SGLD Diffusion Limit Critical Review

Wesley Chung and Andrew Cheng

November 2019

In this report, we discuss a paper by Teh et al. on the theoretical properties of the Stochastic Gradient Langevin Dynamics (SGLD) algorithm [1], which was first proposed by Welling and Teh [2]. We start with a brief history of SGLD along with a description of the algorithm. Then, the relevant prerequisites are introduced. Next, we present the theorem stating its diffusion limit and its proof, filling in details when necessary and adding our own remarks. Finally, we summarize our opinions and briefly discuss certain followup works on SGLD.

1 Introduction

In the early 20^{th} century, the French physicist Paul Langevin proposed a new approach in approximating the dynamics of nonequilibrium systems. In particular, he proposed the *Langevin Equation*, a stochastic differential equation of motion which describes the time evolution of a set of degree of freedoms, like Brownian motion, that obeys properties of the Markov process. One can view the equation as tracing the position of a Brownian particle with respect to time as it is subjected to both frictional and random forces.

In another line of research, in the 1950s, researchers were exploring Markov Chain Monte Carlo (MCMC) methods, a generic sampling technique which could be used to generate samples from complicated distributions. A key component of MCMC methods is the proposal distribution, which dictates how the next value is sampled in the Markov chain. While the original Metropolis-Hastings algorithm proposes a step by adding Gaussian noise (with fixed variance) to the current state, it can be inefficient for generating samples since it makes no use of information about the target distribution in its proposal distribution. Better methods for proposing steps in MCMC were pursued.

Inspired by other methods in statistical physics utilizing the Langevin equation, in 1994, Grenander et al. [3] first proposed incorporating Langevin dynamics in MCMC, yielding the Metropolis-Adjusted Langevin Algorithm (MALA), also known as Langevin Monte Carlo. Two years later, the followup work by Roberts et al. [4] put this algorithm on solid theoretical ground and investigated various conditions for convergence of the Langevin diffusion and discrete approximations of it. By incorporating the gradient of the target distribution

into the computation of the proposal distribution, the algorithm would hopefully be more effective than simpler methods.

In the 21^{st} century, with the growing sizes of datasets, classic MCMC methods such as MALA became too expensive to use as they required processing the full dataset at every proposal step. This would motivate Welling and Teh to propose Stochastic Gradient Langevin Dynamics (SGLD) in 2011, which aimed to improve the computational efficiency of MALA for use with Bayesian models by considering only small subsets of a dataset at each step of the Markov chain. This paper empirically showed SGLD was indeed as effective as MALA with greatly reduced computation and also gave intuitive/back-of-the-envelope calculations to argue that SGLD would generate samples from the desired distribution. It was only in 2016 that Teh et al. would confirm the previous intuitions and rigorously show various convergence properties of the SGLD algorithm.

2 Stochastic Gradient Langevin Dynamics

In this section, we introduce the Stochastic Gradient Langevin dynamics (SGLD) algorithm.

2.1 Motivation

We first introduce two classes of algorithms—gradient-based *stochastic optimization* and MCMC *Langevin dynamics*. Stochastic optimization is a class of algorithms that updates its parameters as follows:

$$\Delta\theta_t = \frac{\delta_t}{2}\left(\nabla \log\ p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log\ p(D_{ti}|\theta_t)\right) \tag{1}$$

where δ_t is a sequence of step sizes, $p(\theta_t)$ is a prior distribution, N is the number of entries in the dataset, n is the size of the minibatch (subset) sampled at each iteration and D_{ti} is the index of the sampled datapoint at time t. The stochasticity is introduced by randomly selecting a subset of the dataset to compute an approximation of the true gradient over the whole dataset. The idea is that over multiple iterations, the whole dataset will be used and we expect an accurate estimation of the true gradient as the noise introduced by using subsets average out. In large datasets where subset gradient approximations are accurate, introducing stochasticity in gradient descent saves substantial computation.
One major requirement to ensure convergence to a local maximum is the following property:

$$\sum_{t=1}^{\infty}\delta_t = \infty \quad \sum_{t=1}^{\infty}\delta_t^2 < \infty$$

The first constraint allows parameters to travel far away from the initial point by taking sufficiently large steps. The second constraint ensures the parameters converge to the mode by decreasing the variance in the updates over time.

The class of MCMC Langevin dynamics algorithms updates its parameters as follows:

$$\Delta\theta_t = \frac{\delta_t}{2}\left(\nabla \log\ p(\theta_t) + \sum_{i=1}^{N} \nabla \log\ p(D_i|\theta_t)\right) + \eta_t$$

$$\eta \sim N(0, \epsilon) \tag{2}$$

The gradient step sizes and the variances of the injected Gaussian noise are chosen such that the variance of the samples correspond with that of the posterior. Gaussian noise is introduced to counteract the collapse to the maximum a posterior (MAP) solution, i.e. a local maximum, which the class of stochastic optimization methods converges to. The problem with estimating only the MAP solution is that they do not capture parameter uncertainty and may potentially overfit data since it is simply a point estimate rather than the full posterior distribution.

The class of stochastic optimization and MCMC Langevin dynamics algorithms are noticeably similar so combining ideas from each is a natural idea. More precisely, we incorporate the noise from Langevin dynamics with the estimates of gradients on subsets of data from stochastic optimization. Hence, we obtain the SGLD updates defined as follows:

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

$$\eta \sim N(0, \delta_t) \tag{3}$$

where stochastic gradients are used as well as step sizes decreasing to zero at a rate satisfying convergence to a local maximum. The motivation behind combining ideas is to counteract the downfalls of stochastic optimization and MCMC methods. In particular, stochastic optimization does not capture parameter uncertainty and may overfit data, while MCMC methods require computations over the entire dataset per iteration, incurring unreasonable computational costs for large datasets.

2.2 SGLD Algorithm

MCMC algorithms typically evolve in continuous state space that could be seen as discretizations of a continuous Markov process $(\theta_t)_{t\geq 0}$. In SGLD we work with such a continuous process, namely the underdamped Langevin diffusion given by the following stochastic differential equation[1]:

$$d\theta_t = \frac{1}{2}\nabla\ \log \pi(\theta_t)dt + dW_t \tag{4}$$

[1]This equation was originally used by the physicist Paul Langevin to describe incremental displacements of a particle undergoing Brownian motion.

where $W_{t\geq 0}$ is a Brownian motion and $\pi : \mathbf{R}^d \to (0, \infty)$ is a probability density. We consider π as the target posterior distribution under a Bayesian model where there are $N >> 1$ i.i.d observations.

The authors assume there exists a computable unbiased estimator $\widehat{\nabla \log \pi}(\theta, \mathcal{U})$ 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 (in practice, this will be stochasticity from sampling subsets of data). Then, for timesteps $m \in \mathbf{N}$, the *SGLD* algorithm is defined through the recursion

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

for a sequence of aysmptotically vanishing time-steps $(\delta_m)_{m\geq 0}$, initial parameter θ_0, i.i.d sequence $\eta_m \sim N(0, I_d)$, and i.i.d sequence $\mathcal{U}_m$ of auxiliary random variables. Note that the computational costs from standard Metropolis-Hasting algorithm comes from computing the proposals and the rejection/acceptance step. SGLD completely avoids the computation of the Metropolis-Hastings ratio. By choosing sufficiently small step sizes $\delta << 1$, and because the Langevin diffusion is ergodic with respect to π, the goal is that the resulting Markov chain has an invariant distribution sufficiently close to π. We will prove that the sample path of the SGLD converges in distribution to the Langevin diffusion and thus we have that SGLD converges to the exact target posterior distribution.

In summary, SGLD is a subsampling based MCMC method combining ideas from the class of stochastic optimizers, in particular using subsets of data to estimate gradients to reduce computational costs, and the class of MCMC methods, with Langevin dynamics, to utilize gradient information to produce improved parameter updates.

3 Preliminaries

In this section, we provide a brief overview of certain relevant concepts.

3.1 Bayesian statistics

In the Bayesian paradigm, probability distributions are placed over parameters θ to represent beliefs over the possible model parameters and the strength of these beliefs. By considering distributions over parameters as opposed to a single point-estimate, the agent can capture uncertainty in its estimates, which may be useful for decision-making.

At first, a *prior* distribution $p(\theta)$ must be specified over the parameters, representing the beliefs before any data is seen. This distribution can be updated as the new data is observed by using Bayes' rule to obtain the *posterior* distribution $p(\theta|D)$.

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

where D represents some data.

$p(D|\theta)$ is known as the likelihood and $p(D)$ as the marginal likelihood. Note that $p(D) = \int p(D|\theta)p(\theta)\, d\theta$, which is usually an intractable integral (except for certain nice pairs of distributions $p(D|\theta)$ and $p(\theta)$).

In many situations, we assume that the data is generated as independent and identically-distributed (i.i.d) draws from some distribution which depends on θ. As such, we can write the likelihood as $p(D|\theta) = \prod_{i=1}^{N} p(D_i|\theta)$, where N is the number of data points and D_i is the i-th example. Despite this simplification, for many choices of $p(D_i|\theta)$ and $p(\theta)$, it is impossible to compute the posterior distribution $p(\theta|D)$ in closed-form. Hence, we must resort to other methods to obtain insight into the posterior distribution.

One approach is to try to sample from $p(\theta|D)$ and use the generated samples as a proxy for the true posterior. More generally, algorithms using samples from a distribution are known as Monte-Carlo methods. Monte-Carlo techniques are a well-studied field, but in general, it can be difficult to sample from an arbitrary distribution, especially in high-dimensions. For these difficult settings, a common family of algorithms are the Markov chain Monte Carlo (MCMC) methods.

The main idea behind MCMC is to generate samples through a Markov chain. By designing the transition probabilities of this Markov chain properly, the limiting distribution will match the distribution we wish to sample from. If this is the case, all we need to do is to run the Markov chain for a long time and treat each of the states visited by the chain as a sample from the distribution of interest.

Another approach to get evade the intractibility of the posterior distribution is to simply try to optimize it and find the parameter value that maximizes the posterior distribution. This maximizing parameter is called the *maximum a posteriori* (MAP) estimator. This optimization problem is often simpler than trying to estimate the full posterior distribution since we only have to be concerned with a single set of parameter values, as opposed to distributions over these parameters. This procedure is also similar to maximum likelihood estimation, except it also incorporates the prior distribution into the calculation instead of only the likelihood.

3.2 Stochastic Differential Equations (SDE)

The world of stochastic calculus is governed by different rules than classical calculus. The Wiener process is a stochastic process which lacks the property of *bounded variation* but has *quadratic variation*, in particular, the quadratic variance of a Wiener process is exactly the amount of time t that has elapsed with probability 1. This is compared to classical functions that 'behave smoothly' and has the bounded variation property. It follows that different operational

rules must be utilized in order to work with SDE's.

Definition (SDE with Drift and Volatility): Let X_t be a continuous stochastic process. If infinitesimal changes in the process X_t can be written as a linear combination of infinitesimal changes in t and infinitesimal changes of the Brownian motion W_t, then we may write

$$dX_t = \mu(t, W_t, X_t)dt + \sigma(t, W_t X_t)dW_t$$

and call it a Stochastic Differential Equation. This differential equation has the integral meaning:

$$X_t = X_0 + \int_0^t \mu(s, W_s, X_s)ds + \int_0^t \sigma(s, W_s, X_s)dW_s$$

where the last integral is taken in the Ito sense. The functions $\mu(t, Wt, Xt)$ and $\sigma(t, Wt, Xt)$ are called *drift rate* and *volatility*, resepctively. In particular, the underdamped Langevin Diffusion we will be working with is a Stochastic Differential Equation with drift term $\mu(s, W_s, \theta_s) = \frac{1}{2}\nabla \log \pi(\theta)$ and with volatility $\sigma = 1$. Furthermore, one can view SGLD as a discretization of a stochastic differential equation of this form.

Definition (Ito's Map): Let $\mathcal{C}([0,T], \mathbf{R}^d)$ be the space of all continuous paths w such that $w : [0,T] \to \mathbf{R}^d$. Then *Ito's map* $\mathcal{I}$: $\mathcal{C}([0,T], \mathbf{R}^d) \to \mathcal{C}([0,T], \mathbf{R}^d)$, sends a continuous path $w \in \mathcal{C}([0,T], \mathbf{R}^d)$ to the unique solution $v = \mathcal{I}(w)$ of the integral equation,

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

where $\pi : \mathbf{R}^d \to (0, \infty)$ is a probability distribution. Since the drift function $\theta \to \frac{1}{2}\nabla \log(\pi)$ is globally Lipschitz on $\mathbf{R}^d$, lemma 3.7 of (Mattingly et al., 2012) [5] shows that Ito's map is well defined and continuous, under the topology over the space $\mathcal{C}([0,T], \mathbf{R}^d)$ induced by the supremum norm $||w|| \equiv \sup\{|w_t| : 0 \leq t \leq T\}$.

3.3 Lyapunov Function

The Lyapunov function and its stability properties is used in assumption 4 of the paper, which plays a critical role in showing sample paths of SGLD converges to the continuous Langevin diffusion. In particular, assumption of existence of Lyapunov functions will control the drift term of Langevin diffusion and impose exponential convergence of the Langevin diffusion to the invariant distribution π. We restrict our definition to autonomous stochastic differential equations. Note that the Langevin diffusion equation is an autonomous SDE.

Definition (Lyapunov Function): The Lyapunov function is a scalar function $V : \mathbf{R}^n \to \mathbf{R}$, that is (1) continuous, (2) positive-definite ($V(x) > 0 \; \forall x \neq 0$) and (3) has first-order continuous partial derivatives $\forall x \in \mathbf{R}^n \setminus \{0\}$.

The derivative of V with respect to the system $x' = f(x)$ is defined as the dot product

$$V^* =< \nabla V, f(x) >$$

Definition:(Lyapunov Stability in Discrete Time): Let (X, d) be a metric space and $f : X \to X$ a continuous function. Then a point $x \in X$ is said to be **Lyapunov stable** if

$$\forall \epsilon > 0 \; \exists \delta > 0 \; \forall y \in X \; d(x, y) < \delta \implies \forall n \in \mathbf{N} \; d(f^n(x), f^n(y)) < \epsilon$$

Lyapunov showed that if a Lyapunov function can be found for a dynamical system then the system is stable in the sense of Lyapunov. The intuition behind Lyapunov Stability's relevance to SGLD is that one can judiciously choose a distance $\epsilon > 0$ in which the solutions will always stay a distance within (with respect center of the state space) contingent on the fact that you start a distance less than δ_ϵ from the center.

4 Theorem

Since the paper we chose is relatively lengthy (34 pages), at the instructor's recommendation, we focus on section 6 of the paper concerning the diffusion limit of the SGLD algorithm. The authors show that the sample paths of the SGLD algorithm converge to those of the continuous-time Langevin diffusion, with the appropriate step sizes.

We define $S^{(r)}(t)$, a continuous version of the discrete sequence of points generated by SGLD, as follows.

First letting

$$\eta_k^{(r)} = (\delta_k^{(r)})^{-1/2} \left(W(T_k^{(r)}) - W(T_{k-1}^{(r)}) \right)$$

and

$$\theta_k^{(r)} = \theta_{k-1}^{(r)} + \frac{1}{2}\delta_k^{(r)} \{ \nabla \log \pi(\theta_{k-1}^{(r)}) + H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)}) \} + (\delta_k^{(r)})^{1/2} \eta_k^{(r)}$$

where $(\mathcal{U}_k^{(r)})_{r \geq 1, k \geq 1}$ form a collection of auxiliary random variables. Note that $\theta_k^{(r)}$ represents the values along a path generated by SGLD. We obtain a continuous version by linearly interpolating between the $\theta_k^{(r)}$.

$$S^{(r)} \left(x T_{k-1}^{(r)} + (1-x) T_k^{(r)} \right) = x \theta_{k-1}^{(r)} + (1-x) \theta_k^{(r)}$$

for $x \in [0, 1]$.

We define $\text{mesh}(\delta^{(r)}) = \max\{\delta_k^{(r)} : 1 \leq k \leq m(r)\}$. We then show that $S^{(r)}(t)$ converges to a path generated by the Langevin diffusion as $r \to \infty$ and $\text{mesh}(\delta_k^{(r)}) \to 0$. The main idea is to express each $S^{(r)}$

Throughout the proof we will assume the following assumptions:

Assumption 4.1. *The step-sizes $\delta = (\delta_m)_{m\geq 1}$ form a decreasing sequence with*

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

For convenience, for two positive functions *f, g*: $\mathbf{R} \to [0, \infty)$, we define $f \lesssim g$ to indicate there exists a positive constant $C > 0$ such that $f(\theta) \leq Cg(\theta)$.

Assumption 4.2. *The drift term $\theta \to \frac{1}{2}\nabla \log \pi(\theta)$ is continuous. There exists a Lyapunov function V: $\mathbf{R}^d \to [1, \infty)$ 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*

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

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

2. *$\forall \theta \in \mathbf{R}^d$, we have*

$$||\nabla\ V(\theta)||^2 + ||\nabla\ \log\ \pi(\theta)||^2 \lesssim V(\theta)$$

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

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

Condition (1) and (2) upperbounds the magnitude of the stochastic drift term whereas (3) ensures, on average, the drift term $\nabla \log \pi(\theta)$ points the centre of the state space. Together they ensure the Langevin diffusion converges towards the equilibrium distribution π.

Lemma 4.1. *(Stability) Let the step-sizes $(\delta_m)_{m\geq 1}$ satisfy Assumption 1 and suppose conditions of assumption 2 hold. For any exponent $0 \leq p \leq p_H$ the following bounds hold almost surely,*

$$\sup_{m\geq 1} \pi_m(V^{p/2}) < \infty \quad \textit{and} \quad \sup_{m\geq 1)} \boldsymbol{E}[V^p(\theta_m)] < \infty$$

Theorem 4.2. *Let assumptions 4.1 and 4.2 hold and suppose that the drift function $\theta \mapsto \frac{1}{2}\nabla \log \pi(\theta)$ is globally Lipschitz on $\mathbf{R}^d$. If mesh$(\delta^{(r)}) \to 0$ as $r \to \infty$, then the sequence of continuous time processes $(S^{(r)})_{r\geq 1}$ converges weakly on $(\mathcal{C}([0, T], \mathbf{R}^d), ||\cdot||_\infty)$ to the Langevin diffusion (4) started at $S_0 = \theta_0$.*

Proof. The overall idea is to use the fact that the Langevin diffusion can be written as the image of a standard Brownian motion under the Itô's map $\mathcal{I} : \mathcal{C}([0,T], \mathbf{R}^d) \to \mathcal{C}([0,T], \mathbf{R}^d)$ (as defined in section 2.2). The authors use Lemma 3.7 of Mattingly et al. [5] to show that the Itô's map $\mathcal{I}$ is continuous under the topology induced by the supremum norm $||f||_\infty = \sup_{0 \le t \le T} |f(t)|$, since the drift term $s \mapsto \frac{1}{2}\nabla \log \pi(s)$ is globally Lipschitz on $\mathbf{R}^d$.

Then, by using the continuous mapping theorem, it is sufficient to show that $S^{(r)}$ can be written as $\mathcal{I}(\widetilde{W}^{(r)}) + e^{(r)}$ where $\widetilde{W}^{(r)}$ is a sequence of stochastic processes that converge weakly in $\mathcal{C}([0,T], \mathbf{R}^d)$ to a standard Brownian motion and $e^{(r)}$ is an error term such that $||e^{(r)}||_\infty$ converges to 0 in probability.

Let $\widetilde{W}^{(r)}$ be a continuous piecewise affine process that satisfies $\widetilde{W}^{(r)}(T_k^{(r)}) = W(T_k^{(r)})$ for all $0 \le k \le m(r)$ and linearly interpolates in between, where $m(r)$ is the number of pieces the interval $[0,T]$ is partitioned into.

Then, for any time $T_{k-1}^{(r)} \le t \le T_k^{(r)}$, by using the fact that $S^{(r)}(t)$ is a linear interpolation of $\theta_k^{(r)}$ and $\theta_{k-1}^{(r)}$, we have that

$$
\begin{aligned}
S^{(r)}(t) &= S^{(r)}(T_{k-1}^{(r)}) + \left[\frac{1}{2}\delta_k^{(r)}\left(\nabla \log \pi(S^{(r)}(T_{k-1}^{(r)})) + H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)})\right) + (W(T_k^{(r)}) - W(T_{k-1}^{(r)}))\right]\frac{t - T_{k-1}^{(r)}}{\delta_k^{(r)}} \\
&= S^{(r)}(T_{k-1}^{(r)}) + \frac{1}{2}\left(\nabla \log \pi(S^{(r)}(T_{k-1}^{(r)})) + H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)})\right)(t - T_{k-1}^{(r)}) + \left(W(T_k^{(r)}) - W(T_{k-1}^{(r)})\right)\frac{t - T_{k-1}^{(r)}}{\delta_k^{(r)}} \\
&= S^{(r)}(T_{k-1}^{(r)}) + \frac{1}{2}\left(\nabla \log \pi(S^{(r)}(T_{k-1}^{(r)})) + H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)})\right)(t - T_{k-1}^{(r)}) + \widetilde{W}^{(r)}(t) - \widetilde{W}^{(r)}(T_{k-1}^{(r)}) \\
&\quad \text{(Using the definition of } \widetilde{W}^{(r)}) \\
&= S^{(r)}(T_{k-1}^{(r)}) + \left(\int_{T_{k-1}^{(r)}}^{t} \frac{1}{2}\nabla \log \pi(S^{(r)}(T_{k-1}^{(r)}) + H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)}))\, du + \widetilde{W}^{(r)}(t) - \widetilde{W}(T_{k-1}^{(r)})\right) \\
&\quad + \frac{1}{2}\int_{T_{k-1}^{(r)}}^{t} H(S^{(r)}(T_{k-1}^{(r)}, \mathcal{U}_k^{(r)})\, du \quad \text{(Noting that the integrand is constant w.r.t. } u)
\end{aligned}
$$

Next, we decompose this expression into three pieces, one which converges to the Langevin diffusion and two error terms. Letting $\widehat{S}^{(r)}$ be a piecewise constant process with $\widehat{S}^{(r)}(T) = S^{(r)}(T_{k-1}^{(r)}) = \theta_{k-1}^{(r)}$ for $t \in [T_{k-1}^{(r)}, T_k^{(r)})$, we get

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

The last equality follows by expressing $S^{(r)}(T_{k-1}^{(r)})$ as an integral from time 0 to $T_{k-1}^{(r)}$ and then combining the integrals by additivity of integration over intervals.

Hence, we have expressed $S^{(r)}$ as $\mathcal{I}(\widetilde{W}^{(r)}) + e_1^{(r)} + e_2^{(r)}$. The authors claimed $\widetilde{W}^{(r)}$ weakly converges to a standard Brownian motion as $\text{mesh}(\delta_k^{(r)}) \to 0$ without proof. Here we prove a (slightly) stronger statement: $\widetilde{W}^{(r)}$ converges in probability to the standard Brownian motion W, thus verifying the authors' claim.

Define $t^k \in [T_{k-1}, T_k]$. It suffices to show for all $1 \le k \le m(r)$ we have

$$\mathbf{Pr}(||\widetilde{W}^{(r)}(T_k) - W(t^k)||_\infty > \epsilon) \xrightarrow{r \to \infty} 0$$

$$\begin{aligned}
&\mathbf{Pr}(||\widetilde{W}^{(r)}(T_k) - W(t^k)||_\infty > \epsilon) \quad \forall 1 \le k \le m(r) \\
&= \mathbf{Pr}(\bigcap_{k=1}^{m(r)} ||\widetilde{W}^{(r)}(T_k) - W(t^k)||_\infty > \epsilon) \\
&= \prod_{k=1}^{m(r)} \mathbf{Pr}(||\widetilde{W}^{(r)}(T_k) - W(t^k)||_\infty > \epsilon) && \text{(Increment independence of brownian motion)} \\
&\le 4 \times \prod_{k=1}^{m(r)} \frac{\mathbf{E}[|\widetilde{W}^{(r)}(T_k) - W(t^k)|^2]}{\epsilon^2} && \text{(Doob's } L^p \text{ inequality for martingales)} \\
&\lesssim \prod_{k=1}^{m(r)} \frac{T_k - t^k}{\epsilon^2} && \text{(Second moment of brownian motion)} \\
&\lesssim \prod_{k=1}^{m(r)} \frac{\text{mesh}(\delta^{(r)})}{\epsilon^2}
\end{aligned}$$

It follows that $\text{mesh}(\delta^{(r)})$ goes to zero as we take $r \to \infty$, thus showing convergence in probability.

Next, we show that $||e_1^{(r)}||_\infty$ and $||e_2^{(r)}||_\infty$ converge to 0 in probability. It suffices to show that $\mathbf{E}\left[||e_i^{(r)}||_\infty^2\right] \to 0$ since convergence in L_2 implies convergence in probability.

Starting with $||e_2^{(r)}||_\infty$, we have that

$$\begin{aligned}
||e_2^{(r)}||_\infty &\leq 4\mathbf{E}\left[||e_2^{(r)}(T)||^2\right] \quad \text{using Doob's } L^p \text{ inequality for martingales} \\
&= 4\sum_{k=1}^{m(r)} (\delta_k^{(r)})^2 \mathbf{E}\left[H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)})^2\right] \\
&\lesssim \sum_{k=1}^{m(r)} (\delta_k^{(r)})^2 \mathbf{E}\left[V(\theta_{k-1}^{(r)})\right] \quad \text{where } V(.) \text{ is given by Lemma 4.1} \\
&\leq \text{mesh}(\delta^{(r)}) \sum_{k=1}^{m(r)} \delta_k^{(r)} \mathbf{E}\left[V(\theta_{k-1}^{(r)})\right] \\
&\leq \text{mesh}(\delta^{(r)}) \times T \times \sup\left\{\mathbf{E}\left[V(\theta_{k-1}^{(r)})\right] : r \geq 1, 1 \leq k \leq m(r)\right\} \\
&\lesssim \text{mesh}(\delta^{(r)})
\end{aligned}$$

where the last inequality follows since the supremum is finite by lemma 4.1. Since $\text{mesh}(\delta^{(r)})$ goes to 0, we have that $||e_2^{(r)}||_\infty$ also goes to 0 in probability. In the paper, the authors write "To prove $\mathbf{E}\left[||e_1^{(r)}||_\infty\right] \to 0$ in probability, ..." although the expectation is not a random quantity. This is probably just a small oversight.

We now show that $\mathbf{E}||e_1^{(r)}||_\infty \to 0$. For the next step, the authors say that use the fact that the drift function $\theta \mapsto \frac{1}{2}\nabla \log \pi(\theta)$ is globally Lipschitz, but we did not see why this was needed. We only need that $\theta \mapsto \nabla \log(\theta)$ is Lipschitz (perhaps this is what the authors meant to say?).

For $T_{k-1}^{(r)} \leq u \leq T_k^{(r)}$,

$$\begin{aligned}
&||\nabla \log(\widehat{S}^{(r)}(u)) - \nabla \log(S^{(r)}(u))|| \\
&= ||\nabla \log(\theta_{k-1}^{(r)}) - \nabla \log(x\theta_{k-1}^{(r)} + (1-x)\theta_k^{(r)})|| && \text{for some } x \in [0,1] \\
&\lesssim ||\theta_{k-1}^{(r)} - (x\theta_{k-1}^{(r)} + (1-x)\theta_k^{(r)})|| && (\nabla \log(\theta) \text{ is Lipschitz}) \\
&= ||(1-x)(\theta_k^{(r)} - \theta_{k-1}^{(r)})|| \\
&\lesssim ||\theta_k^{(r)} - \theta_{k-1}^{(r)}|| \\
&\lesssim ||\nabla \log \pi(\theta_{k-1}^{(r)})||\delta_k^{(r)} + ||H(\theta_{k-1}^{(r)}, \mathcal{U}_k^{(r)})||\delta_k^{(r)} + \sqrt{\delta_k^{(r)}}||\eta_k^{(r)}||
\end{aligned}$$

The last line follows from the definitions of $\theta_k^{(r)}$ and $\theta_{k-1}^{(r)}$ and applying the triangle inequality.

Next, by considering the whole path and applying the inequality to each piece, we have that

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

Using assumption 4.2 and lemma 4.1, we have that the suprema

$$\sup\{\mathbf{E}\left[||\nabla \log \pi(\theta_k^{(r)})\right] : r \geq 1, 1 \leq k \leq m(r)\} < C_1\}$$
$$\sup\{\mathbf{E}\left[||H(\theta_k^{(r)}, \mathcal{U}_k^{(r)}\right] : r \geq 1, 1 \leq k \leq m(r) < C_2\}$$

for $C_1, C_2 \in \mathbf{R}$ i.e. the suprema are finite. Also, $||\eta_k^{(r)}||$ is bounded by some constant C_3 since the $\eta_k^{(r)}$ are standard normal random variables. Using these bounds, we have

$$\begin{aligned}\mathbf{E}||e_1^{(r)}||_\infty &\lesssim \mathbf{E}\left[\sum_{k=1}^{m(r)} \delta_k^{(r)} \left(C_1\delta_k^{(r)} + C_2\delta_k^{(r)} + C_3\sqrt{\delta_k^{(r)}} \right)\right] \\ &\leq C \sum_{k=1}^{m(r)} (\text{mesh}(\delta^{(r)}))^{\frac{3}{2}}\end{aligned}$$

for $C = \max\{C_1, C_2, C_3\}$ and assuming $\delta_k^{(r)}$ is sufficiently small. In the paper, the authors forget to add an expectation in the previous equation.

Since $\text{mesh}(\delta^{(r)}) \to 0$, we simply need $m(r)$ not to grow too fast for the right-hand side to go to 0. We were not able to find an assumption in the paper that assured the growth of $m(r)$ was appropriately bounded but if we take $\text{mesh}(\delta^{(r)})^{-1}$ to be $O(m(r))$, then this is sufficient to show that the right-hand side converges to 0. For example, if $\delta_k^{(r)} = \frac{T}{r}$for all $1 \leq k \leq m(r)$, a uniformly-spaced partition, then $m(r) = r$ and $\text{mesh}(\delta^{(r)}) = \frac{T}{r}$ so the RHS indeed goes to 0.

□

5 Discussion

The paper was well-motivated and the introduction did a good job presenting the relevant background concepts as well as outlining the work's contributions. The theory section was also well-organized. The authors gave an outline of the proof and gave good explanations to motivate the various assumptions. The proof of the diffusion limit theorem was rather concise, often skipping certain steps or omitting details. This made it more challenging to follow although we imagine that a reader well-versed in these subjects would not have these difficulties. Fortunately, all the key parts were present and we were able to fill in the details ourselves. We found a few minor errors that did not impact the

overall proof. The most "major" mistake was the lack of an assumption on the growth of $m(r)$. As we mentioned in the previous section, we were stumped by a certain step in the proof and, as far as we can tell, it seems like an additional assumption is needed. Of course, this is a minor point as the assumption can be added without detracting from the overall significance of the theorem.

The SGLD algorithm seems to be highly influential in the machine learning community as the original paper [2] has accumulated over 800 citations and the followup theory paper [1] has over 100 citations. Many works have built upon SGLD or further examined the algorithm.

As an aside, we find it interesting that apart from the original paper on MALA combining Langevin dynamics and MCMC [3], the papers we looked at almost never mention the physical interpretation of Langevin dynamics aside from mentioning the Langevin equation by name. It seems like the physics aspect has been "lost" over time.

Even though we have shown that SGLD's diffusion limit converges weakly to the Langevin diffusion (and thus to the target distribution π), in order to do so, a few assumptions were made. A natural question to ask is whether the diffusion limit of the algorithm still holds if any one of the assumptions were not made. We focus on assumption 4.1 which enforces that step sizes must form a decreasing sequence. This condition is usually violated in practice, as practitioners often set constant step sizes with stochastic gradient descent algorithms. Heuristically, if the step sizes were constant then the two error terms e_1, e_2 may pose a problem for the diffusion limit. Thus the error terms may prevent SGLD from converging to the target distribution π.
In fact, several works confirm this intuition [6, 7]. Brosse et al. [7] show that, under the constant step size assumption, SGLD resembles stochastic gradient descent more and more closely as the size of the dataset grows (more precisely, the Kantorovich-2 distance[2] between the marginal distributions of the parameters at the k^{th} iteration is bounded). In other words, SGLD fails to sample from the posterior distribution properly. The authors find that the MALA algorithm does not suffer from this problem and that an improved version of SGLD utilizing control variates is also able to capture the posterior distribution (while retaining computational benefits).

Other works have explored using SGLD simply as an optimization algorithm [8, 9, 10]. SGLD updates can be viewed as a regular stochastic gradient descent (SGD) updates summed with an additional independent Gaussian noise. This additional noise can be beneficial as a type of regularizer, preventing overfitting. Additionally, it could be helpful to escape local minima or avoid saddle points, which can be problematic for vanilla SGD in nonconvex optimization.

References

[1] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of*

[2] The authors unfortunately call this the Wasserstein distance.

Machine Learning Research, 17(1):193–225, 2016.

[2] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[3] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

[4] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

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

[6] Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

[7] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

[8] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

[9] Alain Durmus and Szymon Majewski. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[10] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018.