CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2024-0039

Vol. **6**, No. 4, pp. 711-759
December 2025

A Sharp Uniform-in-Time Error Estimate for Stochastic Gradient Langevin Dynamics

Lei Li[1,2] and Yuliang Wang[3,*]

[1] *School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[2] *Shanghai Artificial Intelligence Laboratory.*
[3] *School of Mathematical Sciences, Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*

Received 26 August 2024; Accepted 9 March 2025

Abstract. We establish a sharp uniform-in-time error estimate for the stochastic gradient Langevin dynamics (SGLD), which is a widely-used sampling algorithm. Under mild assumptions, we obtain a uniform-in-time $\mathcal{O}(\eta^2)$ bound for the Kullback-Leibler divergence between the SGLD iteration and the Langevin diffusion, where η is the step size (or learning rate). Our analysis is also valid for varying step sizes. Consequently, we are able to derive an $\mathcal{O}(\eta)$ bound for the distance between the invariant measures of the SGLD iteration and the Langevin diffusion, in terms of Wasserstein or total variation distances. Our result can be viewed as a significant improvement compared with existing analysis for SGLD in related literature.

AMS subject classifications: 65C20, 68Q25, 60H30.

Key words: Random batch, Euler-Maruyama scheme, Fokker-Planck equation, log-Sobolev inequality.

1 Introduction

The stochastic gradient Langevin dynamics, first proposed by Welling and Teh [54] in 2011, has drawn great attention of researchers from various areas, and it shows outstanding performance when dealing with sampling tasks [1, 35, 43]. As an online algorithm, SGLD incorporates independent white noise into the well-known stochastic gradient descent (SGD), making it effective for sampling tasks. Equivalently, the SGLD algorithm can be also viewed as adding a random batch to the drift term of the Euler-Maruyama

*Corresponding author. *Email addresses:* leili2010@sjtu.edu.cn (L. Li), YuliangWang_math@sjtu.edu.cn (Y. Wang)

scheme for the (overdamped) Langevin diffusion, which is a time-continuous stochastic process that can converge to a target distribution π under suitable assumptions. In this paper, we give an optimal estimate for time-discretization error (the distance between SGLD and the Langevin diffusion), and a sharp bound for the sampling error (distance between SGLD and the target distribution π in the sampling task) in terms of Wasserstein or total variation distance as a corollary. In detail, letting η be the constant time step (or learning rate), we prove that under mild assumptions, the time-discretization error in terms of Kullback-Leibler (KL) divergence is $\mathcal{O}(\eta^2)$, which is sharp and enhances the results of most existing analyses [7, 16, 39, 46, 56, 58]. The result is also valid for varying step sizes. Moreover, the techniques involved in our analysis can effectively address challenges from the random batch and time discretization (see a more detailed discussion below and in Section 3). These techniques have the potential for further applications to analyze other stochastic processes and algorithms, for instance, a follow-up work for the sharp error estimate of the random batch method (RBM) for large interacting particle system [23].

Let us first explain the details of the SGLD method. Suppose we aim to generate samples from the target distribution $\pi \propto e^{-\beta U}$, where $U:\mathbb{R}^d \to \mathbb{R}$ is the free energy and $\beta > 0$ is the inverse temperature. One well-known and effective way to sample from π is using overdamped Langevin diffusion, whose invariant measure is exactly π. It is described by the following stochastic differential equation (SDE) in Itô's sense:

$$dX = -\nabla U(X)dt + \sqrt{2\beta^{-1}}dW, \quad X|_{t=0} = X_0, \tag{1.1}$$

where W is the Brownian motion in $\mathbb{R}^d$. The practical sampling method is then to solve the SDE above via suitable numerical schemes. After running the numerical simulation for relatively long time, one treats the obtained numerical solution for (1.1) as an approximation for π. Consider the classical Euler-Maruyama scheme for (1.1). Given the time step (or learning rate) η_k at k-th iteration, and denoting $T_k := \sum_{i=0}^{k-1} \eta_i$, the Euler-Maruyama scheme for (1.1), which is also called the unadjusted Langevin algorithm (ULA), iterates as follows:

$$\hat{X}_{T_{k+1}} = \hat{X}_{T_k} - \eta_k \nabla U(\hat{X}_{T_k}) + \sqrt{2\beta^{-1}}(W_{T_{k+1}} - W_{T_k}). \tag{1.2}$$

Based on ULA (1.2), the key idea of SGLD is to reduce the computation cost by using the random batch when calculate the drift term $-\nabla U$. In various practical tasks such as the Bayesian inference [54], people deal with the potential $U(\cdot)$ coming from high dimensional large-scaled data with size N, which is usually a large number. In these applications, $U(\cdot)$ is often of the form

$$U(\cdot) = \mathbb{E}_{\xi}[U^{\xi}(\cdot)], \tag{1.3}$$

which is the expected value of a function depending on some random variable $\xi \in \mathcal{S}$. Motivated by the "random mini-batch" idea from the stochastic gradient descent algorithm proposed by Robbins and Monre [47] decades ago, the SGLD algorithm replaces the drift $\nabla U(\cdot) = \mathbb{E}_{\xi}[\nabla U^{\xi}(\cdot)]$ by a random drift $\nabla U^{\xi}(\cdot)$, which is an unbiased estimate