \( \pi \). Assume that \( K \) satisfies

\[ W_d(K(x, \cdot), K(y, \cdot)) \leq (1 - \alpha)d(x, y) \]  

(3.3)

for all \( x, y \in \mathcal{X} \) for some \( \alpha > 0 \) and set \( \mathcal{X} \). Also assume that

\[ \sup_{1 \leq t \leq T} W_d(K_t(x, \cdot), K(x, \cdot)) < \delta \]

for some \( \delta > 0 \) and all \( x \in \mathcal{X} \). Then, with \( X_0 = x \in \mathcal{X} \),

\[ W_d(\mathcal{L}(X_T), \pi) \leq \frac{\delta}{\alpha} + (1 - \alpha)^{-1}E(x) + \sum_{t=1}^{T} \left( K^{t-1}(x, \mathcal{X}^c) + \prod_{s=1}^{t-1} K_s(x, \mathcal{X}^c) \right). \]  

(3.4)

**Proof of Lemma 3.2.** By the triangle inequality, for any \( t \in [1, T] \) and any \( x, y \in \mathcal{X} \),

\[ W_d(K(x, \cdot), K_t(y, \cdot)) \leq W_d(K(x, \cdot), K(y, \cdot)) + W_d(K(y, \cdot), K_t(y, \cdot)) \]

\[ \leq (1 - \alpha)d(x, y) + \delta. \]

We couple our chain \( X_s \) driven by kernel \( K_s \) and started at \( X_1 = x \) with a chain \( Y_s \) driven by kernel \( K \) and started at \( Y_1 \sim \pi \) so that, at every step,

\[ \mathbb{E}[d(X_{s+1}, Y_{s+1})|X_s, Y_s] \leq \gamma + W_d(K(X_s, \cdot), K_s(Y_s, \cdot)). \]

Then we have

\[ W_d(\mathcal{L}(X_T), \pi) \leq \mathbb{E}[d(X_T, Y_T)] \]

\[ \leq \mathbb{E}[\delta + \gamma + (1 - \alpha)\mathbb{E}[d(X_{T-1}, Y_{T-1})]] + (1 - \mathbb{P}[X_{T-1}, Y_{T-1} \in \mathcal{X}]) \]

\[ \leq \ldots \]

\[ \leq \frac{\delta + \gamma}{1 - \alpha} + (1 - \alpha)^{T-1}E(x) + \sum_{t=1}^{T} \left( K^{t-1}(x, \mathcal{X}^c) + \prod_{s=1}^{t-1} K_s(x, \mathcal{X}^c) \right). \]

Since this holds for all \( \gamma > 0 \), the claim follows. \qed

Lemma 3.2 has the immediate corollary:

**Corollary 1.** Assume that \( K \) satisfies

\[ \|K(x, \cdot) - K(y, \cdot)\|_{TV} \leq (1 - \alpha) \]

for some \( \alpha > 0 \). Also assume that

\[ \sup_{1 \leq t \leq T} \|K_t(x, \cdot) - K(x, \cdot)\|_{TV} < \delta \]

for some \( \delta > 0 \). Then:

\[ \|\mathcal{L}(X_T) - \pi\|_{TV} \leq \frac{\delta}{\alpha} + (1 - \alpha)^{T-1}. \]  

(3.5)
Remark 3.3. We mention that this corollary is also immediately implied by Corollary 3.1 of [Mit05]; the constants are essentially the same for $\delta < \alpha$ small. Although our Lemma 3.2 and the results in [Mit05] both imply the same result in this restricted setting, they have different emphases. Their result is based on linear algebra; ours is purely probabilistic. Their result gives superior results for general uniformly ergodic chains (though, by taking powers of the kernel, our result also trivially implies inequalities for general uniformly ergodic chains that give similar bounds for both bias and errors of MCMC estimates); our results apply to chains that are not uniformly ergodic. Finally, their result only applies for convergence in Total Variation, while our results explicitly allow the use of many other Wasserstein metrics; this flexibility can lead to bounds that are effectively much sharper if the metric is chosen carefully.

This last difference is most easily seen in situations, such as Example 8, where the Markov chain satisfies inequality (3.3) for some fixed $\alpha > 0$ throughout a non-compact state space, and for which the eccentricity satisfies $E(x) < \infty$ for each $x \in \Omega$ but $\sup_{x \in \Omega} E(x) = \infty$. These chains are generally geometrically ergodic but not uniformly ergodic; Lemma 3.2 provides direct bounds on their finite-time bias while Corollary 3.1 of [Mit05] does not apply.

The following example shows that the bound given by Corollary 1 is sharp up to constants, as a function of $\delta$ and $\alpha$.

Example 2. Consider the family of birth and death chains on $[n] = \{0, 1, 2, \ldots, n\}$ with transition kernels given by

\[
K_{\alpha,\delta}(x, x + 1) = \delta \\
K_{\alpha,\delta}(x, x) = 1 - \alpha - \delta \\
K_{\alpha,\delta}(x, x - 1) = \delta
\]

when $x - 1, x + 1 \in [n]$, and 0 otherwise. Let $\pi_{\alpha,\delta}$ be the associated stationary distribution. For $\delta < \alpha$ and $n > 2$, it can be shown that

\[
\sup_{x \in [n]} \|K_{\alpha,\delta}(x, \cdot) - K_{\alpha,0}(x, \cdot)\|_{TV} = \alpha
\]

\[
\|K_{\alpha,\delta}^T(x, \cdot) - \pi_{\alpha,0}\|_{TV} = O(n(1 - \alpha)^T)
\]

\[
\|\pi_{\alpha,\delta} - \pi_{\alpha,0}\|_{TV} \geq \frac{\delta}{2\alpha} + O\left(\frac{\delta^2}{\alpha}\right).
\]

3.2. Convergence of Approximate Chains Under Drift and Minorization Assumptions. In this subsection, we consider a Metropolis-Hastings chain $K$ on state space $\Omega$ with proposal kernel $L$ and stationary distribution $\pi$, as well as an approximating chain $K_\delta$ on state space $\Omega$ with proposal kernel $L$ and stationary distribution $\pi_\delta$. We make the following assumptions throughout this subsection:

1. There exists a function $V : \Omega \rightarrow \mathbb{R}^+$ and constants $0 < a < 1, b < \infty$ so that a chain $Y_t$ evolving according to $K$ satisfies

\[
\mathbb{E}[V(Y_{t+1})|Y_t = y] \leq (1 - a)V(y) + b.
\]
(2) Define \( V_\varepsilon(x) = V(x)^{\frac{1}{1+\varepsilon}} \). Assume that there exists some \( \varepsilon_0 \geq 0 \) so that, for all \( \varepsilon > \varepsilon_0 \), we have:

\[
\frac{1}{V_\varepsilon(x)} \int_z (V_\varepsilon(x) - V_\varepsilon(z)) L(x, dz) < C_\varepsilon < \infty
\]  

(3.7)

For all compact sets \( \mathcal{X} \subset \Omega \), there exists some \( \varepsilon(\mathcal{X}) > 0 \) and measure \( \mu_\mathcal{X} \) with support equal to \( \mathcal{X} \) so that, for all \( x \in \mathcal{X} \) and some measure \( r_x \) associated with each point \( x \), we have:

\[
K(x, \cdot) = \varepsilon(\mathcal{X}) \mu_\mathcal{X}(\cdot) + (1 - \varepsilon(\mathcal{X})) r_x(\cdot).
\]  

(3.8)

**Remark 3.4.** We briefly discuss the strength of these assumptions.

(1) This assumption is fairly standard; see e.g. [RT96] for useful sufficient conditions.

(2) This holds for most reasonable chains. In particular, this inequality holds if our target and proposal distributions are both Gaussian; it holds with \( \varepsilon_0 = 0 \) if the proposal distribution has smaller variance.

(3) This holds if \( L(x, \cdot) \) has density bounded away from 0 on all compact sets, and in particular it is easy to force any chain to satisfy this for small values of \( \varepsilon \).

Under these assumptions, the drift condition is transferred to the approximate chain \( K_\delta \):

**Lemma 3.5 (Drift and Minorization of Approximate Chains).** Let \( K \) be a kernel satisfying assumptions 3.2, and let \( K_\delta \) be a kernel satisfying

\[
||K(x, \cdot) - K_\delta(x, \cdot)||_{TV} < \delta.
\]

Then a chain \( X_t \) evolving according to \( K_\delta \) satisfies a drift condition of the form

\[
E[V_\varepsilon(X_{t+1})|X_t = x] \leq (1 - a_\delta,\varepsilon) V_\varepsilon(x) + b_\varepsilon,
\]

for any \( \varepsilon > \varepsilon_0 \), where

\[
a_{\delta, \varepsilon} = 1 - \left( \left( 1 - \frac{a}{2} \right)^{\frac{1}{1+\varepsilon}} + C_\varepsilon \delta \right)
\]  

(3.9)

\[
b_\varepsilon = b^{\frac{1}{1+\varepsilon}} \left( 1 + \frac{2(1 - a)}{a} \right)^{\frac{1}{1+\varepsilon}}
\]

and we note that for any \( \varepsilon > 0 \), \( a_{\delta, \varepsilon} \) is strictly greater than 0 for \( \delta \) sufficiently small.

**Proof.** By Jensen’s inequality and inequality (3.6),

\[
E[V_\varepsilon(Y_{t+1})|Y_t = y] \leq E[V(Y_{t+1})|Y_t = y]^{\frac{1}{1+\varepsilon}}
\]

\[
\leq ((1 - a) V(y) + b)^{\frac{1}{1+\varepsilon}}
\]

If \( \frac{a}{2} V(y) > b \), we have \( (1 - a) V(y) + b \leq (1 - \frac{a}{2}) V(y) \). If \( \frac{a}{2} V(y) \leq b \), we have \( (1 - a) V(y) + b \leq b \left( 1 + \frac{2(1 - a)}{a} \right) \). Thus, continuing the above calculation,

\[
E[V_\varepsilon(Y_{t+1})|Y_t = y] \leq \left( 1 - \frac{a}{2} \right)^{\frac{1}{1+\varepsilon}} V_\varepsilon(y) + b^{\frac{1}{1+\varepsilon}} \left( 1 + \frac{2(1 - a)}{a} \right)^{\frac{1}{1+\varepsilon}}.
\]  

(3.10)
Define $a_\epsilon = 1 - \left(1 - \frac{a}{2}\right)^{\frac{1}{1+\epsilon}}$ and $b_\epsilon = b^{\frac{1}{1+\epsilon}} \left(1 + \frac{2(1-a)}{a}\right)^{\frac{1}{1+\epsilon}}$. We then calculate

$$\mathbb{E}[V_\epsilon(X_{t+1})|X_t = x] = \mathbb{E}[V_\epsilon(Y_{t+1})|Y_t = x] + \mathbb{E}[V_\epsilon(X_{t+1}) - V_\epsilon(Y_{t+1})|Y_t = X_t = x]$$

$$\leq (1 - a_\epsilon)V_\epsilon(x) + b_\epsilon + \mathbb{E}[V_\epsilon(X_{t+1}) - V_\epsilon(Y_{t+1})|Y_t = X_t = x].$$

Denote by $\alpha(x,y)$ the acceptance probability associated with $K$ and $\alpha_\delta(x,y)$ the acceptance probability associated with $K_\delta$. By Assumption 3.7, we have for $\epsilon > \epsilon_0$:

$$|\mathbb{E}[V_\epsilon(X_{t+1}) - V_\epsilon(Y_{t+1})]| = \left| \int_{z:\alpha(x,z) > \alpha_\delta(x,z)} (\alpha(x,z) - \alpha_\delta(x,z)) (V(x) - V(z))\ell(x,dz) \right|$$

$$+ \left| \int_{z:\alpha(x,z) < \alpha_\delta(x,z)} (-\alpha(x,z) + \alpha_\delta(x,z)) (V(z) - V(x))\ell(x,dz) \right|$$

$$\leq ||K(x, \cdot) - K_\delta(x, \cdot)||_{TV} \left| \int_{z} (V(x) - V(z))\ell(x,dz) \right|$$

$$\leq C_\epsilon \delta V_\epsilon(x). \quad (3.11)$$

Putting together inequalities (3.10) and (3.11), we have:

$$\mathbb{E}[V_\epsilon(X_{t+1})|X_t = x] \leq (1 - a_\epsilon)V_\epsilon(x) + b_\epsilon + C_\epsilon \delta V_\epsilon(x).$$

$$= \left(1 - \frac{a}{2}\right)^{\frac{1}{1+\epsilon}} + C_\epsilon \delta \right) V_\epsilon(x) + b_\epsilon$$

and the proof is finished. \qed

For a Markov chain $X_t$ evolving according to an approximate chain $K_\delta$, define

$$\pi_{T,\delta}(f) = \frac{1}{T} \sum_{t=1}^{T} f(X_t).$$

Then this Lemma allows us to prove the following concentration result when $K$ is uniformly ergodic and $\Omega$ is countable:

**Lemma 3.6** (Error Bounds for Approximate Chains). Let $K$ satisfy Assumptions 3.2, and also assume that $\sup_{x} V(x) \equiv D < \infty$ and $\Omega$ is countable. Let $K_\delta$ be a kernel satisfying

$$||K(x, \cdot) - K_\delta(x, \cdot)||_{TV} < \delta.$$ 

Then for any function $f$ with $||f||_\infty \leq 1$ and $q, r > 0$,

$$\mathbb{P}[|\pi_{T,\delta}(f) - \pi(f)| > \frac{r}{\sqrt{T}} + \frac{\delta q}{G(1-\theta)^q} + (1 - G(1-\theta)^q)\frac{T}{r}] \leq e^{-\frac{(1-\theta)^2r^2}{2G^2}},$$

where $G, \theta$ are as defined in Equation (3.12) below.

**Proof.** Applying Theorem 5 of [Ros95], with bounds given by Lemma 3.5 and assumptions 3.2, for any $\epsilon > \epsilon_0$ the kernel $K_\delta$ is $(G, \theta)$-uniformly ergodic with

$$G = 2 \left(1 + \frac{b_\epsilon}{a_\delta} + D\right)$$

$$\frac{(1 - \theta)^2}{\epsilon} = \max \left(\epsilon \left(\sup\{x : V_\epsilon(x) \leq \frac{2b_\epsilon}{a_\delta}\} \right), \frac{(a_\delta + 4b_\epsilon - 3a_\delta b_\epsilon)(a_\delta + 6b_\epsilon - 2a_\delta b_\epsilon)}{a_\delta (a_\delta + 2b_\epsilon)} \right)$$

(3.12)
where \(a_{\delta, \epsilon}, b_{\epsilon}\) are given by Equation (3.9). Combining this bound with Theorem 1 of [KW13], we have for these values of \((G, \theta)\) that:

\[
P[|\pi_{T, \delta}(f) - \pi_\delta(f)| > \frac{r}{\sqrt{T}}] \leq 2e^{-\frac{(1-\theta)^2r^2}{2G^2}}.
\] (3.13)

Finally, combining Corollary 1 with Lemma 3.5, we have for all \(1 \leq q \leq T\):

\[
|\pi(f) - \pi_\delta(f)| \leq \frac{\delta q}{G(1 - \theta)^q} + (1 - G(1 - \theta)^q)^\frac{T}{q}.
\] (3.14)

Combining inequalities (3.12), (3.13) and (3.14) gives the result.

We then prove similar, weaker, bounds for geometrically ergodic but not uniformly ergodic chains. We begin by noting that the following follows immediately from Theorem 1 of [KCW13]:

**Theorem 3 (Concentration for Geometrically Ergodic Chains).** Fix a subset \(\mathcal{X}\) of a countable space \(\Omega\) and assume that the kernel \(K\) restricted to \(\mathcal{X}\) satisfies

\[
\|K^t(x, \cdot) - \pi\|_{TV} \leq G_{\mathcal{X}}\theta^{t-1}
\] (3.15)

for some \(G_{\mathcal{X}} < \infty\). Then for all \(r > 0\) and \(x_0 \in \mathcal{X}\),

\[
P[|\pi_{T, \delta}(f) - \pi(f)| > \frac{r}{\sqrt{T}}] \leq e^{-\frac{(1-\theta)^2r^2}{2G_{\mathcal{X}}^2}} + (1 - \mathbb{P}[\{X_t\}_{t=1}^T \subset \mathcal{X}]) .
\]

We then have the following analogue to Lemma 3.6:

**Lemma 3.7 (Error Bounds for Approximate Chains 2).** Fix a countable state space \(\Omega\) and kernel \(K\) that satisfies Assumptions 3.2 and inequality (3.15). Also assume that for all finite sets \(\mathcal{X} \subset \Omega\), we have \(\sup_{x \in \mathcal{X}} V(x) \equiv D_{\mathcal{X}} < \infty\). Let \(K_\delta\) be a kernel satisfying

\[
\|K(x, \cdot) - K_\delta(x, \cdot)\|_{TV} < \delta.
\]

Then for any function \(f\) with \(\|f\|_{\infty} \leq 1\) and \(q, r > 0\),

\[
P\left[|\pi_{T, \delta}(f) - \pi(f)| > \frac{r}{\sqrt{T}} + \frac{\delta q}{G(1 - \theta)^q} + (1 - G(1 - \theta)^q)^\frac{T}{q}\right]
\]

\[
\leq e^{-\frac{(1-\theta)^2r^2}{2G^2}} + (1 - \mathbb{P}[\{X_t\}_{t=1}^T \subset \mathcal{X}]) .
\]

where for all \(\epsilon > \epsilon_0\) fixed, \(G\) and \(\theta\) are given by:

\[
G = 2 \left(1 + \frac{b_{\epsilon}}{a_{\delta, \epsilon}} + D_{\mathcal{X}}\right)
\]

\[
(1 - \theta)^2 = \max \left(\epsilon \left\{ x : V_\epsilon(x) \leq \frac{2b_{\epsilon}}{a_{\delta, \epsilon}}\right\}, \frac{(a_{\delta, \epsilon} + 4b_{\epsilon} - 3a_{\delta, \epsilon}b_{\epsilon})(a_{\delta, \epsilon} + 6b_{\epsilon} - 2a_{\delta, \epsilon}b_{\epsilon})}{a_{\delta, \epsilon}(a_{\delta, \epsilon} + 2b_{\epsilon})}\right).
\]

**Proof.** The proof of this Lemma is identical to that of Lemma 3.6 after changing equality (3.12) to equality (3.17) and propagating this change through the other computations. \(\square\)
3.3. Moderate-Time Concentration Bound for Walks That Almost Have Positive Curvature. In this section, we provide bounds on the convergence of both estimates
\[ \pi_T(f) = \frac{1}{T} \sum_{t=1}^{T} f(X_t) \]
and also empirical measures \( F_t = \mathbb{U}(\{X_t\}_{i=1}^{T}) \) based on samples \( \{X_t\}_{i=1}^{T} \) drawn from a sequence of kernels \( \{K_t\}_{i=1}^{T} \) approximating a desired kernel \( K \). These results generalize convergence results found in [AFEB14], in the sense that those results deal with convergence of the distribution of a single point \( L(X_t) \), while the results in this section deal with convergence of the entire sample \( F_t \). We believe that these bounds are useful in and of themselves, and they are also used to prove tradeoffs in the setting of [KCW13].

**Theorem 4.** Consider a kernel \( K \) with curvature \( \kappa > 0 \) defined on a (possibly non-compact) subset of \( \mathbb{R}^d \) and a collection of approximating kernels \( \{K_t\}_{i=0}^{\infty} \) that satisfy
\[ \sup_{x \in \Omega} W_d(K_t(x, \cdot), K(x, \cdot)) < \delta. \]

Denote by \( F_t \) the empirical measure of the sequence \( \{X_s\}_{s=1}^{t} \) drawn from \( \{K_s\}_{s=1}^{t} \). Assume that there exists a Lipschitz function \( S(x) \geq \frac{\sigma(x)^2}{\eta_x \kappa} \). Then
\[ \mathbb{P}\left[ |F_t(f) - \pi(f)| \geq r + \frac{\delta}{\kappa} + (1 - \kappa)^{t-1} E(X_0) \right] \leq 2e^{-\kappa tr} 4 \max(2\|S\|_{\text{Lip}}, 3\sigma_\infty). \]

For \( r < \frac{4}{3\sigma_\infty} \sup_x \frac{\sigma(x)^2}{\eta_x \kappa} \), we have instead:
\[ \mathbb{P}\left[ |F_t(f) - \pi(f)| \geq r + \frac{\delta}{\kappa} + (1 - \kappa)^{t-1} E(X_0) \right] \leq 2e^{-\frac{\sigma_x^2}{16} \inf_x \frac{\eta_x \kappa}{\sigma(x)^2}}. \]

**Proof.** Fix \( \alpha > 0 \) and let \( Y_t \) be a sequence drawn from \( K \) that satisfies
\[ \mathbb{E}[d(X_t, Y_t)] \leq \alpha + \frac{\delta}{\kappa} + (1 - \kappa)^{t-1} E(X_0). \]

Such a coupling exists by inequality (3.4). Denote by \( F_t \) and \( G_t \) the empirical measures associated with the points \( \{X_s\}_{s=1}^{t} \) and \( \{Y_s\}_{s=1}^{t} \) respectively. By the triangle inequality, we note that
\[ \mathbb{P}[|F_t(f) - \pi(f)| \geq r + W_d(G_t, F_t)] \leq \mathbb{P}[|G_t(f) - \pi(f)| \geq r]. \quad (3.18) \]

Combining Inequality (3.18) with Theorem 5 of [JO10] and (3.4), we have
\[ \mathbb{P}[|F_t(f) - \pi(f)| \geq r + \alpha + \frac{\delta}{\kappa} + (1 - \kappa)^{t-1} E(X_0)] \leq 2e^{-\kappa tr} 4 \max(2\|S\|_{\text{Lip}}, 3\sigma_\infty). \]

Letting \( \alpha \) go to 0 completes the proof of the large-\( r \) bound. The proof for \( r \) small is essentially the same, citing the small-\( r \) bound from Theorem 5 of [JO10] instead of the large-\( r \) version cited above. \( \square \)

3.4. Tradeoff Lemma. In this section, we prove a simple book-keeping lemma that will be used in Section 4. We consider simulating from a kernel \( K_\epsilon \) with stationary distribution \( \pi_\epsilon \), where simulating a step of \( K_\epsilon \) requires \( c(\epsilon) \) units of computational time. We assume that \( c(\epsilon) \) is monotone decreasing in \( \epsilon \). For fixed total computational resources \( M \), we then look at the error of the estimate \( \pi_{T_\epsilon}(f) = \frac{1}{T_\epsilon} \sum_{t=1}^{T_\epsilon} f(X_t) \), where \( X_t \) is a Markov chain driven by \( K_\epsilon \) and \( T_\epsilon = \lceil \frac{M}{c(\epsilon)} \rceil \).

**Lemma 3.8 (General Tradeoff).** Let \( \mathcal{F} \) be a class of functions. Suppose that for some collection of constants \( A_i, c_i > 0 \), for all \( f \in \mathcal{F} \) the following hold:
(1) The bias $|\pi_\epsilon(f) - \pi(f)| \leq A_1 \sup_{x \in \Omega} W_d(K(x, \cdot), K_\epsilon(x, \cdot))^{c_1}$.
(2) The kernel approximation error is bounded by $\sup_{x \in \Omega} W_d(K(x, \cdot), K_\epsilon(x, \cdot)) \leq A_2 c(\epsilon)^{-c_2}$.
(3) For some $c_3 > 0$ and some function $S(\cdot)$ with $\lim_{r \to \infty} S(r) = 0$,
$$P\left[|\pi_{T, \epsilon}(f) - \pi_\epsilon(f)| > T_\epsilon^{-c_3} r\right] \leq S(r).$$
Then, choosing the smallest $\epsilon$ that satisfies $c(\epsilon) \leq M^{c_3 + c_1 + c_2}$, we have for all $f \in F$:
$$P\left[|\pi_{T, \epsilon}(f) - \pi(f)| > M^{-c_1}(A_1 A_2^2 + r)\right] \leq S(r),$$
where $c_4 = \frac{c_1 c_2 c_3}{c_1 + c_2 + c_3}$. If instead we have
(1) For some function $S(\cdot)$ with $\lim_{r \to \infty} S(r) = 0$,
$$P\left[|\pi_{T, \epsilon}(f) - \pi(f)| > T_\epsilon^{-c_1} r\right] \leq S(r) + A_1 T_\epsilon^{-c_1} W_d(K(x, \cdot), K_\epsilon(x, \cdot))^{c_2}.$$
(2) The kernel approximation error is bounded by $\sup_{x \in \Omega} W_d(K(x, \cdot), K_\epsilon(x, \cdot)) \leq A_2 c(\epsilon)^{-c_2}$.
Then, choosing $\epsilon$ so that $c(\epsilon) = M^{-c_1 + c_2 + c_3}$, we have:
$$P\left[|\pi_{T, \epsilon}(f) - \pi(f)| > M^{-c_1}(r + A_1^2 + A_2)\right] \leq S(r),$$
where $c_4 = \frac{c_1 c_2 c_3}{c_1 + c_2 + c_3}$.

Proof of Lemma 3.8. Applying the three initial assumptions above in each step, we compute:
$$P[|\pi_{T, \epsilon}(f) - \pi_\epsilon(f)| > M^{-c_1}(A_1 A_2^2 + r)]$$
$$\leq P[|\pi_{T, \epsilon}(f) - \pi_\epsilon(f)| > r M^{-c_1}] + 1_{|\pi_\epsilon(f)| > A_1 A_2^2 M^{-c_4}}$$
$$\leq S(r) + 1_{M^{-c_4} < T_\epsilon^{-c_3}} + 1_{W_d(K(x, \cdot), K_\epsilon(x, \cdot))^{c_2} > A_1^2 M^{-c_4}}$$
$$\leq S(r) + 1_{M^{-c_4} < T_\epsilon^{-c_3}} + 1_{c(\epsilon)^{-c_1 c_2} > M^{-c_4}}$$
$$= S(r),$$
where the last line is simple algebra. This completes the proof of the first version of the lemma; the second is essentially the same. \qed

4. Application 1: Austerity Framework

In [KCW13], the authors consider the problem that, as one accumulates more and more data, a standard Metropolis-Hastings sampler targeting the associated posterior distribution will become more and more expensive to sample from. This stems from the fact that such a sampler evaluates the full posterior twice in every step, and in particular this normally requires every data point to be used at every step. The authors propose several algorithms that might avoid this problem. In this section, we provide one formalization of the austerity framework and prove various tradeoff results with the application of Corollary 1 and Lemma 3.8.

To set notation, we consider a Metropolis-Hastings kernel $K$ with proposal kernel $L$, stationary distribution
$$\pi(\theta) \equiv \pi(\theta|\{x_i\}_{i=1}^N) = p(\theta) \prod_{i=1}^N \pi(\theta|x_i).$$
and acceptance ratio $Q$ as well as an approximating kernel $K_\epsilon$ with the same proposal kernel $L$ but different stationary distribution $\pi_\epsilon$ and acceptance ratio $Q_\epsilon$. While $K$ is a Metropolis-Hastings chain and so $Q(x, y) = \min \left(1, \frac{\pi(y)L(y, x)}{\pi(x)L(x, y)} \right)$ is a deterministic function, $K_\epsilon$ is not a Metropolis-Hastings chain. Instead we view $Q_\epsilon$ as a random variable generated by some algorithm to be specified later; we denote by its expected value $E[Q_\epsilon]$ the actual probability of a proposal being accepted. For the approximating chain $K_\epsilon$, denote by $\ell_{t+1}$ the proposed point at time $t$ and by $u_t$ the $U[0, 1]$-distributed random variable used to determine if the proposal is accepted. Assume that the full data set has $N$ points, and that at every step the acceptance probability $Q_\epsilon$ is based on a random sample of $n = n(t, X_t, \ell_{t+1}, \epsilon, u_t)$ data points, a (possibly random) function of the chain’s current position and time. To formalize the austerity framework, assume that we have a total computational budget of $M$ data points to view in total, over any number of steps of the Markov chain. We would like to choose a function $n = n(t, X_t, \ell_{t+1}, \epsilon, u_t)$ that (approximately) minimizes the deviation of our estimator

$$\pi_T(f) = \sum_{t=1}^T f(X_t),$$

where $T = \sup\{t : \sum_{s=1}^t E[n(s, X_s, \ell_{s+1}, \epsilon, u_t)] \leq M\}$. Without specifying the construction of the approximate acceptance ratio $Q_\epsilon$, our generic construction of $K_\epsilon$ is given in Algorithm 1.

**Algorithm 1 Austerity Framework**

Initialize $X_1 = x$.

For $t = 1$ to $N$ do

Generate $\ell_{t+1}$ from the distribution $L(X_t, \cdot)$.

Generate $u_t$ from $U[0, 1]$.

Generate the (random) acceptance probability $Q_\epsilon$ using $n = n(t, X_t, \ell_{t+1}, \epsilon, u_t)$ independent draws from the data.

if $u_t \leq Q_\epsilon(X_t, \ell_{t+1})$ then

set $X_{t+1} = \ell_{t+1}$.

else

set $X_{t+1} = X_t$.

end if

In [KCW13], they use the following terms:

$$s = \sqrt{\frac{\bar{t}^2 - (\bar{t})^2}{n-1}} \sqrt{1 - \frac{n-1}{N-1}}, \quad (4.1)$$

$$\mu_0 = \frac{1}{N} \log \left( u_t \frac{\pi(X_t)L(X_t, \ell_{t+1})}{\pi(\ell_{t+1})L(\ell_{t+1}, X_t)} \right).$$

**Remark 4.1.** All of our results depend on the expected number of points required to simulate an approximate acceptance ratio $Q_\epsilon$ that satisfies $\epsilon \geq \sup_{x,y} |E[Q_\epsilon(x, y)] - Q(x, y)|$. We don’t
Algorithm 2 Austerity Framework II

Fix a constant $m$ and initialize $n = 0$, $S = \{x_1, \ldots, x_N\}$, $\mathcal{X} = \emptyset$.
While $\text{done} \neq \text{TRUE}$, do:

- Draw a mini-batch $\mathcal{X}'$ of size $\min(m, |S|)$ without replacement from $S$ and then set $S = S \setminus \mathcal{X}'$, $\mathcal{X} = \mathcal{X} \cup \mathcal{X}'$ and $n = |\mathcal{X}|$.
- Compute the sample mean $\bar{\ell}$ and sample variance $\bar{\ell}^2$ for sample $\mathcal{X}$.
- Estimate $s$ and $\mu_0$ using Equation (4.1).
- Compute $\delta = 1 - \phi_{m-1}\left(\frac{\bar{\ell} - \mu_0}{s}\right)$; $\phi_k$ is the CDF of Student's distribution with $k$ degrees of freedom.
- if $\delta < \epsilon$ then
  - Set $\text{done} = \text{TRUE}$.
- end if
- if $\bar{\ell} > \mu_0$ then
  - Set $Q_\epsilon = 1$.
- else
  - Set $Q_\epsilon = 0$.
- end if

give new algorithms for this approximation, but summarize here the growing literature on the topic.

- For any subset $S \subset \{x_i\}_{i=1}^N$, define
  $$\hat{\rho}_S(\theta) = p(\theta) \prod_{x \in S} \pi(\theta|x)^{N_S}.$$ 

  Let $S$ be a sample of size $n$; if we calculate $Q_\epsilon$ using
  $$Q_\epsilon = \min\left(1, \frac{\hat{\rho}_S(X_t)L(\ell_{t+1}, X_t)}{\hat{\rho}_S(\ell_{t+1})L(X_t, \ell_{t+1})}\right) \quad (4.2)$$

  instead of Algorithm 4, we note that under sufficiently nice conditions (the target distribution being uniformly bounded away from 0 on its support suffices) Hoeffding's inequality guarantees an error of at most $O(\epsilon)$ using a computational budget of $n = O\left(-\epsilon^{-2} \log(\epsilon)\right)$ samples per step.

- As discussed in [BDH14], the specialization in [KCW13] can fail to give good approximations under certain circumstances (obviously using equation (4.2) can also be a poor choice).

- In [BDH14], the authors propose another choice of $Q_\epsilon$ and give bounds both on the error of the approximation and the expected computation time. These bounds are quite general, and come with a minimal loss in efficiency: the average computational budget per step is shown to be $O\left(\epsilon^{-2} \log(\epsilon^{-1})\right)$. See that paper for a more thorough discussion of the choice of $Q_\epsilon$.

All of the results below are insensitive to the details of the choice of approximation, and in particular apply to $Q_\epsilon$ chosen from either algorithm 4 or equation (4.2).

The main theoretical result of [KCW13] is their Theorem 1 on the bias of an approximate Markov chain, as follows:
Theorem. For a Metropolis-Hastings algorithm with transition kernel $K$ satisfying the contraction condition

$$\|\mu K - \pi\|_{TV} \leq (1 - \alpha)\|\mu - \pi\|_{TV},$$

we have

$$\|\pi - \pi_\epsilon\|_{TV} \leq \frac{\Delta}{\alpha},$$

where $\Delta = \sup_{x, y \in \Omega} |Q(x, y) - E[Q_\epsilon(x, y)]|$. This result is strictly weaker than our Lemma 3.2 and is extended by Theorem 4. While the bound in Theorem 4 is qualitatively useful, it doesn’t provide any finite-time bounds for estimates based on the chain. In particular, it doesn’t show that there exist computational budgets $M$ for which there exists some $\epsilon > 0$ for which the kernel $K_\epsilon$ gives more accurate estimates than the kernel $K_0 = K$. In this section, we give a number of tradeoff results showing that $K_\epsilon$ can indeed give better results, and giving asymptotic results describing the optimal choice of $\epsilon$ as $M$ gets large. Our simplest result in this direction is:

Theorem 5 (Austerity Tradeoff). Fix a 1-Lipschitz function $f$, a starting point $X_0$ with finite eccentricity, and a computational budget $M$. Assume that the base Markov chain $K$ has curvature $\kappa > 0$. Assume that there exists a Lipschitz function $S(x) \geq \frac{\sigma(x)^2}{\eta_0 \sigma}$. Assume that the random estimates $Q_\epsilon$ based on an algorithm with an expected cost of $n$ computational units satisfy

$$W_d(K(x, \cdot), K_\epsilon(x, \cdot)) \leq C n^{-\frac{1}{2}}. \quad (4.3)$$

Then for $\frac{C}{M^{\frac{1}{4}}} < \kappa$, an estimate of $Q_\epsilon$ with a mean cost of $\frac{1}{\sqrt{M}}$ satisfies:

$$\mathbb{P}[|\pi_\sqrt{M, \epsilon}(f) - \pi(f)| > r M^{-\frac{1}{2}}] \leq A(r),$$

for some function satisfying $\lim_{r \to \infty} A(r) = 0$ that does not depend on $m, M$ or $N$.

Remark 4.2. If we replace inequality (4.3) with the bound $C n^{-\frac{1}{2}} \log(n)$, as per bounds discussed in Remark 4.1, the same conclusion applies with only a loss of a logarithmic factor in the final bound.

Remark 4.3. The main point is that, as one’s computational budget $M$ goes to infinity, approximate chains are more efficient than the precise chain under some conditions. It is possible to obtain an error rate of $O(M^{-\frac{1}{4}})$ using approximate kernels independently of the amount of data $N$ available; running the precise kernel gives an error rate of $O(\sqrt{N}/M)$. If the computational resources grow sub-quadratically in the amount of data, this is a substantial improvement.

We give an example showing that, under assumptions of the form given by inequality (4.3), the $O(M^{-\frac{1}{4}})$ rate we obtain is sharp; this remains true even when $K_\epsilon$ gives i.i.d. samples from $\pi_\epsilon$. In particular, the mixing rate of the underlying Markov chain is essentially irrelevant to our conclusions. We also give a simple resampling algorithm for which a much stronger asymptotic rate holds. These examples serve to illustrate the sharpness of our theorem and also the fact that point-wise bounds on the difference between kernels, such as those given
by inequality (4.3), are extremely unstructured assumption and give correspondingly weak conclusions.

**Example 6.** We begin by looking at situations that give rise to other rates. Consider the posterior $\pi(\theta|\{x_i\}_{i=1}^N) = N\left(\frac{\sum_{i=1}^N x_i}{N}, \frac{1}{N}\right)$ with approximation $\pi_S(\theta) = N\left(\frac{\sum_{x \in S} x}{|S|}, \frac{1}{N}\right)$ for any subset (possibly with repetition) $S$ of $\{x_i\}_{i=1}^N$. For any given computational budget $M$ and approximation level $n \leq M, N$, we approximate $\theta$ as follows. Choose $n$ points uniformly at random with replacement from $\{x_i\}_{i=1}^N$. We then draw $\theta_t$ from $\pi_S(\theta)$. By the usual decomposition of variance formula,

\[
\text{Var}\left(\frac{n}{M} \sum_{t=0}^M \theta_t\right) = \mathbb{E}\left(\text{Var}\left(\frac{n}{M} \sum_{t=0}^M \theta_t \big| S\right)\right) + \text{Var}\left(\left(\frac{n}{M} \sum_{t=0}^M \theta_t \big| S_t\right)\right)
\]

\[= \frac{n}{MN} + \frac{1}{n}.
\]

In this setting, choosing $n = \min(M, \sqrt{MN})$ is optimal, giving the usual $O\left(\sqrt{M}\right)$ convergence rate as $M$ becomes large much more slowly than $N$. Although the details change, similar conclusions hold if the set $S$ is resampled at each time $t$, and also if sampling is done without replacement.

**Example 7.** To find simple examples for which the $M^{-\frac{1}{4}}$ rate is correct, fix again a computational budget $M$ and approximation level $n \leq M, N$. We define the measure $\mu_n = \left(1 - \frac{1}{\sqrt{n}}\right) \mathbb{U}[0, 1] + \frac{1}{\sqrt{n}} \delta_0$, and consider a sequence of i.i.d. samples $\theta_1, \ldots, \theta_M$ from this distribution. We then have

\[
\mathbb{E}\left(\left(\frac{n}{M} \sum_{t=1}^M \theta_t - \frac{1}{2}\right)^2\right) = \frac{1}{4n} + \frac{n}{3M}.
\]

The optimal choice is $n = \sqrt{\frac{2M}{3}}$, giving a decay rate of $O\left(M^{-\frac{1}{4}}\right)$. We can see here that our approximation assumption does not allow for the sort of cancellation that occurs in the previous example.

One major difference between these two examples is that, in the first, the approximate measures resulted in unbiased estimates; in the second, the approximate measures were also biased. Assumptions of the form given in equation (4.3) are standard in the literature (see e.g., [KCW13, AFEB14, BDH14]), but cannot detect this difference. It may be interesting to try to understand which subsampling algorithms have a useful structure that inequalities such as (4.3) cannot capture.

**Example 8.** As mentioned in remark 3.3, the curvature assumption used in Theorem 5 is strictly weaker than the assumption made in [KCW13]. Our assumption is slightly weaker than the uniform ergodicity assumption used in [AFEB14]. In particular, if a kernel $K$ is uniformly ergodic, a small power of it will have positive curvature in the Total Variation distance, and so we can obtain (slightly weaker) bounds in that case. In the other direction, our result applies to many chains for which no finite power of the transition kernel is uniformly ergodic.
Several examples that are geometrically but not uniformly ergodic, and for which our bounds apply, can be found in [Oll09]. In that paper’s Example 9, the discretization of the Ornstein-Uhlenbeck process \( dX_t = -\alpha X_t dt + dB_t \) with time-steps of size \( \delta \) has curvature \( 1 - e^{-\alpha \delta} \) and eccentricity \( E(x) = O(|x|) \) under the Euclidean metric; thus, our results apply. In the other direction, this kernel is clearly not uniformly ergodic.

We now prove Theorem 5:

**Proof.** We plug earlier estimates into Lemma 3.8, with the family \( \mathcal{F} \) being the collection of 1-Lipschitz functions. The first estimate comes from Lemma 3.2. The second estimate follows from inequality (4.3). The third estimate follows from Theorem 4. □

In the following subsections, we prove tradeoff results similar to Theorem 5, under different assumptions.

### 4.1. Drift and Minorization Conditions

When the state space \( \Omega \) is countable, we have the following related results for chains that are uniformly ergodic:

**Theorem 9** (Austerity Tradeoff: Uniform Ergodicity). Fix a function \( f \) satisfying \( \|f\|_\infty = 1 \) and a computational budget \( M \). Assume that the base kernel \( K \) satisfies Assumptions 3.2 and that the random estimates \( Q_\epsilon \) based on an algorithm with an expected cost of \( n \) computational units satisfy

\[
|\mathbb{E}[Q_\epsilon] - Q| \leq Cn^{-\frac{1}{2}}.
\]

Then for some \( M_0 \) that does not depend on \( N \), and \( M > M_0 \) sufficiently large,

\[
\mathbb{P}[|\pi_{\sqrt{M}_\epsilon}(f) - \pi(f)| > rM^{-\frac{1}{2}}] \leq A(r),
\]

for some function satisfying \( \lim_{r \to \infty} A(r) = 0 \) that does not depend on \( n, M \) or \( N \).

For chains that are geometrically ergodic, we have the marginally weaker conclusion:

**Theorem 10** (Austerity Tradeoff: Geometric Ergodicity). Fix a function \( f \) satisfying \( \|f\|_\infty = 1 \) and a computational budget \( M \).

Assume that, for all finite sets \( \mathcal{X} \), we have \( \sup_{x \in \mathcal{X}} V(x) \equiv D_X < \infty \). Assume that the random estimates \( Q_\epsilon \) based on an algorithm with an expected cost of \( n \) computational units satisfy

\[
|\mathbb{E}[Q_\epsilon] - Q| \leq Cn^{-\frac{1}{2}}.
\]

Then for some \( M_0 \) that does not depend on \( N \), and \( M > M_0 \) sufficiently large, we have for all \( q > 0 \),

\[
\mathbb{P}[|\pi_{\sqrt{M}_\epsilon}(f) - \pi(f)| > rM^{-\frac{1}{2}(1-\frac{1}{q})}] \leq A(r),
\]

where \( A(r) \) is some function satisfying \( \lim_{r \to \infty} A(r) = 0 \) that does not depend on \( n, M \) or \( N \).

**Example 11.** This Theorem applies to the Metropolis-Hastings kernel with target distribution \( \pi(x) \propto e^{-cx} \) on \( \mathbb{N} \) and proposal kernel \( L(x, y) = \frac{1}{3} I_{|x-y| \leq 1} \).

We begin by proving Theorem 9 and then discussing the small modifications needed to prove Theorem 10.
Proof. We must first show that $K$ satisfies condition (3.15) of Lemma 3.7 for all sets $\mathcal{X}$ of the form

$$\mathcal{X} = \mathcal{X}_C = \{ x \in \Omega : V(x) < C \}.$$ 

To see this, let $V$ be a drift function for the original kernel $K$ with constants $a, b$, and let $\hat{X}_t$ be a chain run according to the kernel $K$ restricted to $\mathcal{X}_C$. Let $X_t$ be a copy of the chain run according to $K$, with $X_t = \hat{X}_t \in \mathcal{X}_C$. Then we note that

$$\mathbb{E}[V(\hat{X}_{t+1})|\hat{X}_t] = \mathbb{E}[V(X_{t+1})1_{V(X_{t+1})\leq C} + V(X_t)1_{V(X_{t+1})> C}|X_t]$$

$$\leq \mathbb{E}[V(X_{t+1})|X_t]$$

$$\leq (1 - a)V(X_t) + b$$

$$= (1 - a)V(\hat{X}_t) + b,$$

where the second line follows from the fact that $V(\hat{X}_t) \leq C$. Thus, $V$ is also a drift function for $K$ restricted to $\mathcal{X}_C$, with the same constants. In particular, condition (3.15) is satisfied for all sets of the form $\mathcal{X}_C$.

To complete the proof, we plug earlier estimates into the second part of Lemma 3.8. The first requirement follows from Lemma 3.7. The second estimate follows from the definition of our algorithm with approximation error $\epsilon = \frac{1}{\sqrt{M}}$. This gives $c_1 = c_2 = \frac{1}{2}$, $c_3 = 1$ and $c_4 = \frac{1}{4}$.

We next prove Theorem 10.

Proof. We begin by restricting our attention to sets of the form

$$\mathcal{X}(s) = \{ x : V_\epsilon(x) < s \}.$$ 

For these sets,

$$\mathcal{D}_{\mathcal{X}(s)} \leq s.$$ 

Next, fix an exponent $m$. By inequality (3.10) above, for $\epsilon > 2m$, we have $\mathbb{E}[V_\epsilon(X_{t+1})^m|X_t] \leq (1 - a_{m, \epsilon})V_\epsilon(X_t)^m + b_{m, \epsilon}$ for some $a_{m, \epsilon} < 1$ and $b_{m, \epsilon} < \infty$. By Markov’s inequality, then, we have:

$$(1 - \mathbb{P}[\{X_t\}_{t=1}^T \subset \mathcal{X}(s)] = O\left(\frac{T + V_\epsilon(X_1)}{s^n}\right).$$

Combining this with inequality (3.16), we have for some constant $C$ depending on the initial point $X_0$, all $1 \leq q \leq n$ and all $s > 0$:

$$\mathbb{P}[|\pi_{T, \delta}(f) - \pi(f)| > \frac{r}{\sqrt{T}} + \frac{\delta q}{(C + s)(1 - \theta)^q} + (1 - (C + s)(1 - \theta)^q)\frac{T}{q}]$$

$$\leq e^{-\frac{(1 - \theta)(r)}{2(C + s)^2}} + \frac{T + C}{s^n}.$$

Choosing $s = T^{-\frac{1}{m+1}}, r = uT^{-\frac{1}{m+1}}$ and $q = \log(T)^2$, this becomes:

$$\mathbb{P}[|\pi_{T, \delta}(f) - \pi(f)| > uT^{-\frac{1}{m+1}} + o\left(T^{-\frac{1}{m+1}}\right)] \leq e^{-\frac{(1 - \theta)^2u^2}{2C\epsilon^2}} + O\left(T^{-\frac{1}{m+1}}\right).$$

(4.5)
We plug earlier estimates into the second part of Lemma 3.8. The first requirement follows from Lemma 3.6. The second estimate follows from the definition of our algorithm with approximation error \( \delta = \frac{1}{\sqrt{M}} \). This gives \( c_1 = \frac{1}{2} + \frac{1}{m-1} \), \( c_2 = \frac{1}{2} \), \( c_3 = 1 \) and \( c_4 = \frac{1}{4} (1 - \frac{1}{m}) \). □

5. Application 2: Exponential Random Graph Models

While the austerity framework has so far been discussed, both above and in [KCW13], for i.i.d. data, it makes sense to search for similar bias-variance tradeoffs for MCMC samplers targeting more complicated posterior distributions. In this section, we briefly discuss this problem in the context of the popular exponential random graph models (ERGM), finding an analogue to Theorem 5. Our main result is quantitative enough to show that subsampling can improve the computational efficiency of MCMC estimates in this situation, even though the model is much more complicated. We emphasize that, while this is useful confirmation of our intuition about subsampling, the bounds themselves are not practical. Due to other computational difficulties associated with the ERGM (see e.g., [CD13]), we find it unlikely that subsampling will play an important role in estimation in the near future. In addition, unlike Theorem 5, we do not believe that the bounds in this section are sharp or that they offer practical guidance in the choice of approximating kernel.

Recall that the probability of observing a given graph \( G \) in an ERGM is given by

\[
p(G) = e^{-\sum_{i=1}^{k} f_i(N) \beta_i T_i(G) - f(N) \phi(N, \beta)},
\]

where \( G \) is a graph on \( N \) nodes, \( \beta = (\beta_i)_{i=1}^{k} \) is a vector of parameters, \( f_i(N) \) is a collection of normalizing values, \( T_i \) is a collection of graph functions such as number of edges, number of triangles, and \( \phi(N, \beta) \) is a normalizing constant. We wish to sample from the posterior distribution of \( (\beta_1, \ldots, \beta_k) \) given an observed graph \( G \). We focus on the common situation in which \( |G| = N \) is extremely large, and specialize to the case in which \( T_i \) counts the number of times the graph \( H_i \) is included as a subgraph of the dense graph \( G \). We denote by \( v_i \) and \( e_i \geq 1 \) the number of vertices and edges of \( H_i \). For example, we might have \( H_1 \) be an edge between two vertices \( (v_1 = 2, e_1 = 1) \) and \( H_3 \) be a triangle \( (v_2 = e_2 = 3) \). Then we set the normalizing constants \( f_i(n) = n^{-v_i} \) and define \( V = \max_i v_i \) and \( E = \max_i e_i \).

We next describe our biased kernel. Fix a reversible proposal kernel \( L \) on the parameter space \( \mathbb{R}^k \) and let \( K \) be the associated Metropolis-Hastings kernel with proposal kernel \( L \) and target distribution

\[
\pi(\beta) \equiv \pi(\beta|G) \propto \rho(\beta)e^{-\sum_{i=1}^{k} f_i(N) \beta_i T_i(G) - f(N) \phi(N, \beta)}.
\]

To define our approximate kernel \( K_\epsilon \), we will follow Algorithm 4 with one small modification. As in Equation (4.2), we will compute \( Q_\epsilon \) via an estimate \( \pi_\epsilon \) of the target distribution. To do so, will fix \( n = n(X_\epsilon, \epsilon) \), draw a random collection of \( n \ll N \) vertices, and define \( G_n \subset G \) to be the induced subgraph. We then define \( T_{i, \epsilon}(G) \) to be the number of times that subgraph \( H_i \) appears in \( G_n \) and define

\[
\pi_\epsilon(\beta) \propto \rho(\beta)e^{-\sum_{i=1}^{k} f_i(n) \beta_i T_{i, \epsilon}(G) - f(n) \phi(n, \beta)}.
\]

Theorem 12 (Austerity Improvement for ERGM). Fix a computational budget \( M \) and a function \( f \) of interest with \( ||f||_\infty = 1 \). Assume that the base Markov chain \( K \) satisfies

\[
||\mu K - \nu K||_{TV} \leq (1 - \alpha) ||\mu - \nu||_{TV}
\]

20
for some $\alpha > 0$ and all distributions $\mu, \nu$. Then, fixing $\epsilon = \frac{2}{E \log_2(M)^{\frac{3}{8}}}$ and choosing an approximation based on $n = n(\epsilon) = \min \left( N, 2^{\frac{3}{8} \sum_{i=1}^{k} \beta_i} \right)$, we have for all $0 < \epsilon < \alpha$ that
\[
\mathbb{P} \left[ \left| \pi_{\min(M, 2^{\frac{3}{8} \epsilon})} f - \pi f \right| > r \epsilon \right] \leq A(r),
\]
for some function satisfying $\lim_{r \to \infty} A(r) = 0$ that does not depend on $N$ or $M$.

Remark 5.1. Before proving this result, it is important to remark on the normalizing constant $\phi(N, \beta)$. This constant is necessary in order to run the usual Metropolis-Hastings chain on the space of parameters, and it is generally very hard to estimate. We will ignore this issue for two reasons. The first is that it is a difficult problem, outside of the scope of this paper. The second is that some interesting work has been done on computing this constant precisely for large $N$ in some settings (see e.g., Theorems 3.1 and 4.1 of [CD13]), and we hope that this will make estimates like those in this Theorem easier to apply in the future.

Remark 5.2. This convergence rate, in terms of $M$, is of course very slow. The argument below can be tightened to give a negligibly better bound, but not one that is useful. We point out only that, for $N$ very large, this convergence bound is better than that obtained by choosing $n = N$ at every step.

Proof. Our proof is essentially identical to that of Theorem 5; the biggest change is that we need to bound $E \left( \min \left( 1, \frac{\pi(x)}{\pi(y)} \right) \right)$ directly, without e.g. using the Berry-Esseen Theorem.

Denote by $\delta_{\square}$ the cut metric on graphs (see e.g. [BCL+06] for a definition). Recall Theorem 5.7 of [BCL+06]:

Theorem 13. Fix $\epsilon > 0$. Then for $n > 2^{\frac{4}{3}}$,
\[
\mathbb{P}[\delta_{\square}(G, G_n) > \epsilon] \leq \epsilon.
\]

By Lemma 5.2 of [BCL+06], this implies that for $n > 2^{\frac{4}{3}}$,
\[
\mathbb{P}[|n^{-v_i}T_i(G) - N^{-v_i}T_i(G)| > e_i \epsilon] \leq \epsilon.
\]

We note that
\[
\log \left( \frac{\pi_\epsilon(\beta)}{\pi(\beta)} \right) = \sum_{i=1}^{k} \beta_i \left( N^{-v_i}T_i(G) - n^{-v_i}T_i,\epsilon(G) \right).\]

By inequality (5.1), we have that for any $\epsilon > 0$ and $n > 2^{\frac{4}{3}}$,
\[
\mathbb{P}[|\log \left( \frac{\pi_\epsilon(\beta)}{\pi(\beta)} \right) | > E \epsilon \sum_{i=1}^{k} \beta_i] \leq \epsilon.
\]

In particular, to obtain a uniform upper bound of $E \epsilon$ on this ratio with probability at least $1 - \epsilon$, it is sufficient to sample $n = \min \left( N, 2^{\frac{3}{8} \sum_{i=1}^{k} \beta_i} \right)$ points from $G$ at each step. This provides an upper bound on $\frac{\pi_\epsilon}{\pi}$, and thus the upper bound
\[
\|K(x, \cdot) - K_\epsilon(x, \cdot)\|_{TV} < \epsilon
\]
(5.3)
for $n = \min\left(N, 2^{(\frac{L}{4} \max(1, \sum_{i=1}^{N} \beta_i))^2}\right)$. We then follow the remainder of the proof of Theorem 5, with inequality (5.3) replacing the estimate of $\|K(x, \cdot) - K_\epsilon(x, \cdot)\|_{TV}$ wherever it appears (including passing it through the calculation in Lemma 3.8).

6. Application 3: Stochastic Gradient Langevin Dynamics

Fix a model with parameter vector $\theta$, prior $p(\theta)$ and likelihood $p(\theta|x)$ for a single data point $x$. We consider the setting of drawing $N$ i.i.d. points $x_i, \ldots, x_N$ points from this model, where $N$ is extremely large. In their paper [WT11], Welling and Teh introduced the following Stochastic Gradient Langevin Dynamics for sampling from such a posterior without looking at all $N$ data points at every step:

$$
\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \left( \nabla \log(p(\theta_t)) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log(p(x_{ti}|\theta_t)) \right) + \eta_t
$$

(6.1)

where $\epsilon, n > 0$ are fixed constants, $\eta_t \sim \mathcal{N}(0, \epsilon^2)$ and $x_{ti}$ are a subset of $n < N$ data points chosen uniformly at random. In [WT11], the authors provide useful heuristics for the convergence of this algorithm for fixed $n$. We will provide rigorous justification and error bounds associated with this convergence, and show that convergence can hold even when the heuristics in [WT11] don’t apply. Although we find convergence estimates, we avoid giving tradeoff theorems such as our Theorem 5. Such results would require careful discussion of the convergence properties of the Langevin dynamics themselves, which is beyond the scope of this paper. Convergence results such as those found in [DT12, RS02, BRH13] could be used to find analogues to our Theorem 5 for the Stochastic Gradient Langevin Dynamics.

We begin by comparing the Stochastic Gradient Langevin Dynamics to the full Langevin Dynamics as introduced in [Nea10]:

$$
\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \left( \nabla \log(p(\theta_t)) + \sum_{i=1}^{N} \nabla \log(p(x_i|\theta_t)) \right) + \eta_t,
$$

where the constants are as above. We also define the Gradient Dynamics to be:

$$
\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \left( \nabla \log(p(\theta_t)) + \sum_{i=1}^{N} \nabla \log(p(x_i|\theta_t)) \right),
$$

(6.2)

where again the constants are as above. We compute:

**Lemma 6.1** (Approximation of Dynamics). Let $K$ and $\tilde{K}$ be the kernels associated with the stochastic gradient Langevin dynamics and the Langevin dynamics respectively, and define $V(\theta) = \sup_i |\nabla \log(p(x_i|\theta))|$. Then for $d$ the usual Euclidean distance, we have the bound:

$$
W_d(K(\theta, \cdot), \tilde{K}(\theta, \cdot)) \leq \frac{\epsilon^2 N^2}{4} \frac{N^2}{n} V(\theta)^2.
$$

(6.3)

**Proof.** Let $\theta_t$ evolve according to the Stochastic Gradient Langevin Dynamics, with noise variable $\eta_t$, and let $\tilde{\theta}_t$ evolve according to the full Langevin Dynamics, with noise variable $\tilde{\eta}_t$. We will couple $\theta_{t+1}, \tilde{\theta}_{t+1}$ started at the same point $\theta_t = \tilde{\theta}_t = \Theta$ by first choosing the data
points \( \{x_{ti}\}_{i=1}^{n} \) and then coupling \( \eta_t, \tilde{\eta}_t \) conditional on the points chosen. Define
\[
E_t = \epsilon \frac{N}{n} \sum_{i=1}^{n} \left( \nabla \log(p(x_{ti}|\theta_t)) - \frac{1}{N} \sum_{i=1}^{N} \nabla \log(p(x_i|\theta_t)) \right)
\]
and
\[
V(\theta) = \sup_i |\nabla \log(p(x_i|\theta))|.
\]
We note that each term in the sum defining \( E_t \) is bounded by \( V(\theta_t) \). Thus, we find:
\[
\mathbb{E}[E_t^2] \leq \frac{\epsilon^2 N^2}{4} \frac{V(\theta_t)^2}{n}.
\]
To find inequality (6.3) we set \( \eta_t = \tilde{\eta}_t \) and the proof is finished. □

**Remark 6.2.** We note immediately:

- If \( V(\theta) \) is uniformly bounded in \( \theta \), this bound together with the fact that the convergence rate of the usual Langevin dynamics is order of \( \epsilon \) can be plugged into Lemma 3.2 (or one of the other, similar, lemmas) to show that Stochastic Gradient Langevin dynamics converges when the full Langevin dynamics do, as \( \epsilon \) goes to 0 sufficiently slowly. This is essentially a rigorous version of the argument in [WT11].

- For most familiar distributions on noncompact state spaces, \( V(\theta) \) is not uniformly bounded. For example, it is not bounded if the model is given by \( p(x|\theta) \sim N(\theta, 1) \).

In these situations, the above bounds cannot be used to show convergence.

In light of this remark, we suggest three types of convergence theorem that might be useful:

1. Restrict our attention to target distributions for which \( V(\theta) \) is bounded. In this situation, we obtain straightforward convergence estimates.
2. We consider additional assumptions under which the stochastic gradient Langevin dynamics converge even under the presence of stepwise error that is not uniformly bounded.
3. Allow the number of points \( n \) evaluated per step to vary with the location in the state space. If \( n = n(\theta, t) \) depends on the state space, this allows us to force the appropriately-scaled error \( \delta = \frac{V(\theta)^2}{n(\theta,t)} \) to be uniformly bounded in \( \theta \). In this setting, the same convergence result will hold, but this will come at a potentially unbounded cost in number of function evaluations. This tension can be resolved by estimating the the mean total number of data points \( \sum_{t=0}^{T} n(\theta, t) \) used over a run of length \( T \).

We consider these three types of results in order. For case 1,

**Theorem 14** (Convergence of Stochastic Gradient Langevin Dynamics for Nice Targets). Assume that the Langevin Dynamics associated with parameter \( \epsilon > 0 \) satisfies:
\[
W_d(K_{\epsilon}(x, \cdot), K_{\epsilon}(y, \cdot)) \leq (1 - a_{\epsilon})d(x, y)
\]
for all \( x, y \) and for some function \( a_{\epsilon} \) that satisfies \( \inf_{0 < \epsilon < \epsilon_0} \frac{a_{\epsilon}}{\epsilon} > c > 0 \) for some \( \epsilon_0 > 0 \). Also assume that \( \sup_{\theta} \sup_{i} |\nabla \log(p(x_i|\theta))| \equiv V_0 < \infty \). Finally, define \( \pi_{\epsilon} \) and \( \tilde{\pi}_{\epsilon} \) to be the
stationary distributions of the Stochastic Gradient Langevin Dynamics and usual Langevin Dynamics. We conclude that, for all $0 < \epsilon < \epsilon_0$,

$$W_d(\pi_\epsilon, \bar{\pi}_\epsilon) \leq \frac{\epsilon N^2 V_0}{4nc}.$$ 

**Proof.** This follows immediately from Lemma 3.2 and Lemma 6.1. \qed

**Remark 6.3.** This Theorem applies if $\pi$ is a log-concave distribution restricted to a compact and convex subset of Euclidean space.

Next, we consider case 2. In many natural examples, having a uniformly good approximation of the kernel $\tilde{K}$ turns out to be unnecessary, just as in Lemma 3.2:

**Theorem 15** (Fixed Sample Size Approximation). Fix $\epsilon > 0$. Assume that, for all possible collection of points $x_1, \ldots, x_m$, the (deterministic) gradient dynamics associated with $p(\theta | x_1, \ldots, x_m)$, as defined in equation (6.2), have a Lyapunov function $L_{x_1, \ldots, x_m}$ with associated constants $0 < a_{x_1, \ldots, x_m} \leq 1$ and $b_{x_1, \ldots, x_m} < \infty$. Furthermore, assume that

$$\lim_{r \to \infty} \sup_{|\theta| = r} \left| \frac{L_{x_1, \ldots, x_m}(\theta)}{L_{x_1', \ldots, x_m'}(\theta)} - 1 \right| = 0$$

(6.4)

for any collection of data points $x_1, \ldots, x_m$ and $x_1', \ldots, x_m'$ and that

$$\sup_{\theta : L_0,0,\ldots,0(\theta) < c} V(\theta)^2 \equiv S(c) < \infty$$

(6.5)

for all $c > 0$.

Also assume that, for the same $\epsilon > 0$ and any $x_1, \ldots, x_m$ and $\delta > 0$, there exists a compact set $\mathcal{X} = \mathcal{X}(\delta, x_1, \ldots, x_m)$ so that for $\theta \notin \mathcal{X}$,

$$\mathcal{N}(\theta, \epsilon)(L_{x_1, \ldots, x_m}) \leq (1 + \delta)L_{x_1, \ldots, x_m}(\theta)$$

(6.6)

We also assume that for any compact set $\mathcal{X}$, $\sup_{\theta \in \mathcal{X}} V(\theta) \equiv V(\mathcal{X}) < \infty$. Assume that the Langevin Dynamics kernel associated with parameter $\epsilon > 0$ satisfies:

$$W_d(K_\epsilon(x, \cdot), K_\epsilon(y, \cdot)) \leq (1 - a_\epsilon)d(x, y)$$

for all $x, y$ and for some function $a_\epsilon$ that satisfies $\inf_{0 < \epsilon < \epsilon_0} \frac{a_\epsilon}{\epsilon} > c > 0$ for some $\epsilon_0 > 0$. Then for $\delta < \frac{1 - a_\epsilon}{10}$,

$$W_d(\theta_T - p(\theta | x_1, \ldots, x_N)) < \frac{\delta}{a_\epsilon} + (1 - a_\epsilon)T E(X_0) + T \frac{L_{0,0,\ldots,0}(X_0)}{S^{-1} \left( \frac{\epsilon M}{N} \right)}.$$ 

**Proof.** This result will follow from Lemmas 3.2 and 6.1. Let $(a, b) = (a_0,0,\ldots,0, b_0,0,\ldots,0)$ be the constants associated with the Lyapunov function $L_{0,0,\ldots,0}$ with $n$ 0’s. Let $\delta < \frac{1 - a}{10}$, and define

$$\mathcal{X}_1 = \cup_{\{y_1,\ldots,y_n\} \subset \{x_1,\ldots,x_N\}} \mathcal{X}(\delta, y_1, \ldots, y_n).$$

Then define:

$$R = \inf \left\{ r : \sup_{|\theta| \geq r} \sup_{\{y_1,\ldots,y_n\} \subset \{x_1,\ldots,x_N\}} \left| \frac{L_{y_1,\ldots,y_n}(\theta)}{L_{0,0,\ldots,0}(\theta)} - 1 \right| < \delta \right\}$$
and set
\[ X_2 = B_R(0). \]
Then set \( \mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2. \) We note that for the Langevin dynamics, and any \( \theta_t \notin \mathcal{X}, \)
\[
\mathbb{E}[L_{0,\ldots,0}(\theta_{t+1})|\theta_t = \theta] = \frac{1}{N} \sum_{\{y_1,\ldots,y_N\} \subseteq \{x_1,\ldots,x_N\}} \mathbb{E}[L_{0,\ldots,0}(\theta_{t+1})|\theta_t = \theta, x_{ti} = y_i]
\]
\[
\leq (1 - a)(1 + \delta)^2 L_{0,\ldots,0}(\theta) + 2b
\]
\[
\leq \left(1 - \frac{1 - a}{2}\right) L_{0,\ldots,0}(\theta) + 2b.
\]

For \( a \geq 0, \) define \( \mathcal{X}_a = \{x : L_{0,0,\ldots,0}(x) \leq a\} \cup \mathcal{X}. \) By the above inequality and Markov’s inequality, we have:
\[
\mathbb{P}[X_t \notin \mathcal{X}_a|X_0] \leq \frac{L_{0,0,\ldots,0}(X_0) + 2b}{a}
\]
and thus
\[ 1 - \mathbb{P}[\{X_s\}_{s=0}^t \in \mathcal{X}_a|X_0] \geq 1 - t \frac{L_{0,0,\ldots,0}(X_0) + 2b}{a}. \tag{6.7} \]
Define the good set \( \mathcal{G}(\delta) = \{\theta : \theta^2 \leq \delta\} \) and then \( c_\delta = \sup\{c : \mathcal{X}_c \subset \mathcal{G}(\delta)\}. \) Also note that the stochastic gradient dynamics satisfy inequality (6.7), using the same proof as given above. Then combining inequality (6.7) evaluated at \( a = c_\delta \) with the bound in Lemma 6.1 and plugging this into Lemma 3.2 with \( \mathcal{G}(\delta)^c \) the set to be avoided, we have:
\[
W_d(\theta_T - p(\theta|x_1,\ldots,x_N)) \leq \delta \frac{\epsilon}{a_\epsilon} + (1 - a_\epsilon)\epsilon^T E(X_0)
\]
\[ + 2T \sup_{x \in \mathcal{X}_a \cup \{X_0\}} L_{0,0,\ldots,0}(x) + 2b \frac{c_\delta}{a_\epsilon}. \tag{6.8} \]
Expanding the definition of \( c_\delta, \) we see that it is given by:
\[
c_\delta = \sup\{c : L(\theta) \leq c \Rightarrow V(\theta)^2 \leq \frac{1}{\epsilon^2 \frac{Nn}{N^2}}\}
\]
and so in particular, for all fixed \( \delta > 0, \) we have by equation (6.5) that \( c_\delta = S^{-1}\left(\frac{\sqrt{2\pi}}{N}\right). \)
Combining this with inequality (6.8) completes the proof. \( \square \)

**Remark 6.4.** The assumptions in Equations (6.4) and (6.6) seem strongest. We point out that the assumption in equation (6.4) holds for most distributions, including e.g. the Normal, Cauchy, exponential, and largely requires that individual data points don’t have too much influence on the distribution. Inequality (6.6) is only slightly stronger; it will hold as long as the density decays no more quickly than \( e^{-\epsilon^2 c_m^2} \) for some \( C < \infty. \)

**Example 16.** We point out that Theorem 15 gives asymptotic convergence for some examples where the assumptions of Theorem 14 fail to hold. In particular, we consider a target distribution \( p(\theta|x_1,\ldots,x_m) = \mathcal{N}\left(\frac{\sum_{i=1}^m x_i}{m}, 1\right) \) with an (improper) flat prior. For all \( D > 0, \)
the Langevin dynamics have a Lyapunov function

\[ L_{x_1, ..., x_m}(\theta) = \left( \theta - \frac{1}{m} \sum_{i=1}^{m} x_i \right)^D \]

with associated constants \( a_{x_1, ..., x_m} = b_{x_1, ..., x_m} = 0 \). This Lyapunov function clearly satisfies equation (6.4). The function \( V(\theta) \) is given by

\[ V(\theta) = \sup_{i \in [N]} |\theta - x_i|. \]

This is not bounded as a function of \( \theta \), and so the conditions of Theorem 14 don’t hold. We have

\[ V(\theta)^2 \leq (L_{x_1, ..., x_m}(\theta))^2 + \sup_{i \in [N]} \left( x_i - \frac{1}{N} \sum_{j=1}^{N} x_j \right)^2 \]

uniformly in data points \((x_1, \ldots, x_m)\) and also in \( m \). Thus, in Equation (6.5), we have

\[ S(c) = c^2 \beta + O(1) < \infty. \]

We also note that

\[ \mathcal{N}(\theta, \epsilon)(L_{x_1, ..., x_m}(x)) = \left( \theta - \frac{1}{m} \sum_{i=1}^{m} x_i \right)^2 + \epsilon^2 \]

\[ = (L_{x_1, ..., x_m}(x))(\theta) + \epsilon^2, \]

so inequality (6.6) is satisfied by defining \( \mathcal{N}(\delta, x_1, \ldots, x_m) = \left\{ \theta : |\theta - \frac{1}{m} \sum_{i=1}^{m} x_i| < \frac{\epsilon}{\sqrt{\delta}} \right\} \).

We conclude that, for fixed \( 0 < D < \frac{1}{2} \) and setting \( \delta = \epsilon^{1+\frac{1}{D+1}} \), we find that the error decays asymptotically at least as quickly as \( \epsilon^{\frac{1}{D+1}} \). Letting \( D \) go to 0, we find that we have asymptotic decay of the bias at the rate of \( \epsilon^c \) for any \( c < 1 \). We note that the Langevin dynamics themselves have bias that decays at an asymptotic rate of \( \epsilon \).

Finally, we consider case 3, looking at dynamics for which \( V(\theta) \) is not bounded. We apply Theorem 14, but set \( n = n(\theta, t) \equiv \min \left( N, \frac{N^2 \epsilon^2 V(\theta)^2}{V_0^2} \right) \) for some fixed \( V_0 > 0 \). The result then applies as stated for this value of \( V_0 \). For many examples, such as that considered immediately above, the convergence result in Theorem 14 combined with the Law of Large Numbers for Markov chains give effective a.s. bounds on \( \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} n(\theta_t, t) \). In particular, since \( V(\theta_t)^2 \) is often small outside of the tails of the target distribution, we can often achieve uniform bounds on the approximation quality with minimal extra computational cost.

7. Global Bias and Mixing Properties

In other sections of this paper, we have discussed the situation in which it is computationally expensive to run even a single step from some desired kernel \( K_0 \), and we construct an approximating kernel \( K_\epsilon \) which is easier to run but gives asymptotically biased results. In that setting, we are trading off bias against a purely local improvement in the MCMC algorithm - decreasing the time it takes to calculate individual steps. Although the approximate kernels are easier to run, they may not have better mixing properties than the underlying kernels; indeed, the mixing properties may be significantly worse. In this section, we briefly
discuss different ways to trade off bias against a more global improvement in the MCMC algorithm by constructing approximating chains $K'_\epsilon$ that have better convergence properties than the initial approximations $K_\epsilon$, at the cost of a small amount of additional bias. The algorithms in this section are all defined through a global estimate $\pi_\epsilon$ of $\pi$, though in principle there is no reason to require this.

This is a large subject; indeed our recent paper [PM+13] is entirely concerned with a family of approximations that falls into this framework. In this section, we discuss only two very simple examples: one related to the stochastic gradient Langevin dynamics (SGLD) already discussed, the other related to the well-known MCMC variant of the approximately Bayesian computation (ABC) algorithm (see [MPRR12] for a survey).

The main observation behind the examples in this section is as follows. For many approximate samplers, including both ABC-MCMC and SGLD, the error $d(K_\epsilon(\theta, \cdot), K(\theta, \cdot))$ of the approximations have a tendency to decay for $\theta$ in the tails of the target distributions. For ABC-MCMC, this is discussed in great detail in [LL12]; for SGLD, it is discussed in Section 6 above. In both of these cases, this decay makes the algorithms more difficult to analyze; in the case of ABC-MCMC, this decay results in very poor mixing properties. The simple solution we discuss here is to bias the target distributions towards a global estimate associated with a rapidly-mixing Markov chain.

We begin by looking at ABC-MCMC. Recall that the ABC-MCMC algorithm requires a proposal kernel $L$, a prior $\pi$ on the parameter $\omega \in \Omega$ of interest, a data-generating model $\pi(\cdot | \omega)$ for each value of the parameter $\omega$, observed data $y_{\text{obs}}$, a pseudo-metric $d$ on the state space, and a tuning parameter $\epsilon > 0$. For any $\omega \in \Omega$ and draw $y' \sim \pi(\cdot | \omega)$, we define the estimate:

$$\hat{\pi}_\epsilon(\omega | y_{\text{obs}}) = \pi(\omega)1_{d(y, y_{\text{obs}}) < \epsilon}.$$  

(7.1)

The ABC-MCMC algorithm is described in Algorithm 7. The marginal distribution of $\omega$

Algorithm 3 ABCMCMC

| Initialize $(\omega_1, y_1)$ |
|--------------------------|
| For $t = 1$ to $N$ do:   |
| Generate $\omega'$ from $L(\omega_t, \cdot)$. |
| Generate $y'$ from $\pi(\cdot | \omega')$. |
| Set $\hat{\pi}_\epsilon(\omega | y_{\text{obs}})$ as in Equation (7.1). |
| Generate $u$ from $U[0, 1]$. |
| if $u \leq \frac{\hat{\pi}_\epsilon(\omega | y_{\text{obs}}, y')L(\omega', \omega)}{\hat{\pi}_\epsilon(\omega | y_{\text{obs}})L(\omega, \omega, \omega)}$ then |
| Set $(\omega_{t+1}, y_{t+1}) = (\omega', y')$. |
| else |
| Set $(\omega_{t+1}, y_{t+1}) = (\omega_t, y_t)$. |
| end if |

under this chain’s stationary distribution is given by:

$$\pi_\epsilon(\omega | y_{\text{obs}}) = \pi(\omega|\{y : d(y, y_{\text{obs}}) < \epsilon\}).$$

As discussed in in [LL12], the Markov chain described by this algorithm often fails to be variance bounding (see [RR08] for definition of variance bounding) and thus also fails to be
geometrically ergodic. This problem occurs even for simple posteriors, as when estimating the mean of a normal distribution with known variance.

The primary problem is that the acceptance probabilities of this kernel often go to 0 in the tails of the posterior. There are several proposed solutions to this problem in the literature; some, such as those in [LL12] and the related paper [GLS+13], are quite sophisticated. We describe here a very simple solution: for a fixed distribution \( \rho \) and constant \( \delta = \delta(\epsilon) \), when running this algorithm replace the estimate \( \hat{\pi}_\epsilon \) in equation (7.1) with the estimate \( \hat{\pi}'_\epsilon = \hat{\pi}_\epsilon + \delta \rho \). The result, which we will call Biased MCMC-ABC (B-MCMC-ABC) for the remainder of this section, is a pseudo-marginal Markov chain in the sense of [AR09], and so in particular is a reversible Markov chain. We note that

\[
\mathbb{E}[\hat{\pi}'_\epsilon(\omega|y_{\text{obs}})] = \mathbb{E}[\hat{\pi}_\epsilon(\omega|y_{\text{obs}}) + \delta \rho(\omega)] \\
= C\pi(\omega|y_{\text{obs}}) + \delta \rho(\omega),
\]

for some constant \( C \) independent of \( \epsilon \) and \( \delta \). Thus, the marginal distribution of \( \omega \) under this chain’s stationary distribution is given by:

\[
\pi(\omega|y_{\text{obs}})' \propto C\pi(\omega|\{y : d(y, y_{\text{obs}}) < \epsilon\}) + \delta \rho(\omega).
\]

In particular, for a B-MCMC-ABC algorithm in \( \mathbb{R}^n \), the bias of \( \pi(\omega|y_{\text{obs}})' \) is \( O(\delta + \epsilon^n) \), and is thus easy to control; the bias for the usual MCMC-ABC algorithm is \( O(\epsilon^n) \).

We claim that, for natural target distributions \( \pi \) and priors \( \rho \), this modified algorithm retains variance bounding. We do this by comparing B-MCMC-ABC to a chain with kernel \( K \) described by a mixture of two kernels, \( K_1 \) and \( K_2 \). Denote by \( Q \) the kernel associated with B-MCMC-ABC, and \( q \) its associated acceptance function. \( K_1 \) is a Metropolis-Hastings algorithm that has the same proposal kernel \( L \) and stationary distribution as B-MCMC-ABC; denote by \( k_1 \) its acceptance function. \( K_2 \) is the chain that doesn’t move. We write \( K = \alpha K_1 + (1 - \alpha)K_2 \), where \( \alpha = \sup\{a : \sup_{x,y} \frac{ak_1(x,y)}{q(x,y)} \leq 1\} \). We have

**Proposition 7.1.** If \( \alpha < 1 \) and \( K \) is variance bounding, then \( Q \) is variance bounding.

**Proof.** By the definition of \( \alpha \), if \( \alpha < 1 \) then \( K(x, A) \leq Q(x, A) \) for all \( x \) and all \( A \) satisfying \( A \cap \{x\} = \emptyset \). Thus, \( Q \) inherits the variance bounding property from \( K \) via Theorem 8 of [RR08]. □

**Remark 7.2.** We note that \( \alpha < 1 \) occurs in great generality. In particular, this works for the prototypical example of \( \pi(\cdot|\omega) = \mathcal{N}(\omega,1) \) for \( \omega \in \mathbb{R} \), \( \pi(\omega) = \mathcal{N}(0, 10) \) and \( L(x, \cdot) = \mathcal{N}(x, 0.1) \)

where we allow for shrinkage towards a slightly stretched version of the actual prior, so that \( \rho(\omega) = \mathcal{N}(0, 20) \). In this situation, we have for all \( \delta \) sufficiently small relative to \( \epsilon \) and all sufficiently large \( \omega \) that \( \frac{k_1(x,y)}{q(x,y)} = \Omega(\delta) \).

We now look at notions of shrinkage for stochastic gradient Langevin dynamics. As discussed in Section 6, the SGLD can be arbitrarily far from the full Langevin dynamics that they are approximating. This can occur even for very simple target posterior distributions, such as a normal distribution with known variance. Even though this algorithm is relatively new, some modifications in the literature already ameliorate this problem; see e.g., [CFG14]. In this section, we discuss a much simpler modification.
We follow the notation in Section 6. For constant $0 < C = C(\theta, \delta, \epsilon, n, N) < 1$, we modify Equation (6.1) as follows:

$$
\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \left( C \nabla \log(p(\theta_t)) + (1 - C) \sum_{i=1}^{N} \nabla \log(p(x_i|\theta_t)) \right) + \eta_t;
$$

(7.2)

we call this “shrunk-SGLD” or SSGLD. If we fix $\delta > 0$ and choose

$$
C = 1 - \min \left( 1, \frac{\sqrt{4\delta n}}{\epsilon N V(\theta)} \right),
$$

we can guarantee that the SSGLD stay within $\delta$ of the associated shrunk Langevin dynamics (SLD) given by

$$
\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \left( C \nabla \log(p(\theta_t)) + (1 - C) \sum_{i=1}^{N} \nabla \log(p(x_i|\theta_t)) \right) + \eta_t,
$$

as measured by Wasserstein distance. In particular, Theorem 14 applies to the comparison of these two ‘shrunk’ dynamics. If the prior $p$ is log-convex and $V(\theta)$ satisfies $\sup_{\theta \in \mathcal{X}} |V(\theta)| < \infty$ for all compact sets $\mathcal{X}$, Theorem 3.1 also implies that the shrunk Langevin dynamics converge to the standard Langevin dynamics as $\epsilon$ goes to 0. Thus, as long as $p$ is log-convex and $\sup_{\theta \in \mathcal{X}} |V(\theta)| < \infty$ for all compact sets $\mathcal{X}$, the SSGLD converges to the SLD as $\delta$ goes to 0 and the SLD converges to the LD as $\epsilon$ goes to 0. In particular, under these conditions the SSGLD converges to the correct stationary distribution when the LD does.

ACKNOWLEDGEMENTS

NSP is partially supported by NSF and ONR grants.

REFERENCES

[AFEB14] P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy monte carlo: Convergence of markov chains with approximate transition kernels. Preprint, 2014.

[AR09] Christophe Andrieu and Gareth Roberts. The pseudo-marginal approach for efficient monte carlo computations. Annals of Statistics, 37(2):1139–1160, 2009.

[BCL+06] Christian Borgs, Jennifer Chayes, László Lovász, Vera Sos, Balazs Szegedy, and Katalina Vesztergombi. Counting Graph Homomorphisms, volume Topics in Discrete Mathematics. Springer, 2006.

[BDH14] Remi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up markov chain monte carlo: an adaptative subsampling approach. Preprint, 2014.

[Bea03] M.A. Beaumont. Estimation of population growth or decline in genetically monitored populations. Genetics, (164):1139–1160, 2003.

[BRH13] N. Bou-Rabee and M. Hairer. Nonasymptotic mixing of the mala algorithm. IMA Journal of Numerical Analysis, (33):80–110, 2013.

[CD13] Sourav Chatterjee and Persi Diaconis. Estimating and understanding exponential random graph models. Annals of Statistics, 41(5):2428–2461, 2013.

[CFG14] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. Preprint, 2014.

[DT12] A. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin. J. Comput. System Sci., 78(5):1423–1443, 2012.
[GLS+13] Mark Girolami, Anne-Marie Lyne, Heiko Strathmann, Daniel Simpson, and Yves Atchade. Playing russian roulette with intractable likelihoods. Preprint, 2013.

[JO10] Alderic Joulin and Yann Ollivier. Concentration, curvature and error estimates for Markov chain Monte Carlo. Ann. Prob., 38:2418–2442, 2010.

[Jor13] Michael Jordan. On statistics, computation and scalability. Preprint, 2013.

[KCW13] Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in mcmc land: Cutting the metropolis-hastings budget. Preprint, 2013.

[KW13] Aryeh Kontorovitch and Roi Weiss. Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. Preprint, 2013.

[LL12] Anthony Lee and Krzysztof Latuszynski. Variance bounding and geometric ergodicity of markov chain monte carlo kernels for approximate bayesian computation. Preprint, 2012.

[Mit05] A.Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. Journal of Applied Probability, 42:1003–1014, 2005.

[MPRR12] Jean-Michel Marin, Pierre Pudlo, Christian Robert, and Robin Ryder. Approximate bayesian computational methods. Statistics and Computing, 22(6):1167–1180, 2012.

[Nea10] Radford Neal. Mcmc using hamiltonian dynamics. Handbook of Markov Chain Monte Carlo, 2010.

[OBB+00] P.D. O’Neil, D.J. Balding, N.G. Becker, M. Eerola, and D. Mollison. Analyzes of infectious disease data from houseing the expected value of ratios, hold outbreaks by markov chain monte carlo methods. Appl. Statist., (49):517–542, 2000.

[Oll09] Yann Ollivier. Ricci curvature of markov chains on metric spaces. J. Funct. Anal., 256(3):810–864, 2009.

[PM+13] Conrad Patrick, Youssef Marzouk, , Natesh Pillai, and Aaron Smith. Accelerating mcmc with local quadratic models. Preprint, 2013.

[Ros95] Jeffrey Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. JASA, 90:558–566, 1995.

[RR08] Gareth O Roberts and Jeffrey S Rosenthal. Variance bounding markov chains. Annals of Applied Probability, 18(3):1201–1214, 2008.

[RS02] G. Roberts and O. Stramer. Langevin diffusions and metropolis-hastings algorithms. Methodology and Computing in Applied Probability, 4:337–357, 2002.

[RT96] Gareth O. Roberts and Richard L. Tweedie. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. Biometrika, 83(1):95–110, 1996.

[Vil08] Cedric Villani. Optimal Transport: Old and New. Springer, 2008.

[WT11] M. Welling and Y. Teh. Bayesian learning via stochastic gradient langevin dynamics. Proceedings of the 28th International Conference on Machine Learning (ICML), pages 681–688, 2011.