Abstract
Applying standard Markov chain Monte Carlo (MCMC) algorithms to large data sets is computationally expensive. Both the calculation of the acceptance probability and the creation of informed proposals usually require an iteration through the whole data set. The recently proposed stochastic gradient Langevin dynamics (SGLD) method circumvents this problem by generating proposals which are only based on a subset of the data, by skipping the accept-reject step and by using decreasing step-sizes sequence \( \delta_m \geq 0 \).

We provide in this article a rigorous mathematical framework for analysing this algorithm. We prove that, under verifiable assumptions, the algorithm is consistent, satisfies a central limit theorem (CLT) and its asymptotic bias-variance decomposition can be characterized by an explicit functional of the step-sizes sequence \( (\delta_m)_{m \geq 0} \). We leverage this analysis to give practical recommendations for the notoriously difficult tuning of this algorithm: it is asymptotically optimal to use a step-size sequence of the type \( \delta_m \approx m^{-1/3} \), leading to an algorithm whose mean squared error (MSE) decreases at rate \( O(m^{-1/3}) \).

Keywords: Markov Chain Monte Carlo, Langevin Dynamics, Big Data
1. Introduction

We are entering the age of Big Data, where significant advances across a range of scientific, engineering and societal pursuits hinge upon the gain in understanding derived from the analyses of large scale data sets. Examples include recent advances in genome-wide association studies (Hirschhorn and Daly, 2005; McCarthy et al., 2008; Wang et al., 2005), speech recognition (Hinton et al., 2012), object recognition (Krizhevsky et al., 2012), and self-driving cars (Thrun, 2010). As the quantity of data available has been outpacing the computational resources available in recent years, there is an increasing demand for new scalable learning methods, for example methods based on stochastic optimization (Robbins and Monro, 1951b; Srebro and Tewari, 2010; Sato, 2001; Hoffman et al., 2010), distributed computational architectures (Ahmed et al., 2012; Neiswanger et al., 2013; Minsker et al., 2014), greedy optimization (Harchaoui and Jaggi, 2014), as well as the development of specialized computing systems supporting large scale machine learning applications (Gonzalez, 2014).

Recently, there has also been increasing interest in methods for Bayesian inference scalable to Big Data settings. Rather than attempting a single point estimate of parameters typical in optimization-based or maximum likelihood settings, Bayesian methods attempt to obtain characterizations of the full posterior distribution over the unknown parameters and latent variables in the model, hence providing better characterizations of the uncertainties inherent in the learning process, as well as providing protection against overfitting. Scalable Bayesian methods proposed in the recent literature include stochastic variational inference (Sato, 2001; Hoffman et al., 2010), which applies stochastic approximation techniques to optimizing a variational approximation to the posterior, parallelized Monte Carlo (Neiswanger et al., 2013; Minsker et al., 2014), which distributes the computations needed for Monte Carlo sampling across a large compute cluster, as well as subsampling-based Monte Carlo (Welling and Teh, 2011; Ahn et al., 2012; Korattikara et al., 2014), which attempt to reduce the computational complexity of Markov chain Monte Carlo (MCMC) methods by applying updates to small subsets of data.

In this paper we study the asymptotic properties of the stochastic gradient Langevin dynamics (SGLD) algorithm first proposed by Welling and Teh (2011). SGLD is a subsampling-based MCMC algorithm based on combining ideas from stochastic optimization, specifically...
using small subsets of data to estimate gradients, with Langevin dynamics, a MCMC method making use of gradient information to produce better parameter updates. [Welling and Teh (2011)] demonstrated that SGLD works well on a variety of models and this has since been extended by Ahn et al. (2012, 2014) and Patterson and Teh (2013b).

The stochastic gradients in SGLD introduce approximations into the Markov chain, whose effect has to be controlled by using a slowly decreasing sequence of step sizes. [Welling and Teh (2011)] provided an intuitive argument that as the step-size decreases the variations introduced by the stochastic gradients gets dominated by the natural stochasticity of Langevin dynamics, the result being that the stochastic gradient approximation should wash out asymptotically and that the Markov chain should converge to the true posterior distribution.

In this paper, we make this intuitive argument more precise by providing conditions under which SGLD converges to the targeted posterior distribution; we describe a number of characterizations of this convergence. Specifically, we show that estimators derived from SGLD are consistent (Theorem 6) and satisfy a central limit theorem (CLT) (Theorem 7); the bias-variance trade-off of the algorithm is discussed in details in Section 5. In Section 6 we prove that, when observed on the right (inhomogeneous) time scale, the sample path of the algorithm converges to a Langevin diffusion (Theorem 8).

Our analysis reveals that for a sequence of step-sizes with algebraic decay \( \delta_m \propto m^{-\alpha} \) the optimal choice, when measured in terms of rate of decay of the mean squared error (MSE), is given for \( \alpha_* = 1/3 \); The choice \( \delta_m \propto m^{-\alpha_*} \) leads to an algorithm that converges at rate \( O(m^{-1/3}) \). This rate of convergence is worse than the standard Monte-Carlo \( m^{-1/2} \)-rate of convergence. This is not due to the stochastic gradients used in SGLD, but rather to the decreasing step-sizes.

These results are asymptotic in the sense that they characterise the behaviour of the algorithm as the number of steps approaches infinity. Therefore they do not necessarily translate into any insight into the behaviour for finite computational budgets which is the regime in which the SGLD might provide computational gains over alternatives. The mathematical framework described in this article show that the SGLD is a sound algorithm, an important result that has been missing in the literature.

In the remainder of this article, the notation \( \mathcal{N}(\mu, \sigma^2) \) denotes a Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \). For two positive function \( f, g : \mathbb{R} \to [0, \infty) \), one writes \( f \lesssim g \) too indicate that there exists a positive constant \( C > 0 \) such that \( f(\theta) \leq C g(\theta) \); we write \( f \asymp g \) if \( f \lesssim g \lesssim f \). For a probability measure \( \pi \) on a measured space \( \mathcal{X} \), a measurable function \( \varphi : \mathcal{X} \to \mathbb{R} \) and a measurable set \( A \subset \mathcal{X} \), we define \( \pi(\varphi; A) = \int_{\theta \in A} \varphi(\theta) \pi(d\theta) \) and \( \pi(\varphi) = \pi(\varphi; \mathcal{X}) \).

2. Stochastic Gradient Langevin Dynamics

Many MCMC algorithms evolving in a continuous state space, say \( \mathbb{R}^d \), can be realised as discretizations of a continuous time Markov process \( (\theta_t)_{t \geq 0} \). An example of such a continuous time process, which is central to SGLD as well as many other algorithms, is
Langevin diffusion, which is given by the following stochastic differential equation,
\[ d\theta_t = \frac{1}{2} \nabla \log \pi(\theta_t) \, dt + dW_t, \tag{1} \]
where \( \pi : \mathbb{R}^d \to (0, \infty) \) is a probability density of interest, and \((W_t)_{t \geq 0}\) is a standard Brownian motion in \( \mathbb{R}^d \). The motivation behind the choice of Langevin diffusion is that, under mild regularity assumptions \([\text{Roberts and Tweedie}, 1996; \text{Stramer and Tweedie}, 1999a,b; \text{Mattingly et al.}, 2002]\), it is ergodic with respect to \( \pi \); specifically, for a bounded and measurable test function \( \varphi : \mathbb{R}^d \to \mathbb{R} \), the following limit hold almost surely,
\[ \lim_{T \to \infty} \frac{1}{T} \int_0^T \varphi(\theta_t) \, dt = \int_{\mathbb{R}^d} \varphi(\theta) \pi(d\theta). \tag{2} \]
In words, the random time average (LHS) converges to the space average (RHS) almost surely.

Given a time-step \( \delta > 0 \) and a current position \( \theta_t \), it is often straightforward to simulate a random variable \( \theta_* \) that is approximately distributed as the law of \( \theta_t + \delta \) given \( \theta_t \). For stochastic differential equations, the Euler-Maruyama scheme \([\text{Maruyama}, 1955]\) might be the simplest approach for approximating the law of \( \theta_t + \delta \). For Langevin diffusion this reads
\[ \theta_* = \theta_t + \frac{1}{2} \delta \nabla \log \pi(\theta_t) + \delta^{1/2} \eta, \quad \eta \overset{\mathcal{D}}{\sim} \mathcal{N}(0, I_d). \tag{3} \]
The smaller the time-step the better the approximation. To fully correct for the discretisation approximation, we can use it simply as a Metropolis-Hastings proposal which can be accepted or rejected. The resulting algorithm is usually referred to as the Metropolis-Adjusted-Langevin algorithm (MALA) \([\text{Roberts and Tweedie}, 1996]\). Other discretizations can be used as proposals. For example, the random walk Metropolis-Hastings algorithm uses the discretization of a standard Brownian motion as the proposal, while Hamiltonian Monte Carlo (HMC) algorithm \([\text{Duane et al.}, 1987]\) is based on discretizations of an Hamiltonian system of differential equations. See the excellent review of \([\text{Neal}, 2010]\) for further information.

In this paper, we shall consider the situation where the target \( \pi \) is the density of the posterior distribution under a Bayesian model where there are \( N \gg 1 \) i.i.d. observations, the so called Big Data regime,
\[ \pi(\theta) \propto p_0(\theta) \prod_{i=1}^N p(y_i \mid \theta). \tag{4} \]
Here, both computing the gradient term \( \nabla \log \pi(\theta_t) \) and evaluating the Metropolis-Hastings acceptance ratio require a computational budget that scales unfeasibly as \( \mathcal{O}(N) \). One approach is to use a standard random walk proposal instead of Langevin dynamics, and to efficiently approximating the Metropolis-Hastings accept-reject mechanism using only a subset of the data \([\text{Korattikara et al.}, 2014; \text{Bardenet et al.}, 2014]\).

This paper is concerned with stochastic gradient Langevin dynamics (SGLD), an alternative approach proposed by \([\text{Welling and Teh}, 2011]\). This follows the opposite route and
chooses to completely avoid the computation of the Metropolis-Hastings ratio. By choosing a discretization of the Langevin diffusion with a sufficiently small step-size $\delta \ll 1$, because the Langevin diffusion is ergodic with respect to $\pi$, the hope is that even if the Metropolis-Hastings accept-reject mechanism is completely avoided, the resulting Markov chain still has an invariant distribution that is close to $\pi$. Choosing a decreasing sequence of step-sizes $\delta_m \to 0$ should even allow us to converge to the posterior distribution. To further make this approach viable in large $N$ settings, the gradient term $\nabla \log p(\theta)$ can be further approximated using a subsampling strategy. For an integer $1 \leq n \leq N$ and a random subset $\tau \overset{\text{def}}{=} (\tau_1, \ldots, \tau_n)$ of $[N] \equiv \{1, \ldots, N\}$ generated by sampling with or without replacement from $[N]$, the quantity

$$\nabla \log p_0(\theta) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{\tau_i} | \theta)$$

is an unbiased estimator of $\nabla \log \pi(\theta)$. Most importantly, this stochastic estimate can be computed with a computational budget that scales as $O(n)$ with $n$ potentially much smaller than $N$. Indeed, the larger the quotient $n/N$, the smaller the variance of this estimate.

Stochastic gradient methods have a long history in optimisation and machine learning and are especially relevant in the large dataset regime considered in this article (Robbins and Monro, 1951a; Bottou, 2010; Hoffman et al., 2013). In this paper we will adopt a slightly more general framework and assume that one can compute an unbiased estimate $\hat{\nabla \log \pi}(\theta, \mathcal{U})$ to the gradient $\nabla \log \pi(\theta)$, where $\mathcal{U}$ is an auxiliary random variable which contains all the randomness involved in constructing the estimate. Without loss of generality we may assume (although this is unnecessary) that $\mathcal{U}$ is uniform on $(0, 1)$. The unbiasedness of the estimator $\hat{\nabla \log \pi}(\theta, \mathcal{U})$ means that

$$\mathbb{E}[H(\theta, \mathcal{U})] = 0 \quad \text{with} \quad H(\theta, \mathcal{U}) \overset{\text{def}}{=} \hat{\nabla \log \pi}(\theta, \mathcal{U}) - \nabla \log \pi(\theta).$$

(5)

In summary, the SGLD algorithm can be described as follows. For a sequence of asymptotically vanishing time-steps $(\delta_m)_{m \geq 0}$ and an initial parameter $\theta_0 \in \mathbb{R}^d$, if the current position is $\theta_{m-1}$, the next position $\theta_m$ is defined though the recursion

$$\theta_m = \theta_{m-1} + \frac{1}{2} \delta_m \hat{\nabla \log \pi}(\theta_{m-1}, \mathcal{U}_m) + \delta_m^{1/2} \eta_m$$

(6)

for an i.i.d. sequence $\eta_m \sim \mathcal{N}(0, I_d)$, and an independent and i.i.d. sequence $\mathcal{U}_m$ of auxiliary random variables. This is the equivalent of the Euler-Maruyama discretization of the Langevin diffusion with a decreasing sequence of step-sizes and a stochastic estimate to the gradient term. Note that the process $(\theta_m)_{m \geq 0}$ is a non-homogeneous Markov chain, and many standard analysis techniques for homogeneous Markov chains do not apply.

For a test function $\varphi: \mathbb{R}^d \to \mathbb{R}$, the expectation of $\varphi$ with respect to the posterior distribution $\pi$ can be approximated by the weighted sum

$$\pi_m(\varphi) \overset{\text{def}}{=} \left\{ \delta_1 \varphi(\theta_0) + \ldots + \delta_m \varphi(\theta_{m-1}) \right\} / T(m), \quad T(m) \overset{\text{def}}{=} \delta_1 + \ldots + \delta_m.$$

(7)

The weighted sum in Equation (7) is the discrete analogue of the ergodic integral in Equation (2). During the course of the proof of our fluctuation Theorem we will need to consider
more general averaging schemes than the one above. Instead, for a general positive sequence of weights $\omega = (\omega_m)_{m \geq 1}$, we define the $\omega$-weighted sum
\[
\pi_{m}^{\omega}(\varphi) \overset{\text{def}}{=} \left\{ \omega_1 \varphi(\theta_0) + \ldots + \omega_m \varphi(\theta_{m-1}) \right\} / \Omega(m), \quad \Omega(m) \overset{\text{def}}{=} \omega_1 + \ldots + \omega_m.
\]
Indeed, $\pi_{m}^{\omega}(\varphi) = \pi_{m}(\varphi)$ in the particular case $(\omega_m)_{m \geq 1} = (\delta_m)_{m \geq 1}$, while we will consider the weight sequence $\omega = \{\delta_m^2\}_{m \geq 1}$ in the proof of Theorem 7.

Let us mention several directions that can be explored to improve upon the basic SGLD algorithm explored in this paper. Langevin diffusions of the type $d\theta_t = \text{drift}(\theta_t) \, dt + M(\theta_t) \, dW_t$, reversible with respect to the posterior distribution $\pi$, can be constructed for various choices of positive definite volatility matrix function $M : \mathbb{R}^d \to \mathbb{R}^{d^2}$. Note nonetheless that, for a non-constant volatility matrix function $\theta \mapsto M(\theta)$, the drift term typically involves derivatives of $M$. Concepts of information geometry (Amari and Nagaoka, 2007) give principled ways (Livingstone and Girolami, 2014) of choosing the volatility matrix function $M$; when the Fisher information matrix is used, this leads to the Riemannian manifold MALA algorithm (Girolami and Calderhead, 2011). This approach has recently been applied to the Latent Dirichlet Allocation model for topic modelling (Patterson and Teh, 2013a). For high-dimensional state spaces $d \gg 1$, one can use a constant volatility function $M$, also known in this case as the preconditioning matrix, for taking into account the information contained in the prior distribution $p_0$ in the hope of obtaining better mixing properties (Beskos et al, 2008; Cotter et al, 2013); infinite dimensional limits are obtained in (Pillai et al, 2012; Hairer et al, 2014). Under a uniform-ellipticity condition and a growth assumption on the volatility matrix function $M : \mathbb{R}^d \to \mathbb{R}^{d^2}$, we believe that our framework could, at the cost of increasing complexity in the proofs, be extended to this setting. To avoid the slow random walk behaviour of Markov chains based on discretization of reversible diffusion processes, one can use instead discretizations of an Hamiltonian system of ordinary differential equations (Duane et al, 1987; Neal, 2010); when coupled with the stochastic estimates to the gradient above described, this leads to the stochastic gradient Hamiltonian Monte Carlo algorithm of (Chen et al, 2014).

In the rest of this paper, we will build a rigorous framework for understanding the properties of this SGLD algorithm, demonstrating that the heuristics and numerical evidences presented in Welling and Teh (2011) were indeed correct.

3. Assumptions and Stability Analysis

This section starts with the basics assumptions we will need for the asymptotic results to follow, and illustrates some of the potential stability issues that may occur, would the SGLD algorithm be applied without care.

3.1 Basic Assumptions

Assumption 1 The step-sizes $\delta = (\delta_m)_{m \geq 1}$ form a decreasing sequence with
\[
\lim_{m \to \infty} \delta_m = 0 \quad \text{and} \quad \lim_{m \to \infty} T(m) = \infty.
\]
The condition $\delta_m \to 0$ ensures that the quality of the Euler-Maruyama discretization improves as $m \to \infty$. The condition $T(m) \to \infty$ ensures that in the limit $m \to \infty$ one recovers an ergodic behaviour of the type given in (2).

For the Law of Large Numbers of Section 4, we require the following basic assumption on the weights $\omega$, which controls the variations of $\omega$.

**Assumption 2** The step-sizes sequence $(\omega_m)_{m \geq 1}$ is such that $\omega_m \to 0$ and $\Omega(m) \to \infty$ and

$$
\lim_{m \to \infty} \sum_{m \geq 1} |\Delta(\omega_m/\delta_m)| / \Omega(m) < \infty \quad \text{and} \quad \sum_{m \geq 1} \omega_m^2 / [\delta_m \Omega^2(m)] < \infty.
$$

where $\Delta(\omega_m/\delta_m) \overset{def}{=} \omega_{m+1}/\delta_{m+1} - \omega_m/\delta_m$.

**Remark 3** Assumption 2 holds if $\delta = (\delta_m)_{m \geq 1}$ satisfies Assumption 1 and the weights are defined as $\omega_m = \delta_m^p$, for some some exponent $p \geq 1$ small enough for $\Omega(m) \to \infty$. This is because the first sum is less than $\sum_{m \geq 1} |\Delta(\omega_m/\delta_m)|/\Omega(1) = \delta_{1}^{p-1}/\Omega(1)$, while the finiteness of the second sum can be seen as follows:

$$
\sum_{m \geq 1} \omega_m^2 / [\delta_m \Omega^2(m)] \lesssim 1 + \sum_{m \geq 2} |\omega_m/\delta_m|^2 [1/\Omega(m-1) - 1/\Omega(m)]
$$

$$
\lesssim 1 + \sum_{m \geq 2} [1/\Omega(m-1) - 1/\Omega(m)] = 1 + 1/\Omega(1).
$$

For any exponents $0 < \alpha < 1$ and $0 < p < 1/\alpha$ the sequences $\delta_m = (m_0 + m)^{-\alpha}$ and $\omega_m = \delta_m^p$ satisfy both Assumption 2 and Assumption 3.

### 3.2 Stability

Under mild assumptions on the posterior density $\pi$, the Langevin diffusion (1) is non-explosive and for any starting position $\theta_0 \in \mathbb{R}^d$ the total-variation distance $d_{TV}(\mathbb{P}(\theta_t \in \cdot), \pi)$ converges to zero as $t \to \infty$. In particular, Theorem 2.1 of (Roberts and Tweedie, 1996) shows that it is sufficient to assume that the drift term satisfies a condition of the type

$$
\frac{1}{2} \nabla \log \pi(\theta, \theta) \leq \alpha \|\theta\|^2 + \beta \quad \text{for some constants } \alpha, \beta > 0.
$$

We refer the interested reader to (Roberts and Tweedie, 1996; Stramer and Tweedie, 1999a,b; Roberts and Stramer, 2002; Mattingly et al., 2002) for a detailed study of the convergence properties of the Langevin diffusion (1).

Unfortunately, stability of the continuous time Langevin diffusion does not always translate into good behaviour for its Euler-Maruyama discretization. For example, even if the drift term points towards the right direction in the sense that $\langle \nabla \log \pi(\theta), \theta \rangle < 0$ for every parameter $\theta$, it might happen that the magnitude of the drift term is too large, so that the Euler-Maruyama discretization overshoots and becomes unstable. In a one dimensional setting, this would lead to a Markov chain that diverges in the sense that the sequence $(\theta_m)_{m \geq 0}$ alternates between taking arbitrarily large positive and negative values. Lemma 6.3 of (Mattingly et al., 2002) gives such an example with a target density $\pi(\theta) \propto \exp(-\theta^4)$. See also Theorem 3.2 of (Roberts and Tweedie, 1996) for examples of the same flavours.

To guarantee stability of the Euler-Maruyama discretization requires stronger Lyapunov type conditions. At a heuristic level, one must ensure that the drift term $\nabla \log \pi(\theta)$
points towards the centre of the state space. In addition, the previous discussion indicates that one must also ensure that the magnitude of this drift term is not too large. The following assumptions satisfy both heuristics, and we will show are enough to guarantee that the SGLD algorithm is consistent, with asymptotically Gaussian fluctuations.

**Assumption 4** The drift term \( \theta \mapsto \frac{1}{2} \nabla \log \pi(\theta) \) is continuous. There exists a Lyapunov function \( V : \mathbb{R}^d \to [1, \infty) \) that tends to infinity as \( \|\theta\| \to \infty \), that is twice differentiable with bounded second derivatives, and that satisfies the following conditions.

1. There exists an exponent \( p_H \geq 2 \) such that
   \[
   \mathbb{E}[\|H(\theta, U)\|^{2p_H}] \lesssim V^{p_H}(\theta). \tag{9}
   \]
   This implies that \( \mathbb{E}[\|H(\theta, U)\|^{2p}] \lesssim V^p(\theta) \) for any exponent \( 0 \leq p \leq p_H \).

2. For every \( \theta \in \mathbb{R}^d \) we have
   \[
   \|\nabla V(\theta)\|^2 + \|\nabla \log \pi(\theta)\|^2 \lesssim V(\theta). \tag{10}
   \]

3. There are constants \( \alpha, \beta > 0 \) such that for every \( \theta \in \mathbb{R}^d \) we have
   \[
   \langle \nabla V(\theta), \frac{1}{2} \nabla \log \pi(\theta) \rangle \leq -\alpha V(\theta) + \beta. \tag{11}
   \]

Equation (11) ensures that on average the drift term \( \frac{1}{2} \nabla \log \pi(\theta) \) points towards the centre of the state space, while equations (9) and (10) provide control on the magnitude of the (stochastic) drift term. The drift condition (11) implies in particular that the Langevin diffusion (1) converges exponentially quickly towards the equilibrium distribution \( \pi \) (Mattingly et al., 2002; Roberts and Tweedie, 1996).

For a posterior density \( \pi \) of the form (4), to establish that the Assumptions (4) hold, one can often verify the following stronger conditions: the prior density \( p_0 \) satisfies \( \|\nabla \log p_0(\theta)\|^2 \lesssim V(\theta) \) and for all indices \( 1 \leq i \leq N \) the likelihood term \( p(y_i | \theta) \) is such that
\[
\|\nabla \log p(y_i | \theta)\|^{2p_H} \lesssim V^{p_H}(\theta). \tag{12}
\]
This is because almost surely \( \|H(\theta, U)\|^{2p_H} \) is less than a constant multiple of \( \sum_{i=1}^{N} \|\nabla \log p(y_i | \theta)\|^{2p_H} \). Several such examples are described in Section 7.

The proof of the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) both exploit the following Stability Lemma:

**Lemma 5** Let the step-sizes \( (\delta_m)_{m \geq 1} \) satisfy Assumption [4] and suppose that the stability Assumptions [3] hold. For any exponent \( 0 \leq p \leq p_H \) the following bounds hold almost surely:
\[
\sup_{m \geq 1} \pi_m(V^{p/2}) < \infty \quad \text{and} \quad \sup_{m \geq 1} \mathbb{E}[V^p(\theta_m)] < \infty. \tag{12}
\]
Moreover, for any exponent \( p \geq 0 \) we have \( \pi(V^p) < \infty \). If the sequence of weights \( (\omega_m)_{m \geq 1} \) satisfies Assumption [2] the following holds almost surely,
\[
\sup_{m \geq 1} \pi_\omega^m(V^{p/2}) < \infty \tag{13}
\]
The technical proof can be found in Section A. The idea is to leverage condition (11) in order to establish that the function \( V^p \) satisfies both discrete and continuous drift conditions.
4. Consistency

The problem of estimating the invariant distribution of a stochastic differential equation by using a diminishing step-size Euler discretization has been well explored in the literature (Lamberton and Pages, 2002, 2003; Lemaire, 2007; Panloup, 2008; Pages and Panloup, 2012), while (Mattingly et al., 2002) studied the bias and variance of similar algorithms when fixed step-sizes are used instead. We leverage some of these techniques and adapt it to our setting where the drift term can only be unbiasedly estimated, and establish in this section that the SGLD algorithm is consistent under Assumptions 1 and 4. More precisely, we prove that almost surely the sequence \((\pi_m)_{m \geq 1}\) defined in Equation (7) converges weakly towards \(\pi\). Specifically, under growth assumptions on a test function \(\phi : \mathbb{R}^d \rightarrow \mathbb{R}\), the following strong law of large numbers holds almost surely,

\[
\lim_{m \to \infty} \left( \delta_1 \phi(\theta_0) + \ldots + \delta_m \phi(\theta_m) \right) = \int_{\mathbb{R}^d} \phi(\theta) \pi(d\theta),
\]

with a similar result for \(\omega\)-weighted empirical averages, under assumptions on the weight sequence \(\omega\).

**Theorem 6 (Consistency)** Let the step-sizes satisfy Assumption (1) and suppose that the stability Assumptions (4) hold for a Lyapunov function \(V : \mathbb{R}^d \rightarrow [1, \infty)\). Let \(0 \leq p < p_H/2\) and \(\phi : \mathbb{R}^d \rightarrow \mathbb{R}\) be a test function such that \(|\phi(\theta)|/V^p(\theta)\) is globally bounded. Then the following limit holds almost surely:

\[
\lim_{m \to \infty} \pi_m(\phi) = \pi(\phi).
\]

If in addition the sequence of weights \(\{\omega_m\}_{m \geq 1}\) satisfies Assumption (2), a similar result holds almost surely for the \(\omega\)-weighted ergodic average:

\[
\lim_{m \to \infty} \pi_{\omega_m}(\phi) = \pi(\phi).
\]

**Proof** We will give a detailed proof of Equation (14) and then briefly describe how the more general Equation (15) can be proven using similar arguments. To prove Equation (14), we first show that the sequence \((\pi_m)_{m \geq 1}\) almost surely converges weakly to \(\pi\). Equation (14) is then proved in a second stage.

In the following, we write \(\mathbb{E}_k[\cdot]\) and \(\mathbb{P}_k(\cdot)\) to denote the conditional expectation \(\mathbb{E}[\cdot | \theta_k]\) and conditional probability \(\mathbb{P}(\cdot | \theta_k)\) respectively. We use the notation \(\Delta \theta_k \defeq (\theta_{k+1} - \theta_k)\). The operator \(\mathcal{A}\) denotes the generator of the Langevin diffusion (1): for a sufficiently regular test function \(\phi : \mathbb{R}^d \rightarrow \mathbb{R}\),

\[
\mathcal{A}\phi(\theta) = \frac{1}{2} \langle \nabla \log \pi(\theta), \nabla \phi(\theta) \rangle + \frac{1}{2} \Delta \phi(\theta),
\]

where \(\Delta \phi \defeq \sum_{i=1}^d \nabla_i^2 \phi\) denotes the standard Laplacian operator. Finally, for notational convenience, we only present the proof in the scalar case \(d = 1\), the multidimensional case being entirely similar.
At several stages in the proof, we need to prove that a quantity of the type
\[
\frac{\sum_{k=1}^{m} \Delta M_k + X_k}{T(m)} \tag{17}
\]
converges almost surely to zero as \( m \to \infty \), for a martingale difference \( \Delta M_k \stackrel{\text{def}}{=} M_{k+1} - M_k \) and a random sequence \((X_k)_{k \geq 1}\). Recall Kornak’s Lemma (Shiryaev 1996, Lemma IV.3.2) that states that for a non-decreasing and positive sequence \( \sum_{k=1}^{\infty} \).

Consequently, for proving that the quantity (17) converges almost surely to zero as \( m \to \infty \), it suffices to show that the sum \( \sum_{k=0}^{\infty} \Delta M_k / T(k) \) and \( \sum_{k=0}^{\infty} X_k / T(k) \) are almost surely finite. This follows once we have proved that \( \sum_{k=1}^{\infty} \mathbb{E}[|\Delta M_k|^2 / T^2(k)] \) is finite (\( L^2 \) martingale convergence Theorem) and \( \sum_{k=1}^{\infty} \mathbb{E}[|X_k| / T(k)] < \infty \).

**Weak convergence of \((\pi_m)_{m \geq 1}\).** To prove that almost surely the sequence \((\pi_m)_{m \geq 1}\) converges weakly towards \( \pi \), it suffices to prove that the sequence is almost surely weakly pre-compact and that any weakly convergent subsequence of \((\pi_m)_{m \geq 0}\) necessarily (weakly) converges towards \( \pi \). By Prokhorov’s Theorem (Billingsley 1995) and Equation (12), because the Lyapunov function \( V \) goes to infinity as \( \|	heta\| \to \infty \), the sequence \((\pi_m)_{m \geq 1}\) is almost surely weakly pre-compact. It thus remains to show that if a subsequence converges weakly to a probability measure \( \pi_\infty \), then \( \pi_\infty = \pi \).

Since the Langevin diffusion \((1)\) has a unique strong solution and its generator \( \mathcal{A} \) is uniformly elliptic, Theorem 9.17 of Chapter 4 of Ethier and Kurtz 1986 yields that it suffices to verify that for any smooth and compactly supported test function \( \varphi : \mathbb{R} \to \mathbb{R} \) and any limiting distribution \( \pi_\infty \) of the sequence \((\pi_m)_{m \geq 1}\) the following holds,
\[
\pi_\infty(\mathcal{A}\varphi) = 0. \tag{18}
\]
To prove Equation (18), we use the following decomposition of \( \pi_m(\mathcal{A}\varphi) \),
\[
\frac{\sum_{k=1}^{m} \mathbb{E}_{k-1}[\varphi(\theta_k) - \varphi(\theta_{k-1})]}{T(m)} - \left\{ \frac{\sum_{k=1}^{m} \mathbb{E}_{k-1}[\varphi(\theta_k) - \varphi(\theta_{k-1})]}{T(m)} - \pi_m(\mathcal{A}\varphi) \right\}. \tag{19}
\]
The first term in the above equation will be proved to converge to zero by a martingale argument, while the second term converges to zero because \( \mathcal{A}\varphi(\theta_{k-1}) \delta_k - [\varphi(\theta_k) - \varphi(\theta_{k-1})] \) is small in expectation by definition of the generator \( \mathcal{A} \), with a control on the second moment.

The numerator of the first term in (19) is equal to the sum of \( \sum_{k=1}^{m} \mathbb{E}_{k-1}[\varphi(\theta_k)] - \varphi(\theta_k) \) and \( \varphi(\theta_m) - \varphi(\theta_0) \). By boundedness of \( \varphi \), the term \( [\varphi(\theta_m) - \varphi(\theta_0)] / T(m) \) converges almost surely to zero. The discussion after Equation (17) shows that it suffices to prove that the martingale
\[
M_m = \sum_{k=1}^{m} \mathbb{E}_{k-1}[\varphi(\theta_k)] - \varphi(\theta_k) / T(k)
\]
...
is bounded in $L^2$, i.e. $\sum_{k \geq 1} E(|M_{k+1} - M_k|^2) < \infty$. Because $\varphi$ is Lipschitz, it suffices to prove that $\sum_{k \geq 1} E(\|\theta_{k+1} - \theta_k\|^2) / T^2(k)$ is finite. The stability Assumption 4 and Lemma 5 imply that the supremum $\sup_{k} E[V(\theta_{m})]$ is finite. Since $E_k[\|\delta_{k+1} - \theta_k\|^2] \lesssim \delta_{k+1}^2 V(\theta) + \delta_{k+1}$, it follows that $E(\|\theta_{k+1} - \theta_k\|^2)$ is less than a constant multiple of $\delta_{k+1}$. Under Assumption 4, because the telescoping sum $\sum_{k \geq 1} T^{-1}(k) - T^{-1}(k+1)$ is finite, the sum $\sum_{k \geq 1} \delta_k / T^2(k)$ is finite. This concludes the proof that the first term in (19) converges almost surely to zero.

The second term of (19) equals $(R_0 + \ldots + R_{m-1}) / T(m)$ with

$$R_k \overset{\text{def}}{=} E_k[\varphi(\theta_{k+1}) - \varphi(\theta_k)] - A\varphi(\theta_k) \delta_{k+1}. \quad (20)$$

We now show that there exists a constant $C$ such that the bound $|R_k| \leq C \delta_{k+1}^{3/2}$ holds almost surely for any $k \geq 0$. To do so, let $K > 0$ be such that the support of the test function $\varphi$ is included in $\Omega = [-K, K]$. We examine two cases separately.

- If $|\theta_k| > K + 1$ then $\varphi(\theta_k) = A\varphi(\theta_k) = 0$ so that $|R_k| \leq \|\varphi\|_{\infty} \times P_k(\theta_{k+1} \in \Omega)$. Since $\theta_{k+1} - \theta_k = [1/2 \nabla \log \pi(\theta_k) + H(\theta_k, U)] \delta_{k+1} + \sqrt{\delta_{k+1}} \eta$ we have

$$P_k(\theta_{k+1} \in \Omega) \leq I\left(\frac{1}{2} \nabla \log \pi(\theta_k) \geq \frac{\text{dist}(\theta_k, \Omega)}{3 \delta_{k+1}} \right) \nabla \log \pi(\theta_k) \geq \frac{\text{dist}(\theta_k, \Omega)}{3 \delta_{k+1}} \right) + P_k\left(\|H(\theta_k, U)\| \geq \frac{\text{dist}(\theta_k, \Omega)}{3 \delta_{k+1}} \right) + P_k(\|\eta\| \geq \frac{\text{dist}(\theta_k, \Omega)}{3 \sqrt{\delta_{k+1}}}).$$

We have used the notation $I(A)$ for denoting the indicator function of the event $A$. Under Assumption 4 we have $|\nabla \log \pi(\theta)| \lesssim V(\theta)^{1/2} \lesssim 1 + \|\theta\|$ so that the quotient $|\nabla \log \pi(\theta)| / \text{dist}(\theta, \Omega)$ is bounded on $\{\theta : |\theta| > K\}$; this shows that the first term equals zero for $\delta_k$ small enough. To prove that the second term is bounded by a constant multiple of $\delta_{k+1}$, it suffices to use Markov’s inequality and the fact that $\mathbb{E}[H(\theta_k, U)] / \text{dist}(\theta, \Omega)$ is bounded on $\{\theta : |\theta| > K\}$; this is because $\mathbb{E}[H(\theta_k, U)^2]$ is less than a constant multiple of $V(\theta)$ and $V(\theta) \lesssim 1 + \|\theta\|^2$ by Assumption 4. The third term is less than a constant multiple of $\delta_{k+1}$ by Markov’s inequality and the fact that $\eta$ has a finite moment of order four.

- If $|\theta_k| \leq K + 1$, we decompose $R_k$ into two terms. A second order Taylor formula yields

$$R_k = \frac{1}{2} \delta_{k+1}^2 \varphi''(\theta_k) \left\{ [\nabla \log \pi(\theta_k)]^2 + \mathbb{E}[H^2(\theta_k, U)] \right\}$$

$$+ \frac{1}{2} \sum_{u=0}^{1} \varphi''(\theta_k + u \Delta \theta_k) (1 - u)^2 du$$

$$= R_{k,1} + R_{k,2}.$$
The middle term is small for large sequence $\pi$ for this reason we only highlight the main differences. The same argument shows that the proof of Equation (15). The approach is very similar to the proof of Equation (14) and $t$ made small for large $\pi$ prove (15), we thus concentrate on proving that

One can then upgrade this almost sure weak convergence to a Law of Large Numbers. To prove (15), we thus concentrate on proving that $\pi_m^{\omega}(A\varphi) = 0$. We use the decomposition

$\sup_{k \geq 0} \mathbb{E}[V^{3/2}(\theta_k)] < \infty$ (see Lemma 5) yield that $\mathbb{E}_k|\Delta \theta_k|^3 \leq 9 \tilde{C} (\delta_{k+1}^3 + \delta_{k+1}^{3/2}) \lesssim \delta_{k+1}^{3/2}$ with

$$\tilde{C} = 1 + \sup_{\theta: |\theta| < K+1} |\nabla \log \pi(\theta)|^3 + \mathbb{E}[|H(\theta, U)|^3].$$

Note that $\tilde{C}$ is finite by Assumption 4 and Lemma 5.

We have thus proved that there is a constant $C$ such that almost surely $|R_k| \leq C \delta_{k+1}^{3/2}$ for $k \geq 0$; it follows that the sum $(R_0 + \ldots + R_{m-1})/T(m)$ is almost surely less than a constant multiple of $(\delta_1^{3/2} + \ldots + \delta_m^{3/2})/T(m)$. Under Assumption 1 this upper bound converges to zero as $m \to \infty$, hence the conclusion. This ends the proof of the almost sure weak convergence of $\pi_m$ towards $\pi$.

**Proof of Equation (14).** By assumption we have $|\varphi(\theta)| \leq C_p V^p(\theta)$ for some constant $C_p > 0$ and exponent $p < p_H/2$. To show that $\pi_m(\varphi) \to \pi(\varphi)$ almost surely, we will use Lemma 5 and the almost sure weak convergence, which guarantees that $\pi_m(\tilde{\varphi}) \to \pi(\tilde{\varphi})$ for a continuous and bounded test function $\tilde{\varphi}$.

For any $t > 0$, the set $\Omega_t \overset{\text{def}}{=} \{\theta: V(\theta) \leq t\}$ is compact and Tietze’s extension theorem (Rudin 1986 Theorem 20.4) yields that there exists a continuous function $\tilde{\varphi}_t$ with compact support that agrees with $\varphi$ on $\Omega_t$ and such that $\|\tilde{\varphi}_t\|_\infty = \sup\{|\varphi(\theta)|: \theta \in \Omega_t\}$. We can indeed also assume that $|\tilde{\varphi}_t(\theta)| \leq C_p V^p(\theta)$. Since Lemma 5 states that $\sup_m \pi_m(V^{p_H/2})$ is almost surely finite, it follows that

$$|\pi_m(\varphi) - \pi_m(\tilde{\varphi}_t)| \leq 2 C_p \pi_m(V^p 1_{V \geq t}) \leq 2 C_p \frac{\sup_m \pi_m(V^{p_H/2})}{t^{p_H/2-p}},$$

where the last inequality follows from the fact that for any probability measure $\mu$, exponents $0 < p < q$ and scalar $t > 0$ we have $\mu(V^p 1_{V \geq t}) \leq \mu(V^q 1_{V \geq t})/t^{q-p}$. Similarly

$$|\pi(\varphi) - \pi(\tilde{\varphi}_t)| \leq 2 C_p (V^{p_H/2})/t^{p_H/2-p}.$$

By the triangle inequality, we thus have,

$$|\pi_m(\varphi) - \pi(\varphi)| \leq 2 C_p \frac{\sup_m \pi_m(V^{p_H/2})}{t^{p_H/2-p}} + |\pi_m(\tilde{\varphi}_t) - \pi(\tilde{\varphi}_t)| + 2 C_p \frac{\pi(V^{p_H/2})}{t^{p_H/2-p}}.$$

The middle term is small for large $m$ due to weak convergence, while the other two can be made small for large $t$. This concludes the proof of Equation (14).

**Proof of Equation (15).** The approach is very similar to the proof of Equation (14) and for this reason we only highlight the main differences. The same argument shows that the sequence $\pi_m^{\omega}$ is tight and it suffices to show that $\pi_m^{\omega}(A\varphi) = 0$ for any weak limit $\pi_m^{\omega}$ of the sequence $(\pi_m^{\omega})_{m \geq 0}$ for obtaining the almost sure weak convergences of $(\pi_m^{\omega})_{m \geq 0}$ towards $\pi$. One can then upgrade this almost sure weak convergence to a Law of Large Numbers. To prove (15), we thus concentrate on proving that $\pi_m^{\omega}(A\varphi) = 0$. We use the decomposition
which is ergodic, and that the following asymptotic variance exists and is finite,

\( \text{discussions on existence and uniqueness of the solution of this Poisson equation.} \)

We refer the interested reader to (Garnier, 2003; Pardoux and Veretennikov, 2001) for

\[ h \]

\[ solution \]

\[ \text{the Langevin diffusion (1), we suppose that the following Poisson equation has a unique} \]

\[ S \]

\[ S \]

\[ S \]

\[ S \]

\[ S \]

\[ S \]

\[ and prove that each converges to zero almost surely. \]

For \( S_1(m) \), because \( \mathbb{E}[(\mathbb{E}_{k-1}[\varphi(\theta_k)]-\varphi(\theta_k))^2] \leq \delta_k \), one can use Kroenecker’s Lemma, the convergence theorem for martingales bounded in \( L^2 \) and the bound \( \sum_{m \geq 0} \omega_m^2 / (\Omega(m)^2 \delta_m) \) \( < \) \( \infty \) to show that it converges almost surely to zero. For \( S_2(m) \), we can write it as

\[ S_2(m) = -\frac{\omega_1}{\delta_1} \varphi(\theta_0) + \frac{\omega_{m+1}}{\delta_{m+1}} \varphi(\theta_m) - \frac{\sum_{k=1}^{m} \varphi(\theta_k)}{\Omega(m)} \Delta(\omega_k / \delta_k). \]

Because \( \Omega(m) \rightarrow \infty \), \( (\omega_{m+1} / \delta_{m+1}) / (\Omega(m) \rightarrow 0 \) and \( \varphi \) is bounded, it follows from Assumption 2 and Kroenecker’s Lemma that \( S_2(m) \) also converges almost surely to zero. Finally, some algebra shows that \( S_3(m) = \Omega(m)^{-1} \sum_{k=1}^{m} (\omega_k / \delta_k) R_{k-1} \) with the quantity \( R_k \) defined in Equation 20. It has been proved that there is a constant \( C \) such that, almost surely, \( |R_k| \leq C \delta_k^{3/2} \) for all \( k \geq 0 \). Since \( \delta_m \rightarrow 0 \), the rescaled sum \( \Omega(m)^{-1} \sum_{k=m} \omega_k \delta_k^{1/2} \) converges to zero as \( m \rightarrow \infty \). It follows that \( S_3(m) \) converges almost surely to zero.

\[ \text{5. Fluctuations, Bias-Variance Analysis, and Central Limit Theorem} \]

Let \( \varphi : \mathbb{R}^d \rightarrow \mathbb{R} \) be a sufficiently regular test function. The previous section shows that under suitable conditions the asymptotic value of \( \pi_m(\varphi) \) is \( \pi(\varphi) \). In this section, we investigate the asymptotic variance of the typical fluctuations of \( \pi_m(\varphi) \) around its asymptotic value \( \pi(\varphi) \), leading to a Central Limit Theorem for SGLD. Before stating our main theorem, we first introduce the necessary notations, and give a brief discussion of the form of the asymptotic variance \( \sigma^2(\varphi) \) and the regimes of convergence. Recalling that \( \mathcal{A} \) is the generator of the Langevin diffusion, we suppose that the the following Poisson equation has a unique solution \( h : \mathbb{R}^d \rightarrow \mathbb{R} \),

\[ \varphi - \pi(\varphi) = \mathcal{A} h. \]

We refer the interested reader to (Garnier, 2003; Pardoux and Veretennikov, 2001) for discussions on existence and uniqueness of the solution of this Poisson equation.

Suppose that the continuous time process \( (\theta_s)_{s \geq 0} \) follows the Langevin diffusion, which is ergodic, and that the following asymptotic variance exists and is finite,

\[ \sigma^2(\varphi) \overset{\text{def}}{=} \lim_{T \rightarrow \infty} \text{Var}\left(T^{-1/2} \int_{0}^{T} [\varphi - \pi(\varphi)](\theta_s) \, ds\right). \]
Assuming sufficient regularity for exchanging expectation and summation, the asymptotic variance \( \sigma^2(\varphi) \) can also be expressed as

\[
\sigma^2(\varphi) = \lim_{T \to \infty} \frac{2}{T} \int_0^T \mathbb{E} \left[ (\varphi - \pi(\varphi))(\theta_s) \int_{t=s}^T \phi - \pi(\varphi)(\theta_t) \, dt \right] \, ds
\]

\[
= - \lim_{T \to \infty} \frac{2}{T} \int_0^T h(\theta_s) Ah(\theta_s) \, ds
\]

\[
= -2 \int_{\mathbb{R}^d} h(\theta) Ah(\theta) \pi(d\theta)
\]

\[
= \int_{\mathbb{R}^d} \|\nabla h\|^2 \pi(d\theta).
\]

The first equality uses the fact that \( t \mapsto h(\theta_t) - \int_0^t Ah(\theta_s) \, ds \) is a martingale and assumes that the Langevin diffusion reaches equilibrium sufficiently rapidly, the second equality follows from the Ergodic Theorem \[6\] and the third follows from an integration by parts.

If the asymptotic variance of \( \pi_m(\varphi) \) exists, one expects it to be the asymptotic variance \( \sigma^2(\varphi) \) of the Langevin diffusion given above. We can derive this, along with the associated rate of convergence, using the following argument which will be made precise in the subsequent Fluctuations Thereom \[7\]. Here and in the proof of the theorem we will assume \( d = 1 \) for notational convenience.

We first write the quantity of interest \( \pi_m(\varphi) - \pi(\varphi) \) explicitly as a sum over the first \( m \) steps of the SGLD algorithm,

\[
\pi_m(\varphi) - \pi(\varphi) = \varphi_m(\varphi) - \pi(\varphi) = \varphi_m(Ah) = \sum_{k=0}^{m-1} \delta_{k+1} Ah(\theta_k)/T(m)
\]

\[
= \sum_{k=0}^{m-1} \delta_{k+1} Ah(\theta_k) - [h(\theta_{k+1}) - h(\theta_k)] + \frac{h(\theta_m) - h(\theta_0)}{T(m)}.
\]

In the last equality, we have introduced the idea that \( Ah(\theta_k) \) can be approximated by the difference \( h(\theta_{k+1}) - h(\theta_k) \). The nature of the approximation can be made precise by using a Taylor series expansion of each \( h(\theta_{k+1}) \) term around \( \theta_k \), followed by a binomial expansion of the polynomials \((\theta_{k+1} - \theta_k)^n\) using the SGLD update Equation \[6\]. We will need a fourth order expansion,

\[
h(\theta_{k+1}) - h(\theta_k) = \sum_{n=1}^{4} \sum_{i=0}^{n} \frac{1}{2i!(n-i)!} \nabla h(\theta_k) \nabla \log \pi(\theta_k, \mathcal{U}_{k+1}) \eta_{k+1}^{n-i} \delta_k^{(n+1)/2}
\]

\[
\cup R_{n,i}^{(k)}
\]

\[
+ \left( \frac{1}{2} \nabla \log \pi(\theta_k, \mathcal{U}_{k+1}) \delta_k + \delta_k^{1/2} \eta_{k+1} \right)^5
\]

\[
\def R_{n}^{(k)}
\]

(23)
where $\xi_k$ lies between $\theta_k$ and $\theta_{k+1}$ in the fifth-order remainder. Plugging the above and the generator Equation (16) into Equation (23), we get,

$$\pi_m(\varphi) - \pi(\varphi) = \frac{h(\theta_m) - h(\theta_0)}{T(m)} + \frac{1}{T(m)} \sum_{k=0}^{m-1} \left( \frac{1}{2} \nabla h(\theta_k) \nabla \log \pi(\theta_k) \delta_{k+1} + \frac{1}{2} \nabla^2 h(\theta_k) \delta_{k+1}
\right.
\left. - \sum_{n=1}^{4} \sum_{i=0}^{n} C^{(k)}_{n,i} \delta_{k+1} - R^{(k)}_5 \right)$$

(24)

Under the assumptions of Theorem 7, we will show that all terms are asymptotically negligible except the following ones, which we group into a “fluctuations” term (involving $C^{(k)}_{1,0}$) and a “bias” term (involving $C^{(k)}_{2,2}$, $C^{(k)}_{3,1}$ and $C^{(k)}_{4,0}$):

**Fluctuations:**

$$- \frac{1}{T(m)} \sum_{k=0}^{m-1} \nabla h(\theta_k) \eta_{k+1} \delta_{k+1}^{1/2}$$

**Bias:**

$$- \frac{1}{T(m)} \sum_{k=0}^{m-1} \left( \frac{1}{8} \nabla^2 h(\theta_k) \nabla \log \pi(\theta_k, U_{k+1})^2 \right.
\left. + \frac{1}{4} \nabla^3 h(\theta_k) \nabla \log \pi(\theta_k, U_{k+1}) \eta_{k+1}^2 \right.
\left. + \frac{1}{24} \nabla^4 h(\theta_k) \eta_{k+1}^4 \right) \delta_{k+1}^2.$$  

(25)

The fluctuation term is $O(T(m)^{-1/2})$ since $(\eta_k)_{k \geq 1}$ are i.i.d. $\mathcal{N}(0,1)$, while the bias term is $O(T(m)^{-1} \sum_{k=0}^{m-1} \delta_k^2)$, so that the asymptotic behaviour of $\pi_m(\varphi) - \pi(\varphi)$ is dictated by the ratio of the typical scales of the bias and fluctuations,

$$\mathbb{B}(m) \overset{\text{def}}{=} \frac{1}{\sqrt{T(m)}} \sum_{k=0}^{m-1} \delta_{k+1}^2.$$  

(26)

If $\mathbb{B}(m) \to 0$ as $m \to \infty$, the difference $\pi_m(\varphi) - \pi(\varphi)$ is dominated by the fluctuations, which are of magnitude $O(T(m)^{-1/2})$, and we can expect that the rescaled quantity $T(m)^{1/2} [\pi_m(\varphi) - \pi(\varphi)]$ to converge in distribution to a *centred* Gaussian distribution with the asymptotic variance $\sigma^2(\varphi)$. If $\mathbb{B}(m) \to \mathbb{B}_{\infty} \in (0, \infty)$, there is an exact balance between the scales of the fluctuations and bias, and we expect the rescaled quantity $T(m)^{1/2} [\pi_m(\varphi) - \pi(\varphi)]$ to converge in distribution to a *non-centred* Gaussian distribution whose variance is the asymptotic variance $\sigma^2(\varphi)$ and mean is given by the asymptotic value of the bias term. Lastly, if $\mathbb{B}(m) \to \infty$, the bias dominates and we expect $[\pi_m(\varphi) - \pi(\varphi)]/[T(m)^{-1} \sum_{k=0}^{m} \delta_k^2]$ to converge in probability to the asymptotic bias.

For the standard choice of step-sizes $\delta_m = (m_0 + m)^{-\alpha}$ the statistical fluctuations dominate in the range $1/3 < \alpha \leq 1$, there is an exact balance between bias and fluctuations for $\alpha = 1/3$, and the bias dominates for $0 < \alpha < 1/3$. The optimal rate of convergence is obtained for $\alpha = 1/3$ and leads to an algorithm that converges at rate $m^{-1/3}$.

**Theorem 7 (Fluctuations)** Let the step-sizes $(\delta_m)_{m \geq 1}$ satisfy Assumption 4 and assume that Assumption 4 holds for an exponent $p_H \geq 5$. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be a test function and assume that the unique solution $h : \mathbb{R}^d \to \mathbb{R}^d$ to the Poisson equation $\varphi - \pi(\varphi) = Ah$ satisfies $\|\nabla^n h(\theta)\| \lesssim V^{p_H}(\theta)$ for $n \leq 4$ and bounded fifth derivative.
We claim that \( \text{Remainder} \) is almost surely finite. in probability. This is because

\[
\text{typical scale, } N \to L
\]

We claim that

\[
\text{Leftover}
\]

We show that when multiplied by \( T \) is dominated asymptotically by either the fluctuations or the bias and thus negligible. Then terms other than the fluctuation and bias terms. Specifically, we show that each of these



We will work with each of the terms in Equation (24) in turn. We start with the

\[
E\left( m \sum_{k} | \right.
\]

In case the bias dominates, i.e. \( \mathbb{B}(m) \to 0 \), the following convergence in distribution holds,

\[
\lim_{m \to \infty} T^{1/2}(m) \{ \pi_{m}(\varphi) - \pi(\varphi) \} = \mathcal{N}(0, \sigma^{2}(\varphi)), \tag{27}
\]

with the asymptotic variance \( \sigma^{2}(\varphi) \) given by Equation (21). This gives a Central Limit Theorem for SGGLD.

In case the fluctuations and the bias are on the same scale, i.e. \( \mathbb{B}(m) \to \mathbb{B}_{\infty} \in (0, \infty) \), the following convergence in distribution holds,

\[
\lim_{m \to \infty} T^{1/2}(m) \{ \pi_{m}(\varphi) - \pi(\varphi) \} = \mathcal{N}(\mu(\varphi), \sigma^{2}(\varphi)), \tag{28}
\]

with the asymptotic bias

\[
\mu(\varphi) = -\mathbb{B}_{\infty} E\left[ \frac{1}{8} \nabla^{2} D\pi(\Theta) \nabla \log \pi(\Theta, \mathcal{U}) \right] + \frac{1}{4} \nabla^{3} D\pi(\Theta) \nabla \log \pi(\Theta) + \frac{1}{24} \nabla^{4} D\pi(\Theta)
\]

where the random variables \( \Theta \sim \pi \) and \( \mathcal{U} \) are independent.

In case the bias dominates, i.e. \( \mathbb{B}(m) \to \infty \), the following limit holds in probability,

\[
\lim_{m \to \infty} \frac{\pi_{m}(\varphi) - \pi(\varphi)}{T(m)^{-1} \sum_{k=1}^{m} \delta_{k}^{2}} = \mu(\varphi). \tag{29}
\]

**Proof** We will work with each of the terms in Equation (24) in turn. We start with the terms other than the fluctuation and bias terms. Specifically, we show that each of these terms, when multiplied by either \( T(m)^{1/2} \) or \( T(m)(\sum_{k=0}^{m-1} \delta_{k+1}^{2})^{-1} \), converges to 0, that is, it is dominated asymptotically by either the fluctuations or the bias and thus negligible. Then we show that when multiplied by \( T(m)^{1/2} \), the fluctuation term converges in distribution to \( \mathcal{N}(0, \sigma^{2}(\varphi)) \). Finally, we show that the bias term converge to \( \mu(\varphi) \) when rescaled by its typical scale, \( T(m)(\sum_{k=0}^{m-1} \delta_{k+1}^{2})^{-1} \). Putting these results together under the three cases of \( \mathbb{B}(m) \to 0 \), \( \mathbb{B}(m) \to \mathbb{B}_{\infty} \in (0, \infty) \) and \( \mathbb{B}(m) \to \infty \) leads to the results of the Theorem.

**[Leftover]** We claim that

\[
\frac{h(\theta_{m}) - h(\theta_{0})}{T(m)^{1/2}} \to 0
\]

in probability. This is because \( |h(\theta)| \lesssim V^{\mu}(\theta) \) and Lemma 5 shows that \( \sup_{m \geq 0} E[V^{\mu}(\theta_{m})] \) is almost surely finite.

**[Remainder]** We claim that

\[
\sum_{k=0}^{m-1} R_{k}^{(k)} \to 0
\]

in \( L_{1} \) sense. We show this using Assumptions 1 and 4 and Lemma 5, writing,

\[
E\left[ \sum_{k=0}^{m-1} R_{k}^{(k)} \right] \lesssim \sum_{k=0}^{m-1} E\left[ |\eta_{k+1}|^{5} \right] \delta_{k+1}^{5/2} + E\left[ \left| \nabla \log \pi(\theta_{k}, \mathcal{U}_{k+1}) \right|^{5} \delta_{k+1}^{5} \right] \lesssim \sum_{k=0}^{m-1} \delta_{k+1}^{5/2} \to 0.
\]

(30)
We have exploited the fact that $\nabla^5 h$ is assumed to be globally bounded.

**[High order terms]** We claim that for each $n$ and $i$ with $n + i \geq 5$, we have,

$$\sum_{k=0}^{m-1} \frac{X^{(k)}_{n,i}}{\sum_{k=0}^{m-1} \delta_{k+1}^{2}} \rightarrow 0$$

in $L_1$ sense. The argument is essentially identical to the one above for the remainder, making use of the fact that the coefficients are uniformly bounded in expectation and that $(\sum_{k=0}^{m-1} \delta_{k+1}^{(n+i)/2})/(\sum_{k=0}^{m-1} \delta_{k+1}^{2}) \rightarrow 0$ for these indices $n$ and $i$, since $(n+i)/2 > 2$ and $\delta_k \rightarrow 0$.

**[Low order terms]** We claim that

$$\sum_{k=0}^{m-1} X^{(k)}_{n,i} \delta_{k+1}^{n+i}/2 \rightarrow 0$$

in $L_2$ sense, for

$$X^{(k)}_{1,1} = \frac{1}{2} \nabla h(\theta_k) \nabla \log \pi(\theta_k) - C_{1,1}^{(k)} = \frac{1}{2} \nabla h(\theta_k) H(\theta_k, \mathcal{U}_{k+1})$$

$$X^{(k)}_{2,0} = \frac{1}{2} \nabla^2 h(\theta_k) \eta_{k+1}^2 - C_{2,0}^{(k)} = \frac{1}{2} \nabla^2 h(\theta_k)(\eta_{k+1}^2 - 1)$$

$$X^{(k)}_{2,1} = -C_{2,1}^{(k)} = -\frac{1}{2} \nabla^2 h(\theta_k) \nabla \log \pi(\theta_k, \mathcal{U}_{k+1}) \eta_{k+1}$$

$$X^{(k)}_{3,0} = -C_{3,0}^{(k)} = -\frac{1}{6} \nabla^3 h(\theta_k) \eta_{k+1}^3.$$  

Because each $X^{(k)}_{n,i}$ has zero mean, we have

$$\mathbb{E} \left( \sum_{k=0}^{m-1} X^{(k)}_{n,i} \delta_{k+1}^{n+i}/2 \right)^2 = \sum_{k=0}^{m-1} \mathbb{E} \left[ (X^{(k)}_{n,i})^2 \right] \frac{\delta_{k+1}^{n+i}}{T(m)} \lesssim \sum_{k=0}^{m-1} \frac{\delta_{k+1}^{n+i}}{T(m)} \rightarrow 0.$$  

We made use of the fact that the expectations $\mathbb{E}[(X^{(k)}_{n,i})^2]$ are uniformly bounded for all $k \geq 0$ by the same arguments as above, and that the final expression converges to 0 since $n + i \geq 2$, $\delta_m \rightarrow 0$ and $T(m) \rightarrow \infty$.

**[Fluctuation]** We claim that

$$\frac{\sum_{k=0}^{m-1} \nabla h(\theta_k) \delta_{k+1}^{1/2} \eta_{k+1}}{T(m)^{1/2}} \rightarrow \mathcal{N}(0, \sigma^2(\varphi))$$

in distribution. Using the standard martingale central limit theorem (e.g. Theorem 3.2, Chapter 3 of [Hall and Heyde 1980]), it suffices to verify that for any $\varepsilon > 0$ the following limits hold in probability,

$$\lim_{m \rightarrow \infty} \sum_{k=0}^{m-1} \mathbb{E} \left[ Z_k^2 \mathbb{I}(Z_k^2 > T(m)\varepsilon) \right]/T(m) = 0 \quad (31)$$

$$\lim_{m \rightarrow \infty} \sum_{k=0}^{m-1} \mathbb{E} \left[ Z_k^2 \right]/T(m) = \sigma^2(\varphi) \quad (32)$$
with $Z_k \overset{\text{def}}{=} \nabla h(\theta_k) \delta_k^{1/2} \eta_{k+1}$. Since $E_k[Z_k^2] = \nabla h(\theta_k)^2 \delta_{k+1}$ and the function $\theta \mapsto \nabla h(\theta)^2$ satisfies the assumptions of Theorem 6, the limit (32) directly follows from Theorem 6. Cauchy-Schwarz’s inequality and the boundedness of $\nabla h$ imply that $E_k[Z_k^2 I(|Z_k| > T(m) \epsilon^2)]$ is less than a constant multiple of $\delta_{k+1}^2 \epsilon^2 \eta_{k+1}^2 > T(m) \epsilon^2$, and Markov’s inequality yields
\[
\sum_{k=0}^{m-1} E_k[Z_k^2 I(Z_k^2 > T(m) \epsilon)] / T(m) \leq \sum_{k=0}^{m-1} \frac{\delta_{k+1}^2}{T(m)^2 \epsilon}.
\]
Again, we have $T(m)\epsilon^{-2} \sum_{k=0}^{m-1} \delta_{k+1}^2 \rightarrow 0$, and limit (31) follows.

[Bias] We claim that
\[
\sum_{k=0}^{m-1} \left( \frac{1}{2} \nabla^2 h(\theta_k) \nabla \log \pi(\theta_k, U_{k+1})^2 + \frac{1}{4} \nabla^3 h(\theta_k) \nabla \log \pi(\theta_k, U_{k+1}) \eta_{k+1}^2 \right) \frac{\delta_{k+1}^2}{\sum_{k=0}^{m-1} \delta_{k+1}^2} \rightarrow \mu(\varphi)
\]
almost surely. This follows directly from applying Theorem 6 to the test function given by the expression inside the parentheses, and with $\omega_k = \delta_k^2$. The assumptions of the theorem are satisfied due to Assumption 4.1 and because $(\delta_m^2)_{m \geq 1}$ satisfies Assumption 2 (see Remark 3).

6. Diffusion limit

In this section we show that, when observed on the right (inhomogeneous) time scale, the sample path of the SGLD algorithm converges to the continuous time Langevin diffusion of Equation (1), confirming the heuristic discussion in Welling and Teh (2011).

The result is based on the continuity properties of the Itô’s map $I : C([0, T], \mathbb{R}^d) \rightarrow C([0, T], \mathbb{R}^d)$, which sends a continuous path $w \in C([0, T], \mathbb{R}^d)$ from $[0, T]$ to $\mathbb{R}^d$ to the unique solution $v = I(w)$ of the integral equation,
\[
v_t = \theta_0 + \frac{1}{2} \int_{s=0}^{t} \nabla \log \pi(v_s) \, ds + w_t \quad \text{for all} \quad t \in [0, T].
\]
If the drift function $\theta \mapsto \frac{1}{2} \nabla \log \pi(\theta)$ is globally Lipschitz, then the Itô’s map $I$ is well defined and continuous. Further, the image $I(W)$ under the Itô map of a standard Brownian motion $W$ on $[0, T]$ can be seen to be described by Langevin diffusion.

The approach, inspired by ideas in Mattingly et al. (2012); Pillai et al. (2012), is to construct a sequence of coupled Markov chains $(\theta^{(r)})_{r \geq 1}$, each started at the same initial state $\theta_0 \in \mathbb{R}^d$ and evolved according to the SGLD algorithm with step-sizes $\delta^{(r)} \overset{\text{def}}{=} (\delta_k^{(r)})_{k=1}^m$ such that
\[
\sum_{k=1}^{m(r)} \delta_k^{(r)} = T
\]
and with increasingly fine mesh sizes, i.e.

\[ \text{mesh}(\delta^{(r)}) \overset{\text{def}}{=} \max\{\delta_k^{(r)} : 1 \leq k \leq m(r)\} \to 0 \quad \text{as} \quad r \to \infty. \]

Define \( T_0^{(r)} = 0 \) and \( T_k^{(r)} = \delta_k^{(1)} + \cdots + \delta_k^{(r)} \) for each \( k \geq 1 \). The Markov chains are coupled to \( W \) as follows:

\[
\eta_k^{(r)} = (\delta_k^{(r)})^{-1/2}(W(T_k^{(r)}) - W(T_{k-1}^{(r)})) \\
\theta_k^{(r)} = \theta_{k-1}^{(r)} + \frac{1}{2} \delta_k^{(r)}(\nabla \log \pi(\theta_{k-1}^{(r)}) + H(\theta_{k-1}^{(r)}, U_k^{(r)})) + (\delta_k^{(r)})^{1/2} \eta_k^{(r)},
\]

for an i.i.d. collection of auxiliary random variables \((U_k^{(r)})_{r \geq 1, k \geq 1}\). Note that \((\eta_k^{(r)})_{k \geq 1}\) form an i.i.d. sequence of \( N(0, 1) \) variables for each \( r \).

We can construct piecewise affine continuous time sample paths \((S^{(r)})_{r \geq 1}\) by linearly interpolating the Markov chains:

\[
S^{(r)}(xT_{k-1}^{(r)} + (1 - x)T_k^{(r)}) = x\theta_{k-1}^{(r)} + (1 - x)\theta_k^{(r)},
\]

for \( x \in [0, 1] \). The approach then amounts to showing that each \( S^{(r)} \) can be expressed as \( \mathcal{I}(\tilde{W}^{(r)}) + e^{(r)} \), where \( \tilde{W}^{(r)} \) is a sequence of stochastic processes converging to \( W \) and \( e^{(r)} \) is asymptotically negligible, and making use of the continuity properties of the Itô map \( \mathcal{I} \).

**Theorem 8** Suppose Assumption 4 holds and that the drift function \( \theta \mapsto (1/2)\nabla \log \pi(\theta) \) is globally Lipschitz on \( \mathbb{R}^d \). If \( \text{mesh}(\delta^{(r)}) \to 0 \) as \( r \to \infty \), then the sequence of continuous time processes \((S^{(r)})_{r \geq 1}\) defined in Equation (35) converges weakly on \((\mathcal{C}([0, T], \mathbb{R}^d), \| \cdot \|_\infty)\) to the Langevin diffusion started at \( S_0 = \theta_0 \) that follows the stochastic differential equation

\[
dS_t = \frac{1}{2} \nabla \log \pi(S_t) \, dt + dW_t,
\]

for \((W_t)_{t \in [0, T]}\) a standard Brownian motion in \( \mathbb{R}^d \).

**Proof** Since the drift term \( s \mapsto (1/2)\nabla \log \pi(s) \) is globally Lipschitz on \( \mathbb{R}^d \), Lemma 3.7 of [Mattingly et al., 2012] shows that the Itô’s map \( \mathcal{I} : \mathcal{C}([0, T], \mathbb{R}^d) \to \mathcal{C}([0, T], \mathbb{R}^d) \) is well-defined and continuous, under the topology over the space \( \mathcal{C}([0, T], \mathbb{R}^d) \) induced by the supremum norm \( \|w\|_\infty \equiv \sup\{|w_t| : 0 \leq t \leq T\} \). By the Continuous Mapping Theorem, because the Langevin diffusion (36) can be seen as the image under the Itô’s map \( \mathcal{I} \) of a standard Brownian motion on \([0, T]\) evolving in \( \mathbb{R}^d \), it suffices to verify that the process \( S^{(r)} \) can be expressed as \( \mathcal{I}(\tilde{W}^{(r)}) + e^{(r)} \) where \( \tilde{W}^{(k)} \) is a sequence of stochastic processes that converge weakly in \( \mathcal{C}([0, T], \mathbb{R}^d) \) to a standard Brownian motion \( W \) and \( e^{(r)} \) is an error term that is asymptotically negligible in the sense that \( \|e^{(r)}\|_\infty \) converges to zero in probability.

For convenience, we define \( \tilde{W}^{(r)} \) as the continuous piecewise affine processes that satisfies

\[
\tilde{W}^{(r)}(T_k^{(r)}) = W(T_k^{(r)}) \quad \text{for all} \quad 0 \leq k \leq m(r),
\]
and that is affine in between. It follows that for any time $T_{k-1}^{(r)} \leq t \leq T_k^{(r)}$ we have

\[
S^{(r)}(t) = S^{(r)}(T_{k-1}^{(r)}) + \left( \int_{T_{k-1}^{(r)}}^{t} \frac{1}{2} \nabla \log \pi(S^{(r)}(T_{k-1}^{(r)})) du + \tilde{W}^{(r)}(t) - \tilde{W}(T_{k-1}^{(r)}) \right) \\
+ \int_{T_{k-1}^{(r)}}^{t} \frac{1}{2} H(S^{(r)}(T_{k-1}^{(r)}), U_k^{(r)}) du \\
= \theta_0 + \left( \int_{0}^{t} \frac{1}{2} \nabla \log \pi(S^{(r)}(u)) du + \tilde{W}^{(r)}(t) \right) \\
+ \int_{0}^{t} \nabla \log \pi(\tilde{S}^{(r)}(u)) - \nabla \log \pi(S^{(r)}(u)) \right) du \\
+ \int_{0}^{t} H(\tilde{S}^{(r)}(u), U_k^{(r)}) du, \\
\]

where $\tilde{S}^{(r)}$ is a piecewise constant (non-continuous) process, $\tilde{S}^{(r)}(t) = S^{(r)}(T_{k-1}^{(r)}) = \theta_{k-1}^{(r)}$ for $t \in [T_{k-1}^{(r)}, T_k^{(r)}]$. The process $S^{(r)}$ can thus be expressed as the sum $I(W^{(r)}) + e_1^{(r)} + e_2^{(r)}$. Since the mesh-size of the partition $\delta^{(r)}$ converges to zero as $r \to \infty$, standard properties of Brownian motions yield that $\tilde{W}^{(r)}$ converges weakly in $(\mathcal{C}([0, T], \mathbb{R}^d), \| \cdot \|_{\infty, [0, T]})$ to $W$, a standard Brownian motion in $\mathbb{R}^d$. To conclude the proof, we need to check that the quantities $\|e_1^{(r)}\|_{\infty}$ and $\|e_2^{(r)}\|_{\infty}$ converge to zero in probability.

To prove $\mathbb{E}\left[\|e_2^{(r)}\|_{\infty}^2\right] \to 0$ in probability, we have,

\[
\mathbb{E}\left[\|e_2^{(r)}\|_{\infty}^2\right] \leq 4 \mathbb{E}\left[\|e_2^{(r)}(T)\|^2\right] \quad \text{(Doob’s martingale inequality)} \\
= 4 \sum_{k=1}^{m^{(r)}} (\delta_k^{(r)})^2 \mathbb{E}\left[H(\theta_{k-1}^{(r)}, U_k^{(r)})^2\right] \\
\lesssim \sum_{k=1}^{m^{(r)}} (\delta_k^{(r)})^2 \mathbb{E}\left[V(\theta_{k-1}^{(r)})\right] \quad \text{(Assumption 4)} \\
< \text{mesh}(\delta^{(r)}) \sum_{k=1}^{m^{(r)}} \delta_k^{(r)} \mathbb{E}\left[V(\theta_{k-1}^{(r)})\right] \\
< \text{mesh}(\delta^{(r)}) \times T \times \sup \left\{ \mathbb{E}\left[V(\theta_{k-1}^{(r)})\right] : r \geq 1, 1 \leq k \leq m^{(r)} \right\} \\
\lesssim \text{mesh}(\delta^{(r)}) \quad \text{(Lemma 5).}
\]

Since $\text{mesh}(\delta^{(r)}) \to 0$, the conclusion follows.
To prove $\mathbb{E}[\|e_1^{(r)}\|_\infty] \to 0$ in probability, we first note that since the drift function $\theta \mapsto \frac{1}{2} \nabla \log \pi(\theta)$ is globally Lipschitz, for each $T^{(r)}_{k-1} \leq u \leq T^{(r)}_k$ we have,

$$\| \nabla \log (S^{(r)}(u)) - \nabla \log (S^{(r)}(u)) \| \leq C \| \theta^{(r)}_k - \theta^{(r)}_{k-1} \|$$

For some constant $C > 0$. By (34),

$$\leq \| \nabla \log \pi(\theta^{(r)}_k) \| \delta_k^{(r)} + \| H(\theta^{(r)}_k, U_k) \| \delta_k^{(r)} + \sqrt{\delta_k^{(r)} \| \eta_k^{(r)} \|}.$$ 

It follows that

$$\mathbb{E}[\| e_1^{(r)} \|_\infty] \leq \sum_{k=1}^{m(r)} \delta_k^{(r)} \left( \| \nabla \log \pi(\theta^{(r)}_k) \| \delta_k^{(r)} + \| H(\theta^{(r)}_k, U_k) \| \delta_k^{(r)} + \sqrt{\delta_k^{(r)} \| \eta_k^{(r)} \|} \right).$$

By Assumption 4 and Lemma 5, $\sup \{ \mathbb{E}[\| \nabla \log \pi(\theta^{(r)}_k) \|] : r \geq 1, 1 \leq k \leq m(r) \}$ and $\sup \{ \mathbb{E}[\| H(\theta^{(r)}_k, U_k) \|] : r \geq 1, 1 \leq k \leq m(r) \}$ are finite from which it readily follows, since $\operatorname{mesh}(\delta^{(r)}) \to 0$, that $\| e_1^{(r)} \|_\infty$ converges to zero in expectation.

7. Numerical Illustrations

In this section we illustrate the use of the SGLD method to a simple Gaussian toy model and to a Bayesian logistic regression problem. We verify that both models satisfy Assumption 4, the main assumption needed for our asymptotic results to hold. Simulations are then performed to empirically confirm our theory; for step-sizes sequences of the type $\delta_m = (m_0 + m)^{-\alpha}$, both the rate of decay of the MSE and the impact of the sub-sampling scheme are investigated.

7.1 Linear Gaussian model

Consider $N$ independent and identically distributed observations $(x_i)_{i=1}^N$ from the two parameters location model given by

$$x_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2_x).$$

We use a Gaussian prior $\theta \sim \mathcal{N}(0, \sigma^2_\theta)$ and assume that the variance hyper-parameters $\sigma^2_\theta$ and $\sigma^2_x$ are both known. The posterior density $\pi(\theta)$ is normally distributed with mean $\mu_\theta$ and variance $\sigma^2_p$ given by

$$\mu_p = \bar{x} \left( 1 + \frac{\sigma^2_p}{N \sigma^2_\theta} \right)^{-1}$$

and

$$\sigma^2_p = \frac{\sigma^2_x}{N} \left( 1 + \frac{\sigma^2_\theta}{N \sigma^2_x} \right)^{-1}$$

where $\bar{x} = (x_1 + \ldots + x_N)/N$ is the sample average of the observations. In this case, we have

$$\nabla \log \pi(\theta) = -\frac{\theta - \mu_p}{\sigma^2_p}$$

and

$$H(\theta, U) = \left\{ \frac{(N/n)}{\sum_{j \in \mathcal{I}_n(U)} x_j - \sum_{1 \leq i \leq N} x_i} / \sigma^2_x \right\}$$

for a random subset $\mathcal{I}_n(U) \subset [N]$ of cardinal $n$. 

21
7.1.1 Verification of Assumption 4

We verify in this section that Assumption 4 is satisfied for the following choice of Lyapunov function,

\[ V(\theta) = 1 + \frac{(\theta - \mu_p)^2}{2\sigma_p^2} \]

Since the error term \( H(\theta, U) \) is globally bounded, the drift \((1/2)\nabla \log \pi\) and the Lyapunov function \( V \) are linear, Assumptions 4.1 and 4.2 are satisfied. Finally, to verify Assumption 4.3, it suffices to note that since \( \nabla \log \pi(\theta) = -(\theta - \mu_p)/\sigma_p^2 \) we have

\[ \langle \nabla V(\theta), \frac{1}{2} \nabla \log \pi(\theta) \rangle = -\frac{(\theta - \mu_p)^2}{2\sigma_p^4} = \frac{1 - V(\theta)}{\sigma_p^2}. \]

In other words, Assumption 4.3 holds with \( \alpha = \beta = 1/\sigma_p^2 \).

7.1.2 Simulations

We chose \( \sigma_\theta = 1, \sigma_x = 5 \) and created a data set consisting of \( N = 100 \) data points simulated from the model. We used \( n = 10 \) as the size of subsets used to estimate the gradients. We evaluated the convergence behaviour of SGLD using the test function \( A\varphi \) with

\[ \varphi = \sin (x - \mu_p - 0.5\sigma_p) \]

\[ A\varphi = -\frac{x - \mu_p}{2\sigma_p^2} \cos (x - \mu_p - 0.5\sigma_p) + \frac{1}{2} \sin (x - \mu_p - 0.5\sigma_p) \]
Figure 2: Rates of decay of the MSE obtained from estimating the asymptotic slopes of the plots in Figure 1, compared to theoretical findings of Theorem 7. The fastest convergence rate is achieved at $\alpha = 1/3$.

Figure 3: Plots of the MSE multiplied by $T(m)$ against the number of steps $m$. The plots are flat for $\alpha \geq 0.33$, demonstrating that the MSE scales as $T(m)^{-1}$ in this regime, while the plots diverge for $\alpha < 0.33$, demonstrating that it decays at a slower rate here.

such that $\pi(A\varphi) = 0$.

We are interested in confirming the asymptotic convergence regimes of Theorem 7 by running SGLD with a range of step sizes, and plotting the mean squared error (MSE) achieved by the estimate $\pi_m(A\varphi)$ against the number of steps $m$ of the algorithm to
determine the rates of convergence. We used step sizes $\delta_m = (m + m_0(\alpha))^{-\alpha}$, for $\alpha \in \{0.1, 0.2, 0.3, 0.33, 0.4, 0.5\}$ where $m_0(\alpha)$ is chosen such that $\delta_1$ is less than the posterior standard deviation. According to the Theorem, the MSE should scale as $T(m)^{-1}$ for $\alpha > 1/3$, and $\sum_{k=1}^{m} \delta_k^2 / T(m)$ for $\alpha \leq 1/3$.

The observed MSE is plotted against $m$ on a log-log plot in Figure 1. As predicted by the theory, the optimal rate of decay is around $\alpha^* = 1/3$. To be more precise, we estimate the rates of decay by estimating the slopes on the log-log plots. This is plotted in Figure 2, which also shows a good match to the theoretical rates given in Theorem 7, where the best rate of decay is $2/3$ achieved at $\alpha = 1/3$. Finally, to demonstrate that there are indeed two distinct regimes of convergence, in Figure 3 we have plotted the MSE multiplied by $T(m)$. For $\alpha > 1/3$, the plots remain flat, showing that the MSE does indeed decay as $T(m)^{-1}$. For $\alpha < 1/3$, the plots diverge, showing that the MSE decays at a slower rate than $T(m)^{-1}$.

For $\alpha = 0.33$, Figure 4 depicts how the MSE decreases as a function of the number of likelihood evaluations for subsample sizes $n = 1, 5, 10, 50, 100$.

### 7.2 Logistic Regression

We verify in this section that Assumption (4) is satisfied for the following logistic regression model. Consider $N$ independent and identically observations $(y_i)_{i=1}^{N}$ distributed as

$$P(y_i = 1 \mid x_i, \theta) = 1 - P(y_i = -1 \mid x_i, \theta) = \logit(\langle \theta, x_i \rangle)$$

(37)

for covariate $x_i \in \mathbb{R}^d$, unknown parameter $\theta \in \mathbb{R}^d$ and function $\logit(z) = e^z / (1 + e^z)$. We assume a centred Gaussian prior on $\theta \in \mathbb{R}^d$ with positive definite symmetric covariance
matrix $C \in \mathbb{R}^{d \times d}$. It follows that

$$
\nabla \log \pi(\theta) = -C^{-1}\theta + \sum_{i=1}^{N} \logit\left(-y_i \langle \theta, x_i \rangle\right) y_i x_i
$$

$$
H(\theta, U) = \left(\frac{N}{n}\right) \sum_{j \in \mathcal{I}_n(U)} \logit\left(-y_j \langle \theta, x_j \rangle\right) y_j x_j - \sum_{1 \leq i \leq N} \logit\left(-y_i \langle \theta, x_i \rangle\right) y_i x_i
$$

for a random subset $\mathcal{I}_n(U) \subset [N]$ of cardinal $n$.

### 7.2.1 Verification of Assumption 4

We verify in this section that Assumption (4) is satisfied for the following choice of Lyapunov function,

$$
V(\theta) = 1 + \|\theta\|^2.
$$

Since $H(\theta, U)$ is globally bounded and $\|\nabla V(\theta)\|^2 = \|\theta\|^2$ and

$$
\|\nabla \log \pi(\theta)\|^2 \lesssim 1 + \|C^{-1}\theta\|^2 \lesssim 1 + \|\theta\|^2 = V(\theta),
$$

it is straightforward to see that Assumption (4).1 and (4).2 are satisfied. Finally,

$$
\langle \nabla V(\theta), \frac{1}{2} \nabla \log \pi(\theta) \rangle = -\frac{1}{2} \langle \theta, C^{-1}\theta \rangle + \frac{1}{2} \sum_{i=1}^{N} \logit\left(-y_i \langle \theta, x_i \rangle\right) y_i \langle \theta, x_i \rangle
$$

$$
\leq -\frac{\lambda_{\text{min}}}{2} \|\theta\|^2 + \frac{\sum_{i=1}^{N} \|x_i\| \|\theta\|}{2} \leq -\frac{\lambda_{\text{min}}}{4} V(\theta) + \beta
$$

with $\lambda_{\text{min}} > 0$ the smallest eigenvalue of $C^{-1}$ and $\beta \in (0, \infty)$ the global maximum over $\theta \in \mathbb{R}^d$ of the function $\theta \mapsto -\frac{\lambda_{\text{min}}}{4} \|\theta\|^2 + \frac{\sum_{i=1}^{N} \|x_i\| \|\theta\|}{2}$.

### 7.2.2 Simulations

We consider a simulated dataset where $d = 3$ and $N = 100$. We set the input covariates $x_i = (x_{i,1}, x_{i,2}, 1)$ with $x_{i,1}, x_{i,2} \sim \mathcal{N}(0, 1)$ for $i = 1 \ldots N$, and use a Gaussian prior $\theta \sim \mathcal{N}(0, I)$. We draw a $\theta_0 \sim \mathcal{N}(0, I)$ and based on it we generate $y_i$ according to the model probabilities (37). Similarly to Section (7.1.2), we plot the MSE against the likelihood evaluations for different choices of subset size $n$ in Figure 5. Notice that these simulations do not indicate any advantage of $n \neq N$ over $n = N$.

### 8. Conclusion

So far, the research on the SGLD algorithm has mainly been focused on extending the methodology. In particular, a parallel version has been introduced in Ahn et al. (2014) and it has been adapted to natural gradients in Patterson and Teh (2013b). This research has been accompanied by promising simulations. In contrast, we have focused in this article on providing rigorous mathematical foundations for the SGLD algorithm by showing that the step-size weighted estimator $\pi_m(f)$ is consistent, satisfies a central limit theorem and its asymptotic bias-variance decomposition can be characterised by an explicit functional $\mathbb{B}(m)$.
of the step-sizes sequence \((\delta_m)_{m\geq 0}\). The consistency of the algorithm is mainly due to the decreasing step-sizes procedure that asymptotically removes the bias from the discretisation and ultimately mitigates the use of an unbiased estimate of the gradient instead of the exact value. Additionally, we have proved a diffusion limit result that establishes that, when observed on the right (inhomogeneous) time scale, the sample paths of the SGLD can be approximated by a Langevin diffusion.

The CLT and bias-variance decomposition can be leveraged to show that it is optimal to choose a step-sizes sequences \((\delta_m)_{m\geq 0}\) that scales as \(\delta_m \asymp m^{-1/3}\); the resulting algorithm converges at rate \(m^{-1/3}\). Note that this recommendation is different from the previously suggested \cite{WellingTeh2011} choice of \(\delta_m \asymp m^{-1/2}\).

Our theory suggests that an optimally tuned SGLD method converges at rate \(O(m^{-1/3})\), and is thus asymptotically less efficient than a standard MCMC procedure. We believe that this result does not necessarily preclude SGLD to be more efficient in the initial transient phase, a result hinted at in Figure \(4\) the detailed study of this (non-asymptotic) phenomenon is an interesting venue of research. The asymptotic convergence rate of SGLD depends crucially on the decreasing step sizes, which is required to reduce the effect of the discretisation bias due to the lack of a Metropolis-Hastings correction. Another avenue of exploration is to determine more precisely the bias resulting from the discretisation of the Langevin diffusion, and to study the effect of the choice of step sizes in terms of the trade-off between bias, variance, and computation.

Appendix A. Proof of Lemma \[5\]

For clarity, the proof is only presented in the scalar case \(d = 1\); the multidimensional setting is entirely similar. Before embarking on the proof, let us first mention some consequences of Assumptions \[4\] that will be repeatedly used in the sequel. Since the second derivative \(V''\) is globally bounded and \((V')^2\) is upper bounded by a multiple of \(V\), we have that

\[\left| (V^p)'(\theta) \right| \lesssim V^{p-1}(\theta) \]

(38)
and that the function $V^{1/2}$ is globally Lipschitz. By expressing the quantity $V^p(\theta + \varepsilon)$ as $(V^{1/2}(\theta) + [V^{1/2}(\theta + \varepsilon) - V^{1/2}(\theta)])^{2p}$, it then follows that

$$V^p(\theta + \varepsilon) \lesssim V^p(\theta) + |\varepsilon|^{2p}. \quad (39)$$

Similarly, Definition (6), the bound $\|\nabla \log p(\theta)\|^2 \lesssim V(\theta)$ and Equation (5) yield that for any exponent $0 \leq p \leq p_H$ the following holds,

$$E_m[|\theta_{m+1} - \theta_m|^{2p}] \lesssim \delta_{m+1}^{2p} V^p(\theta) + \delta_{m+1}^p. \quad (40)$$

For clarity, the proof of Lemma (5) is separated into several steps. First, we establish that the process $m \mapsto V^p(\theta_m)$ satisfies a Lyapunov type condition; see Equation (41) below. We then describe how Equation (12) follows from this Lyapunov condition. The fact that $\pi(V^p)$ is finite can be seen as a consequence of Theorem 2.2 of [Roberts and Tweedie 1996].

- **Discrete Lyapunov condition.**

  Let us prove that there exists an index $m_0 \geq 0$ and constants $\alpha_p, \beta_p > 0$ such that for any $m \geq m_0$ we have

$$E_m[V^p(\theta_{m+1}) - V^p(\theta_m)] / \delta_{m+1} \leq -\alpha_p V^p(\theta_m) + \beta_p. \quad (41)$$

Since for any $\varepsilon$ there exists $C_\varepsilon$ such that $V^{-1}(\theta) \leq C_\varepsilon + \varepsilon V^p(\theta)$, for proving (41) it actually suffices to verify that we have

$$E_m[V^p(\theta_{m+1}) - V^p(\theta_m)] / \delta_{m+1} \leq -\tilde{\alpha}_p V^p(\theta_m) + \tilde{\beta}_p V^{-1}(\theta_m) \quad (42)$$

for some constants $\tilde{\alpha}_p, \tilde{\beta}_p > 0$ and index $m \geq 1$ large enough. A second order Taylor expansion yields that the left hand side of (42) is less than

$$E_m[(V^p)'(\theta_m)(\theta_{m+1} - \theta_m)] / \delta_{m+1} + \frac{1}{2} E_m[(V^p)''(\xi)(\theta_{m+1} - \theta_m)^2] / \delta_{m+1} \quad (43)$$

for a random quantity $\xi$ lying between $\theta_m$ and $\theta_{m+1}$. Since $E_m[|\theta_{m+1} - \theta_m|] = \frac{1}{2} \nabla \log p(\theta_m)$, the drift condition (11) yields that the first term of (43) is less than

$$p V^{-1}(\theta_m) [- \alpha V(\theta_m) + \beta] \quad (44)$$

for $\alpha, \beta > 0$ given by Equation (11). Consequently, for proving Equation (41), it remains to bound the second term of (43). Equation (58) shows that $|(V^p)'(\xi)|$ is upper bounded by a multiple of $|V^p-1(\xi)|$; the bound (39) then yields that $|V^p-1(\xi)|$ is less than a constant multiple of $|V^p-1(\theta_m)| + |\theta_{m+1} - \theta_m|^{2(p-1)}$. It follows from the bound (40) on the difference $(\theta_{m+1} - \theta_m)$ and the assumption $E[\|H(\theta, U)\|^{2p_H}] \lesssim V^p(\theta)$ that for any $\varepsilon > 0$ one can find an index $m_0 \geq 1$ large enough such that for any index $m \geq m_0$ the second term of (42) is less than a constant multiple of

$$\varepsilon V^p(\theta_m) + \beta_{p, \varepsilon} V^{-1}(\theta) \quad (45)$$

for a constant $\beta_{p, \varepsilon} > 0$. Equations (44) and (45) directly yield to Equation (42), which in turn implies to Equation (41).
• **Proof that** \( \sup_{m \geq 1} \mathbb{E}[V^p(\theta_m)] < \infty \) for any \( p \leq p_H \).

Equations (39) and (40) show that if \( \mathbb{E}[V^p(\theta_m)] \) is finite then so is \( \mathbb{E}[V^p(\theta_{m+1})] \). Under the conditions of Lemma 5, this shows that \( \mathbb{E}[V^p(\theta_m)] \) is finite for any \( m \geq 0 \). An inductive argument based on the discrete Lyapunov Equation (41) then yields that for any index \( m \geq m_0 \) the expectation \( \mathbb{E}[V^p(\theta_m)] \) is less than

\[
\max \left( \beta_p/\alpha_p, \max \left\{ \mathbb{E}[V^p(\theta_m)] : 0 \leq m \leq m_0 \right\} \right).
\]

(46)

It follows that \( \sup_{m \geq 1} \mathbb{E}[V^p(\theta_m)] \) is finite.

• **Proof that** \( \sup_{m \geq 1} \pi_m(V^p) < \infty \) for any \( p \leq p_H/2 \).

One needs to prove that the sequence \( (1/T(m)) \sum_{k=m_0}^{m} \delta_{k+1} V^p(\theta_k) \) is almost surely bounded. The discrete Lyapunov Equation (41) yields that \( \delta_{k+1} V^p(\theta_k) \) is less than \( \delta_{k+1} \beta_p/\alpha_p - \mathbb{E}_k[V^p(\theta_{k+1}) - V^p(\theta_k)]/\alpha_p \); this yields that \( (1/T(m)) \sum_{k=m_0}^{m} \delta_{k+1} V^p(\theta_k) \) is less than a constant multiple of

\[
1 + \frac{V^p(\theta_m)}{T(m)} + \frac{1}{T(m)} \sum_{k=m_0}^{m} \left\{ V^p(\theta_{k+1}) - \mathbb{E}_k[V^p(\theta_{k+1})] \right\}.
\]

To conclude the proof, we prove that the last term in the above displayed Equation almost surely converges to zero; By Kronecker’s Lemma, it suffices to prove that the martingale

\[
M_m = \sum_{k=m_0}^{m} \frac{V^p(\theta_{k+1}) - \mathbb{E}_k[V^p(\theta_{k+1})]}{T(k)}
\]

is almost surely bounded; to this end, we establish that the sum \( \sum_{m \geq 0} \mathbb{E}\left[ |M_{m+1} - M_m|^2 \right] \) is finite. We have \( \mathbb{E}\left[ |V^p(\theta_{k+1}) - \mathbb{E}_k[V^p(\theta_{k+1})]|^2 \right] \leq 2 \times \mathbb{E}\left[ |V^p(\theta_{k+1}) - V^p(\theta_k)|^2 \right] \) and the mean value Theorem yields that \( |V^p(\theta_{k+1}) - V^p(\theta_k)| \leq V^{p-1}(\xi) V'(\xi) (\theta_{k+1} - \theta_k) \) for some \( \xi \) lying between \( \theta_k \) and \( \theta_{k+1} \). The bound \( |V'(\theta)| \leq V^{1/2}(\theta) \) and Equation (39) then yield that \( |V^p(\theta_{k+1}) - V^p(\theta_k)| \leq V^{p-1/2}(\theta_k) |\theta_{k+1} - \theta_k| + |\theta_{k+1} - \theta_k|^{2p} \). From this bound (40) and the assumption that \( \mathbb{E}[H(\theta, U)^{2pH}] \leq V^{pH}(\theta) \) it follows that

\[
\sum_{m \geq 0} \mathbb{E}\left[ |M_{m+1} - M_m|^2 \right] \leq \sum_{m \geq 0} \mathbb{E}\left[ V^{2p}(\theta_m) \right] \times \delta_m.
\]

Since \( \mathbb{E}[V^{2p}(\theta_m)] \) is uniformly bounded for any \( p \leq p_H/2 \) and \( \sum_{m \geq m_0} \delta_m/T^2(m) < \infty \) (because the sum \( \sum_m T^{-1}(m+1) - T^{-1}(m) \) is finite), the conclusion follows.

• **Proof of** \( \pi(V^p) < \infty \) for any \( p \geq 0 \).

Since \( V(\theta) \leq 1 + \|\theta\|^2 \), the drift condition (11) yields that Theorem 2.1 of (Roberts and Tweedie, 1996) holds. Moreover, the bound \( V^{p-1}(\theta) \leq C_\varepsilon + \varepsilon V^p(\theta) \) implies that there are constants \( \alpha_{p,*}, \beta_{p,*} > 0 \) such that

\[
A V^p(\theta) \leq -\alpha_{p,*} V^p(\theta) + \beta_{p,*}
\]

where \( A \) is the generator of the Langevin diffusion (1). Theorem 2.2 of (Roberts and Tweedie, 1996) gives the conclusion.
Proof that $\sup_{m \geq 1} \pi_m^\omega(V^p) < \infty$ for any $p \leq p_H/2$.

One needs to prove that the sequence $[1/\Omega(m)] \times \sum_{k=m_0}^m \omega_{k+1} V^p(\theta_k)$ is almost surely bounded. The bound $\delta_{k+1} V^p(\theta_k) \leq \delta_{k+1} \beta_p/\alpha_p - E_k[V^p(\theta_{k+1}) - V^p(\theta_k)]/\alpha_p$ yields that $\pi_m^\omega(V^p)$ is less than a constant multiple of

$$1 + \frac{(\omega_{m_0}/\delta_{m_0}) V^p(\theta_{m_0})}{T(m)} + \Omega^{-1}(m) \sum_{k=m_0+1}^m (\omega_k/\delta_k) \left\{ V^p(\theta_{k+1}) - E_k[V^p(\theta_{k+1})] \right\}$$

$$+ \Omega^{-1}(m) \sum_{k=m_0}^{m-1} \Delta(\omega_k/\delta_k) V^p(\theta_k).$$

To conclude the proof, we establish that the following limits hold almost surely,

$$\lim_{m \to \infty} \Omega^{-1}(m) \sum_{k=m_0+1}^m (\omega_k/\delta_k) \left\{ V^p(\theta_{k+1}) - E_k[V^p(\theta_{k+1})] \right\} = 0 \quad (49)$$

$$\lim_{m \to \infty} \Omega^{-1}(m) \sum_{k=m_0}^{m-1} \Delta(\omega_k/\delta_k) V^p(\theta_k) = 0. \quad (50)$$

To prove Equation (49) it suffices to use the assumption that $\sum_{m \geq 0} \omega_m^2/\delta_m \Omega^2(m) < \infty$ and then follow the same approach used to establish that the martingale (47) is bounded in $L^2$. To prove Equation (50), Kronecker’s Lemma shows that it suffices to verify that

$$E \left[ \sum_{m \geq 0} \left| \Delta(\omega_m/\delta_m) \right| V^p(\theta_m)/\Omega(m) \right] < \infty$$

This directly follows from the assumption that $\sum_{m \geq 0} \left| \Delta(\omega_m/\delta_m) \right| /\Omega(m) < \infty$ and the fact that $\sup_{m \geq 0} E[V^p(\theta_m)]$ is finite.

References

Amr Ahmed, Moahmed Aly, Joseph Gonzalez, Shravan Narayananamurthy, and Alexander J Smola. Scalable inference in latent variable models. In Proceedings of the ACM international conference on Web search and data mining, pages 123–132. ACM, 2012.

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In Proceedings of the International Conference on Machine Learning, 2012.

Sungjin Ahn, Babak Shahbaba, and Max Welling. Distributed stochastic gradient MCMC. In Proceedings of the International Conference on Machine Learning, 2014.

Shun-ichi Amari and Hiroshi Nagaoka. Methods of information geometry, volume 191. American Mathematical Society, 2007.

R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up MCMC: an adaptive sub-sampling approach, accepted in Proceedings of the International Conference on Machine Learning (ICML), 2014.

Alexandros Beskos, Gareth Roberts, Andrew Stuart, and Jochen Voss. MCMC methods for diffusion bridges. Stochastics and Dynamics, 8(03):319–350, 2008.

Patrick Billingsley. Probability and Measure. Wiley-Interscience, 3 edition, 1995.
Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer, 2010.

Tianqi Chen, Emily B Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. *arXiv preprint arXiv:1402.4102*, 2014.

SL Cotter, GO Roberts, AM Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986. Characterization and convergence.

Josselin Garnier. *Scattering, spreading, and localization of an acoustic pulse by a random medium*. Walter de Gruyter, Berlin, 2003.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(2):123–214, 2011. With discussion and a reply by the authors.

Joseph Gonzalez. Emerging systems for large-scale machine learning. ICML Tutorial, 2014.

Martin Hairer, Andrew Stuart, and Sebastian Vollmer. Spectral gaps for a metropolis-hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, (to appear), 2014.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press New York, 1980.

Zaid Harchaoui and Martin Jaggi. Frank-Wolfe and greedy optimization for learning with big data. ICML Tutorial, 2014.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, and Tara N Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

Matthew D Hoffman, David M Blei, and Francis R Bach. Online learning for latent dirichlet allocation. In *NIPS*, volume 2, page 5, 2010.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
Analysis of the Stochastic Gradient Langevin Dynamics

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In Proceedings of the International Conference on Machine Learning, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, volume 1, page 4, 2012.

Damien Lamberton and Gilles Pages. Recursive computation of the invariant distribution of a diffusion. Bernoulli, 8(3):367–405, 2002.

Damien Lamberton and Gilles Pages. Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. Stochastics and dynamics, 3 (04):435–451, 2003.

Vincent Lemaire. An adaptive scheme for the approximation of dissipative systems. Stochastic Processes and their Applications, 117(10):1491–1518, 2007.

Samuel Livingstone and mark Girolami. Information-geometric markov chain monte carlo methods using diffusions. arXiv preprint arXiv:1403.7957, 2014.

Gisiro Maruyama. Continuous markov processes and stochastic equations. Rendiconti del Circolo Matematico di Palermo, 4(1):48–90, 1955.

Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. Stochastic processes and their applications, 101(2):185–232, 2002.

Jonathan C Mattingly, Natesh S Pillai, and Andrew M Stuart. Diffusion limits of the random walk metropolis algorithm in high dimensions. The Annals of Applied Probability, 22(3): 881–930, 2012.

Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics, 9(5):356–369, 2008.

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.

R.M. Neal. MCMC using hamiltonian dynamics. Handbook of Markov Chain Monte Carlo: Methods and Applications, page 113, 2010.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel mcmc. arXiv preprint arXiv:1311.4780, 2013.

Gilles Pages and Fabien Panloup. Ergodic approximation of the distribution of a stationary diffusion: rate of convergence. The Annals of Probability, 22(3):1059–1100, 2012.
Fabien Panloup. Recursive computation of the invariant measure of a stochastic differential equation driven by a lévy process. *The Annals of Applied Probability*, 18(2):379–426, 2008.

E. Pardoux and A. Yu. Veretennikov. On the Poisson equation and diffusion approximation. I. *The Annals of Probability*, 29(3):1061–1085, 2001. URL [http://dx.doi.org/10.1214/aop/1015345596](http://dx.doi.org/10.1214/aop/1015345596).

S. Patterson and Y. W. Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, 2013a.

Sam Patterson and Yee Whye Teh. Stochastic Gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, 2013b.

Natesh S Pillai, Andrew M Stuart, and Alexandre H Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012.

H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22:400–407, 1951a. ISSN 0003-4851.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951b.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Walter Rudin. *Real and complex analysis (3rd)*. New York: McGraw-Hill Inc, 1986.

Masa-Aki Sato. Online model selection based on the variational bayes. *Neural Computation*, 13(7):1649–1681, 2001.

Albert N Shiryaev. *Probability*. Graduate Texts in Mathematics, 1996.

Nathan Srebro and Ambuj Tewari. Stochastic optimization for machine learning. ICML Tutorial, 2010.

O Stramer and RL Tweedie. Langevin-type models I: Diffusions with given stationary distributions and their discretizations*. *Methodology and Computing in Applied Probability*, 1(3):283–306, 1999a.

O Stramer and RL Tweedie. Langevin-type models II: Self-targeting candidates for MCMC algorithms. *Methodology and Computing in Applied Probability*, 1(3):307–328, 1999b.

Sebastian Thrun. Toward robotic cars. *Communications of the ACM*, 53(4):99–106, 2010.

William YS Wang, Bryan J Barratt, David G Clayton, and John A Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.
Max Welling and Yee Whye Teh. Bayesian learning via Stochastic Gradient Langevin Dynamics. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 681–688, 2011.