Stochastics and Dynamics
Vol. 24, No. 5 (2024) 2450035 (38 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0219493724500357

Geometric ergodicity of SGLD via reflection coupling

Lei Li

School of Mathematical Sciences, Institute of Natural Sciences,
MOE-LSC, Qing Yuan Research Institute,
Shanghai Artificial Intelligence Laboratory,
Shanghai Jiao Tong University,
Shanghai 200240, P. R.China
leili2010@sjtu.edu.cn

Jian-Guo Liu

Department of Mathematics, Department of Physics,
Duke University, Durham, NC 27708, USA
jliu@math.duke.edu

Yuliang Wang*

School of Mathematical Sciences,
Shanghai Jiao Tong University,
Shanghai 200240, P. R. China
YuliangWang_math@sjtu.edu.cn

Received 30 January 2023
Revised 15 April 2024
Accepted 22 August 2024
Published 7 October 2024

We consider the geometric ergodicity of the Stochastic Gradient Langevin Dynamics (SGLD) algorithm under nonconvexity settings. Via the technique of reflection coupling, we prove the Wasserstein contraction of SGLD when the target distribution is log-concave only outside some compact sets. The time discretization and the minibatch in SGLD introduce several difficulties when applying the reflection coupling, which are addressed by a series of careful estimates of conditional expectations. As a direct corollary, the SGLD with constant step size has an invariant distribution and we are able to obtain its geometric ergodicity in terms of W_1 distance. The generalization to non-gradient drifts is also included.

Keywords: Reflection coupling; random batch; Wasserstein distance; Euler–Maruyama scheme.

AMS Subject Classification: 60H35, 65C40, 37A25

*Corresponding author.

1. Introduction

The Stochastic Gradient Langevin Dynamics (SGLD), first introduced by Welling and Teh [31], has attracted a lot of attention in various areas [4, 23, 33]. The SGLD algorithm and its variants have shown exceptional performance when dealing with many practical sampling or optimization tasks. As an online algorithm, SGLD can be viewed as adding independent white noise to the well known the classical machine learning algorithm, Stochastic Gradient Descent (SGD), making it useful for sampling tasks. The goal here is to generate samples from a target distribution π. In fact, SGLD is a Markov process that approximates the overdamped Langevin diffusion whose invariant measure is the target distribution π in the sampling task. Here the approximation is realized by using random batch to compute the drift at discrete time $T_k := k\eta$, and η is the constant time step (or learning rate). In this paper, our primary focus is on the theoretical study of SGLD's convergence to the invariant measure and its convergence rate.

Let us first explain the SGLD method. Suppose that the distribution of interest is $\pi \propto \exp(-\beta U)$, where $U : \mathbb{R}^d \to \mathbb{R}$ is the free energy and $\beta > 0$ is a positive constant describing the inverse temperature of the system. One effective way to sample from the target π is through the following overdamped Langevin diffusion, whose invariant measure is exactly π:

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

where W is the Brownian motion in $\mathbb{R}^d$. To numerically compute the sampling procedure, one often uses the Euler–Maruyama scheme. Given the time step (or learning rate) η_k at kth iteration, and denote $T_k := \sum_{i=0}^{k-1} \eta_i$, the Euler–Maruyama scheme iterates as follows:

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

The key idea of SGLD is to reduce the computation cost by using the random batch. In fact, in various sampling and optimization tasks from machine learning and data science, people deal with the potential $U(\cdot)$ coming from high dimensional large-scaled data with size N. Often $U(\cdot)$ is of the form $U(\cdot) = \mathbb{E}_\xi[U^\xi(\cdot)]$, which is the expected value of a function depending on a random variable $\xi \in \mathcal{S}$. However, usually we do not have any knowledge of the data's distribution, and the only realistic approach to computing $U(\cdot)$ is through the random batch of a fixed small size $S \ll N$ repeatedly and independently used at each T_k (see (1.1) for the details). When k goes large such that $kS \approx N \gg 1$, the random batch approximation for $U(\cdot) = \mathbb{E}_\xi[U^\xi(\cdot)]$ is then realized accumulatively due to the law of large numbers, and meanwhile the computational cost at each step is significantly reduced since $S \ll N$. In practice, one often has $U(x) = U_0(x) + \frac{1}{N}\sum_{i=1}^{N} \ell_i(x)$ and, as in the stochastic gradient descent algorithm [15, 26], ξ often represents the minibatch of $\{1, \ldots, N\}$ (In this case, for fixed batch-size S (a determined constant), ξ belongs to the set $\mathcal{S} = \{(a_1, \ldots, a_S) : a_i (1 \le i \le S)$ are S different random numbers uniformly chosen from $\{1, \ldots, N\}\}$. For $\xi = (a_1, \ldots, a_S)$, the corresponding unbiased estimate

U^{ξ} is $U^{\xi}(x) = U_0(x) + \frac{1}{S}\sum_{i=1}^{S} \ell_{a_i}(x)$.) The general form of SGLD iteration can be written in the following form.

$$\bar{X}_{T_{k+1}} = \bar{X}_{T_k} - \eta_k \nabla U^{\xi_k}(\bar{X}_{T_k}) + \sqrt{2\beta^{-1}}(W_{T_{k+1}} - W_{T_k}). \tag{1.1}$$

Here, U^{ξ_k} is an unbiased estimate for U, and thus ∇U^{ξ_k} is also an unbiased estimate for ∇U. As mentioned above, ξ_k often represents the random mini-batch of some fixed batch size S at time t_k, and $\{\xi_k\}_{k=0}^{\infty}$ are i.i.d. Also, In our analysis, we also consider the following continuous version, which is a continuous-time Markov process with continuous path:

$$\bar{X}_t = \bar{X}_{T_k} - \int_{T_k}^{t} \nabla U^{\xi_k}(\bar{X}_{T_k})ds + \int_{T_k}^{t} \sqrt{2\beta^{-1}}dW_s,$$
$$t \in [T_k, T_{k+1}), \ \ k = 0, 1, \ldots \tag{1.2}$$

and the corresponding differential form

$$d\bar{X}_t = -\nabla U^{\xi_k}(\bar{X}_{T_k})dt + \sqrt{2\beta^{-1}}dW,$$
$$\bar{X}_t|_{t=T_k} = \bar{X}_{T_k}, \ \ t \in [T_k, T_{k+1}) \ k = 0, 1, \ldots. \tag{1.3}$$

Note that the value of (1.2) at time grid T_k is exactly that of (1.1), so it is enough to study the continuous version to obtain estimate for SGLD at $t = T_k$.

Recent decades have witnessed great development of theoretical research for sampling error bound of SGLD [6, 14, 19, 23, 32, 33]. With SGLD considered a numerical scheme for the overdamped Langevin diffusion, one is naturally motivated to study the algorithm's approximation accuracy. Specifically, when comparing the densities $\bar{\rho}_t$, ρ_t of time marginal distributions of SGLD and overdamped Langevin diffusion, respectively, the authors of [19] proved that $H(\bar{\rho}_t \,\|\, \rho_t) \leq C\eta^2$, where $H(\cdot \,\|\, \cdot)$ is the relative entropy (or KL-divergence) and recall that η is the constant learning rate. Consequently, using ergodicity of the overdamped Langevin diffusion which can be derived provided that its invariant measure π satisfies the log-Sobolev inequality, one can estimate the Wasserstein or total variation distance between $\bar{\rho}_t$ and the target π : $W_p(\bar{\rho}_t, \pi)$, $TV(\bar{\rho}_t, \pi) \leq Ce^{-Ct} + C\eta^{\alpha}$ for some rates $\alpha \leq 1$ and $p = 1, 2$. Notably, recently the authors of [19] obtained the optimal rate $\alpha = 1$ while in some other literature like [6, 14, 23, 32, 33] α is no larger than $\frac{1}{2}$. Moreover, under the global strongly-log-concaveness assumption for the target π, using the synchronous coupling method, it can be proved that the SGLD algorithm itself as a Markov chain has an invariant measure $\tilde{\pi}$, and $\bar{\rho}_t$ converges to $\tilde{\pi}$ exponentially in time in terms of Wasserstein-2 distance [4]. However, the stringent requirement of global strong-log-concaveness potentially restricts the broader applicability of these results. For instance, this result would not give a good theoretical guarantee of convergence when one is sampling from Gaussian mixture distributions. The question of the existence and uniqueness of $\tilde{\pi}$, as well as the algorithm's ergodicity when one only assumes strong-log-concaveness of the target distribution π outside some compact sets, remains an open area for future research. The primary objective of this paper is to resolve such problem; specifically, we aim to study the geometric ergodicity of SGLD, assuming strong-log-concaveness of the target distribution π outside some compact sets and some other regular Lipschitz conditions (see Sec. 2.1 for more details).

Now in order to study the geometric ergodicity of SGLD, we use the classical coupling method [7], and in particular we apply the method of reflection coupling [10, 11, 20], which was originally designed to study the contraction property of many continuous SDEs. Here, we give a brief summary of how the reflection coupling method is adopted to study the geometric ergodicity of SGLD. Consider the two time marginal distributions μ_t, ν_t of SGLD (1.2), starting from the initial distributions μ_0, ν_0, respectively. We aim to prove the contraction property under the Wassertein-1 distance: $W_1(\mu_t, \nu_t) \lesssim e^{-ct} W_1(\mu_0, \nu_0)$. The coupling method then reduces this goal to find some paired dynamics $(\bar{X}_t, \bar{Y}_t)$ satisfying the laws of $\bar{X}_t$, $\bar{Y}_t$ are μ_t, ν_t, respectively, and the Lyapunov exponent

$$\gamma := \limsup_{t\to\infty} \frac{1}{t} \log \mathbb{E}|\bar{X}_t - \bar{Y}_t| \le -c,$$

is negative for this paired dynamics $(\bar{X}_t,\ \bar{Y}_t)$. Note that the geometric ergodicity arises from strong convexity of the potential $U(\cdot)$ outside some compact sets. This strong convexity becomes strong monotonicity property for any two points (x, y) far away, as in Lemma 2.1. Therefore, any such pair $(\bar{X}_t,\ \bar{Y}_t)$ would attract each other if they are sufficiently far away.

Next, in order to construct such paired dynamics $(\bar{X}_t,\ \bar{Y}_t)$, we use the key technique — reflection coupling equipped with a specific Lyapunov function $f(\cdot)$. This technique was originally designed by Lindvall and Rogers in 1986 and was developed by Eberle, etc. to study the geometric ergodicity of many continuous dynamics. Here, the Lyapunov function $f(\cdot)$ defined in (2.7) in our result is an increasing, concave function. Correspondingly, we consider the Kantorovich-Rubinstein distance $W_f(\cdot,\cdot)$ with cost function $f(\cdot)$ defined in (2.6) below. The reflection coupling methods begins with choosing the pair of initial points $(\bar{X}_0, \bar{Y}_0)$ such that $\mathbb{E}f(|\bar{X}_0 - \bar{Y}_0|) = W_f(\mu_0, \nu_0)$. Then we choose a realization $\bar{X}_t$ of SGLD (1.2) such that the law of X_t is μ_t and the law of X_0 is μ_0. The key step in the reflection coupling method is that we construct a companion process $\bar{Y}_t$ with $\bar{Y}_0$ coupled above with $\bar{X}_0$ and satisfies: (i) $\bar{Y}_t$ shares the same random batch and Brownian motion with $\bar{X}_t$, and has an additional reflection term in its diffusion part (see (3.1) below); (ii) $\bar{Y}_t$ is also a realization of SGLD (1.2) and the law of $\bar{Y}_t$ is ν_t (see Lemma 3.1). Then the contraction property mentioned above is reduced to estimation of the negative Lyapunov exponent for the paired dynamics $(\bar{X}_t,\ \bar{Y}_t)$. In fact, with this specially designed diffusion in the paired dynamics $(\bar{X}_t, \bar{Y}_t)$, we can actually prove the exponential decay in time of $\mathbb{E}f(\bar{X}_t - \bar{Y}_t)$ and therefore obtain the W_f-contraction (see Theorem 2.1):

$$W_f(\mu_t, \nu_t) \le \mathbb{E}f(|\bar{X}_t - \bar{Y}_t|) \le Ce^{-Ct}\mathbb{E}f(|\bar{X}_0 - \bar{Y}_0|) = Ce^{-CT_k} W_f(\mu_0, \nu_0).$$

Notably, the key contribution of the reflection coupling is as follows: different from synchronous coupling method where $\bar{X}_t$, $\bar{Y}_t$ shares exactly the same Brownian motion, in the reflection coupling, the process $\bar{X}_t - \bar{Y}_t$ is still a diffusion process. In particular its diffusion is an anisotropic one, see the expression in (3.6). Consequently, the existence of this diffusion leads to a $f''(\cdot)$ term after Itô's calculus, see (3.15). Then the contraction property can be obtained based on the following concave property of the constructed Lyapunov function in (2.7):

$$f''(r) \lesssim -r,$$

for all r in a bounded set. After proving the contraction property, one can directly obtain the geometric ergodicity of SGLD (see Corollary 2.1) using the Banach's contraction mapping theorem. Moreover, our choice of the $f(\cdot)$ makes the two distances $W_f(\cdot,\cdot)$, $W_1(\cdot,\cdot)$ equivalent, enabling one to obtain the geometric ergodicity under the Wasserstein-1 distance. Further details regarding the formulation of such paired dynamics $(\bar{X}_t, \bar{Y}_t)$ and the construction of the Lyapunov function $f(\cdot)$ will be elaborated upon in Sec. 3.

These years, the reflection coupling has been instrumental in establishing the geometric ergodicity of various random dynamic systems including overdamped/underdamped Langevin diffusion [11, 12, 21], Hamiltonian Monte Carlo [2, 3], first-order interacting particle systems [8, 13], etc. Recently, in [21], the authors constructed a reflection coupling for the discrete Euler–Maruyama scheme directly and obtained the contraction and ergodicity in Wasserstein-1 and Wasserstein-2 distances without random batch. Moreover, their method also gives some estimates for the long-time behavior of SGLD, but there is an $O(\eta)$ remainder in the control coming from the variance of the random batch (see [21, Theorem 2.16]), so intuitively the ergodicity of SGLD could not be proved directly through this estimate. In [17], the authors studied the ergodicity of the time-continuous random batch dynamics for the interacting particle systems, using a variant of the reflection coupling. The model studied resembles SGLD but we remark that the proof there makes use of the external confining potential and regards the random batch version of the interaction as perturbation. In our setting below, we only assume the confining property of *the expected drift* with no external potential to help, and we will consider the freezing drift dynamics instead and show that the distance between laws of two SGLD copies will vanish to zero in Wasserstein-1 distance exponentially in time.

However, due to existence of the time discretization and the random mini-batch, there are several difficulties arising when applying the reflection coupling to analyze the SGLD algorithm, detailed as follows. The first difficulty arises from the numerical discretization. In the time continuous interpolation (1.2), the drifts are evaluated at T_k but the dynamics is evolving and the hitting time (defined in (3.3)) in the reflection coupling could be between $[T_k, T_{k+1})$. This dismatch brings extra difficulty compared with the reflection coupling for time-continuous processes. Furthermore, when dealing with this difficulty, one needs to conduct careful estimate for the tail behavior of the multiplicative noise ζ_t (see (3.6) below). In fact, although the diffusion part $\hat{W}_t$ in (3.1) or Lemma 3.1 below is a Brownian motion, it correlates with the original Brownian motion W_t. Therefore, in (3.6) below, $\zeta_t = \int_{T_k}^t d\hat{W}_s - dW_s$ is a multiplicative noise. Estimate for this multiplicative noise is not trivial and we will overcome this via tools including the Burkholder–Davis–Gundy inequality in Lemma 3.2. So far, to the best of our knowledge, there is scant literature addressing the ergodicity of discrete algorithms under such mild assumptions. These difficulties shall be addressed carefully using a series of conditional expectation estimates, detailed in Sec. 3.

The second difficulty arises from how to make use of the consistency of the random batch $\mathbb{E}_\xi[U^\xi(\cdot)] = U(\cdot)$ to prove the geometric ergodicity of SGLD. In our result, we only assumed the confining property for the expected potential. On each time subinterval $[T_k, T_{k+1})$ of the SGLD algorithm, one only sees the behavior of the process associated with U^{ξ_k} rather than U. One has to consider the averaged dynamics so that our assumptions for the averaged potential U can be used. So more technical details will be required to obtain the ergodicity, see Proposition 3.1. Here we give a brief summary of Proposition 3.1 regarding the estimate for the random batch. Recall the paired dynamics $(\bar{X}_t, \bar{Y}_t)$ discussed above. After Itô's calculation, one needs to estimate

$$\mathbb{E}[\phi(\bar{X}_t, \bar{Y}_t)(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))],$$

for some function $\phi(\cdot,\cdot)$ and $t \in [T_k, T_{k+1})$. The key step is to use conduct the following splitting:

$$\begin{aligned}
&\mathbb{E}[\phi(\bar{X}_t, \bar{Y}_t)(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))] \\
&\quad = \mathbb{E}[\phi(\bar{X}_{T_k}, \bar{Y}_{T_k})(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))(\mathbf{1}_A + \mathbf{1}_B)] \\
&\qquad + \mathbb{E}[\phi(\bar{X}_t, \bar{Y}_t) - \phi(\bar{X}_{T_k}, \bar{Y}_{T_k})(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))(\mathbf{1}_A + \mathbf{1}_B)],
\end{aligned}$$

where $A := \{|\bar{X}_{T_k} - \bar{Y}_{T_k}| > R\}$ and $B := A^c$ for some $R > 0$. Since $\bar{X}_{T_k}$, $\bar{Y}_{T_k}$, $\mathbf{1}_A$ are all independent of the random batch ξ_k, we are able to use the consistency of random batch $\mathbb{E}_\xi[U^\xi(\cdot)] = U(\cdot)$ and obtain the following for the first term:

$$\begin{aligned}
&\mathbb{E}[\phi(\bar{X}_{T_k}, \bar{Y}_{T_k})(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))] \\
&\quad = \mathbb{E}[\phi(\bar{X}_{T_k}, \bar{Y}_{T_k})(\nabla U(\bar{X}_{T_k}) - \nabla U(\bar{Y}_{T_k}))].
\end{aligned}$$

Actually, this equality above reveals the consistency between SGLD and the overdamped Langevin diffusion, since it remains true if we replace $\bar{Y}_t$ above with some solutions to the overdamped Langevin diffusion. Moreover, for the second term, under the event A, we use the uniform-in-batch Lipschitz condition in Assumption 2.1 to bound $\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k})$, and the tail estimate obtained in Lemma 3.2 to estimate $(\phi(\bar{X}_t, \bar{Y}_t) - \phi(\bar{X}_{T_k}, \bar{Y}_{T_k})$. So eventually obtain an estimate for

$$\mathbb{E}[\phi(\bar{X}_t, \bar{Y}_t)(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\mathbf{1}_A],$$

under the event A in Proposition 3.1. For estimate under the event B, one cannot directly apply tail estimate in Lemma 3.2 to estimate the remainder term in the splitting above. So the consistency of random batch is no longer used, but the uniform-in-batch Lipschitz condition in Assumption 2.1 is required in our derivation, scattered throughout the proof (see (3.22) and (3.38) for instance). This uniform-in-batch Lipschitz condition is natural: Intuitively, the average of a family of non-smooth functions could be smooth, while in this case one cannot guarantee the convergence since the SGLD dynamics can evolve with non-smooth drift in all time.

The rest of the paper is organized as follows. In Sec. 2, we list our main assumptions and main results. The detailed proof will be given in Sec. 3, where a series of key estimates for the conditional expectations will be given. Section 4 is for the generalization to drifts that are not necessarily gradients. In Appendix A, some missing proofs will be given.

2. Assumptions and Main Result

2.1. *Local nonconvexity assumption*

We will use the reflection coupling to show the ergodicity under the following locally nonconvex setting, which is common in many practical tasks.

Assumption 2.1. (a) (locally nonconvex). The Hessian matrix of U is uniformly positive definite outside $B(0, R_0)$, namely, there exist $R_0 > 0$, $\kappa_0 > 0$ such that

$$\nabla^2 U(x) \succeq \kappa_0 I_d, \quad \forall\, x \in \mathbb{R}^d \backslash B(0, R_0); \tag{2.1}$$

(b) (global uniform-in-batch Lipshitz). There exists $K > 0$ such that $\forall\, x, y \in \mathbb{R}^d$, $\forall\, \xi \in \mathcal{S}$,

$$|\nabla U^{\xi}(x) - \nabla U^{\xi}(y)| \le K|x - y|. \tag{2.2}$$

Moreover, $\sup_{\xi} |\nabla U^{\xi}(0)| < \infty$.

Remark 2.1. In many applications in data science, people study the empirical risk with a penalty [14, 23]. Particularly, one may consider

$$\tilde{U} = \frac{1}{N} \sum_{i=0}^{N} \ell_i(x) + \frac{\lambda}{2}|x|^2.$$

If ℓ_i's have certain decay property as $|x| \to \infty$, the function $\tilde{U}$ satisfies Assumption 2.1. For instance, one may consider $\tilde{U}$ being the cross-entropy loss with some additional l^2-penalties as in some machine learning tasks; one may also compare this with some analogous examples in the interaction particle systems, where one has suitable bounded interactions and some external force U_0 [16, 17].

From the locally nonconvex setting in Assumption 2.1, it is not hard to derive the following strong monotonicity property for the pair (x, y), which is useful in our analysis.

Lemma 2.1. *Suppose Assumption* 2.1 *holds, then there exists* $R \geq 2$, $\kappa > 0$ *such that*

$$(x-y)\cdot(\nabla U(x)-\nabla U(y)) \geq \kappa|x-y|^2, \quad \forall\, x, y \in \mathbb{R}^d,\ |x-y| > R. \tag{2.3}$$

The proof is deferred to Appendix A. Another useful observation from Assumption 2.1 is that, we are able to control pth moment ($p > 1$) of the SGLD iteration X_t defined in (1.2). See the detailed proof in Appendix A.

Lemma 2.2 (Moment control for SGLD). *Consider the SGLD iteration* (1.2). *Suppose Assumption* 2.1 *holds.*

(1) *For any* $p \geq 1$, *any* $T > 0$ *and any step size* $\eta_k > 0$,

$$\sup_{0\leq t\leq T} \mathbb{E}\left[\sup_{0\leq s\leq t} |\bar{X}_s|^p\right] < +\infty. \tag{2.4}$$

The upper bound may depend on p, T, β *and the dimension* d.

(2) *Let* $p \geq 2$. *If* $\exists\,\delta > 0$ *such that* $\eta_k \leq \kappa/(2(p-1)K^2) - \delta$ *for all* k, *then*

$$\sup_{t\geq 0} \mathbb{E}|\bar{X}_t|^p < +\infty. \tag{2.5}$$

The upper bound may depend on p, β *and the dimension* d.

2.2. *Geometric ergodicity of SGLD*

By considering the behavior of Z_t, we aim to obtain a contraction result in terms of the Kantorovich–Rubinstein distance defined by

$$W_f(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} f(|x-y|)d\gamma. \tag{2.6}$$

One aims to find some suitable increasing, concave function f such that $f(|\cdot|)$ is equivalent to $|\cdot|$ and hence one is able to control Wasserstein-1 distance using W_f. The specific function we consider in this work is given by

$$f(r) := \int_0^r e^{-c_f(s\wedge R_1)}ds, \quad r \geq 0. \tag{2.7}$$

Here, $R_1 > 3R/2$ and $c_f > 0$ are constants to be determined. Clearly, f is concave and increasing. Moreover, for $r \geq 0$,

$$e^{-c_f R_1} r \leq f(r) \leq r. \tag{2.8}$$

We will discuss more on the motivation of the construction for such paired dynamics $(\bar{X}_t, \bar{Y}_t)$ and the Lyapunov function in Sec. 3.

Next, we will take $c_f > 0$ such that

$$\frac{1}{2}\sqrt{2\beta^{-1}}c_f R^{-1} - K \geq 0 \tag{2.9}$$

and fix $R_1 := 2R$. Moreover, we consider small steps and require the upper bound for the time step $h := \sup_k \eta_k$ to satisfy

$$\begin{aligned} &h^{\frac{1}{2}}|\log h|^{\frac{1}{2}} \leq \min\left(\frac{1}{6c'}e^{-2c_f R}\kappa, \sqrt{\bar{c}}\right), \\ &h \leq \bar{c}^{-\frac{1}{2}}h^{\frac{1}{2}}|\log h|^{\frac{1}{2}}(KR)^{-1}, \quad h \leq \min(1/(2K), R^2/9, 1) \end{aligned} \tag{2.10}$$

and

$$h \leq \min\left(\frac{\bar{c}\beta R^2}{128\log 5}, \frac{\bar{c}\beta R^2}{128(2c_f R + \log(18/\kappa))}, \frac{\bar{c}\beta R^2/32}{\log(45(1+\sqrt{2\beta}c_f^{-1}e^{2c_f R}KR)/2)}\right), \tag{2.11}$$

where $\bar{c}$, c' are two positive constants independent of k and the dimension d coming from Lemma 3.2 and Proposition 3.1, respectively.

We will establish in this work the following Wasserstein contraction results of SGLD. We leave the detailed proof to Sec. 3.

Theorem 2.1 (Wasserstein contraction for SGLD). *Suppose Assumption* 2.1 *holds. For any two initial distributions* μ_0 *and* ν_0, *denote* μ_t *and* ν_t *to be the corresponding time marginal distributions for the time continuous interpolation of SGLD algorithm* (1.2). *Denote* $h := \sup_k \eta_k$. *Let* f *be the Lyapunov function defined in* (2.7). *Assume that* h *and the parameters satisfy conditions in* (2.9)–(2.11), *then the following Wasserstein contraction result holds*:

$$W_f(\mu_{T_k}, \nu_{T_k}) \leq e^{-cT_k}W_f(\mu_0, \nu_0), \quad k \in \mathbb{N}, \tag{2.12}$$

where

$$c = \frac{1}{3}e^{-2c_f R}\min(\sqrt{2\beta^{-1}}c_f R/2, \kappa).$$

Consequently,

$$W_1(\mu_{T_k}, \nu_{T_k}) \leq c_0 e^{-cT_k}W_1(\mu_0, \nu_0), \quad k \in \mathbb{N}, \ \ c_0 := e^{2c_f R}. \tag{2.13}$$

The contraction rate is not necessarily the optimal one, which we believe is dimension-free (see the discussion in Remark 3.2). We have listed many restrictions on the step size. For the second restriction, the most essential one is that we need $h < 1/K$ for the contraction to hold. Here, we required $h \leq 1/(2K)$ instead for the formulas of contraction rate to be of reasonable order. Other restrictions on the step size can be relaxed somehow (for example $3R/2$ can be replaced by a number close to R and the numerates are loose). They are chosen just to make the formula of contraction rate appear clean. However, the dependence of β in the upper bound of the third restriction is essential. Besides, only Lemma 2.2 (a) is needed for the proof, so the restriction in Lemma 2.2 (b) is not included.

Moreover, if the step size (or learning rate) is constant $\eta_k \equiv \eta$ such that the discrete chain is time-homogeneous, then the SGLD as a discrete time Markov chain has an invariant measure $\tilde{\pi}$ by the Banach contraction mapping theorem [18]. In particular, we have the following corollary.

Corollary 2.1 (Wasserstein ergodicity of SGLD). *Consider the SGLD with constant step size $\eta_k \equiv \eta$. Assume that the step size η satisfies the restrictions in Theorem* 2.1, *for any initial distribution $\rho_0 \in W_1$, the SGLD iteration has a unique invariant measure $\tilde{\pi}$, and the time marginal distribution $\bar{\rho}_t$ of* (1.2) *satisfies for the constants c_0, c in Theorem* 2.1 *that*

$$W_1(\bar{\rho}_{n\eta}, \tilde{\pi}) \le c_0 e^{-cn\eta} W_1(\rho_0, \tilde{\pi}). \tag{2.14}$$

Proof. By Theorem 2.1, there exists $k_0 \in \mathbb{N}_+$ such that

$$W_1(\mu_{T_{k_0}}, \nu_{T_{k_0}}) \le \frac{1}{2} W_1(\mu_0, \nu_0). \tag{2.15}$$

Denote the corresponding transition kernel for nth iteration by P_n. Then, $\mu \mapsto \mu P_{k_0}$ is contractive. By contraction mapping theorem, there exists a fixed point π_* satisfying

$$\pi_* = \pi_* P_{k_0}. \tag{2.16}$$

Then, by Markov property, $\tilde{\pi} := \frac{1}{k_0}\sum_{n=0}^{k_0-1} \pi_* P_n$ is the invariant measure of the SGLD iteration. Moreover, $\tilde{\pi} = \tilde{\pi} P_{k_0}$ for any invariant measure so that the invariant measure is unique by the contraction property of P_{k_0}. Besides, $\tilde{\pi} = \pi_*$.

Letting $\nu_{n\eta} \equiv \tilde{\pi}$ in Theorem 2.1, (2.14) then follows. □

Under Assumption 2.1, $\pi \propto e^{-U}$ satisfies the log-Sobolev inequality, and one can get a uniform-in-time error estimate using KL divergence in [19]. We are then able to estimate the W_1 distance between the target distribution π and the invariant measure $\tilde{\pi}$ of the SGLD algorithm. In fact, for constant step size η, by [19, Theorem 3.2], the discretization error in terms of relative entropy (or KL-divergence) is given by

$$H(\bar{\rho}_{n\eta} \,\|\, \rho_{n\eta}) \le A_0 \eta^2, \quad \forall\, n \in \mathbb{N}, \tag{2.17}$$

where $\bar{\rho}_{n\eta}$, $\rho_{n\eta}$ correspond to the SGLD iteration and the overdamped Langevin diffusions, respectively. As a remark, the constant A_0 scales almost linearly with the dimension d under certain assumptions. The reason for the improved error bound in (2.17) (from $O(\sqrt{\eta})$ to $O(\eta)$ in terms of Wasserstein or total vatiantion distance, in comparison with existing results like [6, 14, 23, 32, 33]) is that, starting from the Fokker–Planck equation for the discrete algorithm, the authors directly considered the distance between distribution instead of other trajectory methods; also, techniques like Girsanov's transform were applied to handle additional difficulties brought by the random batch. As a consequence of (2.17), since π satisfies the log-Sobolev inequality, the Wasserstein-1 distance can be controlled by the square root of the KL-divergence by some classical transportation inequalities [24, 28], enabling one to derive an improved sampling error bound $W_1(\bar{\rho}_{n\eta}, \pi)$ for SGLD [19, Corollary 5.1]. We conclude the result in the following corollary.

Corollary 2.2. *Consider the SGLD with constant step size η and denote its density at time $n\eta$ by $\bar{\rho}_{n\eta}$. Under Assumption* 2.1, *for the step size η small enough* (*with the restrictions in Theorem* 2.1 *and in* [19, *Theorem* 3.2]), *for some positive A, C_1, C_2 independent of ρ_0, η we have*

$$W_1(\bar{\rho}_{n\eta}, \pi) \le C_0\eta + C_1 e^{-C_2 n\eta}. \tag{2.18}$$

Moreover, the SGLD iteration has a unique invariant measure $\tilde{\pi}$ satisfying

$$W_1(\tilde{\pi}, \pi) \le A\eta, \tag{2.19}$$

where $\pi \propto e^{-\beta U}$ is the target distribution.

3. Proof of Theorem 2.1

In this section, we prove Theorem 2.1 — the contraction property under the W_f distance. In the following, we will apply the technique of reflection coupling discussed in the introduction to analyze SGLD. See Appendix B for more details on the construction of the reflection coupling and the Lyapunov function.

We summarize here several challenges we would overcome in the analysis. The first difficulty arises from how to make use of the consistency of the random batch $\mathbb{E}_\xi[U^\xi(\cdot)] = U(\cdot)$ to prove the geometric ergodicity of SGLD. In our result, we only assumed the confining property for the expected potential. On each time subinterval $[T_k, T_{k+1})$ of the SGLD algorithm, one only sees the behavior of the process associated with U^{ξ_k} rather than U. Therefore, one has to consider the averaged dynamics so that our assumptions for the averaged potential U can be used. So more technical details will be required to obtain the ergodicity, see Proposition 3.1. Second, we look into the issues that come with numerical discretization — given the discrete nature of the scheme, the drift term for SGLD is evaluated at X_{T_k} instead of X_t, introducing additional challenging elements into our analysis. Furthermore, when dealing with this difficulty coming from time discretization, one needs to carefully estimate the tail behavior of the multiplicative noise ζ_t in (3.14). In fact, although the diffusion part $\hat{W}_t$ in (3.1) or Lemma 3.1 is a Brownian motion, it correlates with the original Brownian motion W_t. Therefore, in (3.6), $\zeta_t = \int_{T_k}^t d\hat{W}_s - dW_s$ is a multiplicative noise. Estimate for this multiplicative noise is not trivial and we will overcome this via tools including the Burkholder–Davis–Gundy (BDG) inequality in Lemma 3.2.

3.1. *Reflection coupling for SGLD*

For any two initial distributions μ_0, ν_0 in the statement of Theorem 2.1, we construct the following reflection coupling:

$$d\bar{X}_t = -\nabla U^{\xi_k}(\bar{X}_{T_k})dt + \sqrt{2\beta^{-1}}dW, \quad t \in [T_k, T_{k+1}), \;\; t < \tau;$$

$$\begin{aligned} d\bar{Y}_t &= -\nabla U^{\xi_k}(\bar{Y}_{T_k})dt + \sqrt{2\beta^{-1}}(I_d - 2e_t \otimes e_t) \cdot dW, \quad t \in [T_k, T_{k+1}), \ \ t < \tau; \\ \bar{X}_t &= \bar{Y}_t, \quad t \geq \tau, \end{aligned} \tag{3.1}$$

where

$$e_t := \frac{\bar{X}_t - \bar{Y}_t}{|\bar{X}_t - \bar{Y}_t|} \tag{3.2}$$

and the stopping time τ is defined by

$$\tau := \inf\{t \geq 0 : \bar{X}_t = \bar{Y}_t\}. \tag{3.3}$$

Moreover, the initials $\bar{X}_0$, $\bar{Y}_0$ of (3.1) should be chosen such that

$$\mathbb{E}f(|\bar{X}_0 - \bar{Y}_0|) = W_f(\mu_0, \nu_0). \tag{3.4}$$

Recall the definition of W_f in (2.6). For any two μ_0, ν_0 in Theorem 2.1, (3.4) can actually be achieved since one can always choose an optimal coupling $\gamma \in \pi(\mu_0, \mu_0)$ such that $\int_{\mathbb{R}^d \times \mathbb{R}^d} f(|x-y|)d\gamma = W_f(\mu_0, \nu_0)$ [30], and in this case, $\bar{X}_0 \sim \mu_0$ and $\bar{Y}_0 \sim \nu_0$.

Note that $\int_0^t (I_d - 2\mathbf{1}_{\{s<\tau\}} e_s \otimes e_s) \cdot dW_s$ is also a Brownian motion. Then, $\bar{Y}_t$ is thus also a copy of the time continuous interpolation of SGLD. Therefore, (3.1) is a well-defined coupling for the SGLD iteration. Similar arguments also appeared in related literature like [10, 11]. We summarize this in the following lemma.

Lemma 3.1. *Under the settings of* (3.1) *and* (3.3), *the process*

$$\hat{W}_t := \int_0^t (I_d - 2\mathbf{1}_{\{s<\tau\}} e_s e_s^T) dW_s,$$

is a Brownian motion in $\mathbb{R}^d$ *with respect to the natural filtration. Consequently,* $\bar{Y}_t$ *is also a realization of SGLD* (1.2).

Proof. Clearly, $\hat{W}_0 = 0$ and $\hat{W}_t$ is a martingale with respect to $\mathcal{F}_t := \sigma(W_s : s \leq t)$. Then by Levy's characterization of Brownian motion, one only needs to verify that for any $t' > t > 0$, $\mathbb{E}[\hat{W}_{t'} \otimes \hat{W}_t] = tI_d$. Indeed, by independent increment of the Brownian motion W_t, one has

$$\begin{aligned} \mathbb{E}[\hat{W}_{t'} \otimes \hat{W}_t] &= \mathbb{E}\left[\left(\int_0^t (I_d - 2\mathbf{1}_{\{s<\tau\}} e_s e_s^T) dW_s\right)\left(\int_0^t (I_d - 2\mathbf{1}_{\{s<\tau\}} e_s e_s^T) dW_s\right)^T\right] \\ &= \int_0^t \mathbb{E}[(I_d - 2\mathbf{1}_{\{s<\tau\}} e_s e_s^T)(I_d - 2\mathbf{1}_{\{s<\tau\}} e_s e_s^T)^T] ds = tI_d, \end{aligned}$$

where the last inequality is due to the fact that

$$e_s^T e_s = \frac{(\bar{X}_s - \bar{Y}_s)^T}{|\bar{X}_s - \bar{Y}_s|} \frac{(\bar{X}_s - \bar{Y}_s)}{|\bar{X}_s - \bar{Y}_s|} = 1, \quad \forall\, s \geq 0.$$

Therefore, the process $\hat{W}_t$ is a Brownian motion in $\mathbb{R}^d$. Consequently, $\bar{Y}_t$ is also a solution of SGLD (1.2). □

Denote $Z_t := \bar{X}_t - \bar{Y}_t$. Then for $t \in [T_k, T_{k+1})$ and $t < \tau$, the process Z satisfies

$$dZ_t = -(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))dt + 2\sqrt{2\beta^{-1}}\frac{Z_t^{\otimes 2}}{|Z_t|^2} \cdot dW \tag{3.5}$$

and $Z_t = 0$ for all $t \geq \tau$.

Clearly, the process Z_t defined in (3.5) satisfies for $t \in [T_k, T_{k+1})$,

$$Z_t = Z_{T_k} - (t \wedge \tau - T_k \wedge \tau)A_{T_k} + 2\sqrt{2\beta^{-1}}\zeta_t, \tag{3.6}$$

where the process ζ_t is defined by

$$\zeta_t := \int_{T_k \wedge \tau}^{t \wedge \tau} \frac{Z_s^{\otimes 2}}{|Z_s|^2} \cdot dW_s \tag{3.7}$$

and

$$A_{T_k} := \nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}). \tag{3.8}$$

Clearly, by optional stopping theorem [9], ζ_t is a martingale. Later in Lemma 3.2 and Corollary 3.1, we will prove some sub-Guassian properties of such martingale. These estimates are very helpful to overcome the challenge brought by numerical discretization. We remark that (3.6) and (3.7) also guarantee that $Z_t \equiv Z_{T_k} = 0$ for $t \geq T_k \geq \tau$, which is consistent with the definition of the coupling.

3.2. *Geometric ergodicity and uniform estimate for SGLD*

Recall that the increasing, concave function f is of the form

$$f(r) = \int_0^r e^{-c_f(s \wedge R_1)} ds, \quad r \geq 0. \tag{3.9}$$

With the construction in (3.1), we are then able to prove the geometric ergodicity of SGLD. In fact, due to the argument at the beginning of Sec. 3, we aim to show that

$$\mathbb{E}f(|Z_t|) \leq e^{-ct}\mathbb{E}f(|Z_0|),$$

which is clearly equivalent to

$$\mathbb{E}f(|Z_{t\wedge\tau}|) \leq e^{-ct}\mathbb{E}f(|Z_0|). \tag{3.10}$$

Introduce the regularization stopping time sequence

$$\tau_j := \inf\{t \geq 0 : |Z_t| \notin (j^{-1}, j)\}, \quad j \in \mathbb{N}_+, \tag{3.11}$$

which is increasing and can be proved to converge to τ as $j \to \infty$ later in Lemma 3.5. Hence to obtain (3.10), by Fatou's lemma, one needs to show that

$$\mathbb{E}f(|Z_{t\wedge\tau_j}|) \leq e^{-ct}\mathbb{E}f(|Z_0|).$$

Therefore, the main goal in the proof of Theorem 2.1 is to give the following uniform estimate:

$$\frac{d}{dt}\mathbb{E}f(|Z_{t\wedge\tau_j}|) \le -c\mathbb{E}f(|Z_{t\wedge\tau_j}|), \tag{3.12}$$

where c is independent of j, η_k and ξ_k.

In the following, we give the proof of our main result, Theorem 2.1. Some auxiliary lemmas and their proofs will be given in Sec. 3.3.

Proof of Theorem 2.1. Recall the definition of τ_j in (3.11). We first fix $T > 0$ and consider those k values such that $T_{k+1} \le T$. Consider the process $Z_t^{\tau_j} := Z_{t\wedge\tau_j}$. Clearly, for $t \in [T_k, T_{k+1})$,

$$Z_t^{\tau_j} = Z_{T_k}^{\tau_j} - (t\wedge\tau_j - T_k\wedge\tau_j)A_{T_k} + 2\sqrt{2\beta^{-1}}\zeta_t^{\tau_j}, \tag{3.13}$$

with

$$\zeta_t^{\tau_j} := \int_{T_k\wedge\tau_j}^{t\wedge\tau_j} \frac{Z_s^{\otimes 2}}{|Z_s|^2}\cdot dW_s. \tag{3.14}$$

In fact, $\tau_j \le \tau$. If $\tau_j \le T_k$, $Z_t^{\tau_j} = Z_{T_k}^{\tau_j}$ and one can focus on the previous subinterval. If $\tau_j \in [T_k, T_{k+1})$, one may verify that this holds.

Corresponding to (3.13), the process Z_t satisfies for $t \in (T_k\wedge\tau_j, T_{k+1}\wedge\tau_j)$,

$$dZ_t = -(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))dt + 2\sqrt{2\beta^{-1}}\frac{Z_t^{\otimes 2}}{|Z_t|^2}\cdot dW.$$

Since

$$\nabla^2 f(|x|) = f''(|x|)\frac{x\otimes x}{|x|^2} + \frac{f'(|x|)}{|x|}\left(I_d - \frac{x\otimes x}{|x|^2}\right),$$

by Dykin's formula and the strong Markov property [9], one has then for $t \in [T_k, T_{k+1})$,

$$\frac{d}{dt}\mathbb{E}[f(|Z_t^{\tau_j}|)] = \mathbb{E}\left[\left(\sqrt{2\beta^{-1}}f''(|Z_t^{\tau_j}|) - f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\right)\mathbf{1}_{\{t<\tau_j\}}\right]. \tag{3.15}$$

Our goal is to obtain an upper bound of the form $-c\mathbb{E}[f(|Z_t^{\tau_j}|)]$ of the right-hand side of (3.15), where A_{T_k} is computed using $U^{\xi}(X_{T_k})$ and $U^{\xi}(Y_{T_k})$. Note that we only assume convexity property for U outside $B(0,R)$ as stated in Lemma 2.1, so we first split the expectation in (3.15) into the following three parts:

$$\frac{d}{dt}\mathbb{E}[f(|Z_t^{\tau_j}|)] = \mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{t<\tau_j\}}\right]$$

$$+ \Bigg(\mathbb{E}[\sqrt{2\beta^{-1}}f''(|Z_t^{\tau_j}|)\mathbf{1}_{\{t<\tau_j\}}]$$

$$+\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R\}}\mathbf{1}_{\{t<\tau_j\}}\right]\Bigg)$$

$$=: I_1(t)+I_2(t). \tag{3.16}$$

Moreover, when estimating the right-hand side of (3.16), we need to further split them like in Taylor's expansion. The essence of this splitting lies in two main actions: (1) Conduct evaluations at $t=T_k$, which enables the utilization of the tower property by taking expectation with respect to the random batch ξ_k first; (2) estimate the residual terms carefully. The main reason for such splitting is that, for $t\in[T_k,T_{k+1})$, Z_t depends on the random batch ξ_k; for instance, $\mathbb{E}_{\xi_k}[Z_t\cdot(\nabla U^{\xi_k}(\bar{X}_{T_k})-\nabla U^{\xi_k}(\bar{Y}_{T_k}))]\neq\mathbb{E}_{\xi_k}[Z_t\cdot(\nabla U(\bar{X}_{T_k})-\nabla(\bar{Y}_{T_k}))]$. With this main idea of splitting, in the following we will separately estimate each term in (3.16), and we will take j sufficiently large such that Lemma 3.3 (serving Proposition 3.1), Lemma 3.4 (serving the term $I_2(t)$) and Proposition 3.1 (serving the term $I_1(t)$) in Sec. 3.3 below hold.

For the term $I_1(t)$, we will make use of the convexity condition along with the tail estimate in Lemma 3.2. By Proposition 3.1 (which is based on Lemma 3.2), for the step size η_k in the range considered, then for $t\in[T_k,T_{k+1})$

$$I_1(t)\le -\big(e^{-c_fR_1}\kappa-c'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}-3e^{-\bar{c}\beta R^2\eta_k^{-1}/128}\big)\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]. \tag{3.17}$$

Since $|Z_t^{\tau_j}|\le|Z_{T_k}^{\tau_j}|+\eta_kK|Z_{T_k}^{\tau_j}|+|\zeta_t^{\tau_j}|$ by Assumption 2.1, we have

$$-\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]\le-\frac{1}{1+\eta_kK}\big(\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]-\mathbb{E}\big[|\zeta_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]\big).$$

Clearly,

$$\mathbb{E}\big[|\zeta_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]\le\sqrt{\eta_k}\mathbb{P}(|Z_{T_k}^{\tau_j}|>R)\le\sqrt{\eta_k}R^{-1}\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]. \tag{3.18}$$

Therefore, for the event $\{|Z_{T_k}^{\tau_j}|>R\}$, one has

$$-\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]\le-\frac{1}{(1+\sqrt{\eta_k}R^{-1})(1+\eta_kK)}\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big].$$

This then implies for $\eta_k\le\min(1/(2K),R^2/9)$ that

$$\begin{aligned}-\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]&\le-\frac{1}{2}\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big]\\&\le-\frac{1}{2}\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{t<\tau_j\}}\big],\end{aligned} \tag{3.19}$$

for large j.

Now combining (3.17) and (3.19), we have the following estimate for the term $I_1(t)$:

$$I_1(t) \le -\frac{1}{2}\big(e^{-c_f R_1}\kappa - c'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}} - 3e^{-\frac{\bar{c}\beta R^2}{128\eta_k}}\big)\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{t<\tau_j\}}\big]. \tag{3.20}$$

For $I_2(t)$, the strategy is to make use of f'' that provides a negative part when $|Z_t|$ is small. This can also be understood as utilization of the convexity from the concave function f when U does not have convexity in $B(0,R)$.

The first part in $I_2(t)$ is given using the definition by

$$\mathbb{E}[\sqrt{2\beta^{-1}}f''(|Z_t^{\tau_j}|)\mathbf{1}_{\{t<\tau_j\}}] = \mathbb{E}\big[-\sqrt{2\beta^{-1}}c_f e^{-c_f|Z_t^{\tau_j}|}\mathbf{1}_{\{|Z_t^{\tau_j}|\le R_1\}}\mathbf{1}_{\{t<\tau_j\}}\big]. \tag{3.21}$$

For the second part in $I_2(t)$, by Assumption 2.1, Lemma 2.1 and definition of f, one has

$$\begin{aligned}
&\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R\}}\mathbf{1}_{\{t<\tau_j\}}\right]\\
&\quad\le \mathbb{E}\big[Kf'(|Z_t^{\tau_j}|)|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R\}}\mathbf{1}_{\{t<\tau_j\}}\big]\\
&\quad\le \mathbb{E}\big[Ke^{-c_f|Z_t^{\tau_j}|}|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R,|Z_t^{\tau_j}|\le R_1\}}\mathbf{1}_{\{t<\tau_j\}}\big]\\
&\qquad+\mathbb{E}\big[Ke^{-c_f R_1}|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R,\,|Z_t^{\tau_j}|>R_1\}}\mathbf{1}_{\{t<\tau_j\}}\big].
\end{aligned} \tag{3.22}$$

In principle, the idea is to use (3.21) to control the terms arising from (3.22). Hence, one may get

$$\begin{aligned}
I_2(t) &\le \mathbb{E}\left[\left(-\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f|Z_t^{\tau_j}|} + Ke^{-c_f|Z_t^{\tau_j}|}|Z_{T_k}^{\tau_j}|\right)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R,|Z_t^{\tau_j}|\le R_1\}}\mathbf{1}_{\{t<\tau_j\}}\right]\\
&\quad+\left[e^{-c_f R_1}\left(-\frac{1}{2}\sqrt{2\beta^{-1}}c_f\,\mathbb{P}\left(|Z_t^{\tau_j}|\le R_1,|Z_{T_k}^{\tau_j}|\le R,t<\tau_j\right)\right.\right.\\
&\quad\left.\left.+KR\mathbb{P}(|Z_{T_k}^{\tau_j}|\le R,|Z_t^{\tau_j}|>R_1,t<\tau_j)\right)\right]\\
&=: J_1(t)+J_2(t).
\end{aligned}$$

Direct estimate yields:

$$\begin{aligned}
J_1(t) &\le \mathbb{E}\left[-e^{-c_f|Z_t^{\tau_j}|}\left(\frac{1}{2}\sqrt{2\beta^{-1}}c_f R^{-1}-K\right)|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R,|Z_t^{\tau_j}|\le R_1\}}\mathbf{1}_{\{t<\tau_j\}}\right]\\
&\le 0,
\end{aligned} \tag{3.23}$$

if

$$\frac{1}{2}\sqrt{2\beta^{-1}}c_f R^{-1}-K\ge 0.$$

Here we need to choose sufficiently large coefficient c_f in the definition of f such that $\frac{1}{2}\sqrt{2\beta^{-1}}c_f R^{-1} - K \geq 0$.

To handle the remaining term $J_2(t)$, we first observe that

$$\begin{aligned} J_2(t) &\leq -\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f R_1} R_1^{-1}\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_t^{\tau_j}|\leq R_1, |Z_{T_k}^{\tau_j}|\leq R, t<\tau_j\}}\big] \\ &\quad + KRR_1^{-1}\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1, |Z_{T_k}^{\tau_j}|\leq R, t<\tau_j\}}\big] \\ &= -\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f R_1} R_1^{-1}\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\leq R, t<\tau_j\}}\big] \\ &\quad + \left(\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f R_1} R_1^{-1} + KRR_1^{-1}\right)\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1, |Z_{T_k}^{\tau_j}|\leq R, t<\tau_j\}}\big]. \end{aligned} \tag{3.24}$$

Based on tail estimate in Lemma 3.2, we prove in Lemma 3.4, for small η_k and large j,

$$\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1, |Z_{T_k}^{\tau_j}|\leq R, t<\tau_j\}}\big] \leq \varepsilon(\eta_k)\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\leq R, t<\tau_j\}}\big], \tag{3.25}$$

where

$$\varepsilon(\eta) := \frac{15R_1 e^{-\bar{c}\beta(R_1-3R/2)^2\eta^{-1}/8}}{R}. \tag{3.26}$$

Therefore, for the conditions given,

$$J_2(t) \leq -\left(\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f R_1} R_1^{-1} - \tilde{\varepsilon}(\eta_k)\right)\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\leq R\}}\mathbf{1}_{\{t<\tau_j\}}\big], \tag{3.27}$$

where

$$\tilde{\varepsilon}(\eta) := \left(\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f R_1} R_1^{-1} + KRR_1^{-1}\right)\frac{15R_1 e^{-\bar{c}\beta(R_1-3R/2)^2\eta^{-1}/8}}{R}.$$

Hence, for $t \in [T_k, T_{k+1})$, one is able to conclude from (3.20), (3.27) that

$$\frac{d}{dt}\mathbb{E}[f(|Z_t^{\tau_j}|)] \leq -c(k)\mathbb{E}[|Z_t^{\tau_j}|\mathbf{1}_{\{t<\tau_j\}}] \leq -c(k)\mathbb{E}[f(|Z_t^{\tau_j}|)\mathbf{1}_{\{t<\tau_j\}}],$$

where

$$\begin{aligned} c(k) := \min\Bigg(&\frac{1}{2}\sqrt{2\beta^{-1}}c_f e^{-c_f R_1} R_1^{-1} - \tilde{\varepsilon}(\eta_k), \\ &\frac{1}{2}(e^{-c_f R_1}\kappa - c'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}} - 3e^{-\bar{c}\beta R^2\eta_k^{-1}/128})\Bigg). \end{aligned}$$

Letting $j \to +\infty$, since $\tau_j \to \tau$ by Lemma 3.5 in Sec. 3.3 and the moment control in (3.61) (recall that $T_{k+1} \leq T$ and $Z_t = Z_t^{\tau}$), one has by dominated convergence theorem that

$$\mathbb{E}[f(|Z_t|)] \leq \mathbb{E}[f(|Z_{T_k}|)] - c(k)\int_{T_k}^{t}\mathbb{E}[f(|Z_s|)\mathbf{1}_{\{s<\tau\}}]ds.$$

Since $Z_t \equiv 0$ for $t \geq \tau$,

$$\mathbb{E}[f(|Z_t|)\mathbf{1}_{\{t<\tau\}}] = \mathbb{E}[f(|Z_t|)].$$

By Grönwall's inequality, one has

$$\mathbb{E}[f(|Z_{T_{k+1}}|)] \leq e^{-c(k)\eta_k}\mathbb{E}[f(|Z_{T_k}|)].$$

Define $h := \sup_k \eta_k$. Choosing small h as stated in Theorem 2.1, we are able to conclude

$$\mathbb{E}[f(|Z_{T_k}|)] \leq e^{-cT_k}\mathbb{E}[f(|Z_0|)] = e^{-cT_k}W_f(\mu_0,\nu_0), \tag{3.28}$$

where

$$c = \frac{1}{3}e^{-c_f R_1}\min(\sqrt{2\beta^{-1}}c_f R_1^{-1}, \kappa). \tag{3.29}$$

Note that this resulted inequality is independent of T. Since T is arbitrary, this holds for all $k \geq 0$. So one eventually has

$$W_f(\mu_{T_k},\nu_{T_k}) \leq e^{-cT_k}W_f(\mu_0,\nu_0), \quad \forall\, k \in \mathbb{N}. \tag{3.30}$$

Above, μ_0 and ν_0 denote any two initial distributions, and μ_t and ν_t are the corresponding time marginal distributions for the time continuous interpolation of SGLD algorithm (1.2). Moreover, since $e^{-c_f R_1}r \leq f(r) \leq r$ for any $r \geq 0$, we have

$$W_1(\mu_{T_k},\nu_{T_k}) \leq c_0 e^{-cT_k}W_1(\mu_0,\nu_0), \quad \forall\, k \in \mathbb{N}, \tag{3.31}$$

with $c_0 := e^{c_f R_1}$. This then ends the proof by choosing $R_1 = 2R$. □

3.3. *Propositions and lemmas used in the proof of Theorem* 2.1

In the following, we present crucial estimations used in the proof above: Lemma 3.2, Corollary 3.1, Lemmas 3.3–3.5 and Proposition 3.1. As can be observed in the proof of Theorem 2.1, the main issue to solve due to the existence of numerical discretization is that, one needs to estimate how far Z_t moves during the time interval $[T_k, t]$ for $t \in [T_k, T_{k+1})$. This then motivates us to estimate the diffusion part $\zeta_t^{\tau_j}$ defined in (3.14) first.

In the next lemma, we estimate the martingale $\zeta_t^{\tau_j}$ defined in (3.14):

$$\zeta_t^{\tau_j} := \int_{T_k\wedge\tau_j}^{t\wedge\tau_j} \frac{Z_s^{\otimes 2}}{|Z_s|^2}\cdot dW_s.$$

Note that although the diffusion part $\hat{W}_t$ in (3.1) or Lemma 3.1 is a Brownian motion, it correlates with the original Brownian motion W_t. Therefore, in (3.6), $\zeta_t = \int_{T_k}^t d\hat{W}_s - dW_s$ is a multiplicative noise. However, since the diffusion coefficient

$\frac{Z_t^{\otimes 2}}{|Z_t|^2}$ has unit norm, we are able to give the following tail estimate for $\zeta_t^{\tau_j}$ using the BDG inequality.

Lemma 3.2. *Recall the definition for* $\zeta_s^{\tau_j}$ *in* (3.14)*. For any* $j \in \mathbb{N}_+$*, for fixed* $t \in [T_k, T_{k+1})$*, the random variable* $\sup_{T_k \le s \le t} |\zeta_s^{\tau_j}|$ *is subgaussian in the sense that*

$$\mathbb{P}\left(\sup_{T_k \le s \le t} |\zeta_s^{\tau_j}| > a \,\middle|\, \mathcal{F}_{T_k} \right) \le 2e^{-\bar{c}\eta_k^{-1}a^2}, \quad \forall\, a > 0, \tag{3.32}$$

where the σ*-algebra* $\mathcal{F}_{T_k}$ *is defined by* $\mathcal{F}_{T_k} := \sigma(\bar{X}_s, \bar{Y}_s; s \le T_k)$*, and* $\bar{c}$ *is a positive constant independent of* t*,* k*,* ξ *and* a*. Consequently,*

$$\mathbb{P}\left(\sup_{T_k \le s \le t} |\zeta_s^{\tau_j}| > \bar{c}^{-\frac{1}{2}} \eta_k^{\frac{1}{2}} |\log \eta_k|^{\frac{1}{2}} \,\middle|\, \mathcal{F}_{T_k} \right) \le 2\eta_k \to 0 \quad \textit{as } \eta_k \to 0. \tag{3.33}$$

Proof. We prove the subgaussian property (3.32) via the well-known ψ_2-condition [29]: there exists $\alpha > 0$ such that

$$\mathbb{E}[e^{\alpha|\theta_t^{\tau_j}|^2} \,|\, \mathcal{F}_{T_k}] \le 2, \tag{3.34}$$

where we denote $\theta_t^{\tau_j} := \sup_{T_k \le s \le t} \zeta_s^{\tau_j}$. Clearly, $\zeta_t^{\tau_j}$ of the form (3.14) is a martingale by optional stopping theorem [9], and its quadratic variation satisfies $\langle \zeta_t^{\tau_j} \rangle \le t \wedge \tau_j - T_k \wedge \tau_j \le \eta_k$. Then it holds by the BDG inequality [1, 5] that

$$\begin{aligned}
\mathbb{E}[e^{\alpha|\theta_t^{\tau_j}|^2} \,|\, \mathcal{F}_{T_k}] &= 1 + \sum_{p=1}^{+\infty} \frac{1}{p!} \alpha^p \mathbb{E}[|\theta_t^{\tau_j}|^{2p} \,|\, \mathcal{F}_{T_k}] \\
&\le 1 + \sum_{p=1}^{+\infty} \frac{1}{p!} \alpha^p C_{2p} \mathbb{E}[\langle \zeta^{\tau_j} \rangle_{T_{k+1}}^p \,|\, \mathcal{F}_{T_k}] \\
&\le 1 + \sum_{p=1}^{+\infty} \frac{1}{p!} C_{2p} (\eta_k \alpha)^p,
\end{aligned} \tag{3.35}$$

where α is a positive parameter to be determined, and C_{2p} is a positive constant satisfying [1, 5, 25]:

$$C_{2p} \le (C\sqrt{2p})^{2p}, \tag{3.36}$$

where C is a positive constant related to the Hilbert space only (in our case $\mathbb{R}^d$). Combining (3.35) and (3.36), we have

$$\mathbb{E}[e^{\alpha|\theta_t^{\tau_j}|^2} \,|\, \mathcal{F}_{T_k}] \le 1 + C \sum_{p=1}^{+\infty} \frac{p^p}{p!} (2\eta_k \alpha)^p.$$

Clearly, $\frac{p^p}{p!} \le e^p p^{-\frac{1}{2}} \le e^p$, which can be derived from an intermediate result in the proof of Stirling's formula [27]: $\log p! > (p + \frac{1}{2}) \log p - p$. Therefore,

$$\mathbb{E}[e^{\alpha|\theta_t^{\tau_j}|^2} \,|\, \mathcal{F}_{T_k}] \le 1 + C \sum_{p=1}^{+\infty} (2e\eta_k \alpha)^p = 1 + C \frac{2e\eta_k \alpha}{1 - 2e\eta_k \alpha} = 2,$$

by choosing $\alpha = \frac{1}{2e(1+C)\eta_k} := \bar{c}\eta_k^{-1}$. Therefore, the ψ_2 condition (3.34) holds.

Finally, using Chernoff's bound [29], for any $a > 0$, it holds that

$$\mathbb{P}(|\theta_t^{\tau_j}| > a \,|\, \mathcal{F}_{T_k}) \le \mathbb{E}[e^{\alpha|\theta_t^{\tau_j}|^2} \,|\, \mathcal{F}_{T_k}]/e^{\alpha a^2} \le 2e^{-\bar{c}\eta_k^{-1}a^2}. \tag{3.37}$$

Consequently, taking $a = \bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log \eta_k|^{\frac{1}{2}}$ gives the last claim. □

We will make use of the subgaussian estimate to control a series of conditional expectations. In particular, later we need the conditional expectations on events like $|Z_{T_k}^{\tau_j}| > R$ and $t < \tau_j$. If these two events are independent, there is little difficulty. The difficulty is that these two are highly correlated. Actually, we will make use of the fact that the former event almost is contained in the second one so that the estimates can carry through as well.

As a start, we prove the following conditional estimate of $|\zeta_t^{\tau_j}|$ as an illustration.

Corollary 3.1. *Let $\eta_k \le \min(\frac{1}{2K}, \frac{\bar{c}\beta R^2}{128\log 8})$, where $\bar{c}$ is the positive constant obtained in Lemma* 3.2. *Suppose that $\mathbb{P}(|Z_{T_k}^{\tau_j}| > R) > 0$. Then, for j large enough, it holds that*

$$\mathbb{E}[|\zeta_t^{\tau_j}| \,|\, |Z_{T_k}^{\tau_j}| > R, t < \tau_j] \le 8\sqrt{2}\sqrt{\eta_k}, \quad \forall\, t \in [T_k, T_{k+1}). \tag{3.38}$$

Proof. For simplicity, we denote the events

$$A := \{|Z_{T_k}^{\tau_j}| > R\}, \quad B := \{t < \tau_j\}.$$

Our goal is then to control $\mathbb{E}[|\zeta_t^{\tau_j}|\mathbf{1}_A\mathbf{1}_B]/P(A \cap B)$.

First, using the BDG inequality for $p \in (0, 2)$ (see, e.g., [22, Theorem 7.3]), one has

$$\mathbb{E}[|\zeta_t^{\tau_j}| \,|\, A] \le 4\sqrt{2}\mathbb{E}[\mathbb{E}[\langle\zeta^{\tau_j}\rangle_{T_k}^{\frac{1}{2}} \,|\, \mathcal{F}_{T_k}] \,|\, A] \le 4\sqrt{2}\sqrt{\eta_k}. \tag{3.39}$$

Next, we estimate $\mathbb{P}(B^c \,|\, A)$. By Markov inequality and the moment control in Lemma 2.2,

$$\mathbb{P}(|Z_{T_k}^{\tau_j}| \ge j_0) \le \frac{\mathbb{E}|Z_{T_k}^{\tau_j}|^2}{j_0^2} \to 0 \quad \text{as } j_0 \to \infty.$$

Hence, for j_0 large enough (independent of j),

$$\mathbb{P}(|Z_{T_k}^{\tau_j}| \ge j_0) \le \frac{1}{4}\mathbb{P}(A).$$

Clearly,

$$\mathbb{P}(B^c \,|\, A) \le \frac{\mathbb{P}(|Z_{T_k}^{\tau_j}| \ge j_0)}{\mathbb{P}(A)} + \frac{\mathbb{P}(B^c \cap \{R < |Z_{T_k}^{\tau_j}| < j_0\})}{\mathbb{P}(\{R < |Z_{T_k}^{\tau_j}| < j_0\})}.$$

Now, since

$$||Z_s^{\tau_j}| - |Z_{T_k}^{\tau_j}|| \le \eta_k K|Z_{T_k}^{\tau_j}| + 2\sqrt{2\beta^{-1}}|\zeta_s^{\tau_j}|,$$

then B^c could happen only if

$$2\sqrt{2\beta^{-1}} \sup_{T_k \le s \le t} |\zeta_s^{\tau_j}| \ge \max\left\{j - \frac{3}{2}|Z_{T_k}^{\tau_j}|, \frac{1}{2}|Z_{T_k}^{\tau_j}| - j^{-1}\right\}.$$

Taking j with $j > \frac{3}{2}j_0 + R/4$ and $j^{-1} < R/4$, then

$$\frac{\mathbb{P}(B^c \cap \{R < |Z_{T_k}^{\tau_j}| < j_0\})}{\mathbb{P}(\{R < |Z_{T_k}^{\tau_j}| < j_0\})} \leq \mathbb{P}\left(2\sqrt{2\beta^{-1}} \sup_{T_k \leq s \leq t} |\zeta_s^{\tau_j}| > R/4 | \mathcal{F}_{T_k}\right).$$

By Lemma 3.2, one obtains that

$$\mathbb{P}\left(2\sqrt{2\beta^{-1}} \sup_{T_k \leq s \leq t} |\zeta_s^{\tau_j}| > R/4 | \mathcal{F}_{T_k}\right) \leq 2e^{-\bar{c}\beta R^2 \eta_k^{-1}/128}.$$

Hence, for $\eta_k \leq \frac{\bar{c}\beta R^2}{128 \log 8}$, one has

$$\mathbb{P}(B^C | A) < \frac{1}{2}. \tag{3.40}$$

Consequently,

$$\frac{\mathbb{E}[|\zeta_t^{\tau_j}| \mathbf{1}_A \mathbf{1}_B]}{\mathbb{P}(A \cap B)} \leq \frac{\mathbb{E}[|\zeta_t^{\tau_j}| \mathbf{1}_A]}{\mathbb{P}(A) - \mathbb{P}(B^C \cap A)} = \frac{\mathbb{E}[|\zeta_t^{\tau_j}| | A]}{1 - \mathbb{P}(B^C | A)} \leq 2\mathbb{E}[|\zeta_t^{\tau_j}| \Big| A]$$

and the claim then follows. □

Next, we will make use the same idea to establish a series of conditional expectations, which is based on the tail estimate in Lemma 3.2.

Lemma 3.3. *Let $\eta_k \in (0, 1/2K)$. Then, for $j \geq j_0 + R/4$, it holds that*

$$\mathbb{E}\big[|Z_{T_k}^{\tau_j}| \mathbf{1}_{\{|Z_{T_k}^{\tau_j}| > R\}} \mathbf{1}_{\{t \geq \tau_j\}}\big] \leq (\epsilon(j_0) + 2e^{-\bar{c}\beta R^2 \eta_k^{-1}/128}) \mathbb{E}\big[|Z_{T_k}^{\tau_j}| \mathbf{1}_{\{|Z_{T_k}^{\tau_j}| > R\}}\big], \tag{3.41}$$

where $\epsilon(j_0) \to 0$ as $j_0 \to \infty$.

Proof. Clearly, (3.41) is trivial if $\mathbb{P}(|Z_{T_k}^{\tau_j}| > R) = 0$. Below, we assume that $\mathbb{P}(|Z_{T_k}^{\tau_j}| > R) > 0$.

On one hand, using the result for moment control (Lemma 2.2),

$$\mathbb{E}\big[|Z_{T_k}^{\tau_j}| \mathbf{1}_{\{|Z_{T_k}^{\tau_j}| > j_0\}} \mathbf{1}_{\{t \geq \tau_j\}}\big] \leq \mathbb{E}\big[|Z_{T_k}^{\tau_j}| \mathbf{1}_{\{|Z_{T_k}^{\tau_j}| > j_0\}}\big] \leq \epsilon(j_0) \mathbb{E}\big[|Z_{T_k}^{\tau_j}| \mathbf{1}_{\{|Z_{T_k}^{\tau_j}| > R\}}\big], \tag{3.42}$$

where $\epsilon_0(j_0) \to 0$ uniformly in j as $j_0 \to \infty$.

Fix j_0 and let $|Z_{T_k}^{\tau_j}| = z \in (R, j_0]$. By similar discussion as in the proof of Corollary 3.1, since $\eta_k K \leq 1/2$, for $j \geq j_0 + R/4$ and $j^{-1} \leq R/4$, for $\tau_j \leq t$, one necessarily needs $2\sqrt{2\beta^{-1}} \sup_{T_k \leq s \leq t} |\zeta_t^{\tau_j}| \geq R/4$. Hence, for such j, one has

$$\mathbb{P}(t \geq \tau_j | |Z_{T_k}^{\tau_j}| = z) \leq 2e^{-\bar{c}\beta R^2 \eta_k^{-1}/128}. \tag{3.43}$$

Hence, letting $\mu_{T_k}(dz)$ be the law of $|Z_{T_k}^{\tau_j}|$, one has

$$\mathbb{E}\big[|Z_{T_k}^{\tau_j}| \mathbf{1}_{\{R < |Z_{T_k}^{\tau_j}| \leq j_0\}} \mathbf{1}_{\{t \geq \tau_j\}}\big]$$

$$= \int_R^{j_0} z \mathbb{E}\big[\mathbf{1}_{\{t \geq \tau_j\}} \, | \, |Z_{T_k}^{\tau_j}| = z\big] \mu_{T_k}(dz)$$

$$\begin{aligned}&\le 2e^{-\bar{c}\beta R^2\eta_k^{-1}/128}\int_R^{j_0} z\mu_{T_k}(dz)\\&\le 2e^{-\bar{c}\beta R^2\eta_k^{-1}/128}\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big].\end{aligned} \tag{3.44}$$

Combining (3.42) and (3.44), the claim (3.41) holds. □

The following result follows from the same idea as above, but more involved. It tells us that the two random variables $|Z_t^{\tau_j}|$ and $|Z_{T_k}^{\tau_j}|$ are roughly the same. Note that this result is also based on the tail estimate in Lemma 3.2.

Lemma 3.4. *Under Assumption* 2.1, *for any* $R_1 > 3R/2$ *with* R *obtained in Lemma* 2.1, *and* $\eta_k < \min(\frac{1}{2K}, \frac{1}{16}\beta\bar{c}R_1(R_1-3R/2), \beta\bar{c}R^2/(128\log 5))$, *there exists* $j_0 > R_1$ *such that for all* $j > j_0$,

$$\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1,|Z_{T_k}^{\tau_j}|\le R,t<\tau_j\}}\big] \le \varepsilon(\eta_k)\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\le R,t<\tau_j\}}\big], \tag{3.45}$$

where

$$\varepsilon(\eta) := \frac{15R_1e^{-\bar{c}\beta(R_1-3R/2)^2\eta^{-1}/8}}{R}. \tag{3.46}$$

Proof. The idea is that for the event $\{|Z_t^{\tau_j}| > R_1, |Z_{T_k}^{\tau_j}| \le R\}$ to happen, $|Z|$ must be R for some time during T_k and t. Conditioning on this, $|Z_t^{\tau_j}|$ should be large (roughly comparable to R), while the moment for $|Z_t^{\tau_j}| > R_1$ is then a small fraction of this conditional moment.

Define the event

$$E := \{|Z_{T_k}^{\tau_j}| \le R, \exists\, s\in[T_k,t], |Z_s^{\tau_j}| = R\}.$$

If $\mathbb{P}(E) = 0$, there is nothing to prove. Below, we assume that $\mathbb{P}(E) > 0$. Again, the event E is also almost contained in $\{t < \tau_j\}$. We will in fact show that

$$\begin{aligned}&\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1,|Z_{T_k}^{\tau_j}|\le R,t<\tau_j\}}\big]\\&\quad=\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1,t<\tau_j\}}\big] \le \epsilon(\eta_k)\mathbb{E}[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{t<\tau_j\}}].\end{aligned} \tag{3.47}$$

For a nonnegative random variable X, one has

$$\mathbb{E}[X] = \mathbb{E}\int_0^\infty \mathbf{1}_{\{X>r\}}dr = \int_0^\infty \mathbb{P}(X>r)dr.$$

Then,

$$\begin{aligned}&\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1,t<\tau_j\}}\big]\\&\quad=\int_0^\infty \mathbb{P}(|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1,t<\tau_j\}}>r)dr\\&\quad=R_1\mathbb{P}(|Z_t^{\tau_j}|>R_1,E,t<\tau_j)+\int_{R_1}^\infty \mathbb{P}(|Z_t^{\tau_j}|>r,E,t<\tau_j)dr.\end{aligned}$$

By applying the strong Markov property for $|Z_s^{\tau_j}|$ hitting R and using the same idea in the proof of Corollary 3.1, one has for j large enough that

$$\mathbb{P}(|Z_t^{\tau_j}| > r, E, t < \tau_j) \le 2\mathbb{P}(|Z_t^{\tau_j}| > r|E)\mathbb{P}(E, t < \tau_j).$$

Using the definition of the event E and applying Lemma 3.2 with the strong Markov property for $|Z_s^{\tau_j}| = R$, one has

$$P(|Z_t^{\tau_j}| > r \,|\, E) \le 2e^{-\bar{c}\beta(r-3R/2)^2\eta_k^{-1}/8}, \quad \forall\, r \ge R_1 > 3R/2. \tag{3.48}$$

Hence, one then has

$$\begin{aligned}
&\mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{|Z_t^{\tau_j}|>R_1, t<\tau_j\}}\big]\\
&\quad \le 2\left[R_1 e^{-\bar{c}\beta(R_1-3R/2)^2\eta_k^{-1}/8} + \int_{R_1}^{\infty} e^{-\bar{c}\beta(r-3R/2)^2\eta_k^{-1}/8}dr\right]\mathbb{P}(E, t<\tau_j)\\
&\quad \le 2(R_1 + 8\eta_k\bar{c}^{-1}\beta^{-1}(R_1-3R/2)^{-1})e^{-\bar{c}\beta(R_1-3R/2)^2\eta_k^{-1}/8}\mathbb{P}(E, t<\tau_j)\\
&\quad \le 3R_1 e^{-\bar{c}\beta(R_1-3R/2)^2\eta_k^{-1}/8}\mathbb{P}(E, t<\tau_j).
\end{aligned}$$

Next, we aim to show that

$$\mathbb{E}[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{t<\tau_j\}}] \ge \frac{R}{5}\mathbb{P}(E, t<\tau_j).$$

In fact,

$$\begin{aligned}
\mathbb{E}[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{t<\tau_j\}}] &= \int_0^{\infty} \mathbb{P}(|Z_t^{\tau_j}| > r, E, t<\tau_j)dr\\
&\ge \int_0^{R/4} \mathbb{P}(|Z_t^{\tau_j}| > r, E, t<\tau_j)dr.
\end{aligned}$$

For $r \in [0, R/4]$,

$$\mathbb{P}(|Z_t^{\tau_j}| > r, E, t<\tau_j) \ge \mathbb{P}(E, t<\tau_j) - \mathbb{P}(|Z_t^{\tau_j}| \le R/4, E, t<\tau_j).$$

Suppose s is the stopping time for $|Z^{\tau_j}|$ hitting R during $[T_k, t]$, then one needs

$$2\sqrt{2\beta^{-1}}\left|\int_{s\wedge\tau_j}^{t\wedge\tau_j} \frac{Z_{t'}^{\otimes 2}}{|Z_{t'}|^2}\cdot dW_{t'}\right| \ge R/4,$$

for $|Z_t^{\tau_j}| \le R/4$ and $t < \tau_j$ to happen (if j is large enough).

By strong Markov property and similar estimate as in the proof of Corollary 3.1, one has

$$\begin{aligned}
\mathbb{P}(|Z_t^{\tau_j}| \le R/4, E, t<\tau_j) &\le 2\mathbb{P}\left(|Z_t^{\tau_j}| \le \frac{R}{4}\,\middle|\, E\right)\mathbb{P}(E, t<\tau_j)\\
&\le 4e^{-\bar{c}\beta\frac{R^2}{128}\eta_k^{-1}}\mathbb{P}(E, t<\tau_j).
\end{aligned} \tag{3.49}$$

Hence,

$$\mathbb{P}(|Z_t^{\tau_j}| > r, E, t < \tau_j) \geq \frac{R}{5}\mathbb{P}(E, t < \tau_j)$$

and the claim (3.47) holds.

Note that the event $E \cap \{t < \tau_j\}$ is smaller than $\{|Z_{T_k}^{\tau_j}| \leq R\} \cap \{t < \tau_j\}$ so that

$$\mathbb{E}[|Z_t^{\tau_j}|\mathbf{1}_E\mathbf{1}_{\{t<\tau_j\}}] \leq \mathbb{E}\big[|Z_t^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|\leq R\}}\mathbf{1}_{\{t<\tau_j\}}\big]. \tag{3.50}$$

With (3.47) in hand, the claim then follows. □

Next, we obtain the following estimate for the term $I_1(t)$ defined in (3.16), which explains how we treat the random batch at discrete time T_k and make use of the far away convexity. Note that this result is also based on the tail estimate in Lemma 3.2.

Before the detailed derivation, we first give a brief summary of proof for Proposition 3.1 regarding the estimate for the random batch. After Itô's calculation, one needs to estimate

$$\mathbb{E}[\phi(\bar{X}_t, \bar{Y}_t)(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\mathbf{1}_A],$$

for some function $\phi(\cdot,\cdot)$, $t \in [T_k, T_{k+1})$ and some event A independent of the random batch ξ_k. The key step in Proposition 3.1 is to use Taylor's expansion and consistency of the random batch:

$$\begin{aligned}
&\mathbb{E}[\phi(\bar{X}_t, \bar{Y}_t)(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\mathbf{1}_A] \\
&\quad = \mathbb{E}[\phi(\bar{X}_{T_k}, \bar{Y}_{T_k})(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\mathbf{1}_A] + \epsilon(\eta) \\
&\quad = \mathbb{E}[\phi(\bar{X}_{T_k}, \bar{Y}_{T_k})(\nabla U(\bar{X}_{T_k}) - \nabla U(\bar{Y}_{T_k}))\mathbf{1}_A] + \epsilon(\eta),
\end{aligned}$$

where the last equality is due to $\mathbb{E}_\xi[U^\xi(\cdot)] = U(\cdot)$ and the fact that ξ_k is independent of $\bar{X}_{T_k}$ and $\bar{Y}_{T_k}$. Moreover, under the event A, the small remainder term $\epsilon(\eta)$ can be estimated through the tail behavior obtained in Lemma 3.2. The details are given as follows.

Proposition 3.1. *Let f be the Lyapunov function defined in* (2.7). *Suppose Assumption* 2.1 *holds. Assume that $\eta_k \leq 1/2K$, $KR\eta_k \leq \bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}} \leq 1$ ($\bar{c}$ is the constant coming from Lemma* 3.2). *Denote*

$$c' := \bar{c}^{-1/2}\left(\frac{2K}{R/2-1} + Kc_f e^{-c_f(R/2-1)}\right) + \frac{4\bar{c}^{-1/2}}{R}. \tag{3.51}$$

Then for j sufficiently large, it holds that

$$\begin{aligned}
&\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{t<\tau_j\}}\right] \\
&\quad \leq -(e^{-c_f R_1}\kappa - c'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}} - 3e^{-\bar{c}\beta R^2\eta_k^{-1}/128})\mathbb{E}\big[|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\big].
\end{aligned} \tag{3.52}$$

Proof. We first show that for $j \in \mathbb{N}_+$, $t \in [T_k, T_{k+1})$,

$$\begin{aligned}
&\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\,\middle|\,\mathcal{F}_{T_k}\right] \\
&\quad \le \Bigg(-f'(|Z_{T_k}^{\tau_j}|)\frac{Z_{T_k}^{\tau_j}}{|Z_{T_k}^{\tau_j}|}\cdot(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k})) \\
&\qquad + c'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}|Z_{T_k}^{\tau_j}|\Bigg)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}.
\end{aligned} \tag{3.53}$$

Recall that for $t \in [T_k, T_{k+1})$,

$$Z_t^{\tau_j} = Z_{T_k}^{\tau_j} - (t\wedge\tau_j - T_k\wedge\tau_j)A_{T_k} + 2\sqrt{2\beta^{-1}}\zeta_t^{\tau_j}. \tag{3.54}$$

Noting that $|f'(r)| \le 1$, by Assumption 2.1 and Lemma 3.2, it follows easily that

$$\begin{aligned}
&\left|\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{|\zeta_t^{\tau_j}|>\bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}\}}\,\middle|\,\mathcal{F}_{T_k}\right]\right| \\
&\quad \le 2\eta_k K|Z_{T_k}^{\tau_j}|\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}.
\end{aligned} \tag{3.55}$$

On the other hand, consider the following:

$$\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{|\zeta_t^{\tau_j}|\le\bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}\}}\,\middle|\,\mathcal{F}_{T_k}\right].$$

Consider the function $g : \mathbb{R}^d \to \mathbb{R}$ defined by

$$g(x) := -f'(|x|)\frac{x}{|x|}\cdot A_{T_k}.$$

Here, $A_{T_k} = \nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k})$ is $\mathcal{F}_{T_k}$ measurable, satisfying $|A_{T_k}| \le K|Z_{T_k}|$ by Assumption 2.1. Clearly, for $x \neq 0$, the gradient $\nabla g(x) = -f'(|x|)A_{T_k}\cdot\frac{1}{|x|}(I_d - \frac{x^{\otimes 2}}{|x|^2}) - f''(|x|)\frac{x^{\otimes 2}}{|x|^2}\cdot A_{T_k}$ is well defined. Hence

$$\nabla g(\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j})\cdot(-A_{T_k}) \le \frac{|A_{T_k}|^2}{|\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j}|} \le \frac{K^2|Z_{T_k}^{\tau_j}|^2}{|\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j}|},$$

where we used $f''(r)A_{T_k}\cdot\frac{x^{\otimes 2}}{|x|^2}\cdot A_{T_k} \le 0$. Then, for $|\zeta_t^{\tau_j}| \le \bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}$,

$$\begin{aligned}
g(Z_t^{\tau_j}) &= g(Z_{T_k}^{\tau_j}) + \left(\int_0^1 \nabla g(\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j})d\lambda\right)\cdot\left(Z_t^{\tau_j} - Z_{T_k}^{\tau_j}\right) \\
&\le g(Z_{T_k}^{\tau_j}) + \eta_k K^2\int_0^1 \frac{|Z_{T_k}^{\tau_j}|^2}{|\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j}|}d\lambda \\
&\quad + \left(\int_0^1 |\nabla g(\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j})|d\lambda\right)\bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}.
\end{aligned}$$

Hence, choosing small η_k such that $K\eta_k \le \frac{1}{2}$ and $\bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}} \le 1$, it is not difficult to control $|\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j}| \ge |Z_{T_k}^{\tau_j}|(1-\eta_k K) - 1 \ge |Z_{T_k}^{\tau_j}|(1/2 - 1/R)$. Moreover,

$$
\begin{aligned}
|\nabla g(\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j})| &\le \frac{|A_{T_k}|}{|\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j}|} + c_f|A_{T_k}|\, e^{-c_f|\lambda Z_{T_k}^{\tau_j} + (1-\lambda)Z_t^{\tau_j}|} \\
&\le \left(\frac{K}{R/2-1} + Kc_f e^{-c_f(R/2-1)}\right)|Z_{T_k}^{\tau_j}|.
\end{aligned}
$$

Hence, for $KR\eta_k \le \bar{c}^{-\frac{1}{2}}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}$,

$$
g(Z_t^{\tau_j})\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}} \le (g(Z_{T_k}^{\tau_j}) + \bar{c}'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}|Z_{T_k}^{\tau_j}|)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}},
$$

where

$$
\bar{c}' := \bar{c}^{-1/2}\left(\frac{2K}{R/2-1} + Kc_f e^{-c_f(R/2-1)}\right).
$$

By Lemma 3.2, this then implies

$$
\begin{aligned}
&\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot A_{T_k}\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{|\zeta_t^{\tau_j}|\le\bar{c}^{-1/2}\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}\}}\,\middle|\,\mathcal{F}_{T_k}\right] \\
&\quad\le (1-2\eta_k)\left(-f'(|Z_{T_k}^{\tau_j}|)\frac{Z_{T_k}^{\tau_j}}{|Z_{T_k}^{\tau_j}|}\cdot A_{T_k} + \bar{c}'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}|Z_{T_k}^{\tau_j}|\right)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}} \\
&\quad\le \left(-f'(|Z_{T_k}^{\tau_j}|)\frac{Z_{T_k}^{\tau_j}}{|Z_{T_k}^{\tau_j}|}\cdot A_{T_k} + 2\eta_k K|Z_{T_k}^{\tau_j}| + \bar{c}'\eta_k^{\frac{1}{2}}|\log\eta_k|^{\frac{1}{2}}|Z_{T_k}^{\tau_j}|\right)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}.
\end{aligned}
\tag{3.56}
$$

Combining (3.55) and (3.56), the claim (3.53) holds.

Next, we prove (3.52). Clearly,

$$
\begin{aligned}
&\mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k}))\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{t<\tau_j\}}\right] \\
&\quad= \mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot\left(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k})\right)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\right] \\
&\qquad- \mathbb{E}\left[-f'(|Z_t^{\tau_j}|)\frac{Z_t^{\tau_j}}{|Z_t^{\tau_j}|}\cdot\left(\nabla U^{\xi_k}(\bar{X}_{T_k}) - \nabla U^{\xi_k}(\bar{Y}_{T_k})\right)\mathbf{1}_{\{|Z_{T_k}^{\tau_j}|>R\}}\mathbf{1}_{\{t\ge\tau_j\}}\right] \\
&\quad= B_1 + B_2.
\end{aligned}
$$

For the first term, using (3.53), the consistency of the random batch ξ, and the convexity condition in Assumption 2.1, one has

$$\begin{aligned}B_1 &\le \mathbb{E}\left[\left(-f'(|Z^{\tau_j}_{T_k}|)\frac{Z^{\tau_j}_{T_k}}{|Z^{\tau_j}_{T_k}|}\cdot\big(\nabla U^{\xi_k}(\bar X_{T_k})-\nabla U^{\xi_k}(\bar Y_{T_k})\big)\right.\right.\\&\qquad\left.\left.+c'\eta_k^{\frac12}|\log\eta_k|^{\frac12}|Z^{\tau_j}_{T_k}|\right)\mathbf{1}_{\{|Z^{\tau_j}_{T_k}|>R\}}\right]\\&=\mathbb{E}\left[\left(-f'(|Z^{\tau_j}_{T_k}|)\frac{Z^{\tau_j}_{T_k}}{|Z^{\tau_j}_{T_k}|}\cdot\big(\nabla U(\bar X_{T_k})-\nabla U(\bar Y_{T_k})\big)\right.\right.\\&\qquad\left.\left.+c'\eta_k^{\frac12}|\log\eta_k|^{\frac12}|Z^{\tau_j}_{T_k}|\right)\mathbf{1}_{\{|Z^{\tau_j}_{T_k}|>R\}}\right]\\&\le-(e^{-c_fR_1}\kappa-c'\eta_k^{\frac12}|\log\eta_k|^{\frac12})\mathbb{E}\big[|Z^{\tau_j}_{T_k}|\mathbf{1}_{\{|Z^{\tau_j}_{T_k}|>R\}}\big],\end{aligned}\tag{3.57}$$

where we used the convexity outside $B(0,R)$ in the last inequality.

For the second term, by Lipschitz condition in Assumption 2.1, we have

$$|B_2|\le K\mathbb{E}\big[|Z^{\tau_j}_{T_k}|\mathbf{1}_{\{|Z^{\tau_j}_{T_k}|>R\}}\mathbf{1}_{\{t\ge\tau_j\}}\big].\tag{3.58}$$

Finally, combining (3.57), (3.58) and (3.41) in Lemma 3.3, we conclude that for j large enough

$$\begin{aligned}&\mathbb{E}\left[-f'(|Z^{\tau_j}_t|)\frac{Z^{\tau_j}_t}{|Z^{\tau_j}_t|}\cdot(\nabla U^{\xi_k}(\bar X_{T_k})-\nabla U^{\xi_k}(\bar Y_{T_k}))\mathbf{1}_{\{|Z^{\tau_j}_{T_k}|>R\}}\mathbf{1}_{\{t<\tau_j\}}\right]\\&\quad\le-(e^{-c_fR_1}\kappa-c'\eta_k^{\frac12}|\log\eta_k|^{\frac12}-3e^{-\bar c\beta R^2\eta_k^{-1}/128})\mathbb{E}\big[|Z^{\tau_j}_{T_k}|\mathbf{1}_{\{|Z^{\tau_j}_{T_k}|>R\}}\big].\end{aligned}\tag{3.59}$$

□

Lemma 3.5. *It holds that τ_j is nondecreasing in j and*

$$\tau_j\to\tau,\quad a.s.,\tag{3.60}$$

where τ is a stopping time defined in (3.3).

Proof of Lemma 3.5. It is clear that τ_j is nondecreasing in j and $\sup_j\tau_j\le\tau$. Fix $T>0$. By Hölder's inequality and Lemma 2.2, for any $T>0$,

$$\begin{aligned}\mathbb{E}\left[\sup_{s\le t}|Z_s|\right]&\le\left(\mathbb{E}\left[\sup_{s\le t}|Z_s|^2\right]\right)^{\frac12}\\&\le\left(2\mathbb{E}\left[\sup_{s\le t}|\bar X_s|^2\right]+2\mathbb{E}\left[\sup_{s\le t}|\bar Y_s|^2\right]\right)^{\frac12}\le C(T),\quad\forall\,t\in[0,T],\end{aligned}\tag{3.61}$$

where $C(T)$ is a positive constant. Then, $\lim_{j\to\infty}\mathbb{P}(|Z_{\tau_j\wedge T}|\geq j)\leq\lim_{j\to\infty}\frac{C(T)}{j}=0$. By continuity and the definition of τ_j, τ, one has

$$\sup_j \tau_j\wedge T=\tau\wedge T.$$

Since T is arbitrary, $\tau_j\to\tau$. □

Remark 3.1. Lemma 3.5 can also be proved based on moment control for the stopped process $\bar{X}_s^{\tau_j}=\bar{X}_{s\wedge\tau_j}$, which is a weaker result compared with Lemma 2.2, and is in fact easier to prove than the moment control of $\sup_{s\leq t}|\bar{X}_s|$ in Lemma 2.2.

Remark 3.2 (Discussion on dimension dependency for the contraction rate). We believe that the contraction rate c in our result is dimension-free. The only place that might be influenced by the dimension d is the positive constant C_{2p} in (3.36) coming from the BDG inequality. So far, we have not found any reference claiming that the constant in the BDG inequality is independent of the dimension d. However, the process ζ_t defined in (3.7) resembles a 1D Brownian Motion, since the rank of the matrix $Z_t^{\otimes 2}/|Z_t|^2$ is one with trace to be 1. If the direction $Z_t/|Z_t|$ does not change, then it is exactly 1D Brownian motion. The difference is that the direction is changing along time.

4. General Drift Case

In many applications, the drift term may not be of the form $-\nabla U$. In this section, we generalize the results to the general diffusion processes where the drifts are no longer gradients, namely, one still has ergodicity for the random batch version of the Euler–Maruyama scheme for diffusion processes.

Consider the time continuous diffusion process

$$dX=b(X)dt+\sqrt{2\beta^{-1}}dW, \tag{4.1}$$

where $b(\cdot)$ is a given drift, which might not be a gradient field (in this case, X is no longer a Langevin diffusion). Let b^{ξ} be an unbiased stochastic estimate of b, $\mathbb{E}(b^{\xi}(\cdot))=b(\cdot)$. The random batch version of the Euler–Maruyama scheme for the continuous SDE (4.1), the correspondence of SGLD iteration, is then given by

$$\bar{X}_{T_{k+1}}=\bar{X}_{T_k}+\eta_k b^{\xi_k}(\bar{X}_{T_k})+\sqrt{2\beta^{-1}}(W_{T_{k+1}}-W_{T_k}). \tag{4.2}$$

Similarly with (1.2), we only need to analyze the following time interpolation:

$$\bar{X}_t=\bar{X}_{T_k}-\int_{T_k}^{t}b^{\xi_k}(\bar{X}_{T_k})ds+\int_{T_k}^{t}\sqrt{2\beta^{-1}}dW_s,\quad t\in[T_k,T_{k+1}),\ \ k=0,1,\ldots. \tag{4.3}$$

In order to obtain similar contraction property, we need the following assumption, which corresponds with Assumption 2.1.

Assumption 4.1. There exist $R > 0$, $\kappa > 0$, $K > 0$ such that the followings hold:

(a) $\forall |x-y| > R$,

$$-(x-y)\cdot(b(x)-b(y)) \ge \kappa |x-y|^2; \tag{4.4}$$

(b) $\forall\, x, y \in \mathbb{R}^d$, $\forall\, \xi \in \mathcal{S}$,

$$|b^{\xi}(x) - b^{\xi}(y)| \le K|x-y|. \tag{4.5}$$

Again, the first assumption can be obtained if for some $\kappa_0 > 0$ and $R_0 > 0$ the following holds:

$$-v\cdot\nabla b(x)\cdot v \ge \kappa_0 |v|^2, \quad \forall\, x \in \mathbb{R}^d \backslash B(0, R_0), \ \ \forall\, v \in \mathbb{R}^d. \tag{4.6}$$

Then, we are able to prove the contraction property for the algorithm (4.2), which naturally implies geometric ergodicity for constant step size $\eta_k \equiv \eta$ by contraction mapping theorem [18].

Theorem 4.1 (Wasserstein contraction with general drift). *Consider the random Euler–Maruyama iteration* (4.3). *For any two initial distributions μ_0 and ν_0, denote μ_{T_k} and ν_{T_k} to be the corresponding laws at T_k. Denote $h := \sup_k \eta_k$. Let f be the Lyapunov function defined in* (2.7). *Then under Assumption* 4.1, *for fixed $R_1 = 2R$ and c_f satisfying $\frac{1}{3}\sqrt{2\beta^{-1}}c_f R^{-1} - K \ge 0$, there exists $\delta > 0$ such that for $h \le \delta$, the following Wasserstein contraction result holds*:

$$W_f(\mu_{T_k}, \nu_{T_k}) \le e^{-cT_k} W_f(\mu_0, \nu_0), \quad k \in \mathbb{N}, \tag{4.7}$$

where

$$c = \frac{1}{3} e^{-2c_f R} \min(\sqrt{2\beta^{-1}} c_f R^{-1}/2, \kappa).$$

Consequently,

$$W_1(\mu_{T_k}, \nu_{T_k}) \le c_0 e^{-cT_k} W_1(\mu_0, \nu_0), \quad k \in \mathbb{N}, \ \ c_0 := e^{2c_f R}. \tag{4.8}$$

Moreover, if $\eta_k \equiv \eta$ is a constant, then for $\eta \le \delta$, the iteration (4.2) *has a unique invariant distribution $\tilde{\pi}$ such that*

$$W_1(\mu_{T_k}, \tilde{\pi}) \le c_0 e^{-cT_k} W_1(\mu_0, \tilde{\pi}), \quad k \in \mathbb{N}. \tag{4.9}$$

The proof is almost the same as that of Theorem 2.1. The only difference is that we use condition (a) in Assumption 4.1 instead of the convexity condition (condition (a) in Assumption 2.1) when estimating the term I_3 near (3.17).

5. Conclusion

In this paper, We proved the geometric ergodicity of the SGLD algorithm under nonconvexity settings. As a popular online sampling algorithm, SGLD has shown exceptional performance when dealing with high-dimensional and large-scaled data.

Via the technique of reflection coupling, in Theorem 2.1 we proved the Wasserstein contraction of SGLD when the target distribution is log-concave only outside some compact sets. In particular, the time discretization and the minibatch in SGLD introduced several difficulties when applying the reflection coupling, which were addressed by a series of careful estimates of conditional expectations. As a direct corollary, we proved that the SGLD with constant step size has an invariant distribution and obtained its geometric ergodicity in terms of W_1 distance. The generalization to non-gradient drifts was also included. We also remarked that the contraction rate c in Theorem 2.1 is intuitively dimension-free, as discussed in Remark 3.2. Remarkably, we believe that many techniques in this article are applicable to other discrete algorithms involving random batches.

Acknowledgments

This work is financially supported by the National Key R&D Program of China, Project Numbers 2021YFA1002800 and 2020YFA0712000. The work of L. Li was partially supported by NSFC 12371400 and 12031013, Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102, Shanghai Science and Technology Commission (Grant No. 21JC1403700, 21JC1402900). The work of J.-G. Liu was partially supported by NSF DMS-2106988.

Appendix A. Missing Proofs

In this section, we give the detailed proofs for Lemmas 2.1 and 2.2.

Proof of Lemma 2.1. Fix $r > 2R_0$, and choose two arbitrary point $x, y \in \mathbb{R}^d$ satisfying $|x-y| = r$. Then it holds that

$$\frac{(x-y)\cdot(\nabla U(x) - \nabla U(y))}{|x-y|^2} = \frac{(x-y)\cdot\int_0^1 \nabla^2 U(tx+(1-t)y)dt\cdot(x-y)}{|x-y|^2}. \tag{A.1}$$

For fixed x, y above, denote

$$A_1 := \{t \in (0,1) : |tx + (1-t)y| \le R_0\}$$

and

$$A_2 := \{t \in (0,1) : |tx + (1-t)y| > R_0\}.$$

Then, by Assumption 2.1,

$$\begin{aligned}\int_0^1 \nabla^2 U(tx+(1-t)y)dt &= \left(\int_{A_1} + \int_{A_2}\right) \nabla^2 U\left(tx+(1-t)y\right) dt \\ &\succeq (-m(A_1)KI_d) + (m(A_2)\kappa_0 I_d),\end{aligned} \tag{A.2}$$

where m denotes the Lebesgue measure in $\mathbb{R}^1$. Clearly, $\{tx + (1-t)y : t \in (0,1)\}$ is a segment connecting the two point x, y in $\mathbb{R}^d$. Then by definition of the ball

$B(0, R_0) := \{x \in \mathbb{R}^d : |x| < R_0\}$, the longest segment contained in $B(0, R_0)$ is of length $2R_0$. Therefore,

$$m(A_1) \le \frac{2R_0}{r}, \quad m(A_2) = 1 - m(A_1) \ge 1 - \frac{2R_0}{r}. \tag{A.3}$$

Combining (A.1)–(A.3), for all $x, y \in \mathbb{R}^d$ satisfying $|x - y| = r$,

$$\frac{(x-y)\cdot(\nabla U(x) - \nabla U(y))}{|x-y|^2} \ge -\frac{2R_0}{r} K + \left(1 - \frac{2R_0}{r}\right)\kappa_0 = \kappa_0 - \frac{2R_0}{r}(K + \kappa_0). \tag{A.4}$$

Choosing $R := \max(4R_0(K + \kappa_0)/\kappa_0, 2)$, the conclusion (2.3) then holds with $\kappa := \kappa_0/2$. □

Next, we prove the result of p-th moment control for SGLD.

Proof of Lemma 2.2. We first control the moments of $\bar{X}_t$, namely $\mathbb{E}|\bar{X}_t|^p$ on $[0, T]$, and then prove the moment control of $\sup_{t \le T} |\bar{X}_t|$.

We take $p \ge 2$ first. By Itô's formula, for $t \in [T_k, T_{k+1})$, we have

$$\begin{aligned} d|\bar{X}_t|^p = {} & -p|\bar{X}_t|^{p-2}\bar{X}_t \cdot \nabla U^{\xi_k}(\bar{X}_{T_k})dt + \beta^{-1}p|\bar{X}_t|^{p-2}\left(I_d + (p-2)\frac{\bar{X}_t^{\otimes 2}}{|\bar{X}_t|^2}\right) : I_d\, dt \\ & - p|\bar{X}_t|^{p-2}\bar{X}_t \cdot \sqrt{2\beta^{-1}}dW. \end{aligned} \tag{A.5}$$

Note that $(I_d + (p-2)\frac{\bar{X}_t^{\otimes 2}}{|\bar{X}_t|^2}) : I_d = p + d - 2$. This implies that

$$\frac{d}{dt}\mathbb{E}|\bar{X}_t|^p = \mathbb{E}[-p|\bar{X}_t|^{p-2}\bar{X}_t \cdot \nabla U^{\xi_k}(\bar{X}_{T_k})] + \beta^{-1}p(p+d-2)\mathbb{E}|\bar{X}_t|^{p-2}. \tag{A.6}$$

By the Lipschitz condition in Assumption 2.1, we can directly obtain

$$\frac{d}{dt}\mathbb{E}|\bar{X}_t|^p \le (p-1)\mathbb{E}|\bar{X}_t|^p + C(p, K)(1 + \mathbb{E}|\bar{X}_{T_k}|^p) + \beta^{-1}p(p+d-2)\mathbb{E}|\bar{X}_t|^{p-2}.$$

This easily yields

$$\sup_{t \le T} \mathbb{E}|\bar{X}_t|^p < \infty, \tag{A.7}$$

where the upper bound depends on p, T, d but is independent of ξ_k.

Next, we prove

$$\sup_{0 \le t \le T} \mathbb{E}\left[\sup_{0 \le s \le t} |\bar{X}_s|^p\right] < +\infty. \tag{A.8}$$

Note that $|\bar{X}_t|^p = |\bar{X}_0|^p + M_t + A_t$, where

$$M_t := \int_0^t \sqrt{2\beta^{-1}}p|\bar{X}_s|^{p-2}\bar{X}_s \cdot dW_s, \quad \forall\, t \ge 0 \tag{A.9}$$

and

$$
\begin{aligned}
A_t &:= -\int_0^t p|\bar{X}_s|^{p-2}\bar{X}_s \cdot b_s ds + \int_0^t \beta^{-1}p(p+d-2)|\bar{X}_s|^{p-2}ds, \\
b_t &:= \nabla U^{\xi_k}(\bar{X}_{T_k}), \quad \forall\, t \in [T_k, T_{k+1}).
\end{aligned} \tag{A.10}
$$

Then,

$$
\mathbb{E}\left[\sup_{0\le s\le t} |\bar{X}_s|^p\right] \le \mathbb{E}|\bar{X}_0|^p + \mathbb{E}\left[\sup_{0\le s\le t} M_s\right] + \mathbb{E}\left[\sup_{0\le s\le t} A_s\right]. \tag{A.11}
$$

Clearly M_t is a martingale. By BDG inequality [22], one has

$$
\mathbb{E}\left[\sup_{0\le s\le t} M_s\right] \le 4\sqrt{2}\mathbb{E}[\langle M\rangle_t^{\frac{1}{2}}] = 8\sqrt{\beta^{-1}}p\mathbb{E}\left[\left(\int_0^t |\bar{X}_s|^{2p-2}ds\right)^{\frac{1}{2}}\right].
$$

Then using Jensen's inequality and (A.7), one has

$$
\mathbb{E}\left[\sup_{0\le s\le t} M_s\right] \le 8\sqrt{\beta^{-1}}p\left(\int_0^t \mathbb{E}|\bar{X}_s|^{2p-2}ds\right)^{\frac{1}{2}} \le C_2. \tag{A.12}
$$

For the term $\mathbb{E}[\sup_{0\le s\le t} A_s]$, using the Lipshitz condition in Assumption 2.1, we first observe that for $t \in [0,T]$,

$$
A_t \le \int_0^t p|\bar{X}_s|^{p-1}\left(K\sup_{0\le u\le s}|\bar{X}_u| + b_0\right)ds + C_3\int_0^t |\bar{X}_s|^{p-2}ds,
$$

where b_0, C_3 are time-independent positive constants. Applying Young's equality, we have

$$
\mathbb{E}\left[\sup_{0\le s\le t} A_s\right] \le (pK+1)\int_0^t \mathbb{E}\left[\sup_{0\le u\le s}|\bar{X}_u|^p\right]ds + C_4. \tag{A.13}
$$

Combining (A.11)–(A.13) together, one has that

$$
\mathbb{E}\left[\sup_{0\le s\le t}|\bar{X}_s|^p\right] \le C_5 + (pK+1)\int_0^t \mathbb{E}\left[\sup_{0\le u\le s}|\bar{X}_u|^p\right]ds, \quad \forall\, t\in[0,T]. \tag{A.14}
$$

Hence, by Grönwall's inequality, for all $t \in [0,T]$, we have

$$
\mathbb{E}\left[\sup_{0\le s\le t}|\bar{X}_s|^p\right] \le C_5 e^{(pK+1)T}. \tag{A.15}
$$

The bound for $p \in [1,2]$ then follows easily by Hölder's inequality.

Next, we aim to establish the uniform moment control of $\bar{X}_t$ for η_k being sufficiently small. Starting with (A.6), the first term on the right-hand side may be written as

$$\begin{aligned}
&\mathbb{E}[-p|\bar{X}_t|^{p-2}\bar{X}_t \cdot \nabla U^{\xi_k}(\bar{X}_{T_k})] \\
&\quad = \mathbb{E}[-p|\bar{X}_{T_k}|^{p-2}\bar{X}_{T_k} \cdot \nabla U^{\xi_k}(\bar{X}_{T_k})] \\
&\qquad + p\mathbb{E}\left[\int_0^1 \nabla h(\lambda\bar{X}_t + (1-\lambda)\bar{X}_{T_k})d\lambda \cdot (\bar{X}_t - \bar{X}_{T_k})\right] \\
&\qquad := p(K_1 + K_2),
\end{aligned} \tag{A.16}$$

where the function h is defined by $h(x) := -|x|^{p-2}x \cdot \nabla U^{\xi_k}(\bar{X}_{T_k})$, and has a well-defined gradient $\nabla h(x) = -\nabla U^{\xi_k}(\bar{X}_{T_k}) \cdot |x|^{p-2}(I_d + (p-2)\frac{x^{\otimes 2}}{|x|^2})$. Since $\bar{X}_t - \bar{X}_{T_k} = -(t-T_k)\nabla U^{\xi_k}(\bar{X}_{T_k}) + \sqrt{2\beta^{-1}}(W_t - W_{T_k})$,

$$\begin{aligned}
K_2 &\le (p-1)\eta_k \int_0^1 \mathbb{E}|\lambda\bar{X}_t + (1-\lambda)\bar{X}_{T_k}|^{p-2}|\nabla U^{\xi_k}(\bar{X}_{T_k})|^2 d\lambda \\
&\quad + (p-1)\sqrt{2\beta^{-1}} \int_0^1 \mathbb{E}|\lambda\bar{X}_t + (1-\lambda)\bar{X}_{T_k}|^{p-2}|\nabla U^{\xi_k}(\bar{X}_{T_k})| \left|\int_{T_k}^t dW\right| d\lambda.
\end{aligned}$$

Note that $p-2 \ge 0$, $|\lambda\bar{X}_t + (1-\lambda)\bar{X}_{T_k}|^{p-2} \le \max(|\bar{X}_t|^{p-2}, |\bar{X}_{T_k}|^{p-2}) \le (|\bar{X}_t|^{p-2} + |\bar{X}_{T_k}|^{p-2})$. Then, one has

$$\begin{aligned}
K_2 &\le (p-1)\eta_k(1+\delta_1)K^2\mathbb{E}(|\bar{X}_t|^{p-2}|\bar{X}_{T_k}|^2 + |\bar{X}_{T_k}|^p) \\
&\quad + C_{\delta_1} b_0^2 (p-1)\eta_k \mathbb{E}(|\bar{X}_t|^{p-2} + |\bar{X}_{T_k}|^{p-2}) \\
&\quad + \sqrt{\beta^{-1}\eta_k} C(p,d)[(\mathbb{E}|\bar{X}_t|^p)^{(p-1)/p} + (\mathbb{E}|\bar{X}_{T_k}|^p)^{(p-1)/p} + 1].
\end{aligned}$$

For the term K_1, by Assumption 2.1 and Lemma 2.1,

$$-x \cdot \nabla U(x) \le -\kappa|x|^2 + b_0|x| + C(R).$$

Using the consistency of the random batch ξ, we have

$$\begin{aligned}
K_1 &= \mathbb{E}[-|\bar{X}_{T_k}|^{p-2}\bar{X}_{T_k} \cdot \nabla U(\bar{X}_{T_k})] \\
&= \mathbb{E}[-|\bar{X}_{T_k}|^{p-2}\bar{X}_{T_k} \cdot \nabla U(\bar{X}_{T_k})] \\
&\le -\kappa\mathbb{E}|\bar{X}_{T_k}|^p + b_0\mathbb{E}|\bar{X}_{T_k}|^{p-1} + C(R)\mathbb{E}|\bar{X}_{T_k}|^{p-2}.
\end{aligned} \tag{A.17}$$

Let $\epsilon_k = (p-1)\eta_k(1+\delta_1)K^2$. Then by Young's inequality, we conclude that

$$\begin{aligned}
p(K_1 + K_2) \le &-\left(\kappa - \epsilon_k\left(1 + \frac{2}{p}\right) - \delta_2\right)\mathbb{E}|\bar{X}_{T_k}|^p \\
&+ \left(\epsilon_k\left(1 - \frac{2}{p}\right) + \delta_2\right)\mathbb{E}|\bar{X}_t|^p + C.
\end{aligned} \tag{A.18}$$

Letting $u(t) := \mathbb{E}|\bar{X}_t|^p$, one then has for $t \in [T_k, T_{k+1}]$ that

$$\dot{u}(t) \le -\left(\kappa - \epsilon_k\left(1 + \frac{2}{p}\right) - \delta_2\right)u(T_k) + \left(\epsilon_k\left(1 - \frac{2}{p}\right) + \delta_2\right)u(t) + C(\delta_1, \delta_2, p, d). \tag{A.19}$$

For $0 < \lambda_1 < \lambda_2$ and $v \ge 0$ satisfying

$$\dot{v} \le \lambda_1 u(t) - \lambda_2 u(T_k) + C,$$

one may obtain by Grönwall's inequality that

$$v(T_{k+1}) \le \left(e^{\lambda_1\eta_k} - \frac{1}{\lambda_1}(e^{\lambda_1\eta_k} - 1)\lambda_2\right)v(T_k) + C\frac{1}{\lambda_1}(e^{\lambda_1\eta_k} - 1).$$

However, since

$$e^{\lambda_1\eta_k} - \frac{1}{\lambda_1}(e^{\lambda_1\eta_k} - 1)\lambda_2 = \left(1 - \frac{\lambda_2}{\lambda_1}\right)e^{\lambda_1\eta_k} + \frac{\lambda_2}{\lambda_1} \le 1 + (\lambda_1 - \lambda_2)\eta_k,$$

one then has

$$v(T_{k+1}) \le [1 + (\lambda_1 - \lambda_2)]v(T_k)C\frac{1}{\lambda_1}(e^{\lambda_1\eta_k} - 1).$$

We apply this elementary derivation for $\lambda_1 = \epsilon_k(1 - 2/p) + \delta_2$ and $\lambda_2 = \kappa - \epsilon_k(1 + 2/p) - \delta_2$, then obtain

$$u(T_{k+1}) \le [1 - (\kappa - 2\epsilon_k - 2\delta_2)\eta_k]u(T_k) + C(p, d, \beta, \delta_1, \delta_2, \eta_k).$$

Since we can choose δ_1 and δ_2 small, by the condition $\eta_k \le \kappa/(2(p-1)K^2) - \delta$ given, $\kappa - 2\epsilon_k - 2\delta_2$ is bounded below by a positive number and $C(p, d, \beta, \delta_1, \delta_2, \eta_k)$ has a uniform upper bound in k. Moreover, since $\kappa \le K$, $(\kappa - 2\epsilon_k - 2\delta_2)\eta_k < 1$. The claim then follows. □

Appendix B. Details for Construction of Reflection Coupling and Lyapunov Function

Here we present more details for the principal method employed in this study — reflection coupling equipped with a specific Lyapunov function $f(\cdot)$, as described in the introduction.

Consider the two time marginal distributions $\rho_t^{(1)}$, $\rho_t^{(2)}$ of some SDEs (in our result, it is (1.2)), starting from the initial distributions $\rho_0^{(1)}$, $\rho_0^{(2)}$, respectively. As has been discussed in the introduction, here we aim to prove the contraction property:

$$W_f(\rho_t^{(1)}, \rho_t^{(2)}) \lesssim e^{-ct}W_f(\rho_0^{(1)}, \rho_0^{(2)}).$$

Here, $f(\cdot)$ is some suitable Lyapunov function and $W_f(\cdot,\cdot)$ is the Kantorovich–Rubinstein distance associated with the cost function $f(\cdot)$. The reflection coupling method begins with choosing the pair of initial points (X_0, Y_0) such that $\mathbb{E}f(|X_0 - Y_0|) = W_f(\rho_0^{(1)}, \rho_0^{(2)})$. Then we choose a realization $\bar{X}_t$ of SGLD (1.2) such that the law of X_t is $\rho_t^{(1)}$ and the law of X_0 is $\rho_0^{(1)}$. The key step in the reflection coupling method is that we construct a companion process Y_t with Y_0 coupled above with X_0 and satisfies: (i) Y_t shares the same Brownian motion with $\bar{X}_t$, and has an additional reflection term in its diffusion part, and Y_t also shares the same random batch ξ_k at each T_k in our SGLD setting (1.2); (ii) Y_t is also a realization of the same SDE for X_t and the law of Y_t is $\rho_t^{(2)}$. Then the contraction property mentioned above is reduced to estimation of the negative Lyapunov exponent for the paired dynamics (X_t, Y_t). Namely, we aim to show that

$$\mathbb{E}f(|X_t - Y_t|) \le Ce^{-Ct}\mathbb{E}f(|X_0 - Y_0|).$$

In the followings, we will first introduce the necessity of using the technique of reflection coupling, and then introduce the motivation of the construction of the reflection coupling and the associated Lyapunov function. Note that the geometric ergodicity arises from the strong convexity of the potential $U(\cdot)$ outside some compact sets. In fact, by strong monotonicity property in Lemma 2.1

$$(x - y) \cdot (\nabla U(x) - \nabla U(y)) \ge \kappa |x - y|^2,$$

any such pair (X_t, Y_t) would attract each other if they are sufficiently far away. Take the following numerical scheme for SDE as a simple illustration:

$$X^{n+1} = X^n - \eta \nabla U(X^n) + \sqrt{\eta}\zeta, \quad \zeta \sim \mathcal{N}(0, 1). \tag{B.1}$$

In the settings of this paper, on one hand, as mentioned above, the strong convexity outside some compact sets of the potential $U(\cdot)$ in the drift implies that any paired iteration (X^n, Y^n) associated with (B.1) would attract each other if they are far away. On the other hand, the external force is weak inside the compact set since in this area the potential $U(\cdot)$ does not have strong convexity. In this case, the diffusion term would dominates the drift, since $c_1\sqrt{\eta} \le c_2\eta$, where c_1, c_1 are of $O(1)$. Therefore, at first glance, one cannot directly prove the contraction when the diffusion overshadows the drift. Nonetheless, the application of reflection coupling [10] offers a resolution by facilitating the closer convergence of two particles X_t, Y_t even within the compact set. Take the following overdamped Langevin diffusion for example:

$$dX_t = b(X_t)dt + dW, \quad X|_{t=0} = X_0.$$

The reflection coupling method for the overdamped Langevin diffusion considers another slave copy of X_t, which shares the same Brownian motion but has a reflection term in the diffusion part:

$$dY_t = b(Y_t)dt + \left(I_d - 2\frac{(X_t - Y_t)^{\otimes 2}}{|X_t - Y_t|^2}\right) \cdot dW, \quad Y|_{t=0} = Y_0.$$

It can be shown that the diffusion with a reflection is still a Brownian motion (see [10] or Lemma 3.1), so Y_t is also a realization of the overdamped Langevin diffusion. With the reflection matrix $(I_d - 2\frac{(X_t - Y_t)^{\otimes 2}}{|X_t - Y_t|^2})$, the two particles X_t, Y_t

would eventually move towards each other when the diffusion dominates the drift. In fact, from the reflection operator $(I_d - 2\frac{(X_t-Y_t)^{\otimes 2}}{|X_t-Y_t|^2})$, the Brownian motion can be either approaching or depart from each other. However, the restored force would prevent X_t, Y_t from going too far away from each other. With this intuitive picture of how the reflected couple (X_t, Y_t) moves, it is then left to prove the contraction via practical calculation, namely, one needs to find some Lyapunov function $f(\cdot)$ satisfying

(1) $C_1 r \le f(r) \le C_2 r$ for all r.
(2) $\mathbb{E}f(|X_t - Y_t|)$ decays exponentially in time.

Below we discuss a bit on our motivation for how to fund such Lyapunov function. One can see from Itô's formula that

$$\begin{aligned}&\frac{d}{dt}\mathbb{E}[f(|X_t - Y_t|)]\\ &\quad= \mathbb{E}\left[f''(|X_t - Y_t|) + f'(|X_t - Y_t|)\frac{X_t - Y_t}{|X_t - Y_t|}\cdot(b(X_t) - b(Y_t))\right]. \qquad \text{(B.2)}\end{aligned}$$

Since the goal is to obtain an estimate of the form

$$\frac{d}{dt}\mathbb{E}[f(|X_t - Y_t|)] \lesssim -\mathbb{E}[f(|X_t - Y_t|)],$$

one naturally requires the following conditions when constructing such f: (1) $C_1 r \le f(r) \le C_2 r$; (2) $|f'(r)| \le L$; (3) $f''(r) \le -C_3 r$ for all $r < R_1$, where R_1 is some positive constant larger than R. If these conditions are satisfied, then one can see from (B.2) that

$$\begin{aligned}&\frac{d}{dt}\mathbb{E}[f(|X_t - Y_t|)\mathbf{1}_{\{|X_t - Y_t| < R\}}]\\ &\quad\le \mathbb{E}[(-C_3|X_t - Y_t| + L\|b'\|_\infty|X_t - Y_t|)\mathbf{1}_{\{|X_t - Y_t| < R\}}]\\ &\quad\lesssim -\mathbb{E}[f(|X_t - Y_t|)\mathbf{1}_{\{|X_t - Y_t| < R\}}],\end{aligned}$$

provided that C_3 is relatively large. This then motivates one to seek a concave increasing Lyapunov function f of the form

$$f(r) := \int_0^r e^{-c_f(s\wedge R_1)}ds, \quad r \ge 0.$$

for some positive c_f, R_1 to be determined (in our result for SGLD, we choose $R_1 = 2R$ and the required condition for c_f is stated in (2.9)). Then one can obtain the contraction property for this reflection coupled continuous dynamics (X_t, Y_t).

ORCID

Lei Li https://orcid.org/0000-0001-5304-8380
Jian-Guo Liu https://orcid.org/0000-0002-9911-4045
Yuliang Wang https://orcid.org/0000-0001-9769-7411

References

1. M. T. Barlow and M. Yor, Semi-martingale inequalities via the Garsia–Rodemich–Rumsey lemma, and applications to local times, *J. Funct. Anal.* **49**(2) (1982) 198–229.
2. N. Bou-Rabee and A. Eberle, Two-scale coupling for preconditioned Hamiltonian Monte Carlo in infinite dimensions, *Stoch. Partial Differ. Equ.: Anal. Comput.* **9**(1) (2021) 207–242.
3. N. B.-Rabee, A. Eberle and R. Zimmer, Coupling and convergence for Hamiltonian monte carlo, *Ann. Appl. Probab.* **30**(3) (2020) 1209–1250.
4. N. Brosse, A. Durmus and E. Moulines, The promises and pitfalls of stochastic gradient Langevin dynamics, *Adv. Neural Inf. Process. Syst.* **31** (2018) 8278–8288.
5. E. Carlen and P. Kree, Lp estimates on iterated stochastic integrals, *Ann. Probab.* **19**(1) (1991) 354–368.
6. N. Huy Chau, É. Moulines, M. Rásonyi, S. Sabanis and Y. Zhang, On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case, *SIAM J. Math. Data Sci.* **3**(3) (2021) 959–986.
7. R. L. Dobrushin, Vlasov equations, *Functional Anal. Appl.* **13** (1979) 115–123.
8. A. Durmus, A. Eberle, A.Guillin and R. Zimmer, An elementary approach to uniform in time propagation of chaos, *Proc. Amer. Math. Soc.* **148**(12) (2020) 5387–5398.
9. R. Durrett, *Stochastic Calculus: A Practical Introduction* (CRC Press, 2018).
10. A. Eberle, Reflection coupling and Wasserstein contractivity without convexity, *Comptes Rendus Math.* **349**(19–20) (2011) 1101–1104.
11. A. Eberle, Reflection couplings and contraction rates for diffusions, *Probab. Theory a Related Fields* **166**(3) (2016) 851–886.
12. A. Eberle, A. Guillin and R. Zimmer, Couplings and quantitative contraction rates for Langevin dynamics, *Ann. Probab.* **47**(4) (2019) 1982–2010.
13. A. Eberle, A. Guillin and R. Zimmer, Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes, *Trans. Amer. Math. Soc.* **371**(10) (2019) 7135–7173.
14. T. Farghly and P. Rebeschini, Time-independent generalization bounds for SGLD in non-convex settings, *Adv. Neural Inf. Process. Syst.* **34** (2021) 19836–19846.
15. Y. Feng, L. Li and J.-G. Liu, Semi-groups of stochastic gradient descent and online principal component analysis: Properties and diffusion approximations, *Comm. Math. Sci.* **16**(3) (2018) 777–789.
16. S. Jin, L. Li and J. -G. Liu, Random batch methods for interacting particle systems, *J. Comput. Phys.* **400** (2020) 108877.
17. S. Jin, L. Li, X. Ye and Z. Zhou, Ergodicity and long-time behavior of the random batch method for interacting particle systems, preprint (2022), arXiv:2202.04952.
18. P. D. Lax, *Functional Analysis*, Vol. 55 (John Wiley & Sons, 2002).
19. L. Li and Y. Wang, A sharp uniform-in-time error estimate for Stochastic Gradient Langevin Dynamics, preprint (2022), arXiv:2207.09304.
20. T. Lindvall and L. C. G. Rogers, Coupling of multidimensional diffusions by reflection, *Ann. Probab.* **14**(3) (1986) 860–872.
21. M. B. Majka, A. Mijatović and Ł. Szpruch, Nonasymptotic bounds for sampling algorithms without log-concavity, *Ann. Appl. Probab.* **30**(4) (2020) 1534–1581.
22. X. Mao, *Stochastic Differential Equations and Applications* (Elsevier, 2007).
23. W. Mou, L.Wang, X. Zhai and K. Zheng, Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints, in *Conf. Learning Theory*, PMLR, 2018, pp. 605–638.
24. F. Otto and C. Villani, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality, *J. Funct. Anal.* **173**(2) (2000) 361–400.

25. I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *Ann. Probab.* **22**(4) (1994) 1679–1706.
26. H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Stat.* **22**(3) (1951) 400–407.
27. W. Rudin *et al.*, *Principles of Mathematical Analysis*, Vol. 3 (McGraw-Hill, New York, 1976).
28. M. Talagrand, A new isoperimetric inequality and the concentration of measure phenomenon, in *Geometric Aspects of Functional Analysis* (Springer, 1991), pp. 94–124.
29. R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47 (Cambridge University Press, 2018).
30. C. Villani *et al.*, *Optimal Transport: Old and New*, Vol. 338 (Springer, 2009).
31. M. Welling and Y. W. Teh, Bayesian learning via stochastic gradient Langevin dynamics, in *Proc. 28th Int. Conf. Machine Learning* (Citeseer, 2011), pp. 671–688.
32. Y. Zhang, Ö. Deniz Akyildiz, T. Damoulas and S. Sabanis, Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization, *Appl. Math. Opt.* **87**(2) (2023) 25.
33. D. Zou, P. Xu and Q. Gu, Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling, in *Uncertainty in Artificial Intelligence* (PMLR, 2021), pp. 1152–1162.