# Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Itô Process

**Issei Sato** SATO@R.DL.ITC.U-TOKYO.AC.JP
The University of Tokyo

**Hiroshi Nakagawa** N3@DL.ITC.U-TOKYO.AC.JP
The University of Tokyo

## Abstract

The stochastic gradient Langevin dynamics (SGLD) algorithm is appealing for large scale Bayesian learning. The SGLD algorithm seamlessly transit stochastic optimization and Bayesian posterior sampling. However, solid theories, such as convergence proof, have not been developed. We theoretically analyze the SGLD algorithm with constant stepsize in two ways. First, we show by using the Fokker-Planck equation that the probability distribution of random variables generated by the SGLD algorithm converges to the Bayesian posterior. Second, we analyze the convergence of the SGLD algorithm by using the Itô process, which reveals that the SGLD algorithm does not strongly but weakly converges. This result indicates that the SGLD algorithm can be an approximation method for posterior averaging.

## 1. Introduction

Bayesian learning is one of the most important fields in machine learning. It captures uncertainty and avoids overfitting. The stochastic gradient Langevin dynamics (SGLD) algorithm (Welling & Teh, 2011) is appealing for large-scale Bayesian learning. It is constructed by the combination of Robbins-Monro type stochastic approximation (H.Robbins & S.Monro, 1951) and Langevin dynamics.

Stochastic approximation such as stochastic gradient descent is one of the most successful techniques in large scale machine learning. It processes mini-batches of data at each iteration and update model parameters. Langevin dynamics injects noise into model parameters in such a way that the trajectory of the parameters will converge to the full posterior distribution. Langevin dynamics requires all data to update model parameters.

The SGLD algorithm applies stochastic approximation to Langevin dynamics, i.e., its updates are processed by mini-batches. With the original SGLD algorithm (Welling & Teh, 2011), the step sizes are annealed to zero at a certain rate. However, this stepsize condition slows mixing rate. After a sufficient burn-in period, the stepsizes are much smaller, thus, the trajectory of the parameters can be local. Ahn et al. (2012) extended the SGLD algorithm so that for large stepsizes it will sample from an approximate normal distribution of the posterior. Petterson and Teh (2013) proposed an SGLD algorithm on a probability simplex space.

**Contributions :** In this paper, we theoretically analyze the SGLD algorithm with constant stepsize in two ways.
(1) We show that the probability distribution of random variables generated by the SGLD algorithm converges to the Bayesian posterior by using the Fokker-Planck (FP) equation.
(2) We analyze the convergence of the SGLD algorithm by using the Itô process, which reveals that the SGLD algorithm does not strongly but weakly converges.

The SGLD algorithm is regarded as a discretization approximation of a stochastic differential equation corresponding to the FP equation of the Bayesian posterior. Therefore, by analyzing the discretization error of the SGLD algorithm, we can analyze the convergence of the SGLD algorithm. To the best of our knowledge, these are the first theoretical analyses of the SGLD algorithm.

Note that our theoretical analysis is based on the one-dimensional FP equation and Itô process for simplicity (mainly for simple notations). Our analysis can be easily extended by using multi-dimensional FP equation and Itô process.

*Proceedings of the $31^{st}$ International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

## 2. Stochastic Gradient Langevin Dynamics (SGLD)

Let $\boldsymbol{x}_{1:n}$ be a data set $\boldsymbol{x}_{1:n} = (x_1, x_2, \cdots, x_n)$ with a generative model $p(\boldsymbol{x}_{1:n}|\theta) = \prod_{i=1}^n p(x_i|\theta)$ parameterized by $\theta$ with prior $p(\theta)$. The aim of Bayesian learning is to compute the posterior $p(\theta|\boldsymbol{x}_{1:n})$ and the predictive distribution for a new data point, $\int p(x^*|\theta)p(\theta|\boldsymbol{x}_{1:n})d\theta$.

We use the notation $\partial_\theta$ for the partial derivatives with respect to $\theta$.

The SGLD algorithm (Welling & Teh, 2011) contains the following update equation:

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\partial_\theta \tilde{L}(\theta_t) + \eta_t,\ \eta_t \sim N(0, \epsilon_t), \tag{1}$$

$$\partial_\theta \tilde{L}(\theta_t) = \partial_\theta \log p(\theta_t) + \frac{n}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} \partial_\theta \log p(x_i|\theta_t), \tag{2}$$

where $\mathcal{S}_t$ is a set of samples randomly selected from $\boldsymbol{x}_{1:n}$, $\epsilon_t$ is the step size, and $N(0, \epsilon_t)$ is a Gaussian distribution with mean 0 and variance $\epsilon_t$. The step size decreases towards zero at rates satisfying

$$\sum_{t=1}^{\infty} \epsilon_t = \infty,\ \sum_{t=1}^{\infty} \epsilon_t^2 < \infty. \tag{3}$$

Typically, the step size is formulated as $\epsilon_t = \tau_0/(\tau_1 + t)^r$ with $r \in (0.5, 1]$.

## 3. Review of Fokker-Planck (FP) Equation

Let $q(t, \theta)$ be the probability density function of $\theta$ at time $t$. Suppose that

$$\lim_{|\theta|\to\infty} U(\theta) = \infty, \tag{4}$$

which means that $\int \exp(-U(\theta))d\theta < \infty$.

The FP equation (Risken & Frank, 1984; Daum, 1994) is a partial differential equation which describes the time evolution of the probability density function given by

$$\partial_t q(t, \theta) = \partial_\theta(\partial_\theta U(\theta) q(t, \theta)) + \partial_\theta^2 q(t, \theta). \tag{5}$$

Let $q(\theta)$ be the stationary distribution of $q(t, \theta)$. Then, it is known that $q(\theta)$ satisfies

$$q(\theta) \propto \exp\left(-U(\theta)\right). \tag{6}$$

The derivation is as follows. As $t \to \infty$, $q(t, \theta) \to q(\theta)$, i.e., $\lim_{t\to 0} \partial_t q(t, \theta) = 0$, and

$$\begin{aligned} &\partial_\theta(\partial_\theta U(\theta) q(\theta))) + \partial_\theta^2 q(\theta) = 0 \\ \Leftrightarrow &\partial_\theta[\partial_\theta U(\theta) q(\theta)) + \partial_\theta q(\theta)] = 0 \\ \Leftrightarrow &\partial_\theta q(\theta)[\partial_\theta U(\theta) + \partial_\theta \log q(\theta)] = 0. \end{aligned} \tag{7}$$

Thus, we have

$$\partial_\theta U(\theta) + \partial_\theta \log q(\theta) = 0 \Leftrightarrow U(\theta) + \log q(\theta) = \text{Const.}\,.$$

## 4. Analysis of SGLD Algorithm by using FP Equation

We analyze the probability density function (pdf) of random variables generated by the SGLD algorithm. We find that under the assumption which is often used in stochastic approximation fields, its stationary distribution is the Bayesian posterior as is the case in the ordinary Langevin dynamics. That is, the stochastic noise can be ignored.

We analyze the following SGLD with constant step size $\epsilon$.

$$\theta_{t+1} = \theta_t + \epsilon\partial_\theta \tilde{L}(\theta_t) + \sqrt{2}\eta_t,\ \eta_t \sim N(0, \epsilon). \tag{8}$$

For theoretical use, we represent Eq. (8) by using stochastic noise $\xi_t$, which is a well-known technique in stochastic approximation fields.

$$\theta_{t+1} = \theta_t + \epsilon(\partial_\theta L(\theta_t) + \xi_t) + \sqrt{2}\eta_t, \tag{9}$$

$$L(\theta) = \log p(\theta) + \sum_{i=1}^n \log p(x_i|\theta), \tag{10}$$

$$\begin{aligned} \xi_t &= \partial_\theta \tilde{L}(\theta_t) - \partial_\theta L(\theta_t), \\ &= \frac{n}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} \log p(x_i|\theta_t) - \sum_{i=1}^n \log p(x_i|\theta_t) \end{aligned} \tag{11}$$

Note that the expectation over sampling set $\mathcal{S}_t$, denoted by $\mathbb{E}_{\mathcal{S}_t}[\xi_t]$, is

$$\mathbb{E}_{\mathcal{S}_t}[\xi_t] = 0. \tag{12}$$

The stochastic noise is typically assumed to be a white noise or Martingale difference noise, i.e., $\xi_t$ and $\xi_s$ $(s \neq t)$ are independent. Moreover, we assume that

$$\mathbb{E}_{\mathcal{S}_t}[\xi_t^\ell] < \infty, \tag{13}$$

for some integer $\ell \geq 2$.

Let $q(t, \theta)$ be the pdf of $\theta$ at time $t$ and $\phi_t(\theta)$ be the characteristic function of $q(t, \theta)$ defined by

$$\phi_t(s) = \int \exp(is\theta) q(t, \theta) d\theta. \tag{14}$$

The characteristic function of $\epsilon\xi_t$ is, from Eq.(12),

$$\begin{aligned} \mathbb{E}[\exp(is\epsilon\xi_t)] = \mathbb{E}\left[\sum_{\ell=0}^{\infty} \frac{1}{\ell!}(is\epsilon\xi_t)^\ell\right] &= \sum_{\ell=0}^{\infty} \frac{1}{\ell!}\mathbb{E}[\xi_t^\ell](is\epsilon)^\ell, \\ &= 1 + \mathcal{O}(\epsilon^2). \end{aligned} \tag{15}$$

The characteristic function of $\theta_t + \epsilon\partial_\theta L(\theta_t) + \epsilon\xi_t$ is

$$\begin{aligned} &\int \exp(is\theta + is\epsilon\partial_\theta L(\theta))(1 + \mathcal{O}(\epsilon^2))q(t, \theta)d\theta \\ &= \int \exp(is\theta + is\epsilon\partial_\theta L(\theta))q(t, \theta)d\theta + \mathcal{O}(\epsilon^2) \end{aligned} \tag{16}$$

The characteristic function of $\sqrt{2}\eta_t$ is $\exp(-\epsilon s^2)$ because $\eta_t \sim N(0, \epsilon)$.

Here, we rewrite $\theta_{t+1}$ as $\theta_{t+\epsilon}$ for theoretical use. Therefore, the characteristic function of $\theta_{t+\epsilon}(= \theta_{t+1})$ is

$$\phi_{t+\epsilon}(s) = \int \exp\left(is\theta + is\epsilon\partial_\theta L(\theta) - \epsilon s^2\right) q(t,\theta)d\theta + \mathcal{O}(\epsilon^2). \tag{17}$$

Using $\exp(x) = 1 + x + \mathcal{O}(x^2)$,

$$\begin{aligned}
&\phi_{t+\epsilon}(s) - \phi_t(s) \\
&= \int \exp(is\theta)\left[\exp\left(is\epsilon\partial_\theta L(\theta) - \epsilon s^2\right) - 1\right] q(t,\theta)d\theta \\
&\qquad + \mathcal{O}(\epsilon^2), \\
&= \int \exp(is\theta)\left[is\epsilon\partial_\theta L(\theta) - \epsilon s^2 + \mathcal{O}(\epsilon^2)\right] q(t,\theta)d\theta, \\
&\qquad + \mathcal{O}(\epsilon^2), \\
&= \int \exp(is\theta)\left[is\epsilon\partial_\theta L(\theta) - \epsilon s^2\right] q(t,\theta)d\theta + \mathcal{O}(\epsilon^2).
\end{aligned} \tag{18}$$

Thus,

$$\begin{aligned}
&\frac{\phi_{t+\epsilon}(s) - \phi_t(s)}{\epsilon} \\
&\quad = (-is)\int \exp(is\theta)\partial_\theta(-L(\theta))q(t,\theta)d\theta \\
&\qquad + (-is)^2 \int \exp(is\theta)q(t,\theta)d\theta + \mathcal{O}(\epsilon),
\end{aligned} \tag{19}$$

Let $\mathcal{F}$ be the Fourier transform defined by, for an integrable function $f$,

$$\mathcal{F}[f(x)](s) = \frac{1}{\sqrt{2\pi}}\int f(x)\exp(isx)dx,$$
$$\mathcal{F}^{-1}[f(x)](s) = f(x) = \frac{1}{\sqrt{2\pi}}\int \mathcal{F}[f(x)](s)\exp(-isx)ds.$$

and the Fourier transform of the derivatives of the $\ell$-th order $f^{(\ell)}(x)$ is

$$\mathcal{F}(f^{(\ell)})(s) = (-is)^\ell(\mathcal{F}(f))(s). \tag{20}$$

Therefore,

$$\begin{aligned}
&\frac{\phi_{t+\epsilon}(s) - \phi_t(s)}{\sqrt{2\pi}\epsilon} \\
&= (-is)\mathcal{F}\partial_\theta(-L(\theta))q(t,\theta) + (-is)^2\mathcal{F}q(t,\theta) + \mathcal{O}(\epsilon), \\
&= \mathcal{F}\partial_\theta(\partial_\theta(-L(\theta))q(t,\theta)) + \mathcal{F}\partial_\theta^2 q(t,\theta) + \mathcal{O}(\epsilon).
\end{aligned} \tag{21}$$

By using $\mathcal{F}^{-1}\phi_t(s) = \sqrt{2\pi}q(t,\theta)$,

$$\begin{aligned}
&\frac{q(t+\epsilon,\theta) - q(t,\theta)}{\epsilon} \\
&\quad = \partial_\theta(\partial_\theta(-L(\theta))q(t,\theta)) + \partial_\theta^2 q(t,\theta) + \mathcal{O}(\epsilon).
\end{aligned} \tag{22}$$

Therefore,

$$\begin{aligned}
\partial_t q(t,\theta) &= \lim_{\epsilon\to 0}\frac{q(t+\epsilon,\theta) - q(t,\theta)}{\epsilon}, \\
&= \partial_\theta(\partial_\theta(-L(\theta))q(t,\theta)) + \partial_\theta^2 q(t,\theta),
\end{aligned} \tag{23}$$

which means that the probability density function of $\theta_t$ generated by the SGLD algorithm also follows the FP equation (5).

When we use $U(\theta) = -L(\theta)$ in Eq. (6), we have $q(\theta) = p(\theta|\boldsymbol{x}_{1:n})$.

## 5. Review of Itô Process

We found in the previous section that the random variables generated by the SGLD algorithm can be samples from the Bayesian posterior when $\epsilon \to 0$. That is, the SGLD algorithm is considered to be the discretization approximation of the stochastic differential equation. Therefore, our interest is its discretization error. In this section, we review the Itô process which we use to analyze the discretization error of the SGLD algorithm in the next section. Our analysis is based on the basic theories of the stochastic differential equation (Gard, 1988; Kloeden & Platen, 1992; Carlsson et al., 2010).

### 5.1. Wiener Process

$W(t)$ represents the one-dimensional Wiener process, also known as the Brownian motion , which has the following properties:

1. with probability one, the mapping $t \to W(t)$ is continuous and $W(0) = 0$,
2. if we devide $[0, T]$ as $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$, then the increments $\Delta W_k = W(t_k) - W(t_{k-1})$ $(k = 1, \cdots, N)$ are independent, and
3. for all $t > s$, the increment $W(t) - W(s)$ has a normal distribution with

$$\mathbb{E}[W(t)] = 0,\ \mathbb{E}[(W(t) - W(s))^2] = t - s. \tag{24}$$

The Gaussian injective noises of the SGLD algorithm corresponds to the increments $\Delta W_k$ of the Wiener process.

### 5.2. Itô Process

The stochastic process $X = \{X(t)\}_{t\geq 0}$ that solves

$$X(t) = X(0) + \int_0^t a(s, X(s))ds + \int_0^t b(X(s), s)dW(s) \tag{25}$$

is called the Itô process (Itô, 1944). $a(t, X(t))$ and $b(t, X(t))$ are the drift and diffusion function, respectively. The stochastic differential equation form of the Itô process is

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t). \tag{26}$$

It has a unique solution if the coefficients $a$ and $b$ are Lipschitz-continuous functions of linear growth.

The first integral in Eq. (25) is an ordinary integral along paths. The second integral in Eq. (25) is the Ito stochastic integral defined by

$$\int_0^t g(\theta(s))dW(s) = \lim_{\Delta t_{\max}\to 0} \sum_{k=0}^{m-1} g(\theta(t_k))\Delta W_k, \tag{27}$$

where we divide $[0, t]$ into $0 = t_0 < t_1 < t_2 < \cdots < t_m = t$, $\Delta t_{\max} = \max_k(t_{k+1} - t_k)$ and $\Delta W_k = W(t_{k+1}) - W(t_k)$. The mode of convergence is in mean square.

A basic property of the Itô stochastic integral used in this paper is

$$\mathbb{E}\left[\int_0^t f(s, \cdot)dW(s)\right] = 0, \tag{28}$$

where $f : [0, T] \times \Omega \to \mathbb{R}$ is the Itô integrable and independent of the increments $\Delta W_k$.

### 5.3. Itô Formula

The Itô formula is one of the most important tools in Itô process. Intuitively, the Itô formula corresponds to a chain rule in the stochastic process. To explain this, we first explain the case of the ordinary differential equation

$$\frac{d}{dt}X(t) = a(t, X(t)). \tag{29}$$

Let $h$ be a function of $X(t)$. The chain rule derives the evolution of the function $h$ as

$$\begin{aligned}\frac{d}{dt}h(t, X(t)) &= \frac{dX(t)}{dt}\frac{\partial}{\partial X(t)}h(t, X(t)) \\ &= a(t, X(t))\frac{\partial}{\partial X}h(t, X(t)).\end{aligned} \tag{30}$$

By defining a linear operator $\mathcal{L}_0 = a\partial_X$, we have

$$dh(t, X(t)) = \mathcal{L}_0 h(t, X(t))dt, \tag{31}$$

$$\text{where } \mathcal{L}_0 h(t, X(t) = a(t, X(t))\partial_X h(t, X(t)). \tag{32}$$

In the stochastic differential equation, we have the following formula.

**Theorem 1** (Itô Formula (Itô, 1944))**.** *$X(t)$ satisfies the stochastic differential equation*

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t). \tag{33}$$

*Let $h(t, X(t))$ be a given bounded function in $C^2((0, \infty) \times \mathbb{R})$. Then, $h(t, X(t))$ satisfies the stochastic differential equation*

$$dh(t, X(t)) = \mathcal{L}_1 h(t, X(t))dt + \mathcal{L}_2 h(t, X(t))dW(t), \tag{34}$$

*where $\mathcal{L}_1$ and $\mathcal{L}_2$ are linear operators defined by*

$$\mathcal{L}_1 = \partial_t + a\partial_X + \frac{1}{2}b^2\partial_X^2, \quad \mathcal{L}_2 = b\partial_X. \tag{35}$$

### 5.4. Basic Theorems

We introduce three theorems for the Itô process. These are used in the next section.

**Assumption 1.** *Suppose that*

*Lipschitz condition: there exists constant $C > 0$ such that*

$$|a(t, x) - a(t, y)| + |b(t, x) + b(t, y)| \leq C|x - y|. \tag{36}$$

*Linear growth condition: there exists constant $C > 0$ such that*

$$|a(t, x)|^2 + |b(t, x)|^2 \leq C^2(1 + |x|^2). \tag{37}$$

*There exists constant $C > 0$ such that*

$$|a(s, x) - a(t, x)| + |b(s, x) - b(t, x)| \leq C(1 + |x|)|s - t|^{\frac{1}{2}}. \tag{38}$$

**Theorem 2** (Theorem 4.5.4 of (Kloeden & Platen, 1992))**.** *Suppose that $\mathbb{E}[|X(0)|^{2\ell}] < \infty$ for some integer $\ell \geq 1$. Then*

$$\begin{aligned}&\mathbb{E}[|X(t)|^{2\ell}] \leq (1 + \mathbb{E}[|X(0)|^{2\ell}])\exp(D_1(t - t_0)), \\ &\mathbb{E}[|X(t) - X(t_0)|^{2\ell}] \\ &\quad \leq D_2(1 + \mathbb{E}[|X(0)|^{2\ell}])(t - t_0)\exp(D_1(t - t_0)),\end{aligned}$$

*for $t \in [t_0, T]$, where $T < \infty$, $D_1 = 2\ell(2\ell + 1)C^2$ and $D_2$ is a positive constant depending only on $\ell$, $C$ and $T - t_0$.*

**Theorem 3** (Gronwall inequality (Lemma 4.5.1 of (Kloeden & Platen, 1992)))**.** *Let $\alpha$, $\beta$: $[t_0, T] \to \mathbb{R}$ be integrable with*

$$0 \leq \alpha(t) \leq \beta(t) + G\int_{t_0}^t \alpha(s)ds, \ t_0 \leq t \leq T,$$

*where $G > 0$. Then*

$$\alpha(t) \leq \beta(t) + G\int_{t_0}^t \exp(G(t - s))\beta(s)ds, \ t_0 \leq t \leq T.$$

**Theorem 4** (Feynman-Kac Formula (Feynman, 1948; Kac, 1948; 1951))**.** *Suppose that $a$, $b$ and $g$ are smooth and bounded functions. Let $X$ be the solution of the stochastic differential equation*

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t),$$
$$\text{and let } u(t,x) = \mathbb{E}[g(X(T))|X(t) = x].$$

*Then $u$ is the solution of the Kolmogorov backward equation*

$$\begin{cases} \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + \frac{1}{2}b^2\frac{\partial^2}{\partial x^2}u = 0,\ t < T \\ u(T,x) = g(x) \end{cases}. \tag{39}$$

# 6. Analysis of SGLD Method by using Itô Process

We analyze the discretization error of the SGLD algorithm from two aspects: strong error and weak error, which correspond to strong convergence and weak convergence, respectively. We show that the SGLD algorithm does not converge in terms of strong convergence but converges in terms of weak convergence.

## 6.1. Problem Setting

For theoretical use, we introduce a virtual time line, $0 = t_0 < t_1 < \cdots < t_N = T$, to use the Itô process, where $t_k$ denotes the $k$-th update time of the SGLD algorithm, i.e., $\theta_{t_k}$ is the $k$-th sample, and $N$ denotes the total number of updates, i.e., the total number of samples. The time interval is constant, i.e., $t_k - t_{k-1} = \epsilon$ $(k = 1, \cdots, N)$. From $\epsilon = T/N$, $T$ indicates a parameter to determine $\epsilon$ in terms of the implementation of the SGLD algorithm.

Let $a(\theta) = \partial_\theta L(\theta)$. Note that since $L(\cdot)$ does not depend on time $t$, $a(\cdot)$ does not depend on $t$. Let $\tilde{a}(\theta) = \partial_\theta \tilde{L}(\theta)$.

We consider the following Itô process

$$\theta(T) - \theta(0) = \int_0^T a(\theta(t))dt + \int_0^T b(\theta(t))dW(t). \tag{40}$$

Let $\tilde{\theta}$ be a random variable generated by the SGLD algorithm. The SGLD update is represented as

$$\tilde{\theta}(t_k) - \tilde{\theta}(t_{k-1}) = \tilde{a}(\tilde{\theta}(t_{k-1}))\epsilon + b(\tilde{\theta}(t_{k-1}))\eta_k, \tag{41}$$
$$\eta_k \sim N(0, \epsilon), b(\bar{\theta}(t_{k-1})) = \sqrt{2}. \tag{42}$$

For deriving more general results, we do not restrict $b(\cdot)$ constant in this paper.

By using stochastic noise $\xi_{t_{k-1}}$, we can rewrite Eq. (41) as

$$\tilde{\theta}(t_k) - \tilde{\theta}(t_{k-1}) = (a(\tilde{\theta}(t_{k-1})) + \xi_{t_{k-1}})\epsilon + b(\tilde{\theta}(t_{k-1}))\eta_k. \tag{43}$$

## 6.2. Convergence in Stochastic Differential Equation

In the stochastic differential equation, we have the following two definitions of convergence: strong and weak.

**Definition 1** (Strong and Weak Convergence)**.** *Let $N$ be an integer $N > 0$ and $X$ be a stochastic process. We say that a time discrete approximation $\bar{X}_{\Delta t}$ over time interval $[0, T]$ with step size $\Delta t = T/N$*

*converges strongly to $X(T)$ at time $T$ if*

$$\lim_{\Delta t \to 0} \mathbb{E}[|X(T) - \bar{X}_{\Delta t}(T)|] = 0, \tag{44}$$

*and converges weakly to $X(T)$ at time $T$ if, for any continuous differentiable and polynomial growth function $h$,*

$$\lim_{\Delta t \to 0} |\mathbb{E}[h(X(T))] - \mathbb{E}[h(\bar{X}_{\Delta t}(T))]| = 0. \tag{45}$$

In the next section, we analyze the strong and weak convergence of the SGLD algorithm.

## 6.3. Strong and Weak Convergence of SGLD

**Assumption 2.** *We assume that*

*Initialize condition: $\tilde{\theta}(0) = \theta(0)$ and is bounded.*

*Lipschitz condition:*

$$|a(\theta(t)) - a(\theta(s))| \le C_1|\theta(t) - \theta(s)|, \tag{46}$$
$$|b(\theta(t)) - b(\theta(s))| \le C_2|\theta(t) - \theta(s)|. \tag{47}$$

*Linear growth condition:*

$$|a(\theta(t))|^2 + |b(\theta(t))|^2 \le C_3^2(1 + |\theta(t)|^2), \tag{48}$$

*and $C_1, \cdots, C_3$ do not depend on $\epsilon$.*

First, we have

**Theorem 5** (Strong error)**.**

$$\mathbb{E}[|\theta(T) - \tilde{\theta}(T)|^2] = \mathcal{O}(\epsilon + \max_k \mathbb{E}[|\xi_{t_k}|^2]). \tag{49}$$

The proof is given in Appendix A. Theorem 5 indicates the pathwise error of the SGLD algorithm is affected by stochastic noise $\xi$. This also shows that the SGLD algorithm does not converge in terms of strong converge. Note that if $\max_k \mathbb{E}[|\xi_{t_k}|^2] = 0$, the order of convergence is the same as ordinary Langevin dynamics.

Next, we can show that

**Theorem 6** (Weak error)**.**

$$|\mathbb{E}[h(\theta(T))] - \mathbb{E}[h(\tilde{\theta}(T))]| = \mathcal{O}(\epsilon), \tag{50}$$

*for any continuous differentiable and polynomial growth function $h$.*

The proof is given in Appendix B. Theorem 6 indicates the statistics error of the SGLD algorithm is not affected by stochastic noise $\xi$. This also shows that the SGLD algorithm converges in terms of weak converge and the order of convergence is the same as ordinary Langevin dynamics.

Theorem 5 is a positive property of the SGLD algorithm because the expectation of some function $\mathbb{E}[h(\theta)]$ is more important for Bayesian learning. One example is Bayes predictive distribution, i.e., $h(\theta) = p(x^*|\theta)$. When calculating some statistics, the SGLD algorithm can be an alternative to ordinary Langevin dynamics.

## 7. Conclusion

We theoretically analyzed the SGLD algorithm with a constant stepsize $\epsilon$ in two ways: using the Fokker-Planck equation and Itô process. These results show the following properties of the SGLD algorithm.

- As stepsize $\epsilon \to 0$, the stationary distribution of random variables generated by the SGLD algorithm converges to the Bayesian posterior.
- Stochastic noise negatively affects the SGLD algorithm in a mean of strong convergence but does not affect in a mean of weak convergence.

These properties suggest that if we use the SGLD algorithm as a posterior averaging method, e.g., Bayesian prediction, it can be an alternative to ordinary Langevin dynamics.

## A. Proof of Theorem 5

Consider the stochastic differential equation of (40)

$$d\theta(t) = a(\theta(t))dt + b(\theta(t))dW(t),\ 0 \le t \le T. \quad (51)$$

For theoretical use, for $t_{k-1} \le t < t_k$, define

$$\tilde{a}(\theta(t)) = \tilde{a}(\tilde{\theta}(t_{k-1})),\ \tilde{b}(\theta(t)) = b(\tilde{\theta}(t_{k-1})) \quad (52)$$

and the stochastic differential equation of (41)

$$d\tilde{\theta}(t) = \tilde{a}(\theta(t))dt + \tilde{b}(\theta(t))dW(t),\ t_{k-1} \le t < t_k.$$

Let $Z(t) = \theta(t) - \tilde{\theta}(t)$ for $t_{k-1} \le t < t_k$, i.e.,

$$dZ(t) = (a - \tilde{a})(\theta(t))dt + (b - \tilde{b})(\theta(t))dW(t), \quad (53)$$

where, for $t_{k-1} \le t < t_k$,

$$(a - \tilde{a})(\theta(t)) = a(\theta(t)) - \tilde{a}(\theta(t)) = a(\theta(t)) - \tilde{a}(\tilde{\theta}(t_k)),$$
$$(b - \tilde{b})(\theta(t)) = b(\theta(t)) - \tilde{b}(\theta(t)) = b(\theta(t)) - b(\tilde{\theta}(t_k)).$$

The Itô formula applied to $Z(t)^2$ shows, for $t_{k-1} \le t < t_k$,

$$\begin{aligned} &Z(t_k)^2 - Z(t_{k-1})^2 \\ &= \int_{t_{k-1}}^{t_k} 2(a - \tilde{a})(\theta(t))(\theta(t) - \tilde{\theta}(t)) + \frac{1}{2}((b - \tilde{b})(\theta(t)))^2 dt \\ &\quad + \int_{t_{k-1}}^{t_k} 2(b - \tilde{b})(\theta(t))(\theta(t) - \tilde{\theta}(t))dW(t). \end{aligned} \quad (54)$$

Since the expectation of the Itô integral is zero (see Eq.(28)), i.e., the second integral of Eq. (54) is

$$\mathbb{E}\left[\int_{t_{k-1}}^{t_k} 2(b - \tilde{b})(\theta(t))(\theta(t) - \tilde{\theta}(t))dW(t)\right] = 0, \quad (55)$$

take the expectation of $Z(t_k)^2$,

$$\begin{aligned} &\mathbb{E}[Z(t_k)^2] - \mathbb{E}[Z(t_{k-1})^2] \\ &= \int_{t_{k-1}}^{t_k} \mathbb{E}[2(a - \tilde{a})(\theta(t))(\theta(t) - \tilde{\theta}(t))]dt \\ &\quad + \frac{1}{2}\int_{t_{k-1}}^{t_k} \mathbb{E}[((b - \tilde{b})(\theta(t)))^2]dt. \end{aligned} \quad (56)$$

Since generally $2xy \le (x + y)^2$,

$$\begin{aligned} &\mathbb{E}[2(a - \tilde{a})(\theta(t))(\theta(t) - \tilde{\theta}(t))] \\ &\le \mathbb{E}[((a - \tilde{a})(\theta(t)) + (\theta(t) - \tilde{\theta}(t)))^2] \\ &\le \mathbb{E}[|(a - \tilde{a})(\theta(t))|^2] + \mathbb{E}[|\theta(t) - \tilde{\theta}(t)|^2]. \end{aligned} \quad (57)$$

Using stochastic noise $\xi$, we have

$$\begin{aligned} &|(a - \tilde{a})(\theta(t))|^2 \\ &= |(a(\theta(t)) - \tilde{a}(\tilde{\theta}(t_{k-1}))|^2 \\ &= |(a(\theta(t)) - a(\tilde{\theta}(t_{k-1})) - \underbrace{(\tilde{a}(\tilde{\theta}(t_{k-1})) - a(\tilde{\theta}(t_{k-1})))}_{=\xi_{t_{k-1}}}|^2 \\ &\le |a(\theta(t)) - a(\tilde{\theta}(t_{k-1}))|^2 + |\xi_{t_{k-1}}|^2. \end{aligned} \quad (58)$$

By the Lipschitz condition,

$$\begin{aligned} &|a(\theta(t)) - a(\tilde{\theta}(t_{k-1}))|^2 \\ &\le |a(\theta(t)) - a(\theta(t_{k-1})) + a(\theta(t_{k-1})) - a(\tilde{\theta}(t_{k-1}))|^2 \\ &\le |a(\theta(t)) - a(\theta(t_{k-1}))|^2 + |a(\theta(t_{k-1})) - a(\tilde{\theta}(t_{k-1}))|^2 \\ &\le C_1^2[|\theta(t) - \theta(t_{k-1})|^2 + |\theta(t_{k-1}) - \tilde{\theta}(t_{k-1})|^2], \end{aligned}$$

and the same for $|(b(\theta(t)) - \tilde{b}(\theta(t_{k-1}))|^2$.

From Theorem 2, for $t_{k-1} \le t < t_k$,

$$\begin{aligned} &\mathbb{E}[|\theta(t) - \theta(t_{k-1})|^2] \\ &\le D_2(1 + \mathbb{E}[|\theta(t_0)|^2])(t - t_{k-1})\exp(D_1(t - t_{k-1})), \\ &\le D_2(1 + \mathbb{E}[|\theta(t_0)|^2])(t - t_{k-1})\exp(D_1(T - t_0)). \end{aligned}$$

This means that there is a constant $D_3$ depending only on $D_1$, $D_2$, $T$ and $\theta(0)$, i.e.,

$$\mathbb{E}[|\theta(t)-\theta(t_{k-1})|^2] \le D_3(t-t_{k-1}) \tag{59}$$

where $D_3 = D_2(1+\mathbb{E}[|\theta(t_0)|^2])\exp(D_1 T)$.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[Z(t_k)^2] - \mathbb{E}[Z(t_{k-1})^2] \\
&\le \int_{t_{k-1}}^{t_k} (C_1^2+C_2^2)\{ \underbrace{\mathbb{E}[|\theta(t)-\theta(t_{k-1})|^2]}_{\le D_3(t-t_{k-1}) \text{ by (59)}} \\
&\qquad + \underbrace{\mathbb{E}[|\theta(t_{k-1})-\tilde{\theta}(t_{k-1})|^2]}_{=\mathbb{E}[|Z(t_{k-1})|^2]}\} \\
&\qquad + \mathbb{E}[|\xi_{t_{k-1}}|^2] + \mathbb{E}[|\theta(t)-\tilde{\theta}(t)|^2]dt. \\
&\le \int_{t_{k-1}}^{t_k} (C_1^2+C_2^2)\{D_3(t-t_{k-1}) + \mathbb{E}[|Z(t_{k-1})|^2]\} \\
&\qquad + \mathbb{E}[|\xi_{t_{k-1}}|^2] + \mathbb{E}[|\theta(t)-\tilde{\theta}(t)|^2]dt. \\
&\le (C_1^2+C_2^2)D_3\epsilon^2 + (C_1^2+C_2^2)\mathbb{E}[|Z(t_{k-1})|^2]\epsilon \\
&\qquad + \mathbb{E}[|\xi_{t_{k-1}}|^2]\epsilon + \int_{t_{k-1}}^{t_k} \mathbb{E}[|\theta(t)-\tilde{\theta}(t)|^2]dt.
\end{aligned} \tag{60}$$

That is,

$$\begin{aligned}
&\mathbb{E}[Z(t_k)^2] \\
&\le (1+F_1\epsilon)\mathbb{E}[Z(t_{k-1})^2] + F_2\epsilon^2 \\
&\qquad + \mathbb{E}[|\xi_{t_{k-1}}|^2]\epsilon + \int_{t_{k-1}}^{t_k} \mathbb{E}[|\theta(t)-\tilde{\theta}(t)|^2]dt,
\end{aligned} \tag{61}$$

where $F_1 = C_1^2+C_2^2$ and $F_2 = (C_1^2+C_2^2)D_3$

The Gronwall inequality (Theorem 3) can be applied as follows.

$$\mathbb{E}[Z(t_k)^2] \le \beta(t_{k-1}) + \int_{t_{k-1}}^{t_k} \mathbb{E}[Z(t)^2]dt, \tag{62}$$

where let $\overline{\mathbb{E}[|\xi_t|^2]} = \max_k \mathbb{E}[|\xi_{t_{k-1}}|^2]$ and $\beta(t_{k-1}) = (1+F_1\epsilon)\mathbb{E}[Z(t_{k-1})^2] + F_2\epsilon^2 + \overline{\mathbb{E}[|\xi_t|^2]}\epsilon$.

Then,

$$\begin{aligned}
\mathbb{E}[Z(t_k)^2] &\le \beta(t_{k-1}) + \int_{t_{k-1}}^{t_k} \exp(t_k-t)\beta(t_{k-1})dt, \\
&= \beta(t_{k-1})\exp(\epsilon).
\end{aligned} \tag{63}$$

Moreover, let $\gamma = (1+F_1\epsilon)\exp(\epsilon)$ and $\beta(t_0) = F_2\epsilon^2 + \overline{\mathbb{E}[|\xi_t|^2]}\epsilon$, i.e., (63) is

$$\mathbb{E}[Z(t_k)^2] \le \gamma\mathbb{E}[Z(t_{k-1})^2] + \beta(t_0)\exp(\epsilon). \tag{64}$$

Iterating Eq. (64) with $\mathbb{E}[Z(t_0)^2] = 0$ leads to

$$\mathbb{E}[Z(t_N)^2] \le \beta(t_0)\exp(\epsilon)\left(\frac{1-\gamma^N}{1-r}\right). \tag{65}$$

Here, note that $N = T/\epsilon$. Moreover, as $\epsilon \to 0$,

$$\gamma^{\frac{1}{\epsilon}} = (1+F_1\epsilon)^{\frac{1}{\epsilon}} e \to e^{F_1+1}, \tag{66}$$

and, by the l'Hospital formula,

$$\frac{\epsilon e^{\epsilon}}{1-\gamma} = \frac{\epsilon}{e^{-\epsilon}-(1+F_1\epsilon)} \to \frac{-1}{1+F_1}. \tag{67}$$

Therefore,

$$\beta(t_0)e^{\epsilon}\left(\frac{1-\gamma^N}{1-r}\right) = (F_2\epsilon + \overline{\mathbb{E}[|\xi_t|^2]})(1-\gamma^{\frac{T}{\epsilon}})\frac{\epsilon e^{\epsilon}}{1-\gamma}, \tag{68}$$

and, by using Eqs. (66) and (67), as $\to 0$,

$$\mathbb{E}[Z(t_N)^2] \le (F_2\epsilon + \overline{\mathbb{E}[|\xi_t|^2]}\frac{e^{F_1+1}-1}{F_1+1}. \tag{69}$$

That is,

$$\mathbb{E}[|\theta(T)-\tilde{\theta}(T)|^2] = \mathbb{E}[|Z(t_N)|^2] = \mathcal{O}(\epsilon + \overline{\mathbb{E}[|\xi_t|^2]}).$$

## B. Proof of Theorem 6

Let

$$u(t,\phi) = \mathbb{E}[h(\theta(T))|\theta(t)=\phi]. \tag{70}$$

Then, we have $\mathbb{E}[h(\theta(T))] = \mathbb{E}[h(\theta(T))|\theta(0)=\theta(0)] = u(0,\theta(0))$ and $\mathbb{E}[h(\tilde{\theta}(T))] = \mathbb{E}[h(\theta(T))|\theta(T)=\tilde{\theta}(T)] = u(T,\tilde{\theta}(T))$.

By using the Feynman-Kac formula (Theorem (4)), $u(t,\phi)$ satisfies

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial \phi} + \frac{1}{2}b^2\frac{\partial^2 u}{\partial \phi^2} = 0,\ t<T, \tag{71}$$

$$u(T,\phi) = h(\phi). \tag{72}$$

The Itô formula applied to $u(t,\tilde{\theta}(t))$ shows, for $t_{k-1} \le t < t_k$,

$$\begin{aligned}
du(t,\tilde{\theta}(t)) &= \left(\frac{\partial u}{\partial t} + \tilde{a}\frac{\partial u}{\partial \phi} + \frac{1}{2}\tilde{b}^2\frac{\partial^2 u}{\partial \phi^2}\right)(t,\tilde{\theta}(t))dt \\
&\quad + \tilde{b}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))dW(t) \\
&\overset{(71)}{=} \left((\tilde{a}-a)\frac{\partial u}{\partial \phi} + \frac{1}{2}(\tilde{b}^2-b^2)\frac{\partial^2 u}{\partial \phi^2}\right)(t,\tilde{\theta}(t))dt \\
&\quad + \tilde{b}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))dW(t),
\end{aligned} \tag{73}$$

where, for $t_{k-1} \le t < t_k$,

$$a\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t)) = a(\tilde{\theta}(t))\left.\frac{\partial u(t,\phi)}{\partial \phi}\right|_{\phi=\tilde{\theta}(t)}, \tag{74}$$

$$\tilde{a}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t)) = \tilde{a}(\tilde{\theta}(t_{k-1}))\left.\frac{\partial u(t,\phi)}{\partial \phi}\right|_{\phi=\tilde{\theta}(t)}, \tag{75}$$

$$b\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t)) = b(\tilde{\theta}(t))\left.\frac{\partial u(t,\phi)}{\partial \phi}\right|_{\phi=\tilde{\theta}(t)}, \tag{76}$$

$$\tilde{b}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t)) = b(\tilde{\theta}(t_{t_{k-1}}))\left.\frac{\partial u(t,\phi)}{\partial \phi}\right|_{\phi=\tilde{\theta}(t)}. \tag{77}$$

Evaluate the integral from 0 to $T$, noting $\tilde{\theta}(0) = \theta(0)$,

$$\begin{aligned} &u(T,\tilde{\theta}(T)) - u(0,\theta(0)) = \\ &\int_0^T \left((\tilde{a}-a)\frac{\partial u}{\partial \phi} + \frac{1}{2}(\tilde{b}^2 - b^2)\frac{\partial^2 u}{\partial \phi^2}\right)(t,\tilde{\theta}(t))dt \\ &+ \int_0^T \tilde{b}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))dW(t). \end{aligned} \tag{78}$$

Take the expectation and use that the expected value of the Itô integral is zero, i.e.,

$$\mathbb{E}\left[\int_0^T \tilde{b}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))dW(t)\right] = 0 \tag{79}$$

and

$$\begin{aligned} &\mathbb{E}[u(T,\tilde{\theta}(T))] - \mathbb{E}[u(0,\theta(0))] \\ &= \int_0^T \mathbb{E}\left[(\tilde{a}-a)\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right]dt \\ &+ \int_0^T \frac{1}{2}\mathbb{E}\left[(\tilde{b}^2 - b^2)\frac{\partial^2 u}{\partial \phi^2}(t,\tilde{\theta}(t))\right]dt. \end{aligned} \tag{80}$$

By (41) and (43),

$$\tilde{a}(\tilde{\theta}(t_{k-1})) = a(\tilde{\theta}(t_{k-1})) + \xi_{t_{k-1}}. \tag{81}$$

Thus,

$$\begin{aligned} &\mathbb{E}\left[(\tilde{a}-a)\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right] \\ &= \mathbb{E}\left[(\tilde{a}(\tilde{\theta}(t_{k-1})) - a(\tilde{\theta}(t)))\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right] \\ &= \mathbb{E}\left[(a(\tilde{\theta}(t_{k-1})) + \xi_{t_{k-1}} - a(\tilde{\theta}(t)))\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right] \\ &= \mathbb{E}\left[(a(\tilde{\theta}(t_{k-1})) - a(\tilde{\theta}(t)))\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right] \\ &\qquad + \mathbb{E}\left[\underbrace{\mathbb{E}_{\mathcal{S}}[\xi_{t_{k-1}}]}_{=0}\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right] \\ &= \mathbb{E}\left[(a(\tilde{\theta}(t_{k-1})) - a(\tilde{\theta}(t)))\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right]. \end{aligned} \tag{82}$$

Let

$$\rho(t,\tilde{\theta}(t)) = (a(\tilde{\theta}(t_{k-1})) - a(\tilde{\theta}(t)))\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t)), \tag{83}$$

for $t_{k-1} \le t < t_k$ and, by the Itô formula,

$$\begin{aligned} d\rho(t,\tilde{\theta}(t)) = &\left(\frac{\partial \rho}{\partial t} + \tilde{a}\frac{\partial \rho}{\partial \phi} + \frac{1}{2}\tilde{b}^2\frac{\partial^2 \rho}{\partial \phi^2}\right)(t,\tilde{\theta}(t))dt \\ &+ \tilde{b}\frac{\partial \rho}{\partial \phi}(t,\tilde{\theta}(t))dW(t). \end{aligned} \tag{84}$$

Since

$$\mathbb{E}\left[\tilde{b}\frac{\partial \rho}{\partial \phi}(t,\tilde{\theta}(t))dW(t)\right] = \mathbb{E}\left[\tilde{b}\frac{\partial \rho}{\partial \phi}(t,\tilde{\theta}(t))\right]\underbrace{\mathbb{E}\left[dW(t)\right]}_{=0} = 0,$$

for $t_{k-1} \le t < t_k$,

$$\begin{aligned} \frac{d\mathbb{E}[\rho(t,\tilde{\theta}(t))]}{dt} &= \mathbb{E}\left[\frac{d\rho(t,\tilde{\theta}(t))}{dt}\right] \\ &= \mathbb{E}\left[\left(\frac{\partial \rho}{\partial t} + \tilde{a}\frac{\partial \rho}{\partial \phi} + \frac{1}{2}\tilde{b}^2\frac{\partial^2 \rho}{\partial \phi^2}\right)(t,\tilde{\theta}(t))\right], \end{aligned} \tag{85}$$

thus, by using Weierstrass theorem, there exists a constant $C_k > 0$ such that

$$\left|\frac{d\mathbb{E}[\rho(t,\tilde{\theta}(t))]}{dt}\right| \le C_k, \text{ for } t_{k-1} \le t < t_k, \tag{86}$$

i.e.,

$$\mathbb{E}\left[\rho(t,\tilde{\theta}(t))\right] = \mathbb{E}\left[(\tilde{a}-a)\frac{\partial u}{\partial \phi}(t,\tilde{\theta}(t))\right] \le C_k\epsilon. \tag{87}$$

Similarly, we have

$$\mathbb{E}\left[(\tilde{b}^2 - b^2)\frac{\partial^2 u}{\partial \phi^2}(t,\tilde{\theta}(t))\right] \le C_k\epsilon, \text{ for } t_{k-1} \le t < t_k.$$

Therefore, using $C_{\max} = \max_k C_k$,

$$\begin{aligned} &|\mathbb{E}[h(\tilde{\theta}(T))] - \mathbb{E}[h(\theta(T))]| \\ &\quad = |\mathbb{E}[u(T,\tilde{\theta}(T))] - \mathbb{E}[u(0,\theta(0))]| \\ &\quad \le \int_0^T C_{\max}\epsilon dt = TC_{\max}\epsilon \end{aligned} \tag{88}$$

## References


Ahn, Sungjin, Balan, Anoop Korattikara, and Welling, Max. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Carlsson, Jesper, Moon, Kyoung-sook, Szepessy, Anders, Zouraris, Georgios, and Tempone, Raúl. *Stochastic Differential Equations: Models and Numerics 1*. 2010.

Daum, Frederick E. *New Exact Nonlinear Filters: Theory and Applications*. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Signal and Data Processing of Small Targets, 1994.

Feynman, R. P. Space-time approach to nonrelativistic quantum mechanics. *Rev. Mod. Phys.*, 20:367–387, 1948.

Gard, T. C. *Introduction to Stochastic Differential Equations*. M. Dekker, 1988.

H.Robbins and S.Monro. A stochastic approximation method. In *Annals of Mathematical Statistics*, pp. 400–407, 1951.

Itô, Kiyoshi. Stochastic integral. In *Proc. Imperial Acad. Tokyo*, pp. 519–524, 1944.

Kac, M. On distributions of certain wiener functionals. *Trans. Amer. Math. Soc.*, 65:1–13, 1948.

Kac, M. On some connections between probability theory and differential and integral equations. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 189–215, 1951.

Kloeden, Peter E and Platen, Eckhard. *Numerical solution of stochastic differential equations*. Springer Verlag, 1992.

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

Risken, Hannes and Frank, Till. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer, 1984.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.