Consistency and fluctuations for stochastic gradient Langevin dynamics

Yee Whye Teh y.w.teh@stats.ox.ac.uk
Department of Statistics
University of Oxford

Alexandre H. Thiery a.h.thiery@nus.edu.sg
Department of Statistics and Applied Probability
National University of Singapore

Sebastian J. Vollmer vollmer@stats.ox.ac.uk
Department of Statistics
University of Oxford

Editor:

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 $(\delta_m)_{m\geq 0}$.

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 $\delta_m \asymp m^{-1/3}$, leading to an algorithm whose mean squared error (MSE) decreases at rate $\mathcal{O}(m^{-1/3})$.

Keywords: Markov Chain Monte Carlo, Langevin Dynamics, Big Data

Contents

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 7) and satisfy a central limit theorem (CLT) (Theorem 8); 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 9).

Our analysis reveals that for a sequence of step-sizes with algebraic decay $\delta_m \asymp m^{-\alpha}$ the optimal choice, when measured in terms of rate of decay of the mean squared error (MSE), is given for $\alpha_\star = 1/3$; the choice $\delta_m \asymp m^{-\alpha_\star}$ leads to an algorithm that converges at rate $\mathcal{O}(m^{-1/3})$. 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 $\mathrm{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. For two positive functions $f, g : \mathbb{R} \to [0, \infty)$, one writes $f \lesssim g$ to 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 π 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 $\pi(\varphi) = \pi(\varphi; \mathcal{X})$. Finally, densities of probability distributions on $\mathbb{R}^d$ are implicitly assumed to be defined with respect to the usual d-dimensional Lebesgue measure.

Acknowledgement

SJV and YWT acknowledge EPSRC for research funding through grant EP/K009850/1 and EP/K009362/1. AHT is grateful for financial support in carrying out this research from a Singaporean MoE grant.

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 the Langevin diffusion, which is given by the 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 and $(W_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The linear operator $\mathcal{A}$ denotes the generator of the Langevin diffusion (1): for a twice continuously differentiable test function $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\mathcal{A}\varphi(\theta) = \frac{1}{2}\langle \nabla \log \pi(\theta), \nabla \varphi(\theta)\rangle + \frac{1}{2}\Delta\varphi(\theta), \tag{2}$$

where $\Delta\varphi \stackrel{\text{def}}{=} \sum_{i=1}^d \nabla_i^2 \varphi$ denotes the standard Laplacian operator. The motivation behind the choice of Langevin diffusions is that, under certain conditions, they are ergodic with respect to the distribution π; for example, (Roberts and Tweedie, 1996; Stramer and Tweedie, 1999a,b; Mattingly et al., 2002) describe drift conditions of the type described in Section 3.2 that ensure that the total variation distance from stationarity of the law at time t of the Langevin diffusion (1) decreases to zero exponentially quickly as $t \to \infty$.

Given a time-step $\delta > 0$ and a current position θ_t, it is often straightforward to simulate a random variable $\theta_\star$ that is approximately distributed as the law of $\theta_{t+\delta}$ given θ_t. For stochastic differential equations, the Euler-Maruyama scheme (Maruyama, 1955) might be the simplest approach for approximating the law of $\theta_{t+\delta}$. For a Langevin diffusion this reads

$$\theta_\star = \theta_t + \frac{1}{2}\delta \nabla \log \pi(\theta_t) + \delta^{1/2}\eta \tag{3}$$

for a standard d-dimensional centred Gaussian random variable η. To fully correct the discretization error, one can adopt a Metropolis-Hastings accept-reject mechanism. The resulting algorithm is usually referred to as the Metropolis-Adjusted-Langevin algorithm (MALA) (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 the Hamiltonian Monte Carlo (HMC) algorithm (Duane et al., 1987) is based on discretizations of an Hamiltonian system of differential equations. See the excellent review of Neal (2010) for further information.

In this paper, we shall consider the situation where the target π 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,

$$\pi(\theta) \propto \mathrm{p}_0(\theta) \prod_{i=1}^N \mathrm{p}\left(y_i \,|\, \theta\right). \tag{4}$$

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 (Korattikara et al., 2014; Bardenet et al., 2014).

This paper is concerned with stochastic gradient Langevin dynamics (SGLD), an alternative approach proposed by 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 (1) with a sufficiently small step-size $\delta \ll 1$, because the Langevin diffusion is ergodic with respect to π, 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 π. Choosing a decreasing sequence of step-sizes $\delta_m \to 0$ should even allow us to converge to the exact posterior distribution. To further make this approach viable in large N settings, the gradient term $\nabla \log \pi(\theta)$ can be further approximated using a subsampling strategy. For an integer $1 \leq n \leq N$ and a random subset $\tau \stackrel{\text{def}}{=} (\tau_1, \ldots, \tau_n)$ of $[N] \equiv \{1, \ldots, N\}$ generated by sampling with or without replacement from $[N]$, the quantity

$$\nabla \log \mathrm{p}_0(\theta) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log \mathrm{p}(x_{\tau_i} \mid \theta) \tag{5}$$

is an unbiased estimator of $\nabla \log \pi(\theta)$. Most importantly, this stochastic estimate can be computed with a computational budget that scales as $\mathcal{O}(n)$ 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 $\widehat{\nabla \log \pi}(\theta, \mathcal{U})$ 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 $\widehat{\nabla \log \pi}(\theta, \mathcal{U})$ means that

$$\mathbf{E}\left[H(\theta, \mathcal{U})\right] = 0 \qquad \text{with} \qquad H(\theta, \mathcal{U}) \stackrel{\text{def}}{=} \widehat{\nabla \log \pi}(\theta, \mathcal{U}) - \nabla \log \pi(\theta). \tag{6}$$

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 θ_{m-1}, the next position θ_m is defined though the recursion

$$\theta_m = \theta_{m-1} + \frac{1}{2} \delta_m \, \widehat{\nabla \log \pi}(\theta_{m-1}, \mathcal{U}_m) + \delta_m^{1/2} \, \eta_m \tag{7}$$

for an i.i.d. sequence $\eta_m \sim \mathrm{N}(0, I_d)$, and an independent and i.i.d. sequence $\mathcal{U}_m$ of auxiliary random variables. This is the equivalent of the Euler-Maruyama discretization (3) of the Langevin diffusion (1) with a decreasing sequence of step-sizes and a stochastic estimate to the gradient term. The analysis presented in this article assumes for simplicity that the

initial position θ_0 of the algorithm is deterministic; in the simulation study of Section 7, the algorithms are started at the MAP estimator. Indeed, more general situations could be analysed with similar arguments at the cost of slightly less transparent proofs. 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 φ with respect to the posterior distribution π can be approximated by the weighted sum

$$\pi_m(\varphi) \stackrel{\text{def}}{=} \frac{\delta_1 \, \varphi(\theta_0) + \ldots + \delta_m \, \varphi(\theta_{m-1})}{T_m} \tag{8}$$

with $T_m = \delta_1 + \ldots + \delta_m$. The quantity $\pi_m(\varphi)$ thus approximates the ergodic average $T_m^{-1} \int_0^{T_m} \varphi(\theta_t) \, dt$ between time zero and $t = T_m$. During the course of the proof of our fluctuation Theorem 8, 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 ω-weighted sum

$$\pi_m^{\omega}(\varphi) \stackrel{\text{def}}{=} \frac{\omega_1 \, \varphi(\theta_0) + \ldots + \omega_m \, \varphi(\theta_{m-1})}{\Omega_m} \tag{9}$$

with $\Omega_m \stackrel{\text{def}}{=} \omega_1 + \ldots + \omega_m$. Indeed, $\pi_m^{\omega}(\varphi) = \pi_m(\varphi)$ in the particular case $(\omega_m)_{m\geq 1} = (\delta_m)_{m\geq 1}$; we will consider the weight sequence $\omega = \{\delta_m^2\}_{m\geq 1}$ in the proof of Theorem 8.

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 π, can be constructed for various choices of positive definite volatility matrix function $M : \mathbb{R}^d \to \mathbb{R}^{d,d}$. 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 an uniform-ellipticity condition and a growth assumption on the volatility matrix function $M : \mathbb{R}^d \to \mathbb{R}^{d,d}$, 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

Throughout this text, we assume that the sequence of step-sizes $\delta = (\delta_m)_{m\geq 1}$ satisfies the following usual assumption.

Assumption 1 *The step-sizes $\delta = (\delta_m)_{m\geq 1}$ form a decreasing sequence with*

$$\lim_{m\to\infty} \delta_m = 0 \qquad \textit{and} \qquad \lim_{m\to\infty} T_m = \infty.$$

Indeed, this assumption is easily seen to also be necessary for the Law of Large Numbers of Section 4 to hold. Furthermore, we will need at several occasions to assume the following assumption on the oscillations of a sequence of step-sizes $(\omega_m)_{m\geq 1}$.

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

$$\lim_{m\to\infty} \sum_{m\geq 1} \left|\Delta(\omega_m/\delta_m)\right| / \Omega_m < \infty \qquad \textit{and} \qquad \sum_{m\geq 1} \omega_m^2/[\delta_m \Omega_m^2] < \infty.$$

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

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 $\sum_{m\geq 1} \left|\Delta(\omega_m/\delta_m)\right|/\Omega_1 = \delta_1^{p-1}/\Omega_1$, while the finiteness of the second sum can be seen as follows:*

$$\begin{aligned}\sum_{m\geq 1} \omega_m^2/\big(\delta_m \Omega_m^2\big) &\lesssim 1 + \sum_{m\geq 2} (\omega_m/\delta_m)^2 \,(1/\Omega_{m-1} - 1/\Omega_m) \\ &\lesssim 1 + \sum_{m\geq 2} (1/\Omega_{m-1} - 1/\Omega_m) = 1 + 1/\Omega_1.\end{aligned}$$

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 1 and Assumption 2.

3.2 Stability

Under assumptions on the tails of the posterior density π, the Langevin diffusion (1) is non-explosive and for any starting position $\theta_0 \in \mathbb{R}^d$ the total-variation distance $d_{\mathrm{TV}}\big(\mathbf{P}(\theta_t \in \cdot), \pi\big)$ converges to zero as $t \to \infty$. For instance, Theorem 2.1 of (Roberts and Tweedie, 1996) shows that it is sufficient to assume that the drift term satisfies the condition $(1/2)\,\langle \nabla \log \pi(\theta), \theta\rangle \leq \alpha\|\theta\|^2 + \beta$ 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 θ, 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.

Guaranteeing stability of the Euler-Maruyama discretization requires stronger Lyapanunov 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 $\|\theta\| \to \infty$, is twice differentiable with bounded second derivatives, and satisfies the following conditions.*

1. *There exists an exponent $p_H \geq 2$ such that*

$$\mathbf{E}\left[\|H(\theta, \mathcal{U})\|^{2p_H}\right] \lesssim V^{p_H}(\theta). \tag{10}$$

   This implies that $\mathbf{E}\left[\|H(\theta, \mathcal{U})\|^{2p}\right] \lesssim V^p(\theta)$ for any exponent $0 \leq p \leq p_H$.

2. *For every $\theta \in \mathbb{R}^d$ we have*

$$\|\nabla V(\theta)\|^2 + \|\nabla \log \pi(\theta)\|^2 \lesssim V(\theta). \tag{11}$$

3. *There are constants $\alpha, \beta > 0$ such that for every $\theta \in \mathbb{R}^d$ we have*

$$\frac{1}{2} \langle \nabla V(\theta), \nabla \log \pi(\theta) \rangle \leq -\alpha V(\theta) + \beta. \tag{12}$$

Equation (12) ensures that on average the drift term $\widehat{\nabla \log \pi}(\theta)$ points towards the centre of the state space, while equations (10) and (11) provide control on the magnitude of the (stochastic) drift term. The drift condition (12) implies in particular that the Langevin diffusion (1) converges exponentially quickly towards the equilibrium distribution π (Mattingly et al., 2002; Roberts and Tweedie, 1996). The proof of the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) both exploit the following Lemma.

Lemma 5 (Stability) *Let the step-sizes $(\delta_m)_{m \geq 1}$ satisfy Assumption 1 and suppose that the stability Assumptions 4 hold. For any exponent $0 \leq p \leq p_H$ the following bounds hold almost surely,*

$$\sup_{m \geq 1} \pi_m(V^{p/2}) < \infty \qquad \text{and} \qquad \sup_{m \geq 1} \mathbf{E}\left[V^p(\theta_m)\right] < \infty. \tag{13}$$

Moreover, for any exponent $0 \leq p \leq p_H$ *we have* $\pi(V^p) < \infty$. *If the sequence of weights* $(\omega_m)_{m\geq 1}$ *satisfies Assumption 2 the following holds almost surely,*

$$\sup_{m\geq 1} \pi_m^{\omega}(V^{p/2}) \; < \; \infty \tag{14}$$

The technical proof can be found in Section B. The idea is to leverage condition (12) in order to establish that the function V^p satisfies both discrete and continuous drift conditions.

3.3 Scope of the analysis

For a posterior density π of the form (4) and the usual unbiased estimate to $\nabla \log \pi$ described in Equation (5), to establish that Equations (10) and (11) hold it suffices to verify that the prior density p_0 is such that $\|\nabla \log \mathrm{p}_0(\theta)\|^2 \lesssim V(\theta)$ and that for any index $1 \leq i \leq N$ the likelihood term $\mathrm{p}\,(\,y_i \,|\, \theta)$ is such that

$$\|\nabla \log \mathrm{p}\,(\,y_i \,|\, \theta)\|^{2\,p_H} \lesssim V^{p_H}(\theta).$$

Indeed, in these circumstances, we have $\|H(\theta,\mathcal{U})\|^{2p_H} \lesssim \sum_{i=1}^N \|\nabla \log \mathrm{p}(y_i \mid \theta)\|^{2p_H}$. Several such examples are described in Section 7.

It is important to note that the drift Condition (12) typically does not hold for distributions with heavy tails such that $\nabla \log \pi(x) \to 0$ as $\|x\| \to \infty$ (Roberts and Tweedie, 1996). For example, the standard MALA algorithm is not geometrically ergodic when $\nabla \log \pi(x)$ converges to zero as $\|x\| \to \infty$ (Theorem 4.3 of (Roberts and Tweedie, 1996)); indeed, the analysis of standard local-move MCMC algorithms when applied to target densities with heavy tails is delicate and typically necessitate other tools Stramer and Tweedie (1999b); Jarner and Roberts (2007); Kamatani (2014) than the approach based on drift conditions of the type (12). The analysis of the properties of the SGLD algorithm when applied to such heavy tail densities is out of the scope of this article. It is important to note that many more complex scenarios involving high-dimensionality, multi-modality, non-parametric settings where the complexity of the target distribution increases with the size of the data, or combination thereof, are examples of interesting and relevant situations where our analysis typically does not apply; analysing the SGLD algorithm when applied to these challenging target distributions is well out of the scope of this article.

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 (8) converges weakly towards π. Specifically, under growth assumptions on a test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, the

following strong law of large numbers holds almost surely,

$$\lim_{m\to\infty} \frac{\delta_1\,\varphi(\theta_0) + \ldots + \delta_m\,\varphi(\theta_m)}{T_m} \;=\; \int_{\mathbb{R}^d} \varphi(\theta)\,\pi(d\theta),$$

with a similar result for ω-weighted empirical averages, under assumptions on the weight sequence ω. The proofs of several results of this paper make use of the following elementary lemma.

Lemma 6 *Let $(\Delta M_k)_{k\geq 0}$ and $(R_k)_{k\geq 0}$ be two sequences of random variables adapted to a filtration $(\mathcal{F}_k)_{k\geq 0}$ and let $(\Gamma_k)_{k\geq 0}$ be an increasing sequence of positive real numbers. The limit*

$$\lim_{m\to\infty} \frac{\sum_{k=0}^{m} \Delta M_k + R_k}{T_m} \;=\; 0 \tag{15}$$

holds almost surely if the following two conditions are satisfied.

1. *The process $M_m = \sum_{k\leq m} \Delta M_k$ is a martingale, i.e. $\mathbf{E}\left[\Delta M_k \,|\, \mathcal{F}_k\right] = 0$ and*

$$\lim_{k\to\infty} \sum_{k\geq 0} \frac{\mathbf{E}\left[|\Delta M_k|\right]^2}{T_k^2} \;<\; \infty. \tag{16}$$

2. *The sequence $(R_k)_{k\geq 0}$ is such that*

$$\lim_{k\to\infty} \sum_{k\geq 0} \frac{\mathbf{E}\left[|R_k|\right]}{T_k} \;<\; \infty. \tag{17}$$

The above lemma, whose proof can be found in the appendix A, is standard; Lamberton and Pages (2002) also follows this route to prove several of their results.

Theorem 7 (Consistency) *Let the step-sizes satisfy Assumption* (1) *and suppose that the stability Assumptions 4 hold for a Lyapunov function $V : \mathbb{R}^d \to [1,\infty)$. Let $0 \leq p < p_H/2$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a test function such that $|\varphi(\theta)|/V^p(\theta)$ is globally bounded. Then the following limit holds almost surely:*

$$\lim_{m\to\infty} \pi_m(\varphi) \;=\; \pi(\varphi). \tag{18}$$

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

$$\lim_{m\to\infty} \pi^{\omega}_m(\varphi) \;=\; \pi(\varphi). \tag{19}$$

Proof In the following, we write $\mathbf{E}_k\left[\,\cdot\,\right]$ and $\mathbf{P}_k\left(\,\cdot\,\right)$ to denote the conditional expectation $\mathbf{E}\left[\,\cdot\,|\,\theta_k\right]$ and conditional probability $\mathbf{P}\left(\,\cdot\,|\,\theta_k\right)$ respectively. We use the notation $\Delta\theta_k \stackrel{\text{def}}{=} (\theta_{k+1} - \theta_k)$. Finally, for notational convenience, we only present the proof in the scalar case $d = 1$, the multidimensional case being entirely similar. We will give a detailed proof of Equation (18) and then briefly describe how the more general Equation (19) can be proven using similar arguments. To prove Equation (18), we first show that the sequence

$(\pi_m)_{m\geq 1}$ almost surely converges weakly to π. Equation (18) is then proved in a second stage.

Weak convergence of $(\pi_m)_{m\geq 1}$. To prove that almost surely the sequence $(\pi_m)_{m\geq 1}$ converges weakly towards π 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 π. By Prokhorov's Theorem (Billingsley, 1995) and Equation (13), because the Lyapunov function V goes to infinity as $\|\theta\| \to \infty$, 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 π_∞ 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 π_∞ of the sequence $(\pi_m)_{m\geq 1}$ the following holds,

$$\pi_\infty(\mathcal{A}\varphi) = 0. \tag{20}$$

To prove Equation (20) we use the following decomposition of $\pi_m(\mathcal{A}\varphi)$,

$$\left\{ \frac{\sum_{k=1}^m \mathbf{E}_{k-1}[\varphi(\theta_k) - \varphi(\theta_{k-1})]}{T_m} \right\} \; - \; \left\{ \frac{\sum_{k=1}^m \mathbf{E}_{k-1}[\varphi(\theta_k) - \varphi(\theta_{k-1})]}{T_m} - \pi_m(\mathcal{A}\varphi) \right\}. \tag{21}$$

- Let us prove that the first term of (21) converges almost surely to zero. The numerator is equal to the sum of $\sum_{k=1}^m \mathbf{E}_{k-1}[\varphi(\theta_k)] - \varphi(\theta_k)$ and $\varphi(\theta_m) - \varphi(\theta_0)$. By boundedness of φ, the term $\{\varphi(\theta_m) - \varphi(\theta_0)\}/T_m$ converges almost surely to zero. By Lemma 6, to conclude is suffices to show that the martingale difference terms $\mathbf{E}_{k-1}\left[\varphi(\theta_k)\right] - \varphi(\theta_k)$ are such that

$$\sum_{k\geq 1} \frac{\mathbf{E}\left[\left|\mathbf{E}_{k-1}\left[\varphi(\theta_k)\right] - \varphi(\theta_k)\right|^2\right]}{T_k^2} \; < \; \infty.$$

  Because φ is Lipschitz, it suffices to prove that $\sum_{k\geq 1} \mathbf{E}\left(\|\theta_{k+1} - \theta_k\|^2\right)/T_k^2$ is finite. The stability Assumption 4 and Lemma 5 imply that the supremum $\sup_m \mathbf{E}\left[V(\theta_m)\right]$ is finite. Since $\mathbf{E}_k\left[\|\theta_{k+1} - \theta_k\|^2\right] \lesssim \delta_{k+1}^2 V(\theta) + \delta_{k+1}$, it follows that $\mathbf{E}\left(\|\theta_{k+1} - \theta_k\|^2\right)$ is less than a constant multiple of δ_{k+1}. Under Assumption 1, because the telescoping sum $\sum_{k\geq 1} T^{-1}(k) - T^{-1}(k+1)$ is finite, the sum $\sum_{k\geq 1} \delta_k/T_k^2$ is finite. This concludes the proof that the first term in (21) converges almost surely to zero.

- The second term of (21) equals $\big(R_0 + \ldots + R_{m-1}\big)/T_m$ with

$$R_k \stackrel{\text{def}}{=} \mathbf{E}_k\left[\varphi(\theta_{k+1}) - \varphi(\theta_k)\right] - \mathcal{A}\varphi(\theta_k)\,\delta_{k+1}. \tag{22}$$

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

– If $|\theta_k| > K+1$ then $\varphi(\theta_k) = \mathcal{A}\varphi(\theta_k) = 0$ so that $|R_k| \leq \|\varphi\|_\infty \times \mathbf{P}_k(\theta_{k+1} \in \Omega)$. Since $\theta_{k+1} - \theta_k = \left\{\frac{1}{2}\nabla \log \pi(\theta_k) + H(\theta_k, \mathcal{U})\right\} \delta_{k+1} + \sqrt{\delta_{k+1}}\,\eta$ we have

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

We have used the notation $\mathbb{I}(A)$ 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 $|\nabla \log \pi(\theta)|/\mathrm{dist}(\theta, \Omega)$ is bounded on the set $\{\theta : |\theta| > K\}$; this shows that the first term equals zero for δ_k small enough. To prove that the second term is bounded by a constant multiple of δ_{k+1}^2, it suffices to use Markov's inequality and the fact that $\mathbf{E}[H(\theta_k, \mathcal{U})^2]/\mathrm{dist}^2(\theta, \Omega)$ is bounded on $\{\theta : |\theta| > K\}$; this is because $\mathbf{E}[H(\theta_k, \mathcal{U})^2]$ 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 δ_{k+1}^2 by Markov's inequality and the fact that η 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

$$\begin{aligned} R_k &= \frac{1}{2}\,\delta_{k+1}^2\,\varphi^{''}(\theta_k)\left\{[\nabla \log \pi(\theta_k)]^2 + \mathbf{E}_k\left[H^2(\theta_k, \mathcal{U})\right]\right\} \\ &\quad + (1/2)\;\mathbf{E}_k\left[(\Delta\theta_k)^3 \int_0^1 \varphi^{'''}(\theta_k + u\,\Delta\theta_k)\,(1-u)^2\,du\right] \\ &= R_{k,1} + R_{k,2}. \end{aligned}$$

Under Assumption 4, the quantities $[\nabla \log \pi(\theta_k)]^2$ and $\mathbf{E}[H^2(\theta_k, \mathcal{U})]$ are upper bounded by a constant multiple of $V(\theta_k)$. Since the function $\theta \mapsto \varphi^{''}(\theta)\,V(\theta)$ is globally bounded (because continuous with compact support) this shows that $R_{k,1}$ is less than a constant multiple of δ_{k+1}^2. Since $|\theta_k| \leq K+1$, the bounds $\mathbf{E}[H^3(\theta, \mathcal{U})] \lesssim V^{3/2}(\theta)$ and $\sup_{k \geq 0} \mathbf{E}[V^{3/2}(\theta_k)] < \infty$ (see Lemma 5) yield that $\mathbf{E}_k\,|\Delta\theta_k|^3 \leq 9\,\overline{C}\,(\delta_{k+1}^3 + \delta_{k+1}^{3/2}) \lesssim \delta_{k+1}^{3/2}$ with

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

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

We have thus proved that there is a constant C such $|R_k| \leq C\,\delta_{k+1}^{3/2}$ for $k \geq 0$; it follows that the sum $(R_0 + \ldots + R_{m-1})/T_m$ is less than a constant multiple of $\left(\delta_1^{3/2} + \ldots + \delta_m^{3/2}\right)/T_m$. 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 π_m towards π.

Proof of Equation (18). 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 $\pi_m(\widetilde{\varphi}) \to \pi(\widetilde{\varphi})$ for a continuous and bounded test function $\widetilde{\varphi}$.

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

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

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

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

By the triangle inequality, we thus have,

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

On the right-hand-side, the term in the middle can be made arbitrarily small as $m \to \infty$ since π_m converges weakly towards π, while the other two terms converges to zero as $t \to \infty$. This concludes the proof of Equation (18).

Proof of Equation (19). The approach is very similar to the proof of Equation (18) and for this reason we only highlight the main differences. The same argument shows that the sequence π_m^ω is tight and it suffices to show that $\pi_\infty^\omega(\mathcal{A}\varphi) = 0$ for any weak limit π_∞^ω of the sequence $(\pi_m^\omega)_{m \geq 0}$ for obtaining the almost sure weak convergences of $(\pi_m^\omega)_{m \geq 0}$ towards π. One can then upgrade this almost sure weak convergence to a Law of Large Numbers. To prove (19), we thus concentrate on proving that $\pi_\infty^\omega(\mathcal{A}\varphi) = 0$. For a smooth and compactly supported test function φ we use the decomposition $\pi_m^\omega(\mathcal{A}\varphi) = S_1(m) + S_2(m) + S_3(m)$ with

$$\left\{ \begin{array}{ll} S_1(m) & = \frac{1}{\Omega_m} \sum_{k=1}^m \frac{\omega_k}{\delta_k} \big(\mathbf{E}_{k-1}[\varphi(\theta_k)] - \varphi(\theta_k) \big) \\ S_2(m) & = \frac{1}{\Omega_m} \sum_{k=1}^m \frac{\omega_k}{\delta_k} \big(\varphi(\theta_k) - \varphi(\theta_{k-1}) \big) \\ S_3(m) & = \pi_m^\omega(\mathcal{A}\varphi) - \frac{1}{\Omega_m} \sum_{k=1}^m \frac{\omega_k}{\delta_k}\, \mathbf{E}_{k-1}[\varphi(\theta_k) - \varphi(\theta_{k-1})] \end{array} \right.$$

and prove that each term converges to zero almost surely. For $S_1(m)$, by Lemma 6 it suffices to show that $\sum_{k \geq 1} (\omega_k/\delta_k)^2\, \mathbf{E}\left[\{\mathbf{E}_{k-1}[\varphi(\theta_k)] - \varphi(\theta_k)\}^2 \right] / \Omega_k^2$ is finite. This follows from the bound $\mathbf{E}\left[\big(\mathbf{E}_{k-1}[\varphi(\theta_k)] - \varphi(\theta_k) \big)^2 \right] \lesssim \delta_k$ and the fact that $\sum_{m \geq 0} \omega_m^2/(\Omega_m^2\, \delta_m)$ is finite. For $S_2(m)$, we can write it as

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

Because $\Omega_m \to \infty$, $(\omega_{m+1}/\delta_{m+1})/\Omega_m \to 0$ and φ is bounded, one can concentrate on proving that $\Omega_m^{-1} \sum_{k=1}^m \varphi(\theta_k)\,\Delta(\omega_k/\delta_k)$ converges almost surely to zero. By Lemma 6, it suffices to verify that $\sum_{k\geq 1} \mathbf{E}\left[|\varphi(\theta_k)\,\Delta(\omega_k/\delta_k)|\right]/\Omega_k$ is finite; this directly follows from the boundedness of φ and Assumption 2. Finally, algebra shows that $S_3(m) = \Omega_m^{-1} \sum_1^m (\omega_k/\delta_k)\, R_{k-1}$ with the quantity R_k defined in Equation (22). It has been proved that there is a constant C such that, almost surely, $|R_k| \leq C\,\delta_{k+1}^{3/2}$ for all $k \geq 0$. Since $\delta_m \to 0$, the rescaled sum $\Omega_m^{-1} \sum_{k\leq m} \omega_k \delta_k^{1/2}$ converges to zero as $m \to \infty$. It follows that $S_3(m)$ converges almost surely to zero. ■

5. Fluctuations, Bias-Variance Analysis, and Central Limit Theorem

The previous section shows that, under suitable conditions, for a test function $\varphi : \mathbb{R}^d \to \mathbb{R}$ the quantity $\pi_m(\varphi)$ converges almost surely to $\pi(\varphi)$ as $m \to \infty$. In this section, we investigate the fluctuations of $\pi_m(\varphi)$ around its asymptotic value $\pi(\varphi)$. We establish that the asymptotic bias-variance decomposition of the SGLD algorithm is dictated by the behaviour of the sequence

$$\mathbb{B}_m \stackrel{\text{def}}{=} T_m^{-1/2} \sum_{k=0}^{m-1} \delta_{k+1}^2. \tag{23}$$

Indeed, the proof of Theorem 8 reveals that the fluctuations of $\pi_m(\varphi)$ are of order $\mathcal{O}\Big(T_m^{-1/2}\Big)$ and its bias is of order $\mathcal{O}\Big(T_m^{-1} \sum_{k=0}^{m-1} \delta_{k+1}^2\Big)$; the quantity $\mathbb{B}_m$ is thus the ratio of the typical scales of the bias and fluctuations. In the case where $\mathbb{B}_m \to 0$, the fluctuations dominate the bias and the rescaled difference $T_m^{1/2} \times (\pi_m(\varphi) - \pi(\varphi))$ converges weakly to a centred Gaussian distribution. In the case where $\mathbb{B}_m \to \mathbb{B}_\infty \in (0, \infty)$, there is an exact balance between the scale of the bias and the scale of the fluctuations; the rescaled quantity $T_m^{1/2} \times (\pi_m(\varphi) - \pi(\varphi))$ converges to a non-centred Gaussian distribution. Finally, in the case where $\mathbb{B}_m \to \infty$, the bias dominates and the rescaled quantity $\big(T_m^{-1} \sum_{k=1}^m \delta_k^2\big)^{-1} \times (\pi_m(\varphi) - \pi(\varphi))$ converges in probability to a quantity $\mu(\varphi) \in \mathbb{R}$ whose exact value is described in the sequel. The strategy of the proof is standard; the solution h of the Poisson equation

$$\varphi - \pi(\varphi) = \mathcal{A}h \tag{24}$$

is introduced so that the additive functional $\pi_m(\varphi)$ of the trajectory of the Markov process $\{\theta_k\}_{k\geq 0}$ can be expressed as the sum of a martingale and a remainder term. A central limit for martingales can then be invoked to describe the asymptotic behaviour of the fluctuations

Theorem 8 (Fluctuations) *Let the step-sizes $(\delta_m)_{m\geq 1}$ satisfy Assumption 1 and assume that Assumption 4 holds for an exponent $p_H \geq 5$. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a test function and assume that the unique solution $h : \mathbb{R}^d \to \mathbb{R}$ to the Poisson Equation (24) satisfies $\|\nabla^n h(\theta)\| \lesssim V^{p_H}(\theta)$ for $n \leq 4$ and has a bounded fifth derivative. Define $\sigma^2(\varphi) = \pi\Big(\|\nabla h\|^2\Big)$.*

- *In case the fluctuations dominate, i.e.* $\mathbb{B}_m \to 0$*, the following convergence in distribution holds,*

$$\lim_{m\to\infty} T_m^{1/2} \left\{ \pi_m(\varphi) - \pi(\varphi) \right\} \;=\; \mathrm{N}\left(0, \sigma^2(\varphi)\right). \tag{25}$$

- *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,*

$$\lim_{m\to\infty} T_m^{1/2} \left\{ \pi_m(\varphi) - \pi(\varphi) \right\} \;=\; \mathrm{N}\left(\mu(\varphi), \sigma^2(\varphi)\right), \tag{26}$$

with the asymptotic bias

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

where the random variables $\Theta \overset{\mathcal{D}}{\sim} \pi$ *and* $\mathcal{U}$ *are independent.*

- *In case the bias dominates, i.e.* $\mathbb{B}_m \to \infty$*, the following limit holds in probability,*

$$\lim_{m\to\infty} \frac{\pi_m(\varphi) - \pi(\varphi)}{T_m^{-1} \sum_{k=1}^m \delta_k^2} \;=\; \mu(\varphi). \tag{27}$$

Proof The proof follows the strategy described in Lamberton and Pages (2002), with the additional difficulty that only unbiased estimates of the drift term of the Langevin diffusion are available. We use the decomposition

$$\pi_m(\varphi) - \pi(\varphi) = \left\{ \frac{\sum_{k=0}^{m-1} \delta_{k+1} \, \mathcal{A}h(\theta_k) - (h(\theta_{k+1}) - h(\theta_k))}{T_m} \right\} + \left\{ \frac{h(\theta_m) - h(\theta_0)}{T_m} \right\}. \tag{28}$$

A fifth order Taylor expansion and Equation (7) yields that

$$h(\theta_{k+1}) - h(\theta_k) = \sum_{n=1}^{4} \left\{ \sum_{i=0}^{n} \mathcal{C}_{n,i}^{(k)} \, \delta_{k+1}^{(n+i)/2} \right\} + \nabla^5 h(\xi_k) \left(\theta_{k+1} - \theta_k\right)^5 / 5!. \tag{29}$$

In the above, we have defined $\mathcal{C}_{n,i}^{(k)} \equiv \left(2^i \, i! \, (n-i)!\right)^{-1} \nabla^n h(\theta_k) \widehat{\nabla \log \pi}(\theta_k, \mathcal{U}_{k+1})^i \eta_{k+1}^{n-i}$; the quantity ξ_k lies between θ_k and θ_{k+1}. It follows from the expression (2) of the generator of the $\mathcal{A}$ of the Langevin diffusion (1) and decomposition (28) that $\pi_m(\varphi) - \pi(\varphi) = \mathscr{F}_m + \mathscr{B}_m + \mathscr{R}_m$ where the fluctuation and bias terms are given by

$$\mathscr{F}_m \equiv -\frac{1}{T_m} \sum_{k=0}^{m-1} \mathcal{C}_{1,0}^{(k)} \delta_{k+1}^{1/2} \quad \text{and} \quad \mathscr{B}_m \equiv -\frac{1}{T_m} \sum_{k=0}^{m-1} \left\{ \mathcal{C}_{2,2}^{(k)} + \mathcal{C}_{3,1}^{(k)} + \mathcal{C}_{4,0}^{(k)} \right\} \delta_{k+1}^2$$

while the remainder term reads

$$\begin{aligned} \mathscr{R}_m \equiv & -\frac{1}{T_m} \sum_{k=0}^{m-1} \left\{ \frac{1}{2} H(\theta_k, \mathcal{U}_{k+1}) \nabla h(\theta_k) + \frac{1}{2} \left(\eta_{k+1}^2 - 1\right) \nabla^2 h(\theta_k) \right\} \delta_{k+1} \\ & - \frac{1}{T_m} \sum_{k=0}^{m-1} \left\{ \sum_{(n,i)\in\mathcal{I}_{\mathscr{R}}} \mathcal{C}_{n,i}^{(k)} \, \delta_{k+1}^{(n+i)/2} \right\} - \frac{1}{T_m} \sum_{k=0}^{m-1} \nabla^5 h(\xi_k) \left(\theta_{k+1} - \theta_k\right)^5 / 5! \\ & + \left\{ \frac{h(\theta_m) - h(\theta_0)}{T_m} \right\} \end{aligned} \tag{30}$$

for $\mathcal{I}_{\mathscr{R}} = \bigcup_{p\in\{3,5,6,7,8\}} \mathcal{I}_{\mathscr{R},p}$ and $\mathcal{I}_{\mathscr{R},p} \equiv \{(n,i) \in [1:4]\times[0:4] \,:\, i \leq n,\, i+n=p\}$. We will show that the remainder term is negligible in the sense that each term on the R.H.S of Equation (30), when multiplied by either $T_m^{1/2}$ or $T_m(\sum_{k=0}^{m-1}\delta_{k+1}^2)^{-1}$, converges in probability to zero; in other words, each one of these terms is dominated asymptotically by either the fluctuations or the bias and is thus negligible. We then show that when multiplied by $T_m^{1/2}$, the fluctuation term converges in distribution to $\mathrm{N}(0,\sigma^2(\varphi))$. 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.

Remainder term: we start by proving that the term $\mathscr{R}_m$ is negligible. The term $\{h(\theta_m)-h(\theta_0)\}/T_m^{1/2}$ converges to zero in probability because $|h(\theta)| \lesssim V^{p_H}(\theta)$ and Lemma 5 shows that $\sup_{m\geq 0}\mathbf{E}[V^{p_H}(\theta_m)]$ is almost surely finite. Similarly, Assumptions 1 and 4 and Lemma 5 yield that

$$\mathbf{E}\left[\nabla^5 h(\xi_k)\,(\theta_{k+1}-\theta_k)^5\right] \lesssim \mathbf{E}\left[|\eta_{k+1}|^5\right]\,\delta_{k+1}^{5/2} + \mathbf{E}\left[\left|\widehat{\nabla\log\pi}(\theta_k,\mathcal{U}_{k+1})\right|^5\right]\delta_{k+1}^5 \lesssim \delta_{k+1}^{5/2}$$

from which it follows that $\left\{\sum_{k=0}^{m-1}\nabla^5 h(\xi_k)\,(\theta_{k+1}-\theta_k)^5\right\}/\left\{\sum_{k=0}^{m-1}\delta_{k+1}^2\right\}$ converges to zero in probability; we have exploited the fact that $\nabla^5 h$ is assumed to be globally bounded. Essentially the same argument yield that the high-order terms are asymptotically negligible: for $(n,i)\in\mathcal{I}_{\mathscr{R},p}$ and $p\in\{5,6,7,8\}$ the limit

$$\lim_{m\to\infty}\ \frac{\sum_{k=0}^{m-1}\mathcal{C}_{n,i}^{(k)}\,\delta_{k+1}^{(n+i)/2}}{\sum_{k=0}^{m-1}\delta_{k+1}^2} \;=\; 0$$

holds in probability because the coefficients $\mathcal{C}_{n,i}^{(k)}$ are uniformly bounded in expectation and the quantity $\left(\sum_{k=0}^{m-1}\delta_{k+1}^{(n+i)/2}\right)/\left(\sum_{k=0}^{m-1}\delta_{k+1}^2\right)$ converges to zero since $(n+i)/2 \geq 5/2$ and $\delta_k \to 0$. To conclude, one needs to verify that the low order terms are also negligible in the sense that the limit

$$\lim_{m\to\infty}\ \frac{\sum_{k=0}^{m-1}X_{n,i}^{(k)}\delta_{k+1}^{(n+i)/2}}{T_m^{1/2}} = 0$$

holds in probability with $X_{1,1}^{(k)} = \nabla h(\theta_k)\,H(\theta_k,\mathcal{U}_{k+1})$ and $X_{2,0}^{(k)} = \nabla^2 h(\theta_k)(\eta_{k+1}^2-1)$ and $X_{2,1}^{(k)} = -\mathcal{C}_{2,1}^{(k)}$ and $X_{3,0}^{(k)} = -\mathcal{C}_{3,0}^{(k)}$. Since $\mathbf{E}\left[X_{n,i}^{(k)}\,\middle|\,\mathcal{F}_k\right] = 0$ where $\mathcal{F}_k = \sigma(\theta_0,\ldots,\theta_k)$ is the natural filtration associated to the process $(\theta_k)_{k\geq 0}$ it follows that

$$\mathbf{E}\left[\left(\frac{\sum_{k=0}^{m-1}X_{n,i}^{(k)}\delta_{k+1}^{(n+i)/2}}{T_m^{1/2}}\right)^2\right] = \frac{\sum_{k=0}^{m-1}\mathbf{E}\left[(X_{n,i}^{(k)})^2\right]\delta_{k+1}^{n+i}}{T_m} \lesssim \frac{\sum_{k=0}^{m-1}\delta_{k+1}^{n+i}}{T_m} \to 0.$$

We made use of the fact that the expectations $\mathbf{E}\left[(X_{n,i}^{(k)})^2\right]$ are uniformly bounded for all $k\geq 0$ by the same arguments as above, and that the final expression converges to 0 since

$n+i \geq 2$, $\delta_m \to 0$ and $T_m \to \infty$. This concludes the proof that the remainder term $\mathscr{R}_m$ is asymptotically negligible.

Fluctuation term: we now prove that the fluctuations term converges in distribution at Monte-Carlo rate towards a Gaussian distribution,

$$T_m^{1/2}\,\mathscr{F}_m \equiv -\frac{\sum_{k=0}^{m-1} \nabla h(\theta_k)\,\delta_{k+1}^{1/2}\,\eta_{k+1}}{T_m^{1/2}} \to \mathrm{N}\left(0, \sigma^2(\varphi)\right).$$

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,

$$\lim_{m\to\infty} \sum_{k=0}^{m-1} \frac{\mathbf{E}_k\left[Z_k^2\,\mathbb{I}\big(Z_k^2 > T_m\varepsilon\big)\right]}{T_m} = 0 \quad \text{and} \quad \lim_{m\to\infty} \frac{\sum_{k=0}^{m-1} \mathbf{E}_k\left[Z_k^2\right]}{T_m} = \sigma^2(\varphi)$$

with $Z_k \stackrel{\text{def}}{=} \nabla h(\theta_k)\,\delta_{k+1}^{1/2}\,\eta_{k+1}$. Since $\mathbf{E}_k\left[Z_k^2\right] = \nabla h(\theta_k)^2\,\delta_{k+1}$ and the function $\theta \mapsto \nabla h(\theta)^2$ satisfies the assumptions of Theorem 7, the second limit directly follows from Theorem 7. For proving the first limit, note that the Cauchy-Schwarz's inequality and the boundedness of ∇h imply that $\mathbf{E}_k\left[Z_k^2\,\mathbb{I}\big(Z_k^2 > T_m\varepsilon\big)\right] \lesssim \delta_{k+1} \times \mathbf{P}\left[\delta_{k+1}\,\|\nabla h\|_\infty^2\,\eta_{k+1}^2 > T_m\,\varepsilon\right]^{1/2}$; the Markov's inequality thus yields that

$$\sum_{k=0}^{m-1} \mathbf{E}_k\left[Z_k^2\,I\big(Z_k^2 > T_m\varepsilon\big)\right]/T_m \lesssim \frac{\sum_{k=0}^{m-1}\delta_{k+1}^2}{T_m^2\,\varepsilon}.$$

Since $T_m^{-2}\,\sum_{k=0}^{m-1}\delta_{k+1}^2 \to 0$, the conclusion follows.

Bias term: we conclude by proving that the bias term is such that the limit

$$\lim_{m\to\infty}\ \frac{\mathscr{B}_m}{\sum_{k=1}^m \delta_k^2/T_m}\ =\ \mu(\varphi)$$

holds in probability. The quantity $\mathscr{B}_m/\big(\sum_{k=1}^m \delta_k^2/T_m\big)^{-1}$ can also be expressed as

$$\frac{\sum_{k=0}^{m-1}\Psi(\theta_k)\,\delta_{k+1}^2}{\sum_{k=0}^{m-1}\delta_{k+1}^2} + \frac{\sum_{k=0}^{m-1}\Delta M_k\,\delta_{k+1}^2}{\sum_{k=0}^{m-1}\delta_{k+1}^2} \tag{31}$$

for a martingale difference term $\Delta M_k \equiv \left(\mathcal{C}_{2,2}^{(k)} + \mathcal{C}_{3,1}^{(k)} + \mathcal{C}_{4,0}^{(k)}\right) - \Psi(\theta_k)$ where $\Psi(\theta_k) \equiv \mathbf{E}\left[\mathcal{C}_{2,2}^{(k)} + \mathcal{C}_{3,1}^{(k)} + \mathcal{C}_{4,0}^{(k)}\,\middle|\,\mathcal{F}_k\right]$ and $\left(\mathcal{C}_{2,2}^{(k)} + \mathcal{C}_{3,1}^{(k)} + \mathcal{C}_{4,0}^{(k)}\right)$ equals

$$\frac{1}{8}\nabla^2 h(\theta_k)\widehat{\nabla\log\pi}(\theta_k,\mathcal{U}_{k+1})^2 + \frac{1}{4}\nabla^3 h(\theta_k)\widehat{\nabla\log\pi}(\theta_k,\mathcal{U}_{k+1})\eta_{k+1}^2 \quad + \frac{1}{24}\nabla^4 h(\theta_k)\eta_{k+1}^4.$$

Under the assumptions of Theorem 8, the function Ψ satisfies the hypothesis of Theorem 7 applied to the weight sequence $\{\delta_k^2\}_{k\geq 0}$; it follows that the first term in Equation (31) converge almost surely to $\mu(\varphi)$. It remains to prove that the second term in Equation (31) also converges almost surely to zero. By Lemma 6, it suffices to prove that the martingale

$$m \mapsto \sum_{k=0}^{m} \frac{\Delta M_k \, \delta_{k+1}^2}{\sum_{j=1}^{k+1} \delta_{j+1}^2}$$

is bounded in L^2. Under the Assumption of Theorem 8, Lemma 5 yields that the martingale difference term ΔM_k is uniformly bounded in L^2 from which the conclusion readily follows. ■

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 $m^{-1/3}$.

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 $\mathcal{I} : \mathcal{C}([0,T], \mathbb{R}^d) \to \mathcal{C}([0,T], \mathbb{R}^d)$, which sends a continuous path $w \in \mathcal{C}([0,T], \mathbb{R}^d)$ to the unique solution $v = \mathcal{I}(w)$ of the integral equation,

$$v_t = \theta_0 + \frac{1}{2} \int_{s=0}^{t} \nabla \log \pi(v_s) \, ds + w_t \qquad \text{for all} \quad t \in [0,T].$$

If the drift function $\theta \mapsto \frac{1}{2} \nabla \log \pi(\theta)$ is globally Lipschitz, then the Itô's map $\mathcal{I}$ is well defined and continuous. Further, the image $\mathcal{I}(W)$ under the Itô map of a standard Brownian motion W on $[0,T]$ can be seen to be described by Langevin diffusion (1).

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 $\delta^{(r)} \stackrel{\text{def}}{=} (\delta_k^{(r)})_{k=1}^{m(r)}$ such that

$$\sum_{k=1}^{m(r)} \delta_k^{(r)} = T$$

and with increasingly fine mesh sizes $\mathrm{mesh}(\delta^{(r)}) \to 0$ with

$$\mathrm{mesh}(\delta^{(r)}) \stackrel{\text{def}}{=} \max \Big\{ \delta_k^{(r)} \ : \ 1 \leq k \leq m(r) \Big\}.$$

Define $T_0^{(r)} = 0$ and $T_k^{(r)} = \delta_1^{(r)} + \cdots + \delta_k^{(r)}$ for each $k \geq 1$. The Markov chains are coupled to W as follows:

$$\left\{ \begin{array}{ll} \eta_k^{(r)} & = (\delta_k^{(r)})^{-1/2} \Big(W(T_k^{(r)}) - W(T_{k-1}^{(r)}) \Big) \\ \theta_k^{(r)} & = \theta_{k-1}^{(r)} + \frac{1}{2}\, \delta_k^{(r)} \Big\{ \nabla \log \pi(\theta_{k-1}^{(r)}) + H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)}) \Big\} \, + \, (\delta_k^{(r)})^{1/2}\, \eta_k^{(r)}, \end{array} \right. \tag{32}$$

for an i.i.d. collection of auxiliary random variables $(\mathcal{U}_k^{(r)})_{r \geq 1, k \geq 1}$. Note that $(\eta_k^{(r)})_{k \geq 1}$ form an i.i.d. sequence of $\mathrm{N}(0,1)$ variables for each r. We can construct piecewise affine continuous time sample paths $(S^{(r)})_{r \geq 1}$ by linearly interpolating the Markov chains,

$$S^{(r)} \Big(x T_{k-1}^{(r)} + (1-x) T_k^{(r)} \Big) = x\, \theta_{k-1}^{(r)} + (1-x)\, \theta_k^{(r)}, \tag{33}$$

for $x \in [0,1]$. The approach then amounts to showing that each $S^{(r)}$ can be expressed as $\mathcal{I}(\widetilde{W}^{(r)}) + e^{(r)}$, where $\widetilde{W}^{(r)}$ 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 9 *Let Assumption 4 holds and suppose that the drift function $\theta \mapsto (1/2) \nabla \log \pi(\theta)$ is globally Lipschitz on $\mathbb{R}^d$. If $\mathrm{mesh}(\delta^{(r)}) \to 0$ as $r \to \infty$, then the sequence of continuous time processes $(S^{(r)})_{r \geq 1}$ defined in Equation (33) converges weakly on $\big(\mathcal{C}([0,T], \mathbb{R}^d), \| \cdot \|_\infty \big)$ to the Langevin diffusion (1) started at $S_0 = \theta_0$.*

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 (1) 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 $\mathcal{I}(\widetilde{W}^{(r)}) + e^{(r)}$ where $\widetilde{W}^{(k)}$ 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 $\widetilde{W}^{(r)}$ as the continuous piecewise affine processes that satisfies $\widetilde{W}^{(r)}(T_k^{(r)}) = W(T_k^{(r)})$ for all $0 \leq k \leq m(r)$ and that is affine in between. It follows that for

any time $T_{k-1}^{(r)} \leq t \leq T_k^{(r)}$ we have

$$\begin{aligned} S^{(r)}(t) &= S^{(r)}(T_{k-1}^{(r)}) + \left(\int_{T_{k-1}^{(r)}}^{t} \frac{1}{2} \nabla \log \pi\big(S^{(r)}(T_{k-1}^{(r)})\big) du + \widetilde{W}^{(r)}(t) - \widetilde{W}(T_{k-1}^{(r)}) \right) \\ &\quad + \frac{1}{2} \int_{T_{k-1}^{(r)}}^{t} H\big(S^{(r)}(T_{k-1}^{(r)}), \mathcal{U}_k^{(r)}\big) du \\ &= \underbrace{\theta_0 + \left(\int_0^t \frac{1}{2} \nabla \log \pi\big(S^{(r)}(u)\big) du + \widetilde{W}^{(r)}(t) \right)}_{\mathcal{I}(\widetilde{W})(t)} \\ &\quad + \underbrace{\int_0^t \frac{1}{2} \left(\nabla \log \pi\big(\widehat{S}^{(r)}(u)\big) - \nabla \log \pi\big(S^{(r)}(u)\big) \right) du}_{e_1^{(r)}(t)} \\ &\quad + \frac{1}{2} \underbrace{\int_0^t H\big(\widehat{S}^{(r)}(u), \mathcal{U}_k^{(r)}\big) du}_{e_2^{(r)}(t)}, \end{aligned}$$

where $\widehat{S}^{(r)}$ is a piecewise constant (non-continuous) process, $\widehat{S}^{(r)}(t) = S^{(r)}(T_{k-1}^{(r)}) = \theta_{k-1}^{(r)}$ for $t \in [T_{k-1}^{(r)}, T_k^{(r)})$. The process $S^{(r)}$ can thus be expressed as the sum $\mathcal{I}(W^{(r)}) + e_1^{(r)} + e_2^{(r)}$. Since the mesh-size of the partition $\delta^{(r)}$ converges to zero as $r \to \infty$, standard properties of Brownian motions yield that $\widetilde{W}^{(r)}$ converges weakly in $\big(\mathcal{C}([0,t], \mathbb{R}^d), \|\cdot\|_{\infty,[0,T]}\big)$ 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 $\mathbf{E}\left[\|e_2^{(r)}\|_\infty^2\right] \to 0$ in probability, we have,

$$\begin{aligned} \mathbf{E}\left[\|e_2^{(r)}\|_\infty^2\right] &\leq 4\, \mathbf{E}\left[\|e_2^{(r)}(T)\|^2\right] = 4 \sum_{k=1}^{m(r)} \big(\delta_k^{(r)}\big)^2\, \mathbf{E}\left[H\Big(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)}\Big)^2\right] \\ &\lesssim \sum_{k=1}^{m(r)} \big(\delta_k^{(r)}\big)^2\, \mathbf{E}\left[V(\theta_{k-1}^{(r)})\right] \leq \mathrm{mesh}(\delta^{(r)}) \sum_{k=1}^{m(r)} \delta_k^{(r)}\, \mathbf{E}\left[V(\theta_{k-1}^{(r)})\right] \\ &\leq \mathrm{mesh}(\delta^{(r)}) \times T \times \sup\left\{ \mathbf{E}\left[V(\theta_{k-1}^{(r)})\right] \,:\, r \geq 1, 1 \leq k \leq m(r) \right\} \lesssim \mathrm{mesh}(\delta^{(r)}). \end{aligned}$$

We have used Doob's martingal inequality, Assumption 4 and Lemma 5. Since $\mathrm{mesh}(\delta^{(r)})$ converges to zero, the conclusion follows. To prove $\mathbf{E}\left[\|e_1^{(r)}\|_\infty\right] \to 0$ in probability, we use Equation (32) and note that since the drift function $\theta \mapsto \frac{1}{2}\nabla \log \pi(\theta)$ is globally Lipschitz, for each $T_{k-1}^{(r)} \leq u \leq T_k^{(r)}$ we have,

$$\begin{aligned} \left\| \nabla \log\big(\widehat{S}^{(r)}(u)\big) - \nabla \log\big(S^{(r)}(u)\big) \right\| &\lesssim \left\| \theta_k^{(r)} - \theta_{k-1}^{(r)} \right\| \\ &\lesssim \|\nabla \log \pi(\theta_{k-1}^{(r)})\|\, \delta_k^{(r)} + \|H\big(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)}\big)\|\, \delta_k^{(r)} + \sqrt{\delta_k^{(r)}}\, \|\eta_k^{(r)}\|. \end{aligned}$$

It follows that

$$\mathbf{E}\left[\|e_1^{(r)}\|_\infty\right] \lesssim \sum_{k=1}^{m(r)} \delta_k^{(r)} \left(\|\nabla \log \pi(\theta_k^{(r)})\|\, \delta_k^{(r)} + \|H\big(\theta_k^{(r)}, \mathcal{U}_k\big)\|\, \delta_k^{(r)} + \sqrt{\delta_k^{(r)}}\, \|\eta_k^{(r)}\| \right).$$

Since $\mathrm{mesh}(\delta^{(r)})$ converges to zero and by Assumption 4 and Lemma 5 the suprema

$$\begin{cases} \sup & \left\{ \mathbf{E}\left[\|\nabla \log \pi(\theta_k^{(r)})\|\right] \,:\, r \geq 1, 1 \leq k \leq m(r) \right\}, \\ \sup & \left\{ \mathbf{E}\left[\|H\big(\theta_k^{(r)}, \mathcal{U}_k\big)\|\right] \,:\, r \geq 1, 1 \leq k \leq m(r) \right\} \end{cases}$$

are finite, it readily follows 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. The main purpose of this article is to establish the missing theoretical foundation of stochastic gradient methods for the approximation of expectations. For more exhaustive simulation studies we refer to Welling and Teh (2011); S. Ahn and Welling (2012); Patterson and Teh (2013a); Chen et al. (2014). By considering a logistic regression model, we demonstrate that the SGLD can be advantageous over the Metropolis-Adjusted-Langevin (MALA) algorithm if the available computational budget only allows a few iterations through the whole data set, see Section 7.2.2.

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

$$x_i \mid \theta \;\sim\; \mathrm{N}(\theta, \sigma_x^2).$$

We use a Gaussian prior $\theta \sim \mathrm{N}(0, \sigma_\theta^2)$ and assume that the variance hyper-parameters σ_θ^2 and σ_x^2 are both known. The posterior density $\pi(\theta)$ is normally distributed with mean μ_p and variance σ_p^2 given by

$$\mu_p = \bar{x}\left(1 + \frac{\sigma_x^2}{N\sigma_\theta^2}\right)^{-1} \qquad \text{and} \qquad \sigma_p^2 = \frac{\sigma_x^2}{N}\left(1 + \frac{\sigma_x^2}{N\sigma_\theta^2}\right)^{-1}$$

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

$$\nabla \log \pi(\theta) = -\frac{\theta - \mu_p}{\sigma_p^2} \quad \text{and} \quad H(\theta, \mathcal{U}) = \Big\{ (N/n) \sum_{j \in \mathcal{I}_n(\mathcal{U})} x_j - \sum_{1 \leq i \leq N} x_i \Big\} / \sigma_x^2$$

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

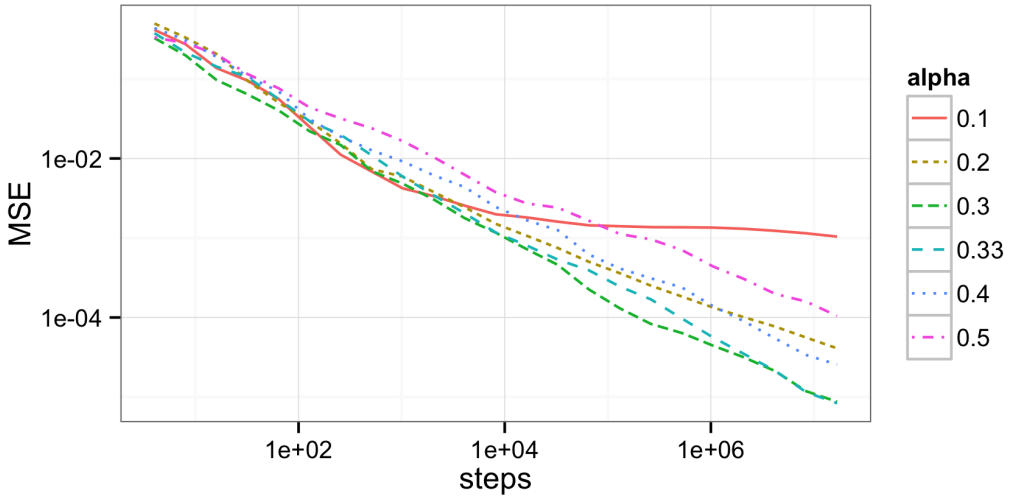

Figure 1: Decay of the MSE for step sizes $\delta_m \asymp m^{-\alpha}$, $\alpha \in \{0.1, 0.2, 0.3, 0.33, 0.4, 0.5\}$. The MSE decays algebraically for all step sizes, with fastest decay at approximately $\alpha = 0.33$.

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 $H(\theta, \mathcal{U})$ 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

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

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 $\mathcal{A}\varphi$ where $\varphi = \sin(x - \mu_p - 0.5\sigma_p)$.

We are interested in confirming the asymptotic convergence regimes of Theorem 8 by running SGLD with a range of step sizes, and plotting the mean squared error (MSE) achieved by the estimate $\pi_m(\mathcal{A}\varphi)$ 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 $\alpha \in$

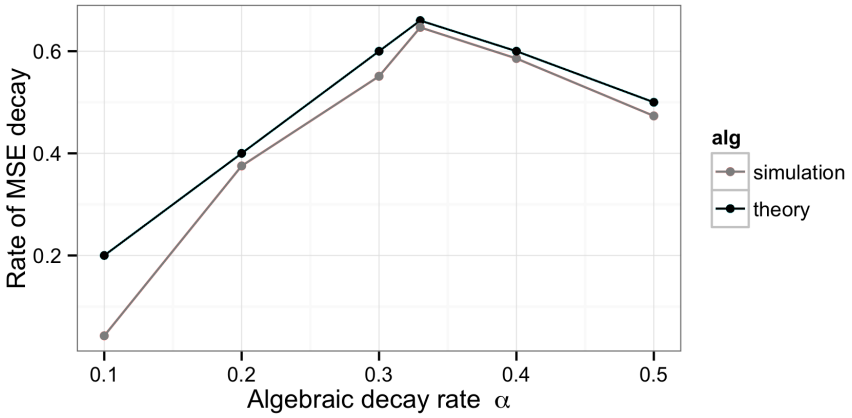

Figure 2: Rates of decay of the MSE obtained from estimating the asymptotic slopes of the plots in Figure 1, compared to theoretical findings of Theorem 8. The fastest convergence rate is achieved at $\alpha = 1/3$.

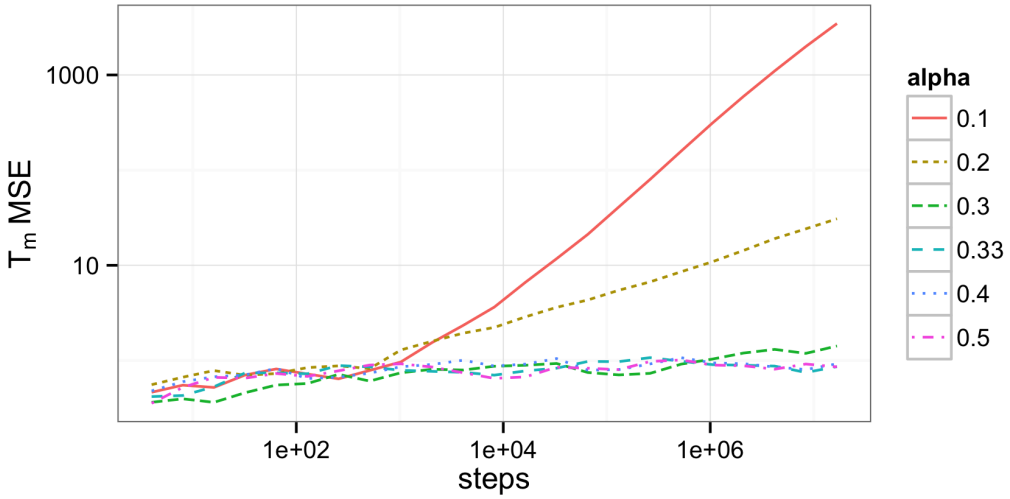

Figure 3: Plots of the MSE multiplied by T_m against the number of steps m. The plots are flat for $\alpha \geq 0.33$, demonstrating that the MSE scales as T_m^{-1} in this regime, while the plots diverge for $\alpha < 0.33$, demonstrating that it decays at a slower rate here.

$\{0.1, 0.2, 0.3, 0.33, 0.4, 0.5\}$ where $m_0(\alpha)$ is chosen such that δ_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 $\alpha_\star = 1/3$. To be more precise, we estimate

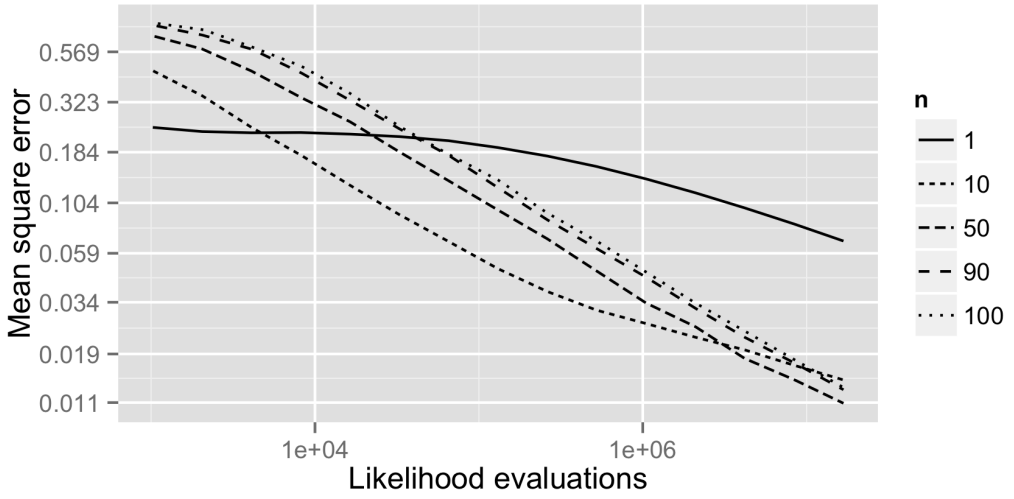

Figure 4: Behaviour of the mean squared error for different subsample sizes n.

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 8, 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

$$\mathbb{P}(y_i = 1 \mid x_i, \theta) \;=\; 1 - \mathbb{P}(y_i = -1 \mid x_i, \theta) \;=\; \mathrm{logit}\big(\langle \theta, x_i \rangle\big) \tag{34}$$

for covariate $x_i \in \mathbb{R}^d$, unknown parameter $\theta \in \mathbb{R}^d$ and function $\mathrm{logit}(z) = e^z/(1+e^z)$. 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

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

$$H(\theta, \mathcal{U}) = (N/n) \sum_{j \in \mathcal{I}_n(U)} \mathrm{logit}\big(-y_j \langle \theta, x_j \rangle\big)\, y_j\, x_j - \sum_{1 \le i \le N} \mathrm{logit}\big(-y_i \langle \theta, x_i \rangle\big)\, y_i\, x_i$$

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

7.2.1 Verification of Assumption 4

We verify in this section that Assumption (4) is satisfied for the Lyapunov function $V(\theta) = 1 + \|\theta\|^2$. Since $H(\theta,\mathcal{U})$ 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,

$$\begin{aligned}\left\langle \nabla V(\theta), \frac{1}{2}\nabla \log \pi(\theta) \right\rangle &= -\frac{1}{2}\langle \theta, C^{-1}\theta\rangle + \frac{1}{2}\sum_{i=1}^{N} \operatorname{logit}\big(-y_i\langle\theta, x_i\rangle\big)\, y_i\, \langle \theta, x_i\rangle \\ &\leq -\frac{\lambda_{\min}}{2}\|\theta\|^2 + \frac{\sum_{i=1}^{N}\|x_i\|}{2}\|\theta\| \leq -\frac{\lambda_{\min}}{4}V(\theta) + \beta\end{aligned}$$

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

7.2.2 Comparison of the SGLD and the MALA for logistic regression

We consider a simulated dataset where $d = 3$ and $N = 1000$. We set the input covariates $x_i = (x_{i,1}, x_{i,2}, 1)$ with $x_{i,1}, x_{i,2} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$ for $i = 1 \dots N$, and use a Gaussian prior $\theta \sim \mathrm{N}(0, I)$. We draw a $\theta_0 \sim \mathrm{N}(0, I)$ and based on it we generate y_i according to the model probabilities (34). In the following we compare MALA in SGLD by comparing their estimate for the variance of the first component.

The findings of this article show that SGLD-based expectation estimates converge at a slower rate of at most $n^{-\frac{1}{3}}$ compared to the standard rate of $n^{-\frac{1}{2}}$ for standard MCMC algorithms such as the MALA algorithm. In the following we demonstrate that in the non-asymptotic regime (allowing only a few passes through the data set) the SGLD can be advantageous. We start both algorithms at the MAP estimator and we ensure that this study is not biased due to different speeds in finding the mode of the posterior. For a fair comparison we tune the MALA to an acceptance rate of approximately 0.564 following the findings of Roberts and Rosenthal (1998). For the SGLD-based variance estimate of the first component for $n = 30$ we choose $\delta_m = (a \cdot m + b)^{-0.38}$ as step sizes and optimise over the choices of a and b. This is achieved by estimating the MSE for choices of a and b on a log-scale grid based on 512 independent runs. The estimates based on 20 and 1000 effective iterations through the data set the averages are visualised in the heat maps in Figure 5. That means we limit the algorithm to 200 and 1000000 likelihood evaluations, respectively. The figures indicate that the range of the good parameter choices seems to be the same in both cases. Using the heat map for the estimated MSE after 20 iterations through the data set, we pick $a = 5.89 \cdot 10^7$ and $b = 7.90 \cdot 10^8$ and compare the time behaviour of the SGLD and the MALA algorithm in Figure 6. The figure is a simulation evidence that the SGLD algorithm can be advantageous in the initial phase for the first few iterations through the data set. This recommends further investigation as the initial phase can be quite different from the asymptotic phase.

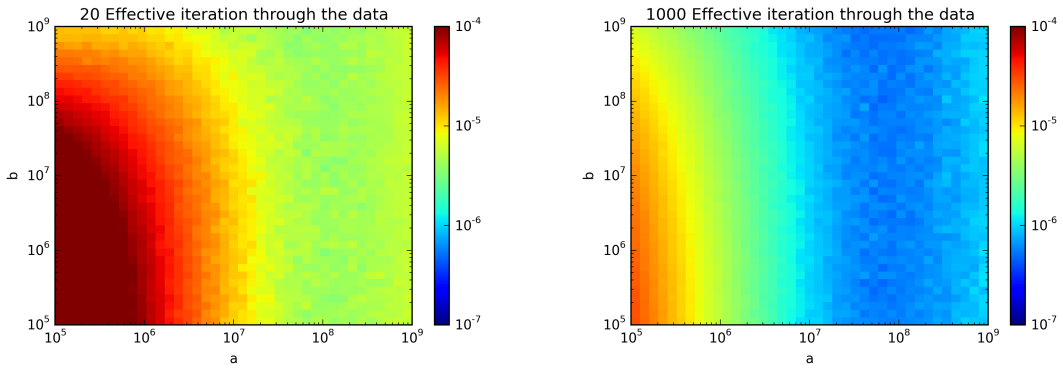

Figure 5: Expected MSE of the SGLD-based estimate variance estimate of the first component for $n = 30$ and step sizes $\delta_m = (a \cdot m + b)^{-0.38}$ after 20 and 1000 iterations through the data set

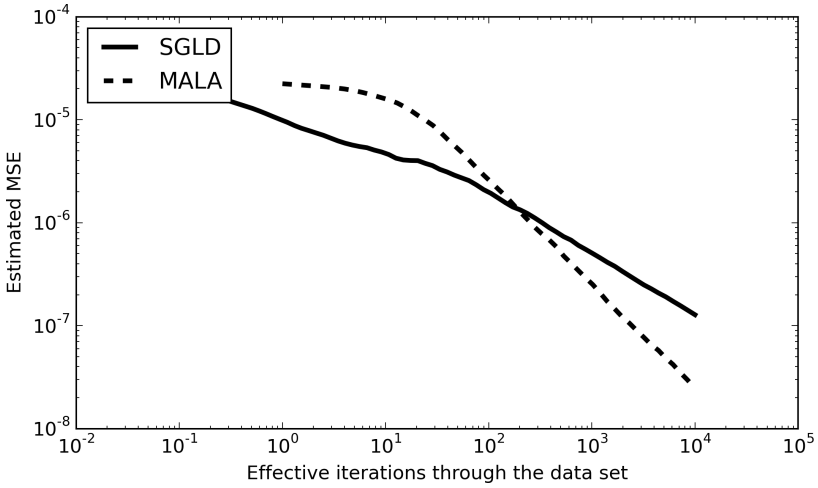

Figure 6: Behaviour of the MSE of estimating the posterior variance of the first component for 3-dimensional logistic regression of MALA and SGLD with tuned parameters

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 discretization 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 Welling and Teh (2011) choice of $\delta_m \asymp m^{-1/2}$.

Our theory suggests that an optimally tuned SGLD method converges at rate $\mathcal{O}(m^{-1/3})$, 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 discretization bias due to the lack of a Metropolis-Hastings correction. Another avenue of exploration is to determine more precisely the bias resulting from the discretization 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 6

Recall Kronecker's Lemma (Shiryaev, 1996, Lemma IV.3.2) that states that for a non-decreasing and positive sequence $b_m \to \infty$ and another real valued sequence $(a_m)_{m\geq 0}$ such that the series $\sum_{m\geq 0} a_m/b_m$ converges the following limit holds,

$$\lim_{m\to\infty} \frac{\sum_{k=0}^{m} a_k}{b_m} = 0.$$

For proving Equation (15) it thus suffices to show that the sums $\sum_{k\geq 0} |\Delta M_k| / T_k$ and $\sum_{k\geq 0} |X_k| / T_k$ are almost surely finite. This follows from Condition (16) (L^2 martingale convergence theorem) and Condition (17).

Appendix B. 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

$$\left|(V^p)''(\theta)\right| \lesssim V^{p-1}(\theta) \tag{35}$$

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

$$V^p(\theta+\varepsilon) \lesssim V^p(\theta) + |\varepsilon|^{2p}. \tag{36}$$

Similarly, Definition (7), the bound $\|\nabla \log \mathrm{p}(\theta)\|^2 \lesssim V(\theta)$ and Equation (10) yield that for any exponent $0 \leq p \leq p_H$ the following holds,

$$\mathbf{E}_m[\,|\theta_{m+1} - \theta_m|^{2p}\,] \lesssim \delta_{m+1}^{2p}\, V^p(\theta) + \delta_{m+1}^p. \tag{37}$$

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 (38) below. We then describe how Equation (13) 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

$$\mathbf{E}_m\left[V^p(\theta_{m+1}) - V^p(\theta_m)\right]/\delta_{m+1} \;\leq\; -\alpha_p\, V^p(\theta_m) + \beta_p. \tag{38}$$

  Since for any ε there exists C_ε such that $V^{p-1}(\theta) \leq C_\varepsilon + \varepsilon V^p(\theta)$, for proving (38) it actually suffices to verify that we have

$$\mathbf{E}_m\left[V^p(\theta_{m+1}) - V^p(\theta_m)\right]/\delta_{m+1} \;\leq\; -\widetilde{\alpha}_p\, V^p(\theta_m) + \widetilde{\beta}_p\, V^{p-1}(\theta_m) \tag{39}$$

  for some constants $\widetilde{\alpha_p}, \widetilde{\beta_p} > 0$ and index $m \geq 1$ large enough. A second order Taylor expansion yields that the left hand side of (39) is less than

$$\mathbf{E}_m\left[(V^p)'(\theta_m)\,(\theta_{m+1} - \theta_m)\right]/\delta_{m+1} + \frac{1}{2}\,\mathbf{E}_m\left[(V^p)''(\xi)\,(\theta_{m+1} - \theta_m)^2\right]/\delta_{m+1} \tag{40}$$

  for a random quantity ξ lying between θ_m and θ_{m+1}. Since $\mathbf{E}_m[\theta_{m+1} - \theta_m] = \frac{1}{2}\nabla \log \mathrm{p}(\theta_m)$, the drift condition (12) yields that the first term of (40) is less than

$$p\, V^{p-1}(\theta_m)\,(-\alpha\, V(\theta_m) + \beta) \tag{41}$$

  for $\alpha, \beta > 0$ given by Equation (12). Consequently, for proving Equation (38), it remains to bound the second term of (40). Equation (35) shows that $|(V^p)''(\xi)|$ is upper bounded by a multiple of $|V^{p-1}(\xi)|$; the bound (36) 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 (37) on the difference $(\theta_{m+1} - \theta_m)$ and the assumption $\mathbf{E}[\,\|H(\theta, \mathcal{U})\|^{2p_H}\,] \lesssim V^{p_H}(\theta)$ 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 (39) is less than a constant multiple of

$$\varepsilon\, V^p(\theta_m) + \beta_{p,\varepsilon}\, V^{p-1}(\theta) \tag{42}$$

  for a constant $\beta_{p,\varepsilon} > 0$. Equations (41) and (42) directly yield to Equation (39), which in turn implies to Equation (38).

- **Proof that $\sup_{m\geq 1}\ \mathbf{E}[V^p(\theta_m)] \ < \ \infty$ for any $p \leq p_H$.**
Equations (36) and (37) show that if $\mathbf{E}[V^p(\theta_m)]$ is finite then so is $\mathbf{E}[V^p(\theta_{m+1})]$. Under the conditions of Lemma 5, this shows that $\mathbf{E}[V^p(\theta_m)]$ is finite for any $m \geq 0$. An inductive argument based on the discrete Lyapunov Equation (38) then yields that for any index $m \geq m_0$ the expectation $\mathbf{E}[V^p(\theta_m)]$ is less than

$$\max\Big(\beta_p/\alpha_p, \max\big\{\,\mathbf{E}[V^p(\theta_m)]:\ \ 0 \leq m \leq m_0\big\}\Big). \tag{43}$$

It follows that $\sup_{m\geq 1}\ \mathbf{E}[V^p(\theta_m)]$ 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 (38) yields that $\delta_{k+1}V^p(\theta_k)$ is less than $\delta_{k+1}\,\beta_p/\alpha_p - \mathbf{E}_k[V^p(\theta_{k+1}) - V^p(\theta_k)]/\alpha_p$; 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

$$1 + \frac{V^p(\theta_{m_0})}{T_m} + \frac{1}{T_m}\sum_{k=m_0}^{m}\Big\{V^p(\theta_{k+1}) - \mathbf{E}_k[V^p(\theta_{k+1})]\Big\}.$$

To conclude the proof, we prove that the last term in the above displayed Equation almost surely converges to zero; by Lemma 6, it suffices to prove that the quantity

$$\sum_{k\geq m_0}\mathbf{E}\left[\left|\frac{V^p(\theta_{k+1}) - \mathbf{E}_k[V^p(\theta_{k+1})]}{T_k}\right|^2\right] \tag{44}$$

is almost surely finite. We have $\mathbf{E}\left[|V^p(\theta_{k+1}) - \mathbf{E}_k[V^p(\theta_{k+1})]|^2\right] \leq 2\times\mathbf{E}[|\,V^p(\theta_{k+1}) - V^p(\theta_k)\,|^2]$ and the mean value theorem yields that $|V^p(\theta_{k+1}) - V^p(\theta_k)| \lesssim V^{p-1}(\xi)\,V'(\xi)\,(\theta_{k+1} - \theta_k)$ for some ξ lying between θ_k and θ_{k+1}. The bound $|V'(\theta)| \lesssim V^{1/2}(\theta)$ and Equation (36) then yield that $|V^p(\theta_{k+1}) - V^p(\theta_k)| \lesssim V^{p-1/2}(\theta_k)\,\big|\theta_{k+1} - \theta_k\big| + \big|\theta_{k+1} - \theta_k\big|^{2p}$. From the bound (37) and the assumption that $\mathbf{E}[H(\theta,\mathcal{U})^{2p_H}] \lesssim V^{p_H}(\theta)$ it follows that the quantity in Equation (44) is less than a constant multiple of

$$\sum_{k\geq m_0}\frac{\mathbf{E}\left[\,V^{2p}(\theta_k)\,\right]\times\delta_k}{T^2(k)}.$$

Since $\mathbf{E}\left[V^{2p}(\theta_k)\right]$ 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 $V(\theta) \lesssim 1 + \|\theta\|^2$, the drift condition (12) 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

$$\mathcal{A}V^p(\theta) \leq -\alpha_{p,*}\,V^p(\theta) + \beta_{p,*} \tag{45}$$

where $\mathcal{A}$ 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 $\delta_{k+1}V^p(\theta_k) \lesssim \delta_{k+1}\,\beta_p/\alpha_p - \mathbf{E}_k[V^p(\theta_{k+1}) - V^p(\theta_k)]/\alpha_p$ yields that $\pi_m^\omega(V^p)$ is less than a constant multiple of

$$1 + \frac{(\omega_{m_0}/\delta_{m_0})\,V^p(\theta_{m_0})}{T_m} \quad +\Omega^{-1}(m)\ \sum_{k=m_0+1}^m(\omega_k/\delta_k)\left\{V^p(\theta_{k+1}) - \mathbf{E}_k[V^p(\theta_{k+1})]\right\}$$
$$+\Omega^{-1}(m)\ \sum_{k=m_0}^{m-1}\Delta(\omega_k/\delta_k)\,V^p(\theta_k).$$

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

$$\lim_{m\to\infty}\ \Omega^{-1}(m)\ \sum_{k=m_0+1}^m(\omega_k/\delta_k)\left\{V^p(\theta_{k+1}) - \mathbf{E}_k[V^p(\theta_{k+1})]\right\} = 0 \tag{46}$$
$$\lim_{m\to\infty}\ \ \Omega^{-1}(m)\ \sum_{k=m_0}^{m-1}\Delta(\omega_k/\delta_k)\,V^p(\theta_k) = 0. \tag{47}$$

To prove Equation (46) it suffices to use the assumption that $\sum_{m\geq 0}\omega_m^2/[\delta_m\Omega_m^2] < \infty$ and then follow the same approach used to establish that the quantity (44) is finite. Lemma 6 shows that to prove Equation (47) it suffices to verify that

$$\mathbf{E}\Big[\sum_{m\geq 0}\ \big|\Delta(\omega_m/\delta_m)\big|\,V^p(\theta_m)/\Omega_m\Big] < \infty.$$

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

References

Amr Ahmed, Moahmed Aly, Joseph Gonzalez, Shravan Narayanamurthy, 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.

Rémi Bardenet, Arnaud Doucet, and Chris C. Holmes. Towards scaling up MCMC: an adaptive subsampling 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.

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.

Søren F Jarner and Gareth O Roberts. Convergence of heavy-tailed monte carlo markov chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815, 2007.

Kengo Kamatani. Rate optimality of random walk metropolis algorithm in high-dimension with heavy-tailed target distribution. *arXiv preprint arXiv:1406.5392*, 2014.

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.

Radford. 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.

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 Mathemaitcal 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.

G. O. Roberts and J. S. Rosenthal. Optimal Scaling of Discrete Approximations to Langevin Diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998. ISSN 1369-7412. doi: 10.1111/1467-9868.00123. URL `http://0-dx.doi.org.pugwash.lib.warwick.ac.uk/10.1111/1467-9868.00123`.

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.

A. Koratticara S. Ahn and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *ICML*, 2012.

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.