Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization

Ying Zhang[1*], Ömer Deniz Akyildiz[2], Theodoros Damoulas[3,4] and Sotirios Sabanis[3,5]

[1*]Division of Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, 637371, Singapore.
[2]Department of Mathematics, Imperial College London, Exhibition Road, London, SW7 2AZ, United Kingdom.
[3]The Alan Turing Institute, 96 Euston Road, London, NW1 2DB, United Kingdom.
[4]Department of Computer Science and Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom.
[5]School of Mathematics, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom.

*Corresponding author(s). E-mail(s): ying.zhang@ntu.edu.sg;
Contributing authors: deniz.akyildiz@imperial.ac.uk;
t.damoulas@warwick.ac.uk; s.sabanis@ed.ac.uk;

Abstract

In this paper, we are concerned with a non-asymptotic analysis of sampling algorithms used in nonconvex optimization. In particular, we obtain non-asymptotic estimates in Wasserstein-1 and Wasserstein-2 distances for a popular class of algorithms called Stochastic Gradient Langevin Dynamics (SGLD). In addition, the aforementioned Wasserstein-2 convergence result can be applied to establish a non-asymptotic error bound for the expected excess risk. Crucially, these results are obtained under a local Lipschitz condition and a local dissipativity condition where we remove the uniform dependence in the data stream. We illustrate the importance of this relaxation by presenting examples from variational inference and from index tracking optimization.

Keywords: Stochastic gradient Langevin dynamics, Non-convex optimization, Non-asymptotic estimates, Local Lipschitz continuous, Local dissipativity, Variational inference

MSC Classification: 60J20 , 60J22 , 65C05 , 65C40 , 62D05

1 Introduction

We consider a nonconvex stochastic optimization problem

$$\text{minimize} \quad U(\theta) := \mathbb{E}[f(\theta, X)],$$

where $\theta \in \mathbb{R}^d$ and X is a random element. We aim to generate an estimate $\hat{\theta}$ such that the expected excess risk $\mathbb{E}[U(\hat{\theta})] - \inf_{\theta \in \mathbb{R}^d} U(\theta)$ is minimized. The optimization problem of minimizing U is closely linked to the problem of sampling from a target distribution which concentrates around the minimizers of U. It is, therefore, important to investigate the Langevin dynamics based algorithms and their sampling behaviour in the context of optimization. The latter is the primary focus of this article.

The Langevin SDE is given by

$$\mathrm{d}Z_t = -h(Z_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \qquad t > 0, \tag{1}$$

with a (possibly random) initial condition θ_0, where $h := \nabla U$, $\beta > 0$, and $(B_t)_{t \geq 0}$ is a d-dimensional Brownian motion. Under mild conditions, it is well-known that SDE (1) admits as a unique invariant measure $\pi_\beta(\theta) \propto \exp(-\beta U(\theta))$. Moreover, π_β concentrates around the minimizers of U when β takes sufficiently large values (see, e.g., [1]). To sample from π_β, a standard approach is to approximate the Langevin SDE (1) by using an Euler discretization scheme, which serves as a sampling algorithm and is known as the unadjusted Langevin algorithm (ULA) or Langevin Monte Carlo (LMC). Theoretical guarantees for the convergence of ULA in Wasserstein distance and in total variation have been obtained under the assumption that U is strongly convex with globally Lipschitz gradient [2–4]. Extensions which include locally Lipschitz gradient and higher order algorithms can be found in [5], [6] and [7].

In practice, however, the gradient h is usually unknown and one only has an unbiased estimate of h. A natural extension of ULA, which was introduced in [8] in the context of Bayesian inference and which has found great applicability in this type of stochastic optimization problems, is the Stochastic Gradient Langevin Dynamics (SGLD) algorithm. More precisely, fix an $\mathbb{R}^d$-valued random variable θ_0 representing its initial value and let $(X_n)_{n \in \mathbb{N}}$ be an i.i.d. sequence, the SGLD algorithm corresponding to SDE (1) is given by, for any $n \in \mathbb{N}$,

$$\theta_0^\lambda := \theta_0, \quad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \tag{2}$$

where $\lambda > 0$ is often called the stepsize or gain of the algorithm, $\beta > 0$ is the so-called inverse temperature parameter, $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a measurable function and $(\xi_n)_{n\in\mathbb{N}}$ is an independent sequence of standard d-dimensional Gaussian random variables. The properties of the i.i.d. process $(X_n)_{n\in\mathbb{N}}$ are given below.

For a strongly convex objective function U, [9], [10], [11], and [6] obtain non-asymptotic bounds in Wasserstein-2 distance between the SGLD algorithm and the target distribution π_β. While [6] assumes the stochastic gradient H is a linear combination of h and $(X_n)_{n\in\mathbb{N}}$, which allows bounded conditional bias, a general form of H with non-Markovian $(X_n)_{n\in\mathbb{N}}$ is considered in [9]. For the case where U is nonconvex, one line of research is to consider a dissipativity condition. The first such non-asymptotic estimate is provided by [12] in Wasserstein-2 distance although its rate of convergence is $\lambda^{5/4}n$ which depends on the number of iterations n. Improved results are obtained in [13], by using a direct analysis of the ergodicity of the overdamped Langevin Monte Carlo algorithms. While a faster convergence rate is achieved in [13] compared to [12], it is still dependent on n. Recently, [14] obtained a convergence rate 1/2 in Wasserstein-1 distance. Its analysis relies on the construction of certain auxiliary continuous processes and the contraction results in [15]. Another line of research is to assume a convexity at infinity condition of U. [16] and [17] obtain convergence results in Wasserstein-1 distance by using the contraction property developed in [18]. In both convex and nonconvex settings, the non-asymptotic analysis of the Langevin diffusion can be extended to a wider class of diffusions under certain conditions, see [19] and references therein.

In this paper, we establish non-asymptotic convergence results in Theorem 2.4 and Corollary 2.5 for the SGLD algorithm (2) in Wasserstein-1 and Wasserstein-2 distances, respectively. Moreover, by using a similar splitting approach as in [12], the Wasserstein-2 convergence result can then be applied to establish a nonasymptotic error bound for the expected excess risk, which is provided in Corollary 2.8. These main results are obtained under the relaxed conditions as stated in Assumptions 2 and 3 below. Crucially, we relax substantially the assumptions of dissipativity and Lipschitz continuity on the stochastic gradient $H(\theta, x)$ by allowing non-uniform dependence in x.

To illustrate the applicability of the proposed algorithm under the local assumptions, examples from variational inference (VI) and from index tracking optimization are considered, which represent key paradigms in statistical machine learning and financial mathematics. In the VI example, a nonconvex objective function is considered, and it can be shown that its stochastic gradient, denoted by $H(\theta, u)$, satisfies the local dissipativity and local Lipschitz conditions. To the best of the authors' knowledge, this is the first time that non-asymptotic guarantees are provided for a concrete variational inference example due to the local nature of the aforementioned dissipativity and Lipschitz conditions which stem from the lack of a uniform bound in u. As for the example from index tracking optimization, the mean squared tracking error is considered as the objective function (see, e.g. [20], [21]). Reparametrization is performed to remove the constraints on the parameter θ, which results in a nonconvex objective function. In addition, as this example can be viewed as an online regression problem, a uniform bound of the data stream is unavailable. However, one can check that the stochastic gradient, denoted by $H(\theta, z)$, satisfies the local dissipativity and local Lipschitz conditions but not the corresponding global ones.

We conclude this section by introducing some notation. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable X. For $1 \leq p < \infty$, L^p is used to denote the usual space of p-integrable real-valued random variables. The L^p-integrability of a random variable X is defined as $\mathbb{E}[|X|^p] < \infty$. Fix an integer $d \geq 1$. For an $\mathbb{R}^d$-valued random variable X, its law on $\mathcal{B}(\mathbb{R}^d)$ (the Borel sigma-algebra of $\mathbb{R}^d$) is denoted by $\mathcal{L}(X)$. For a positive real number a, we denote by $\lfloor a \rfloor$ its integer part. For a vector $b \in \mathbb{R}^d$, denote by b^T its transpose. Scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $| \cdot |$ standing for the corresponding norm. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable function. Denote by ∇f, $\nabla^2 f$ and Δf the gradient of f, the Hessian of f and the Laplacian of f, respectively. For any integer $q \geq 1$, let $\mathcal{P}(\mathbb{R}^q)$ denote the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ, ν. For two probability measures μ and ν, the Wasserstein distance of order $p \geq 1$ is defined as, for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $W_p(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p}$.

2 Main results and comparisons

Let $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ be a measurable function. It satisfies $\mathbb{E}[|f(\theta, X)|] < \infty$ for all $\theta \in \mathbb{R}^d$, where X is a random variable with probablity law $\mathcal{L}(X)$. Let $U : \mathbb{R}^d \to \mathbb{R}$ defined by $U(\theta) := \mathbb{E}[f(\theta, X)]$ be a continuously differentiable function with gradient denoted by $h := \nabla U$. Moreover, define

$$\pi_\beta(A) := \frac{\int_A e^{-\beta U(\theta)} \, \mathrm{d}\theta}{\int_{\mathbb{R}^d} e^{-\beta U(\theta)} \, \mathrm{d}\theta}, \quad A \in \mathcal{B}(\mathbb{R}^d), \tag{3}$$

with $\int_{\mathbb{R}^d} e^{-\beta U(\theta)} \, \mathrm{d}\theta < \infty$.

Denote by $(\mathcal{G}_n)_{n \in \mathbb{N}}$ a given filtration representing the flow of past information, and denote by $\mathcal{G}_\infty := \sigma(\bigcup_{n \in \mathbb{N}} \mathcal{G}_n)$. Fix $m \geq 1$. Let $(X_n)_{n \in \mathbb{N}}$ be an $\mathbb{R}^m$-valued, $(\mathcal{G}_n)$-adapted process with $X_n \sim \mathcal{L}(X)$ for all $n \in \mathbb{N}$. It is assumed throughout the paper that θ_0, $\mathcal{G}_\infty$ and $(\xi_n)_{n \in \mathbb{N}}$ are independent. Next, we introduce our main assumptions.

Fix $\beta > 0$. For each $\lambda > 0$, the SGLD algorithm is given by, for any $n \in \mathbb{N}$,

$$\theta_0^\lambda := \theta_0, \quad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}} \xi_{n+1}, \tag{4}$$

where $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a measurable function and $(\xi_n)_{n \in \mathbb{N}}$ is an independent sequence of standard d-dimensional Gaussian random variables.

Then, we present our assumptions. The first assumption describes the requirement on the moment of the initial parameter θ_0. Moreover, it is stated that stochastic gradient $H(\theta, \cdot)$ is assumed to be unbiased.

Assumption 1. $|\theta_0| \in L^4$. *The process* $(X_n)_{n\in\mathbb{N}}$ *is i.i.d.. Moreover, it holds that* $\mathbb{E}[H(\theta, X_0)] = h(\theta)$.

Our second assumption describes the requirement on the moment of the initial data X_0 and on the regularity of the stochastic gradient with respect to its first and second arguments. As a result, growth estimates are derived.

Assumption 2. *There exist* $\eta : \mathbb{R}^m \to [1, \infty)$ *with* $(1 + |X_0|)\eta(X_0) \in L^4$ *and positive constants* L_1, L_2 *such that, for all* $x, x' \in \mathbb{R}^m$ *and* $\theta, \theta' \in \mathbb{R}^d$,

$$\begin{aligned}
&|H(\theta, x) - H(\theta', x)| \leq L_1\eta(x)|\theta - \theta'|,\\
&|H(\theta, x) - H(\theta, x')| \leq L_2(\eta(x) + \eta(x'))(1 + |\theta|)|x - x'|.
\end{aligned}$$

Remark 2.1. *Assumption 2 implies, for all* $\theta, \theta' \in \mathbb{R}^d$,

$$|h(\theta) - h(\theta')| \leq L_1\mathbb{E}[\eta(X_0)]|\theta - \theta'|\,. \tag{5}$$

Also, Assumption 2 implies

$$|H(\theta, x)| \leq L_1\eta(x)|\theta| + L_2\bar{\eta}(x) + H_\star, \tag{6}$$

where $\bar{\eta}(x) = (\eta(x) + \eta(0))|x|$ *and* $H_\star := |H(0, 0)|$. *Moreover, under Assumptions 1 and 2, the gradient* $h(\theta) = \mathbb{E}[H(\theta, X_0)]$ *for all* $\theta \in \mathbb{R}^d$, *is well-defined.*

The proof of the statements in Remark 2.1 is postponed to Appendix C.

Our next assumption is a *dissipativity* condition for stochastic gradients. We note that this and the previous assumption significantly relax the analogous requirements found in the literature, e.g. see [12], [14] and references therein.

Assumption 3. *There exist a measurable (symmetric matrix-valued) function* $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$ *and a measurable function* $\hat{b} : \mathbb{R}^m \to \mathbb{R}$ *such that for any* $x \in \mathbb{R}^m$, $y \in \mathbb{R}^d$, $\langle y, A(x)y\rangle \geq 0$ *and for all* $\theta \in \mathbb{R}^d$ *and* $x \in \mathbb{R}^m$,

$$\langle H(\theta, x), \theta\rangle \geq \langle \theta, A(x)\theta\rangle - \hat{b}(x).$$

The smallest eigenvalue of $\mathbb{E}[A(X_0)]$ *is a positive real number* $a > 0$ *and* $E[\hat{b}(X_0)] := b > 0$.

Remark 2.2. *By Assumptions 1 and 3, one obtains a dissipativity condition of* h, *i.e., for any* $\theta \in \mathbb{R}^d$, $\langle h(\theta), \theta\rangle \geq a|\theta|^2 - b$.

Remark 2.3. *We emphasize that we call the process $(X_n)_{n\geq 0}$ as 'data', following the convention [9, 14]. This can represent 'data' in the classical meaning but also can represent, e.g., samples from variational approximations in the Bayesian inference setting (see Section 3.1). In the latter case, its statistical properties are straightforward to assess (since the variational approximation is a design choice) and our assumptions are easier to verify as shown in Sec. 3.1.*

We next state our main result, which fully characterises the convergence in Wasserstein-1 distance of the law of the SGLD at its n-th iteration, which is denoted by $\mathcal{L}(\theta_n^\lambda)$, to the target measure π_β. Define first

$$\lambda_{\max} := \min\left\{ \frac{\min\{a, a^{1/3}\}}{16(1+L_1)^2 \left(\mathbb{E}\left[(1+\eta(X_0))^4\right]\right)^{1/2}}, \frac{1}{a} \right\}, \tag{7}$$

where L_1 and a are defined in Assumptions 2 and 3, respectively.

Theorem 2.4. *Let Assumptions 1, 2 and 3 hold. Then, there exist constants $\dot{c}, C_1, C_2, C_3 > 0$ such that, for every $\beta > 0$, $0 < \lambda \leq \lambda_{\max}$, and $n \in \mathbb{N}$,*

$$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C_1 \mathrm{e}^{-\dot{c}\lambda n/2}(\mathbb{E}[|\theta_0|^4] + 1) + (C_2 + C_3)\sqrt{\lambda},$$

where $\dot{c}$ is given in (23), C_1, C_2, C_3 are given explicitly in (28). Moreover, for any $\varepsilon > 0$, if we choose $\lambda \leq \frac{\varepsilon^2}{4(C_2+C_3)^2} \wedge \lambda_{\max}$, and $n \geq \frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon^2 \dot{c}} \left(1 + \frac{1}{(1-e^{-\dot{c}})^2}\right) \ln\left(\frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon}\left(1 + \frac{1}{1-e^{-\dot{c}}}\right)\right)$ with $C_\star > 0$ independent of d, β, n, then $W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq \varepsilon$.

By further observing a trivial functional inequality, one can state an analogous result in Wasserstein-2 distance, which matches the rate obtained in [17] but which is known to be suboptimal.

Corollary 2.5. *Let Assumptions 1, 2 and 3 hold. Then, there exist constants $\dot{c}, C_4, C_5, C_6 > 0$ such that, for every $\beta > 0$, $0 < \lambda \leq \lambda_{\max}$, and $n \in \mathbb{N}$,*

$$W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C_4 \mathrm{e}^{-\dot{c}\lambda n/4}(\mathbb{E}[|\theta_0|^4] + 1) + (C_5 + C_6)\lambda^{1/4},$$

where $\dot{c}$ is given in (23), C_4, C_5, C_6 are given explicitly in (29). Moreover, for any $\varepsilon > 0$, if we choose $\lambda \leq \frac{\varepsilon^4}{16(C_5+C_6)^4} \wedge \lambda_{\max}$, and $n \geq \frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon^4 \dot{c}} \left(1 + \frac{1}{(1-e^{-\dot{c}/2})^4}\right) \ln\left(\frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right)$ with $C_\star > 0$ independent of d, β, n, then $W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq \varepsilon$.

Remark 2.6. *By (28) and (29), one notes that the constants C_2 and C_5 are of order $\sqrt{d/\beta}$, and this implies that large values of d on the upper bound can be controlled by large values of β. However, the constants C_1, C_3, C_4, C_6 have* *exponential dependence on d and β, rather than d/β, due to the contraction result in [15, Theorem 2.2]. Furthermore, one notes that the undesirable dependence on d, which is derived under a geometric drift condition (and which, in turn, is implied by a dissipativity condition such as Assumption 3), can be found, typically, in extreme cases, i.e. pathological examples of theoretical nature. This appears not to be the case in many practical applications.*

Remark 2.7. *In the case where $H(\theta, x) = h(\theta)$ for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$, i.e. when the stochastic gradient coincides with the full gradient, Theorem 2.4 and Corollary 2.5 provide the full non-asymptotic convergence results of the unadjusted Langevin algorithm (ULA) under dissipativity and Lipschitz continuity assumptions.*

Let $\hat{\theta} := \theta_n^\lambda$, where θ_n^λ denotes the n-th iteration of the SGLD algorithm (4). Then, an upper bound for the expected excess risk $\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta\in\mathbb{R}^d} U(\theta)$ can be obtained by using the following splitting: $\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta\in\mathbb{R}^d} U(\theta) = \big(\mathbb{E}[U(\theta_n^\lambda)] - \mathbb{E}[U(Z_\infty)]\big) + (\mathbb{E}[U(Z_\infty)] - \inf_{\theta\in\mathbb{R}^d} U(\theta))$, where $Z_\infty \sim \pi_\beta$ with π_β defined in (3). By using Corollary 2.5, an upper bound for the first term on the RHS of the above equality can be obtained as explained in [12, Lemma 3.5]. The second term on the RHS of the above equality can be upper bounded by applying [12, Proposition 3.4]. The precise statement with explicit constants is provided below.

Corollary 2.8. *Let Assumptions 1, 2 and 3 hold. Then, there exist constants $\dot{c}, C_1^\sharp, C_2^\sharp, C_3^\sharp > 0$ such that, for every $\beta > 0$, $0 < \lambda \leq \lambda_{\max}$, $n \in \mathbb{N}$,*

$$\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta\in\mathbb{R}^d} U(\theta) \leq C_1^\sharp e^{-\dot{c}\lambda n/4} + C_2^\sharp \lambda^{1/4} + C_3^\sharp,$$

where $\dot{c}$ is given in (23), $C_1^\sharp, C_2^\sharp, C_3^\sharp$ are given explicitly in (31) and (32). Moreover, for any $\varepsilon > 0$, if we choose $\beta \geq \beta_\varepsilon \vee \frac{3d}{\varepsilon} \log\left(\frac{eL_1\mathbb{E}[\eta(X_0)]}{ad}(b+1)(d+1)\right)$ with β_ε denoting the root of the function $f^\sharp(\beta) = \frac{\log(\beta+1)}{\beta} - \frac{\varepsilon}{3d}$, $\lambda \leq \frac{\varepsilon^4}{81(C_2^\sharp)^4} \wedge \lambda_{\max}$, and $n \geq \frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon^4 \dot{c}}\left(1 + \frac{1}{(1-e^{-\dot{c}/2})^4}\right) \ln\left(\frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon}\left(1+\frac{1}{1-e^{-\dot{c}/2}}\right)\right)$ with $C_\star > 0$ independent of d, β, n, then $\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta\in\mathbb{R}^d} U(\theta) \leq \varepsilon$.

Remark 2.9. *By (32), one observes that $C_3^\sharp$ vanishes as β tends to infinity. This implies that π_β converges to a distribution which concentrates on the minimizers of U for large enough β. This result provides a nonasymptotic bound for this concentration – thus, sampling from π_β solves the optimization problem of minimizing U. However, for a large β, it would require a large number of iterations for the SGLD algorithm to reach a given precision level measured using expected excess risk, see the expression for the lower bound of n in Corollary 2.8. This is due to a slow convergence of the*

Langevin dynamics (1) *to the target distribution* π_β *as the contraction constant* $\dot{c}$ *is inversely related to* β*, see* (23)*, which, in turn, lead to a slow convergence of SGLD to* π_β*. There is thus a trade-off between the precision of the approximation and the efficiency of the algorithm. Consequently, one may set* $\beta = \beta_\varepsilon \vee \frac{3d}{\varepsilon} \log\left(\frac{eL_1\mathbb{E}[\eta(X_0)]}{ad}(b+1)(d+1)\right)$ *so as to achieve a given precision level for the expected excess risk while ensuring that the SGLD algorithm takes the smallest possible number of iterations to sample approximately from* π_β *(also within the given precision level). Furthermore, for any precision level, once the value for* β *is specified, one may calculate the upper bound of* λ *using the expression given in Corollary 2.8 and then set the upper bound as the value of* λ *for the efficiency of the algorithm as* λ *is negatively related to* n*.*

The proofs of Theorem 2.4, Corollary 2.5, 2.8 are postponed to Section 4. Furthermore, the explicit expressions for the constants in the main results are summarised in Table D1 and D2.

2.1 Related work and discussions

Table 1 Comparison of Theorem 2.4 and Corollary 2.5 with [12, Proposition 3.3], [17, Theorem 1.4] and [14, Theorem 2.5].

	Smoothness	Contractivity	$\mathrm{Var}(H(\theta, X_0))$	Data	Results
[12]	Globally Lipschitz H in θ uniformly in x	Uniform in x dissipativity of H	Bounded by $C\|\theta\|^2$	i.i.d.	$W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C\lambda^{5/4}n$
[17]	Globally Lipschitz h	Convexity at infinity of h	Bounded by $C\|\theta\|^2\lambda^\alpha$, $\alpha > 0$	i.i.d.	$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C\lambda^{\alpha/2}$, $W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C\lambda^{\alpha/4}$
[14]	Globally Lipschitz H in θ and x	Uniform in x dissipativity of H	—	L-mixing	$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C\lambda^{1/2}$
This paper	Locally Lipschitz H in θ and x	Local dissipativity of H	—	i.i.d.	$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C\lambda^{1/2}$, $W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C\lambda^{1/4}$

Under the preceding Assumptions 1, 2 and 3, a convergence result in W_1 distance with rate $1/2$ is given in Theorem 2.4, while in Corollary 2.5, a convergence result in W_2 distance with rate $1/4$ is provided. [9, Theorem 3.10] provides a convergence result in W_2 distance under similar assumptions in the convex setting, i.e. with Assumption 3 replaced by a strong convexity requirement. Moreover, the analysis of Theorem 2.4 follows a similar approach as in [14], while its framework is crucially extended by assuming local Lipschitz continuity of H in Assumption 2, and non-uniform estimates with respect to the x variable in Assumption 3.

Next, we mainly focus on the comparison of our work with [12] and [17]. In [12, Proposition 3.3], a finite-time convergence result of the SGLD algorithm (4) in Wasserstein-2 distance is provided, and the rate of convergence is shown to be $\lambda^{5/4}n$ with n the number of iterations. To obtain this result, a dissipativity condition [12, Assumption (A.3)] is proposed. In [12, Assumption (A.1)], the quantities $f(0, \cdot)$ and $H(0, \cdot)$ are assumed to be bounded, where $U(\theta) = \mathbb{E}[f(\theta, X_0)]$ and $H(\cdot, \cdot) = \nabla_\theta f(\cdot, \cdot)$, $\theta \in \mathbb{R}^d$. In addition, it requires the finiteness of an exponential moment of the initial value [12, Assumption (A.5)] and the Lipschitz continuity of H in θ [12, Assumption (A.2)]. While Corollary 2.5 improves the convergence rate provided in [12] in the sense that our rate of convergence does not depend on the number of iterations, we further require a local Lipschitz continuity of $H(\theta, x)$ in x. However, compared to [12, Assumption (A.3)], we allow the dissipativity condition without imposing the uniformity in x in Assumption 3, and we require only polynomial moments of the initial value θ_0. Furthermore, in Assumption 2, we relax the (global) Lipschitz condition of H in θ by allowing the Lipschitz constant to depend on x. We note that [12, Assumption (A.4)] can be obtained by using Assumptions 1 and 2.

Further, we compare our results with those in [17]. Compared to [17, Theorem 1.4] with $\alpha = 1$, Theorem 2.4 achieves the same rate in W_1 without assuming that the variance of the stochastic gradient is controlled by the stepsize [17, Assumption 1.3]. To obtain Theorem 2.4, we assume a Lipschitz continuity of H in Assumption 2, while [17, Theorem 1.4] requires a Lipschitz continuity of h [17, Assumption 1.1]. This latter condition on h is implied by our Assumption 2 as indicated in Remark 2.1. It is typical though, for many real applications, that the full gradient h is unknown, and crucially one can check conditions only for H (as in Assumption 2) and not for h. This is also apparent in Section 3 where Assumptions 2 and 3 are easily checkable for various examples. The same cannot be said for the corresponding conditions regarding h. In a recent update of [17], it is noted in [17, Section 5.3] that the requirements for [17, Theorem 1.4] can be potentially relaxed by replacing the convexity at infinity condition with a uniform dissipativity condition. This observation coincides with the results obtained in [14, Theorem 2.5], when one considers i.i.d. data, while we further generalise this framework by requiring only a local dissipativity condition, i.e. Assumption 3.

3 Applications

In this section, we use $x^\mathsf{T} y$ for any $x, y \in \mathbb{R}^d$ for the inner product (instead of $\langle x, y \rangle$) to make the notation compact.

3.1 Variational inference for Bayesian logistic regression

Variational inference (VI) aims at approximating a posterior distribution $p(w|x)$, where w is the quantity of interest and x is the data, using a parameterized family of distributions which is often called the *variational family* and is denoted by $q_\theta(w)$ [22]. Thus, the VI approach converts an inference problem into an optimization problem, where the objective function is typically nonconvex. This enables us to apply our results to the VI problem and come up with theoretical guarantees. We present one such example below and provide guarantees for the application of the VI approach by using the conclusions of our main theorem.

Consider a probabilistic model consisting of a likelihood $p(w|x)$ and a prior $p(w)$. Here we implicitly assume the existence of probability density functions which are used to identify the corresponding probability distributions. To ease the notation, for a distribution p, its marginal, conditional and joint distributions are also denoted by p, however dependencies on appropriate state variables are explicitly declared. Furthermore, recall that given a joint distribution $p(w, x)$, one observes that for any distribution $q(w)$, $x \in \mathbb{R}^{\bar{m}}$, $\bar{m} \geq 1$, the following holds

$$\begin{aligned}\log p(x) &= \int_{\mathbb{R}^d} q(w) \log\left(\frac{p(w,x)}{q(w)}\right) dw + \int_{\mathbb{R}^d} q(w) \log\left(\frac{q(w)}{p(w|x)}\right) dw \\ &= \mathbb{E}_{\mathbf{w}\sim q} \log \frac{p(\mathbf{w}, x)}{q(\mathbf{w})} + \mathrm{KL}(q(w) \| p(w|x)),\end{aligned} \tag{8}$$

where the first term in (8) is usually denoted in VI literature by ELBO(q). The aim is to choose a suitable approximating family q_θ parameterized by θ, so as to minimize the KL divergence of the two distributions $q_\theta(w)$ and $p(w|x)$, for a given x, over θ. This turns out to be equivalent to maximizing ELBO(q_θ) since $\log p(x)$ is fixed. One can decompose ELBO$(q_\theta) = l_1(\theta) + l_2(\theta)$ where $l_1(\theta) = \mathbb{E}_{\mathbf{w}\sim q_\theta}[\log p(\mathbf{w}, x)]$ and $l_2(\theta)$ is the entropy of q_θ. Moreover, we suppose there exists a transformation $\mathcal{T}_\theta$ such that $\mathcal{T}_\theta(\mathbf{u}) \overset{\mathrm{d}}{=} \mathbf{w}$. As a result, one obtains $l_1(\theta) := \mathbb{E}_{\mathbf{u}\sim s}[\log p(\mathcal{T}_\theta(\mathbf{u}), x)]$, and similarly, $l_2(\theta) := -\mathbb{E}_{\mathbf{u}\sim s}[\log q(\mathcal{T}_\theta(\mathbf{u}))]$. This is called *reparameterization trick* in VI literature [23–26]. By using this technique, one can obtain stochastic estimates of $\nabla_\theta(l_1(\theta) + l_2(\theta))$ and then use SGLD algorithm to maximize ELBO(q_θ).

We consider an example from Bayesian logistic regression [22]. Suppose a collection of data points $\mathcal{X} = \{(z_i, y_i)\}_{i=1,\dots,n}$ is given, where $z_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ for all i. Denote by $\mathcal{Z}_i = (z_i, y_i)$ for all i, then $\mathcal{X} = \{\mathcal{Z}_i\}_{i=1,\dots,n}$. Assume Gaussian mixture prior to define a multimodal distribution characterized by $p(w, \mathcal{X}) = \pi_0(w) \prod_{i=1}^n p(\mathcal{Z}_i|w)$, where $\pi_0(w)$ is the prior given by $\pi_0(w) \propto \exp(-\bar{f}(w)) = e^{-|w-\hat{a}|^2/2} + e^{-|w+\hat{a}|^2/2}$ with $\bar{f}(w) = |w - \hat{a}|^2/2 - \log(1 + \exp(-2\hat{a}^\mathsf{T} w))$, $\hat{a} \in \mathbb{R}^d$, $|\hat{a}|^2 > 1$ and $p(\mathcal{Z}_i|w) = (1/(1 + e^{-z_i^\mathsf{T} w}))^{y_i}(1 - 1/(1 + e^{-z_i^\mathsf{T} w}))^{1-y_i}$ is the likelihood function. Moreover, take a variational distribution parameterized by θ, which is given as $q_\theta(w) \propto e^{-|w-\theta|^2/2} + e^{-|w+\theta|^2/2}$. Then, maximizing $l_1(\theta) + l_2(\theta) = \mathbb{E}_{\mathbf{w}\sim q_\theta}[\log p(\mathbf{w}, \mathcal{X}) - \log q_\theta(\mathbf{w})]$ in θ is equivalent to maximizing the following:

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}\sim q_\theta}\Big[-|\mathbf{w} - \hat{a}|^2/2 + \log(1 + \exp(-2\hat{a}^\mathsf{T}\mathbf{w})) \\ &\qquad + \sum_{i=1}^n (-y_i \log(1 + e^{-z_i^\mathsf{T}\mathbf{w}}) + (y_i - 1)\log(1 + e^{z_i^\mathsf{T}\mathbf{w}}))\Big] \\ &+ \mathbb{E}_{\mathbf{w}\sim q_\theta}\left[|\mathbf{w} - \theta|^2/2 - \log(1 + \exp(-2\theta^\mathsf{T}\mathbf{w}))\right].\end{aligned} \tag{9}$$

Further, the reparameterization technique is applied by considering the mapping $\mathcal{T}_\theta(u) := \mathbb{1}_{\{v=1\}}(Cu+m)+\mathbb{1}_{\{v=0\}}(Cu-m)$ where $v \sim \mathrm{Ber}(q)$ and $\theta = (C, m, q)$. Here, we fix $C = \mathbf{I}_d/4$, $q = 7/8$, and thus $\mathcal{T}_\theta(u) = \mathbb{1}_{\{v=1\}}(u/4+\theta)+\mathbb{1}_{\{v=0\}}(u/4-\theta)$. Then, for $\mathbf{u} \sim s$ where s is the standard Gaussian distribution, the expression in (9) becomes

$$
\begin{aligned}
l_1(\theta) + l_2(\theta) = {} & \frac{7}{8}\mathbb{E}_{\mathbf{u}\sim s}\Big[-|\mathbf{u}/4+\theta-\hat{a}|^2/2 + \log(1+\exp(-2\hat{a}^\mathsf{T}(\mathbf{u}/4+\theta))) \\
& + \sum_{i=1}^{n}(-y_i\log(1+e^{-z_i^\mathsf{T}(\mathbf{u}/4+\theta)}) + (y_i-1)\log(1+e^{z_i^\mathsf{T}(\mathbf{u}/4+\theta)}))\Big] \\
& + \frac{1}{8}\mathbb{E}_{\mathbf{u}\sim s}\Big[-|\mathbf{u}/4-\theta-\hat{a}|^2/2 + \log(1+\exp(-2\hat{a}^\mathsf{T}(\mathbf{u}/4-\theta))) \\
& + \sum_{i=1}^{n}(-y_i\log(1+e^{-z_i^\mathsf{T}(\mathbf{u}/4-\theta)}) + (y_i-1)\log(1+e^{z_i^\mathsf{T}(\mathbf{u}/4-\theta)}))\Big] \\
& + \frac{7}{8}\mathbb{E}_{\mathbf{u}\sim s}\left[|\mathbf{u}/4+\theta-\theta|^2/2 - \log(1+\exp(-2(\theta^\mathsf{T}\mathbf{u}/4+|\theta|^2)))\right] \\
& + \frac{1}{8}\mathbb{E}_{\mathbf{u}\sim s}\left[|\mathbf{u}/4-\theta-\theta|^2/2 - \log(1+\exp(-2(\theta^\mathsf{T}\mathbf{u}/4-|\theta|^2)))\right].
\end{aligned}
\tag{10}
$$

In what follows, we derive the stochastic gradient expression for the cost function defined in Eq. (10). We note that, stochasticity in this example comes from sampling $\mathbf{u}$ variables and constructing empirical expectation estimates, rather than subsampling data points.

Proposition 3.1. *Let the objective function of the VI example be defined in* (10)*. Moreover, let* $\mathbf{u}$ *be a standard d-dimensional Gaussian random variable, and let* $(\mathbf{u}_n)_{n\in\mathbb{N}}$ *be a sequence of i.i.d. standard d-dimensional Gaussian random variables. In addition, assume* $|\theta_0| \in L^4$*. Let* $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ *be the stochastic gradient of* (10) *given by*

$$
\begin{aligned}
H(\theta, u) = {} & \frac{\theta}{2} + \frac{u}{4} - \frac{3}{4}\hat{a} + \frac{\hat{a}}{4}\left(\frac{7}{1+e^{2\hat{a}^\mathsf{T}(u/4+\theta)}} - \frac{1}{1+e^{2\hat{a}^\mathsf{T}(u/4-\theta)}}\right) \\
& + \frac{1}{8}\sum_{i=1}^{n}\left(-6z_iy_i + \frac{7z_i}{1+e^{-z_i^\mathsf{T}(u/4+\theta)}} - \frac{z_i}{1+e^{-z_i^\mathsf{T}(u/4-\theta)}}\right) \\
& - \frac{7(u+8\theta)}{16(1+e^{2(\theta^\mathsf{T}u/4+|\theta|^2)})} - \frac{u-8\theta}{16(1+e^{2(\theta^\mathsf{T}u/4-|\theta|^2)})}.
\end{aligned}
\tag{11}
$$

Then, H *satisfies Assumptions 1, 2, and 3. More precisely, Assumption 2 holds with* $L_1 = 1$*,* $L_2 = 1/4$*,* $\eta(u) = 9/2 + 8e^{|u|^2/32} + \sum_{i=1}^{n}|z_i|^2 + 4|\hat{a}|^2 + 3|u|^2/8$*, and moreover,* $\mathbb{E}[(1+|\mathbf{u}|)^4\eta^4(\mathbf{u})] < \infty$*. Assumption 3 holds with* $A(u) = \mathbf{I}_d/4$ *and* $\hat{b}(u) = (9|u|^2/4 + 121|\hat{a}|^2/4) + 49n\sum_{i=1}^{n}|z_i|^2/8 + 7n/4$*.*

The proof of Proposition 3.1 is postponed to Appendix B. Moreover, the meaning of this result is that we can verify, in a practical variational inference example, that our local assumptions are satisfied, therefore in this instance of nonconvex optimization problem, the theoretical guarantees we provide in this paper regarding SGLD provably hold. This is a significant improvement over previous results which are based on global Lipschitz assumptions, which fail to hold for this simple, yet very illustrative, example.

3.2 Index tracking

We consider the problem of index tracking, which can be formulated precisely as follows (see, e.g., [21, Eqn. (3), (4), and (8)]):

$$\min_\theta U(\theta) := \min_\theta \left(\mathbb{E}\left[\left(Y - \sum_{i=1}^N g_i(\theta) X_i \right)^2 \right] + \hat{\eta}|\theta|^2 \right), \tag{12}$$

where $\theta \in \mathbb{R}^N$, $U : \mathbb{R}^N \to \mathbb{R}$, and $Z = (Y, X_1, \dots, X_N)$ is an $\mathbb{R}^{N+1}$-valued random variable with $Y \in \mathbb{R}$ denoting the return of the target index, $X_i \in \mathbb{R}, i = 1, \dots, N$ denoting the return of the i-th asset. Moreover, $\hat{\eta} > 0$ is the regularization constant, and $g_i(\theta)$ represents the weight of asset i in the portfolio given explicitly by $g_i(\theta) = \frac{e^{\theta_i}}{\sum_{k=1}^N e^{\theta_k}}$, for all $\theta \in \mathbb{R}^N$. One notes that for any $\theta \in \mathbb{R}^N$, $i = 1, \dots, N$, $g_i(\theta) \in (0, 1)$. Moreover, for any $\theta \in \mathbb{R}^N$, $m = 1, \dots, N$, one obtains

$$\partial_{\theta_m} U(\theta) = 2\hat{\eta}\theta_m + 2\mathbb{E}\left[\left(Y - \sum_{i=1}^N g_i(\theta) X_i \right) g_m(\theta) \sum_{i \neq m}^N g_i(\theta)(X_i - X_m) \right]. \tag{13}$$

Proposition 3.2. *Let the objective function U be defined in (12). Denote by $\mathcal{L}(Z)$ the probability law of Z. Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with probability law $\mathcal{L}(Z)$. Assume $|\theta_0| \in L^4$, and $|Z| \in L^{12}$. Moreover, let $H : \mathbb{R}^N \times \mathbb{R}^{N+1} \to \mathbb{R}^N$ be the stochastic gradient of (12) given by*

$$H(\theta, z) := (H_1(\theta, z), \dots, H_N(\theta, z)), \tag{14}$$

where $z := (y, x_1, \dots, x_N) \in \mathbb{R}^{N+1}$, $H_m : \mathbb{R}^N \times \mathbb{R}^{N+1} \to \mathbb{R}$, $m = 1, \dots, N$. The explicit expressions for H_m, $m = 1, \dots, N$ are given as follows:

$$H_m(\theta, z) = 2\hat{\eta}\theta_m + 2\left(y - \sum_{i=1}^N g_i(\theta) x_i \right) g_m(\theta) \sum_{i \neq m}^N g_i(\theta)(x_i - x_m). \tag{15}$$

Then, the following holds:

(i) *The function U is in general nonconvex.*

(ii) *The stochastic gradient H satisfies Assumptions 1, 2, and 3. More precisely, Assumption 2 holds with $L_1 = 6N$, $L_2 = 4\sqrt{N}(N+1)$, $\eta(z) = \hat{\eta} + \left(1 + |y| + \sum_{i=1}^{N}|x_i|\right)\left(1 + \sum_{i\neq m}^{N}(|x_i| + |x_m|)\right)$. Assumption 3 holds with $A(z) = \hat{\eta}\mathbf{I}_N$ and $\hat{b}(z) = \hat{\eta}^{-1}N\left(\left(|y| + \sum_{i=1}^{N}|x_i|\right)\sum_{i\neq m}^{N}(|x_i| + |x_m|)\right)^2$.*

The proof of Proposition 3.2 is postponed to Appendix B. One notes that the stochastic gradient $H(\theta, z)$ fails to satisfy the global conditions as the index tracking optimization (12) can be viewed as an online optimization problem where a uniform bound of the data z is unavailable. However, it is shown in Proposition 3.2 that, for any $\theta \in \mathbb{R}^N, z \in \mathbb{R}^{N+1}$, $H(\theta, z)$ satisfies the local Lipschitz condition (Assumption 2) and the local dissipativity condition (Assumption 3). Moreover, Proposition 3.2 implies that our main results hold for the nonconvex optimization problem (12), which provide theoretical guarantees for the SGLD algorithm to find the approximate minimizers.

4 Proof Overview

In this section, we explain the main idea of proving Theorem 2.4 and Corollary 2.5. We proceed by introducing suitable Lyapunov functions for the analysis of moment estimates of the SGLD algorithm (4). This is done with the help of a continuous-time interpolation of the original recursion (4), which results in a continuous-time process whose laws at discrete times are the same as those of the SGLD. We further introduce a couple of auxiliary continuous-time process, which are used for the derivation of preliminary results. The proof of the main results then follows. We defer all proofs to Appendix C and focus on the exposition of main ideas.

4.1 Introduction of suitable Lyapunov functions and auxiliary processes

We start by defining, for each $p \geq 1$, the Lyapunov function V_p by $V_p(\theta) := (1 + |\theta|^2)^{p/2}$, $\theta \in \mathbb{R}^d$, and similarly $\mathrm{v}_p(\omega) := (1 + \omega^2)^{p/2}$, for any real $\omega \geq 0$. Notice that these functions are twice continuously differentiable and

$$\sup_{\theta}(|\nabla V_p(\theta)|/V_p(\theta)) < \infty, \quad \lim_{|\theta|\to\infty}(\nabla V_p(\theta)/V_p(\theta)) = 0. \tag{16}$$

Let $\mathcal{P}_{V_p}$ denote the set of $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} V_p(\theta)\,\mu(\mathrm{d}\theta) < \infty$.

Consider the Langevin SDE $(Z_t)_{t\in\mathbb{R}_+}$ given by

$$\mathrm{d}Z_t := -h(Z_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t \tag{17}$$

with $Z_0 := \theta_0 \in \mathbb{R}^d$, where $h := \nabla U$ and $(B_t)_{t\geq0}$ is a standard d-dimensional Brownian motion. Denote by $(\mathcal{F}_t)_{t\geq0}$ the natural filtration of $(B_t)_{t\geq0}$, and we assume that $(\mathcal{F}_t)_{t\geq 0}$ is independent of $\mathcal{G}_\infty \vee \sigma(\theta_0)$. Moreover, denote by $\mathcal{F}_\infty := \sigma(\bigcup_{t\geq 0} \mathcal{F}_t)$.

We next introduce the auxiliary processes which are used in our analysis. For each $\lambda > 0$, $Z_t^\lambda := Z_{\lambda t},\ t \in \mathbb{R}_+$, where the process $(Z_t)_{t\in\mathbb{R}_+}$ is defined in (17). We also define $\tilde{B}_t^\lambda := B_{\lambda t}/\sqrt{\lambda}$, $t \geq 0$. We note that $(\tilde{B}_t^\lambda)_{t\geq 0}$ is a Brownian motion and

$$\mathrm{d}Z_t^\lambda := -\lambda h(Z_t^\lambda)\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda, \quad Z_0^\lambda := \theta_0.$$

The natural filtration of $(\tilde{B}_t^\lambda)_{t\geq 0}$ is denoted by $(\mathcal{F}_t^\lambda)_{t\geq 0}$ with $\mathcal{F}_t^\lambda := \mathcal{F}_{\lambda t}$, $t \in \mathbb{R}_+$. Note that $(\mathcal{F}_t^\lambda)_{t\geq 0}$ is independent of $\mathcal{G}_\infty \vee \sigma(\theta_0)$.

Then, define the continuous-time interpolation of the SGLD algorithm (4) as

$$\mathrm{d}\bar{\theta}_t^\lambda := -\lambda H(\bar{\theta}_{\lfloor t\rfloor}^\lambda, X_{\lceil t\rceil})\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda, \quad \bar{\theta}_0^\lambda := \theta_0. \tag{18}$$

In addition, one notes the law of the interpolated process coincides with the law of the SGLD algorithm (4) at grid-points, i.e. $\mathcal{L}(\bar{\theta}_n^\lambda) := \mathcal{L}(\theta_n^\lambda)$, for each $n \in \mathbb{N}$. Hence, crucial estimates for the SGLD can be derived by studying equation (18).

Furthermore, consider a continuous-time process $\zeta_t^{s,v,\lambda}$, $t \geq s$, which denotes the solution of the SDE

$$\mathrm{d}\zeta_t^{s,v,\lambda} := -\lambda h(\zeta_t^{s,v,\lambda})\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda, \quad \zeta_s^{s,v,\lambda} := v \in \mathbb{R}^d.$$

Definition 4.1. *Fix $n \in \mathbb{N}$. For any $t \geq nT$, define $\bar{\zeta}_t^{\lambda,n} := \zeta_t^{nT,\bar{\theta}_{nT}^\lambda,\lambda}$, where $T := \lfloor 1/\lambda \rfloor$.*

Intuitively, $\bar{\zeta}_t^{\lambda,n}$ is a process started from the value of the SGLD process (18) at time nT and run until time $t \geq nT$ with the continuous-time Langevin dynamics.

4.2 Preliminary estimates

It is a classic result that SDE (17) has a unique solution adapted to $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$, since h is Lipschitz-continuous by (5). Note that the second moments of the SDE (17) and of the target distribtuion π_β are finite, see [12, Lemma 3.2] and [4, Proposition 1-(ii)]. For results of SDEs under local Lipschitz conditions, see, e.g. [27] and [28].

In order to obtain the convergence results, we first establish the moment bounds of the process $(\bar{\theta}_t^\lambda)_{t\geq 0}$. This result proves that second and fourth moments of the process $(\bar{\theta}_t^\lambda)_{t\geq 0}$ are uniformly bounded, therefore well-behaved.

Lemma 4.2. *Let Assumptions 1, 2 and 3 hold. For any $0 < \lambda < \lambda_{\max}$ given in (7), $n \in \mathbb{N}$, $t \in (n, n+1]$,*

$$\mathbb{E}\left[|\bar{\theta}_t^\lambda|^2\right] \leq (1 - a\lambda(t-n))(1-a\lambda)^n \mathbb{E}\left[|\theta_0|^2\right] + c_1(\lambda_{\max} + a^{-1}),$$

where

$$c_1 := c_0 + 2d/\beta, \quad c_0 := 4\lambda_{\max} L_2^2 \mathbb{E}\left[\bar{\eta}^2(X_0)\right] + 4\lambda_{\max} H_\star^2 + 2b. \tag{19}$$

In addition, $\sup_t \mathbb{E}\left[|\bar{\theta}_t^\lambda|^2\right] \leq \mathbb{E}\left[|\theta_0|^2\right] + c_1(\lambda_{\max} + a^{-1}) < \infty$*. Similarly, one obtains*

$$\mathbb{E}\left[|\bar{\theta}_t^\lambda|^4\right] \leq (1 - a\lambda(t-n))(1-a\lambda)^n \mathbb{E}\left[|\theta_0|^4\right] + c_3(\lambda_{\max} + a^{-1}),$$

where

$$\begin{aligned} c_3 &:= (1 + a\lambda_{\max})c_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9a^{-1}), \\ c_2 &:= 4bM^2 + 152(1+\lambda_{\max})^3 \\ &\quad \times \left((1+L_2)^4 \mathbb{E}\left[(1+\bar{\eta}(X_0))^4\right] + (1+H_\star)^4\right)(1+M)^2, \\ M &:= \max\{(8ba^{-1} + 48a^{-1}\lambda_{\max}(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2))^{1/2}, \\ &\qquad (128a^{-1}\lambda_{\max}^2(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3))^{1/3}\}. \end{aligned} \tag{20}$$

Moreover, this implies $\sup_t \mathbb{E}|\bar{\theta}_t^\lambda|^4 < \infty$*.*

The uniform bound achieved in Lemma 4.2 for the fourth moment of the process $(\bar{\theta}_t^\lambda)_{t\geq 0}$ enables us to obtain a uniform bound for $V_4(\theta_t^\lambda)$ as presented in the following corollary.

Corollary 4.3. *Let Assumptions 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{\max}$ *given in* (7)*,* $n \in \mathbb{N}$*,* $t \in (n, n+1]$*,*

$$\mathbb{E}[V_4(\bar{\theta}_t^\lambda)] \leq 2(1-a\lambda)^{\lfloor t \rfloor}\mathbb{E}[V_4(\theta_0)] + 2c_3(\lambda_{\max} + a^{-1}) + 2,$$

where c_3 *is given in* (20)*.*

Next, we turn our attention to the process $(\bar{\zeta}_t^{\lambda,n})_{t\in\mathbb{R}}$. We first present a drift condition associated with the SDE (17), which will be used to obtain the moment bounds of the process $\bar{\zeta}_t^{\lambda,n}$.

Lemma 4.4 ([14, Lemma 3.6])**.** *Let Assumptions 1 and 3 hold. Then, for each* $p \geq 2$*,* $\theta \in \mathbb{R}^d$*,*

$$\Delta V_p(\theta)/\beta - \langle h(\theta), \nabla V_p(\theta)\rangle \leq -\bar{c}(p)V_p(\theta) + \tilde{c}(p),$$

where $\bar{c}(p) := ap/4$ *and* $\tilde{c}(p) := (3/4)ap\mathrm{v}_p(\overline{M}_p)$ *with* $\overline{M}_p := (1/3 + 4b/(3a) + 4d/(3a\beta) + 4(p-2)/(3a\beta))^{1/2}$*.*

The following lemma provides explicit upper bounds for $V_p(\bar{\zeta}_t^{\lambda,n})$ in the case $p = 2$ and $p = 4$.

Lemma 4.5. *Let Assumptions 1, 2 and 3 hold. For any $0 < \lambda < \lambda_{\max}$ given in (7), $t \geq nT$, $n \in \mathbb{N}$, one obtains the following inequality*

$$\mathbb{E}[V_2(\bar{\zeta}_t^{\lambda,n})] \leq e^{-a\lambda t/2}\mathbb{E}[V_2(\theta_0)] + 3\mathrm{v}_2(\overline{M}_2) + c_1(\lambda_{\max} + a^{-1}) + 1,$$

where the process $\bar{\zeta}_t^{\lambda,n}$ is defined in Definition 4.1 and c_1 is given in (19). Furthermore,

$$\mathbb{E}[V_4(\bar{\zeta}_t^{\lambda,n})] \leq 2e^{-a\lambda t}\mathbb{E}[V_4(\theta_0)] + 3\mathrm{v}_4(\overline{M}_4) + 2c_3(\lambda_{\max} + a^{-1}) + 2,$$

where c_3 is given in (20).

4.3 Proof of the main theorems

We are now ready to establish our main results. Recall that, our goal is to establish a non-asymptotic bound for $W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta)$.

We first split $W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta)$ as follows by using triangle inequality: $W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq W_1(\mathcal{L}(\bar{\theta}_n^\lambda), \mathcal{L}(Z_n^\lambda)) + W_1(\mathcal{L}(Z_n^\lambda), \pi_\beta)$. We aim at bounding the two terms on the right hand side separately. To achieve this, we first introduce a functional which is crucial to obtain the convergence rate in W_1. For any $p \geq 1$, $\mu, \nu \in \mathcal{P}_{V_p}$,

$$w_{1,p}(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} [1 \wedge |\theta - \theta'|](1 + V_p(\theta) + V_p(\theta'))\zeta(\mathrm{d}\theta \mathrm{d}\theta'), \quad (21)$$

and it satisfies trivially $W_1(\mu, \nu) \leq w_{1,p}(\mu, \nu)$. The case $p = 2$, i.e. $w_{1,2}$, is used throughout the section. The result below states a contraction property of $w_{1,2}$.

Proposition 4.6. *Let Assumptions 1, 2 and 3 hold. Let Z_t', $t \in \mathbb{R}_+$ be the solution of (17) with initial condition $Z_0' = \theta_0'$ which is independent of $\mathcal{F}_\infty$ and satisfies $|\theta_0'| \in L^2$. Then,*

$$w_{1,2}(\mathcal{L}(Z_t), \mathcal{L}(Z_t')) \leq \hat{c}e^{-\dot{c}t} w_{1,2}(\mathcal{L}(\theta_0), \mathcal{L}(\theta_0')),$$

where the constants $\dot{c}$ and $\hat{c}$ are given in Lemma 4.11.

Proof One notes that [15, Assumption 2.1] holds with $\kappa = L_1\mathbb{E}[\eta(X_0)]$ due to Remark 2.1. [15, Assumption 2.2] holds with $V = V_2$ due to Lemma 4.4. Moreover, [15, Assumptions 2.4 and 2.5] hold due to (16). Thus, [15, Theorem 2.2, Corollary 2.3] hold under Assumptions 1, 2 and 3. Then, the desired result can be obtained by using the same argument as in the proof of [14, Proposition 3.14]. □

Now recall $T := \lfloor 1/\lambda \rfloor$ defined in Definition 4.1. By using the contraction property provided in Proposition 4.6, one can construct the non-asymptotic bound between $\mathcal{L}(\bar{\theta}_t^\lambda)$ and $\mathcal{L}(Z_t^\lambda)$ in W_1 distance by decomposing the error using the auxiliary process $\bar{\zeta}_t^{\lambda,n}$:

$$W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) \leq W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})) + W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^\lambda)). \tag{22}$$

By the definition of $\lambda_{\max}$ given in (7), we have that $0 < \lambda \leq \lambda_{\max} \leq 1$, which implies $1/2 < \lambda T \leq 1$. An upper bound for the first term in (22) is obtained below.

Lemma 4.7. *Let Assumptions 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{\max}$ *given in* (7), $t \in (nT, (n+1)T]$,

$$W_2(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})) \leq \sqrt{\lambda}(e^{-an/4}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)] + \bar{C}_{2,2})^{1/2},$$

where $\bar{C}_{2,1}$ *and* $\bar{C}_{2,2}$ *are given in* (C12).

Then, the following Lemma provides the bound for the second term in (22).

Lemma 4.8. *Let Assumptions 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{\max}$ *given in* (7), $t \in (nT, (n+1)T]$,

$$W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^\lambda)) \leq \sqrt{\lambda}(e^{-\dot{c}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)] + \bar{C}_{2,4}),$$

where $\bar{C}_{2,3}$, $\bar{C}_{2,4}$ *are given in* (C13).

By using similar arguments as in Lemma 4.8, one can obtain the non-asymptotic estimate in W_2 distance between $\mathcal{L}(\bar{\zeta}_t^{\lambda,n})$ and $\mathcal{L}(Z_t^\lambda)$, which is given in the following corollary.

Corollary 4.9. *Let Assumptions 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{\max}$ *given in* (7), $t \in (nT, (n+1)T]$,

$$W_2(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^\lambda)) \leq \lambda^{1/4}(e^{-\dot{c}n/4}\bar{C}_{2,3}^*\mathbb{E}^{1/2}[V_4(\theta_0)] + \bar{C}_{2,4}^*),$$

where $\bar{C}_{2,3}^*$, $\bar{C}_{2,4}^*$ *are given in* (C14).

Finally, by using the inequality (22) and the results from previous lemmas, one can obtain the non-asymptotic bound between $\mathcal{L}(\bar{\theta}_t^\lambda)$ and $\mathcal{L}(Z_t^\lambda)$, $t \in (nT, (n+1)T]$, in W_1 distance.

Lemma 4.10. *Let Assumptions 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{\max}$ *given in* (7), $t \in (nT, (n+1)T]$,

$$W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) \leq (\bar{C}_{2,1}^{1/2} + \bar{C}_{2,2}^{1/2} + \bar{C}_{2,3} + \bar{C}_{2,4})\sqrt{\lambda}(e^{-\dot{c}n/2}\mathbb{E}[V_4(\theta_0)] + 1),$$

where $\bar{C}_{2,1}$, $\bar{C}_{2,2}$ are given in (C12) *(Lemma 4.7), and $\bar{C}_{2,3}$, $\bar{C}_{2,4}$ are given in* (C13) *(Lemma 4.8).*

Before proceeding to the proofs of the main results, the constants $\dot{c}$ and $\hat{c}$ from Proposition 4.6 are given in an explicit form.

Lemma 4.11. *The contraction constant $\dot{c} > 0$ in Proposition 4.6 is given by*

$$\dot{c} := \min\{\bar{\phi}, \bar{c}(2), 4\tilde{c}(2)\epsilon\bar{c}(2)\}/2, \tag{23}$$

where $\bar{c}(2) = a/2$, $\tilde{c}(2) = (3/2)a\mathrm{v}_2(\overline{M}_2)$ with $\overline{M}_2$ given in Lemma 4.4, $\bar{\phi}$ is given by

$$\bar{\phi} := \left(\bar{b}\sqrt{8\pi/(\beta K_1)}\exp\left(\left(\bar{b}\sqrt{\beta K_1/8} + \sqrt{8/(\beta K_1)}\right)^2\right)\right)^{-1}, \tag{24}$$

and moreover, $\epsilon > 0$ can be chosen such that following inequality is satisfied

$$\epsilon \leq 1 \wedge \left(4\tilde{c}(2)\sqrt{2\beta\pi/K_1}\int_0^{\tilde{b}} \exp\left(\left(s\sqrt{\beta K_1/8} + \sqrt{8/(\beta K_1)}\right)^2\right) \mathrm{d}s\right)^{-1}, \tag{25}$$

where $K_1 := L_1\mathbb{E}[\eta(X_0)]$, $\tilde{b} := 2\sqrt{2\tilde{c}(2)/\bar{c}(2) - 1}$ and $\bar{b} := 2\sqrt{4\tilde{c}(2)(1+\bar{c}(2))/\bar{c}(2) - 1}$.

The constant $\hat{c} > 0$ is given by $\hat{c} := 2(1+\bar{b})\exp(\beta K_1\bar{b}^2/8 + 2\bar{b})/\epsilon$.

Now, we are ready to prove our first main result, namely Theorem 2.4.

Proof of Theorem 2.4 One notes that, by using $\lambda T > 1/2$, Lemma 4.10 and Proposition 4.6, for $t \in (nT, (n+1)T]$

$$\begin{aligned}
W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \pi_\beta) &\leq W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) + W_1(\mathcal{L}(Z_t^\lambda), \pi_\beta) \\
&\leq (\bar{C}_{2,1}^{1/2} + \bar{C}_{2,2}^{1/2} + \bar{C}_{2,3} + \bar{C}_{2,4})\sqrt{\lambda}(e^{-\dot{c}n/2}\mathbb{E}[V_4(\theta_0)] + 1) \\
&\quad + \hat{c}e^{-\dot{c}\lambda t} w_{1,2}(\theta_0, \pi_\beta) \\
&\leq (\bar{C}_{2,1}^{1/2} + \bar{C}_{2,2}^{1/2} + \bar{C}_{2,3} + \bar{C}_{2,4})\sqrt{\lambda}(e^{-\dot{c}n/2}\mathbb{E}[V_4(\theta_0)] + 1) \\
&\quad + \hat{c}e^{-\dot{c}\lambda t}\left[1 + \mathbb{E}[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right] \\
&\leq 2e^{-\dot{c}n/2}(\lambda_{\max}^{1/2}(\bar{C}_{2,1}^{1/2} + \bar{C}_{2,2}^{1/2} + \bar{C}_{2,3} + \bar{C}_{2,4}) + \hat{c})(1 + \mathbb{E}[|\theta_0|^4]) \\
&\quad + \hat{c}e^{-\dot{c}n/2}\left[1 + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right](1 + \mathbb{E}[|\theta_0|^4]) \\
&\quad + \sqrt{\lambda}(\bar{C}_{2,1}^{1/2} + \bar{C}_{2,2}^{1/2} + \bar{C}_{2,3} + \bar{C}_{2,4}).
\end{aligned} \tag{26}$$

The above result implies, for any $n \in \mathbb{N}$,

$$W_1(\mathcal{L}(\bar{\theta}_{(n+1)T}^\lambda), \pi_\beta) \leq C_1 e^{-\dot{c}(n+1)/2}(1 + \mathbb{E}[|\theta_0|^4]) + (C_2 + C_3)\sqrt{\lambda}, \tag{27}$$

where

$$\begin{aligned}
C_1 &:= 2e^{\dot{c}/2}\left[(\lambda_{\max}^{1/2}(\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2}+\bar{C}_{2,3}+\bar{C}_{2,4})+\hat{c})+\hat{c}\left(1+\int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)\right] \\
&= O\left(e^{C_\star(1+d/\beta)(1+\beta)}\left(1+\frac{1}{1-e^{-\dot{c}}}\right)\right), \\
C_2 &:= \bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2} = O\left(1+\sqrt{\frac{d}{\beta}}\right), \\
C_3 &:= \bar{C}_{2,3}+\bar{C}_{2,4} = O\left(e^{C_\star(1+d/\beta)(1+\beta)}\left(1+\frac{1}{1-e^{-\dot{c}}}\right)\right)
\end{aligned} \tag{28}$$

with $\dot{c}, \hat{c}$ given in Lemma 4.11, $\bar{C}_{2,1}$, $\bar{C}_{2,2}$ given in (C12) (Lemma 4.7), $\bar{C}_{2,3}$, $\bar{C}_{2,4}$ given in (C13) (Lemma 4.8), $C_\star > 0$ independent of d, β, n. One notes that the above estimate (27) is established for $(\bar{\theta}^\lambda_{(n+1)T})_{n\in\mathbb{N}}$. To obtain a non-asymptotic error bound for $(\bar{\theta}^\lambda_{(n+1)})_{n\in\mathbb{N}}$, we set $(n+1)T$ to $n+1$ on the LHS of (27), and set $n+1$ to $(n+1)/T$ on the RHS of (27). By using $\lambda(n+1) \leq (n+1)/T$, it follows that, for any $n \in \mathbb{N}$,

$$W_1(\mathcal{L}(\theta^\lambda_{n+1}), \pi_\beta) \leq C_1 e^{-\dot{c}\lambda(n+1)/2}(1+\mathbb{E}[|\theta_0|^4]) + (C_2+C_3)\sqrt{\lambda}.$$

Moreover, for $\varepsilon > 0$, if we choose λ and n such that $\lambda \leq \lambda_{\max}$, $C_1 e^{-\dot{c}\lambda n/2}(1+\mathbb{E}[|\theta_0|^4]) \leq \varepsilon/2$, $(C_2+C_3)\sqrt{\lambda} \leq \varepsilon/2$, where $\lambda_{\max}$ is given in (7), then $W_1(\mathcal{L}(\theta^\lambda_n), \pi_\beta) \leq \varepsilon$. This implies $\lambda \leq \frac{\varepsilon^2}{4(C_2+C_3)^2} \wedge \lambda_{\max}$, $\lambda n \geq \frac{2}{\dot{c}} \ln \frac{2C_1(1+\mathbb{E}[|\theta_0|^4])}{\varepsilon}$. More precisely, by using (28), one obtains $n \geq \frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon^2 \dot{c}}\left(1+\frac{1}{(1-e^{-\dot{c}})^2}\right)\ln\left(\frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon}\left(1+\frac{1}{1-e^{-\dot{c}}}\right)\right)$, where $\dot{c}$ is the contraction constant of the Langevin diffusion (17) given explicitly in Lemma 4.11. □

Next, we prove our second result Corollary 2.5, i.e., a uniform bound in W_2 distance.

Proof of Corollary 2.5 By using (C11) in Lemma 4.7, Corollary 4.9 and Proposition 4.6, one obtains

$$\begin{aligned}
W_2(\mathcal{L}(\bar{\theta}^\lambda_t), \pi_\beta) &\leq W_2(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(Z^\lambda_t)) + W_2(\mathcal{L}(Z^\lambda_t), \pi_\beta) \\
&\leq W_2(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(\bar{\zeta}^{\lambda,n}_t)) + W_2(\mathcal{L}(\bar{\zeta}^{\lambda,n}_t), \mathcal{L}(Z^\lambda_t)) + W_2(\mathcal{L}(Z^\lambda_t), \pi_\beta) \\
&\leq \sqrt{\lambda}(e^{-an/4}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)] + \bar{C}_{2,2})^{1/2} \\
&\quad + \lambda^{1/4}(e^{-\dot{c}n/4}\bar{C}^*_{2,3}\mathbb{E}^{1/2}[V_4(\theta_0)] + \bar{C}^*_{2,4}) \\
&\quad + \sqrt{2w_{1,2}(\mathcal{L}(Z^\lambda_t), \pi_\beta)} \\
&\leq \lambda^{1/4}(\lambda_{\max}^{1/4}\bar{C}_{2,1}^{1/2} + \lambda_{\max}^{1/4}\bar{C}_{2,2}^{1/2} + \bar{C}^*_{2,3} + \bar{C}^*_{2,4})(e^{-\dot{c}n/4}\mathbb{E}[V_4(\theta_0)] + 1) \\
&\quad + \hat{c}^{1/2}e^{-\dot{c}\lambda t/2}\sqrt{2w_{1,2}(\theta_0, \pi_\beta)},
\end{aligned}$$

Further calculations yield,

$$\begin{aligned}
&W_2(\mathcal{L}(\bar{\theta}^\lambda_t), \pi_\beta) \\
&\leq \lambda^{1/4}(\lambda_{\max}^{1/4}\bar{C}_{2,1}^{1/2} + \lambda_{\max}^{1/4}\bar{C}_{2,2}^{1/2} + \bar{C}^*_{2,3} + \bar{C}^*_{2,4})(e^{-\dot{c}n/4}\mathbb{E}[V_4(\theta_0)] + 1)
\end{aligned}$$

$$
\begin{aligned}
&+\sqrt{2}\hat{c}^{1/2}e^{-\dot{c}\lambda t/2}\left(1+\mathbb{E}[V_2(\theta_0)]+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right)^{1/2}\\
\leq\;& 2e^{-\dot{c}n/4}(\lambda_{\max}^{1/2}(\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2})+\lambda_{\max}^{1/4}(\bar{C}_{2,3}^{*}+\bar{C}_{2,4}^{*})+\sqrt{2}\hat{c}^{1/2})(1+\mathbb{E}[|\theta_0|^4])\\
&+\sqrt{2}\hat{c}^{1/2}e^{-\dot{c}n/4}\left[1+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right]\\
&+\lambda^{1/4}(\lambda_{\max}^{1/4}\bar{C}_{2,1}^{1/2}+\lambda_{\max}^{1/4}\bar{C}_{2,2}^{1/2}+\bar{C}_{2,3}^{*}+\bar{C}_{2,4}^{*}),
\end{aligned}
$$

where the last inequality holds due to $\lambda T > 1/2$. Thus, for any $n \in \mathbb{N}$, it follows that

$$
W_2(\mathcal{L}(\theta_n^\lambda),\pi_\beta) \leq C_4\mathrm{e}^{-\dot{c}\lambda n/4}\mathbb{E}[|\theta_0|^4+1]+(C_5+C_6)\lambda^{1/4}
$$

where

$$
\begin{aligned}
C_4 &:= 2\left(\lambda_{\max}^{1/2}(\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2})+\lambda_{\max}^{1/4}(\bar{C}_{2,3}^{*}+\bar{C}_{2,4}^{*})+\sqrt{2}\hat{c}^{1/2}\right)\\
&\quad+\sqrt{2}\hat{c}^{1/2}\left(1+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right)\\
&\;=O\left(e^{C_\star(1+d/\beta)(1+\beta)}\left(1+\frac{1}{1-e^{-\dot{c}/2}}\right)\right) \qquad (29)\\
C_5 &:= \lambda_{\max}^{1/4}\bar{C}_{2,1}^{1/2}+\lambda_{\max}^{1/4}\bar{C}_{2,2}^{1/2}=O\left(1+\sqrt{\frac{d}{\beta}}\right),\\
C_6 &:= \bar{C}_{2,3}^{*}+\bar{C}_{2,4}^{*}=O\left(e^{C_\star(1+d/\beta)(1+\beta)}\left(1+\frac{1}{1-e^{-\dot{c}/2}}\right)\right),
\end{aligned}
$$

with $\dot{c},\hat{c}$ given in Lemma 4.11, $\bar{C}_{2,1}$, $\bar{C}_{2,2}$ given in (C12) (Lemma 4.7), $\bar{C}_{2,3}^{*}$, $\bar{C}_{2,4}^{*}$ given in (C14) (Lemma 4.9), $C_\star > 0$ independent of d,β,n.

Moreover, for $\varepsilon > 0$, if we choose λ and n such that, $\lambda \leq \lambda_{\max}$, $C_4\mathrm{e}^{-\dot{c}\lambda n/4}\mathbb{E}[|\theta_0|^4+1] \leq \varepsilon/2$, $(C_5+C_6)\lambda^{1/4} \leq \varepsilon/2$, where $\lambda_{\max}$ is given in (7), then $W_2(\mathcal{L}(\theta_n^\lambda),\pi_\beta) \leq \varepsilon$. This implies $\lambda \leq \frac{\varepsilon^4}{16(C_5+C_6)^4} \wedge \lambda_{\max}$, $\lambda n \geq \frac{4}{\dot{c}}\ln\frac{2C_4(1+\mathbb{E}[|\theta_0|^4])}{\varepsilon}$. More precisely, by using (29), one obtains $n \geq \frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon^4\dot{c}}\left(1+\frac{1}{(1-e^{-\dot{c}/2})^4}\right)\ln\left(\frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon}\left(1+\frac{1}{1-e^{-\dot{c}/2}}\right)\right)$, where $\dot{c}$ is the contraction constant of the Langevin diffusion (17) given explicitly in Lemma 4.11. □

Finally, we move on to prove our result on nonconvex optimization, namely, Corollary 2.8.

Proof of Corollary 2.8 To obtain an upper bound for the expected excess risk $\mathbb{E}[U(\theta_n^\lambda)]-\inf_{\theta\in\mathbb{R}^d}U(\theta)$, one considers the following splitting

$$
\mathbb{E}[U(\theta_n^\lambda)]-\inf_{\theta\in\mathbb{R}^d}U(\theta)=\left(\mathbb{E}[U(\theta_n^\lambda)]-\mathbb{E}[U(Z_\infty)]\right)+\left(\mathbb{E}[U(Z_\infty)]-\inf_{\theta\in\mathbb{R}^d}U(\theta)\right), \quad (30)
$$

where $Z_\infty \sim \pi_\beta$ with π_β defined in (3). By using [12, Lemma 3.5], Remark 2.1, Lemma 4.2, and Corollary 2.5, the first term on the RHS of (30) can be bounded by

$$
\begin{aligned}
&\mathbb{E}[U(\theta_n^\lambda)]-\mathbb{E}[U(Z_\infty)]\\
&\leq\left(L_1\mathbb{E}[\eta(X_0)](\mathbb{E}[|\theta_0|^2]+c_1(\lambda_{\max}+a^{-1}))+L_2\mathbb{E}[\bar{\eta}(X_0)]+H_\star\right)W_2(\mathcal{L}(\theta_n^\lambda),\pi_\beta)\\
&\leq\left(L_1\mathbb{E}[\eta(X_0)](\mathbb{E}[|\theta_0|^2]+c_1(\lambda_{\max}+a^{-1}))+L_2\mathbb{E}[\bar{\eta}(X_0)]+H_\star\right)
\end{aligned}
$$

$$
\begin{aligned}
&\quad \times \left(C_4 \mathrm{e}^{-\dot{c}\lambda n/4} \mathbb{E}[|\theta_0|^4 + 1] + (C_5 + C_6)\lambda^{1/4} \right) \\
&\leq C_1^\sharp e^{-\dot{c}\lambda n/4} + C_2^\sharp \lambda^{1/4},
\end{aligned}
$$

where

$$
\begin{aligned}
C_1^\sharp &:= C_4 \left(L_1 \mathbb{E}[\eta(X_0)](\mathbb{E}[|\theta_0|^2] + c_1(\lambda_{\max} + a^{-1})) + L_2 \mathbb{E}[\bar{\eta}(X_0)] + H_\star \right) \mathbb{E}[|\theta_0|^4 + 1], \\
C_2^\sharp &:= (C_5 + C_6) \left(L_1 \mathbb{E}[\eta(X_0)](\mathbb{E}[|\theta_0|^2] + c_1(\lambda_{\max} + a^{-1})) + L_2 \mathbb{E}[\bar{\eta}(X_0)] + H_\star \right),
\end{aligned} \tag{31}
$$

with $\dot{c}$ given in (23), C_4, C_5, C_6 given in (29) and c_1 given in (19). Moreover, the second term on the RHS of (30) can be estimated by using [12, Proposition 3.4], which gives, $\mathbb{E}[U(Z_\infty)] - \inf_{\theta\in\mathbb{R}^d} U(\theta) \leq C_3^\sharp$, where

$$
C_3^\sharp := \frac{d}{2\beta} \log \left(\frac{eL_1 \mathbb{E}[\eta(X_0)]}{a} \left(\frac{b\beta}{d} + 1 \right) \right). \tag{32}
$$

Finally, one obtains $\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta\in\mathbb{R}^d} U(\theta) \leq C_1^\sharp e^{-\dot{c}\lambda n/4} + C_2^\sharp \lambda^{1/4} + C_3^\sharp$, where $\dot{c}$ is given in (23), and $C_1^\sharp = O\left(e^{C_\star(1+d/\beta)(1+\beta)} \left(1 + \frac{1}{1-e^{-\dot{c}/2}} \right) \right)$, $C_2^\sharp = O\left(e^{C_\star(1+d/\beta)(1+\beta)} \left(1 + \frac{1}{1-e^{-\dot{c}/2}} \right) \right)$, $C_3^\sharp = O((d/\beta)\log(C_\star(\beta/d+1)))$ with $C_\star > 0$ a constant independent of n, d, β.

Moreover, for $\varepsilon > 0$, if we choose β such that $C_3^\sharp \leq \varepsilon/3$, then choose λ such that $\lambda \leq \lambda_{\max}$ with $\lambda_{\max}$ given in (7) and $C_2^\sharp \lambda^{1/4} \leq \varepsilon/3$, and finally choose n such that $C_1^\sharp e^{-\dot{c}\lambda n/4} \leq \varepsilon/3$, consequently, we obtain $\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta\in\mathbb{R}^d} U(\theta) \leq \varepsilon$. This implies $\beta \geq \beta_\varepsilon \vee \frac{3d}{\varepsilon} \log \left(\frac{eL_1\mathbb{E}[\eta(X_0)]}{ad} (b+1)(d+1) \right)$, where β_ε is the root of the function $f^\sharp(\beta) = \frac{\log(\beta+1)}{\beta} - \frac{\varepsilon}{3d}$, $\beta > 0$, i.e. $f^\sharp(\beta_\varepsilon) = 0$. Indeed, since

$$
C_3^\sharp \leq \frac{d}{2\beta} \log \left(\frac{eL_1 \mathbb{E}[\eta(X_0)]}{ad} (b+1)(d+1)(\beta+1) \right),
$$

by setting $\frac{d}{2\beta} \log \left(\frac{eL_1\mathbb{E}[\eta(X_0)]}{ad} (b+1)(d+1) \right) \leq \varepsilon/6$ and $\frac{d}{2\beta} \log(\beta+1) \leq \varepsilon/6$, one obtains $C_3^\sharp \leq \varepsilon/3$. Noticing that $\frac{\log(\beta+1)}{\beta}$ is decreasing in β yields the desired result. Furthermore, calculations yield $\lambda \leq \frac{\varepsilon^4}{81(C_2^\sharp)^4} \wedge \lambda_{\max}$, and $\lambda n \geq \frac{4}{\dot{c}} \ln \frac{3C_1^\sharp}{\varepsilon}$. More precisely, $n \geq \frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon^4 \dot{c}} \left(1 + \frac{1}{(1-e^{-\dot{c}/2})^4} \right) \ln \left(\frac{C_\star e^{C_\star(1+d/\beta)(1+\beta)}}{\varepsilon} \left(1 + \frac{1}{1-e^{-\dot{c}/2}} \right) \right)$, where $\dot{c}$ is the contraction constant of the Langevin diffusion (17) given explicitly in Lemma 4.11. □

5 Conclusions

We have provided non-asymptotic estimates for the SGLD which explicitly bounds the error between the target measure and the law of the SGLD in Wasserstein-1 and 2 distances. These results further allow us to establish a non-asymptotic error bound for the expected excess risk. Moreover, the theoretical findings enable us to obtain theoretical guarantees for fundamental problems in machine learning and in financial mathematics: Nonasymptotic error bounds for *nonconvex optimization* problems. We have shown that our assumptions are verifiable for a large class of practical problems. In particular, we demonstrate this by providing two important applications: (i) *variational inference for Bayesian logistic regression* (VI), (ii) *index tracking optimization.* We believe that our results provide a detailed understanding of the sampling behaviour of SGLD even when it is examined within the context of nonconvex optimization.

Appendix A Additional Lemmata

Lemma A.1. *Let Assumptions 1, 2 and 3 hold. For any $t \in (nT, (n+1)T]$, $n \in \mathbb{N}$ and $k = 1, \dots, K+1$, $K+1 \leq T$, one obtains*

$$\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n}) - H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k})\right|^2\right] \leq e^{-a\lambda t/2}\bar{\sigma}_Z \mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Z,$$

where

$$\begin{aligned}\bar{\sigma}_Z &:= 8L_2^2\hat{\sigma}_Z, \quad \tilde{\sigma}_Z := 8L_2^2\hat{\sigma}_Z(3\mathrm{v}_2(\overline{M}_2) + c_1(\lambda_{\max} + a^{-1}) + 1),\\ \hat{\sigma}_Z &:= \mathbb{E}[(\eta(X_0) + \eta(\mathbb{E}[X_0]))^2|X_0 - \mathbb{E}[X_0]|^2].\end{aligned} \tag{A1}$$

Proof Recall $\mathcal{H}_t = \mathcal{F}_\infty^\lambda \vee \mathcal{G}_{\lfloor t \rfloor}$. One notices that

$$\begin{aligned}&\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n}) - H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k})\right|^2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n}) - H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k})\right|^2\middle|\mathcal{H}_{nT}\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\left|\mathbb{E}\left[H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k})\middle|\mathcal{H}_{nT}\right] - H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k})\right|^2\middle|\mathcal{H}_{nT}\right]\right]\\ &\leq 4\mathbb{E}\left[\mathbb{E}\left[\left|H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k}) - H(\bar{\zeta}_t^{\lambda,n}, \mathbb{E}\left[X_{nT+k}\middle|\mathcal{H}_{nT}\right])\right|^2\middle|\mathcal{H}_{nT}\right]\right]\\ &\leq 4L_2^2\hat{\sigma}_Z\mathbb{E}\left[\left(1 + \left|\bar{\zeta}_t^{\lambda,n}\right|\right)^2\right],\end{aligned}$$

where the first inequality holds due to Lemma A.3 and $\hat{\sigma}_Z := \mathbb{E}[(\eta(X_0) + \eta(\mathbb{E}[X_0]))^2|X_0 - \mathbb{E}[X_0]|^2]$. Then, by using Lemma 4.5, one obtains

$$\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n}) - H(\bar{\zeta}_t^{\lambda,n}, X_{nT+k})\right|^2\right] \leq 8L_2^2\hat{\sigma}_Z\mathbb{E}\left[V_2(\bar{\zeta}_t^{\lambda,n})\right] \leq e^{-a\lambda t/2}\bar{\sigma}_Z\mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Z,$$

where $\bar{\sigma}_Z := 8L_2^2\hat{\sigma}_Z$ and $\tilde{\sigma}_Z := 8L_2^2\hat{\sigma}_Z(3\mathrm{v}_2(\overline{M}_2) + c_1(\lambda_{\max} + a^{-1}) + 1)$. □

Lemma A.2. *Let Assumptions 1, 2 and 3 hold. For any $t > 0$, one obtains*

$$\mathbb{E}\left[\left|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda\right|^2\right] \leq \lambda(e^{-a\lambda\lfloor t \rfloor}\bar{\sigma}_Y\mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Y),$$

where

$$\begin{aligned}\bar{\sigma}_Y &:= 2\lambda_{\max} L_1^2 \mathbb{E}[\eta^2(X_0)] \\ \tilde{\sigma}_Y &:= 2\lambda_{\max} L_1^2 \mathbb{E}[\eta^2(X_0)] c_1(\lambda_{\max} + a^{-1}) + 4\lambda_{\max} L_2^2 \mathbb{E}[\bar{\eta}^2(X_0)] \\ &\quad + 4\lambda_{\max} H_\star^2 + 2d\beta^{-1}.\end{aligned} \tag{A2}$$

Proof For any $t > 0$, we write the difference $\left|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda\right|$ and use $(a+b)^2 \leq 2a^2 + 2b^2$ which yields

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda\right|^2\right] &= \mathbb{E}\left[\left|-\lambda \int_{\lfloor t \rfloor}^t H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil}) \mathrm{d}s + \sqrt{\frac{2\lambda}{\beta}} (\tilde{B}_t^\lambda - \tilde{B}_{\lfloor t \rfloor}^\lambda)\right|^2\right] \\ &\leq \lambda^2 \mathbb{E}\left[\left(L_1 \eta(X_{\lceil t \rceil}) |\bar{\theta}_{\lfloor t \rfloor}^\lambda| + L_2 \bar{\eta}(X_{\lceil t \rceil}) + H_\star\right)^2\right] + 2d\lambda\beta^{-1},\end{aligned}$$

where the inequality holds due to Remark 2.1 and by applying Lemma 4.2, one obtains

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda\right|^2\right] &\leq 2\lambda^2 L_1^2 \mathbb{E}[\eta^2(X_0)] \mathbb{E}[|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^2] + 4\lambda^2 L_2^2 \mathbb{E}[\bar{\eta}^2(X_0)] + 4\lambda^2 H_\star^2 + 2d\lambda\beta^{-1} \\ &\leq \lambda((1 - a\lambda)^{\lfloor t \rfloor} \bar{\sigma}_Y \mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Y),\end{aligned}$$

where $\bar{\sigma}_Y := 2\lambda_{\max} L_1^2 \mathbb{E}[\eta^2(X_0)]$ and $\tilde{\sigma}_Y := 2\lambda_{\max} L_1^2 \mathbb{E}[\eta^2(X_0)] c_1(\lambda_{\max} + a^{-1}) + 4\lambda_{\max} L_2^2 \mathbb{E}[\bar{\eta}^2(X_0)] + 4\lambda_{\max} H_\star^2 + 2d\beta^{-1}$. □

Lemma A.3. *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be sigma-algebras. Let $p \geq 1$. Let X, Y be $\mathbb{R}^d$-valued random vectors in L^p such that Y is measurable with respect to $\mathcal{H} \vee \mathcal{G}$. Then, $\mathbb{E}^{1/p}\left[|X - \mathbb{E}[X|\mathcal{H} \vee \mathcal{G}]|^p | \mathcal{G}\right] \leq 2\mathbb{E}^{1/p}\left[|X - Y|^p | \mathcal{G}\right]$.*

Proof See [29, Lemma 6.1]. □

Appendix B Proofs of the results in 3

Proof of Proposition 3.1 By using (10), it can be shown by direct calculations that H defined in (11) satisfies Assumption 1.

One notes that (11) can be rewritten as

$$\begin{aligned}H(\theta, u) &= \sum_{i=1}^n H_i(\theta, u) \\ &= \sum_{i=1}^n \left(\frac{1}{n}\left(\frac{\theta}{2} + \frac{u}{4} - \frac{3}{4}\hat{a} + \frac{\hat{a}}{4}\left(\frac{7}{1 + e^{2\hat{a}^\mathsf{T}(u/4+\theta)}} - \frac{1}{1 + e^{2\hat{a}^\mathsf{T}(u/4-\theta)}}\right)\right.\right. \\ &\qquad \left. - \frac{7(u + 8\theta)}{16(1 + e^{2(\theta^\mathsf{T} u/4 + |\theta|^2)})} - \frac{u - 8\theta}{16(1 + e^{2(\theta^\mathsf{T} u/4 - |\theta|^2)})}\right) \\ &\qquad \left. + \frac{1}{8}\left(-6z_i y_i + \frac{7z_i}{1 + e^{-z_i^\mathsf{T}(u/4+\theta)}} - \frac{z_i}{1 + e^{-z_i^\mathsf{T}(u/4-\theta)}}\right)\right),\end{aligned}$$

where $H_i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ for each $i = 1, \dots, n$. To verify Assumption 3, which is the (local) dissipativity condition, one calculates, for any $\theta \in \mathbb{R}^d$, $u \in \mathbb{R}^d$

$$\begin{aligned}
\theta^\mathsf{T} H_i(\theta,u) = {} & \frac{1}{n}\left(\frac{|\theta|^2}{2} + \frac{u^\mathsf{T}\theta}{4} - \frac{3}{4}\hat{a}^\mathsf{T}\theta + \frac{\hat{a}^\mathsf{T}\theta}{4}\left(\frac{7}{1+e^{2\hat{a}^\mathsf{T}(u/4+\theta)}} - \frac{1}{1+e^{2\hat{a}^\mathsf{T}(u/4-\theta)}}\right)\right. \\
& \left. - \frac{7(u^\mathsf{T}\theta + 8|\theta|^2)}{16(1+e^{2(\theta^\mathsf{T}u/4+|\theta|^2)})} - \frac{u^\mathsf{T}\theta - 8|\theta|^2}{16(1+e^{2(\theta^\mathsf{T}u/4-|\theta|^2)})}\right) \\
& + \frac{1}{8}\left(-6y_i z_i^\mathsf{T}\theta + \frac{7z_i^\mathsf{T}\theta}{1+e^{-z_i^\mathsf{T}(u/4+\theta)}} - \frac{z_i^\mathsf{T}\theta}{1+e^{-z_i^\mathsf{T}(u/4-\theta)}}\right) \\
\geq {} & \frac{1}{n}\left(\frac{|\theta|^2}{2} - \frac{3|u^\mathsf{T}\theta|}{4} - \frac{11}{4}|\hat{a}^\mathsf{T}\theta|\right) - \frac{7}{4}|z_i^\mathsf{T}\theta| - \frac{7}{4} \\
\geq {} & \frac{1}{4n}|\theta|^2 - \frac{1}{n}\left(\frac{9|u|^2}{4} + \frac{121}{4}|\hat{a}|^2\right) - \frac{49n}{8}|z_i|^2 - \frac{7}{4},
\end{aligned}$$

which implies

$$\theta^\mathsf{T} H(\theta,u) \geq \frac{1}{4}|\theta|^2 - \left(\frac{9|u|^2}{4} + \frac{121}{4}|\hat{a}|^2\right) - \frac{49n}{8}\sum_{i=1}^n |z_i|^2 - \frac{7n}{4}.$$

Thus the (local) dissipativity condition holds with $A(u) = \mathbf{I}_d/4$ and $\hat{b}(u) = (9|u|^2/4 + 121|\hat{a}|^2/4) + 49n\sum_{i=1}^n |z_i|^2/8 + 7n/4$.

As for the Lipschitz conditions in Assumption 2, one notices that $1 + e^{2(\theta^\mathsf{T}u/4+|\theta|^2)} = e^{-|u|^2/32}(e^{|u|^2/32} + e^{2|\theta+u/8|^2})$, then

$$\begin{aligned}
\nabla_\theta H_i(\theta,u) = {} & \frac{1}{n}\left(\frac{\mathbf{I}_d}{2} - \frac{\hat{a}\hat{a}^\mathsf{T}}{2}\left(\frac{7e^{2\hat{a}^\mathsf{T}(u/4+\theta)}}{(1+e^{2\hat{a}^\mathsf{T}(u/4+\theta)})^2} + \frac{e^{2\hat{a}^\mathsf{T}(u/4-\theta)}}{(1+e^{2\hat{a}^\mathsf{T}(u/4-\theta)})^2}\right)\right. \\
& - \left(\frac{7\mathbf{I}_d}{2(1+e^{2(\theta^\mathsf{T}u/4+|\theta|^2)})} - \frac{\mathbf{I}_d}{2(1+e^{2(\theta^\mathsf{T}u/4-|\theta|^2)})}\right) \\
& + \left(\frac{7e^{|u|^2/32}e^{2|\theta+u/8|^2}(u+8\theta)(u^\mathsf{T}+8\theta^\mathsf{T})}{32(e^{|u|^2/32}+e^{2|\theta+u/8|^2})^2}\right. \\
& \left.\left. + \frac{e^{|u|^2/32}e^{2|\theta-u/8|^2}(u-8\theta)(u^\mathsf{T}-8\theta^\mathsf{T})}{32(e^{|u|^2/32}+e^{2|\theta-u/8|^2})^2}\right)\right) \\
& + \frac{z_i z_i^\mathsf{T}}{8}\left(\frac{7e^{-z_i^\mathsf{T}(u/4+\theta)}}{(1+e^{-z_i^\mathsf{T}(u/4+\theta)})^2} + \frac{e^{-z_i^\mathsf{T}(u/4-\theta)}}{(1+e^{-z_i^\mathsf{T}(u/4-\theta)})^2}\right),
\end{aligned}$$

which implies Assumption 2 holds with $L_1 = 1$ and $\eta(u) = 9/2 + 8e^{|u|^2/32} + \sum_{i=1}^n |z_i|^2 + 4|\hat{a}|^2 + 3|u|^2/8$. On the other hand,

$$\begin{aligned}
\nabla_u H_i(\theta,u) = {} & \frac{1}{n}\left(\frac{1}{4}\mathbf{I}_d - \frac{\hat{a}\hat{a}^\mathsf{T}}{8}\left(\frac{7e^{2\hat{a}^\mathsf{T}(u/4+\theta)}}{(1+e^{2\hat{a}^\mathsf{T}(u/4+\theta)})^2} - \frac{e^{2\hat{a}^\mathsf{T}(u/4-\theta)}}{(1+e^{2\hat{a}^\mathsf{T}(u/4-\theta)})^2}\right)\right. \\
& - \left(\frac{7\mathbf{I}_d}{16(1+e^{2(\theta^\mathsf{T}u/4+|\theta|^2)})} + \frac{\mathbf{I}_d}{16(1+e^{2(\theta^\mathsf{T}u/4-|\theta|^2)})}\right) \\
& + \left(\frac{7e^{|u|^2/32}e^{2|\theta+u/8|^2}(u+8\theta)(-u^\mathsf{T}/16 + (u^\mathsf{T}/2 + 4\theta^\mathsf{T})/8)}{16(e^{|u|^2/32}+e^{2|\theta+u/8|^2})^2}\right.
\end{aligned}$$

$$+\frac{e^{|u|^2/32}e^{2|\theta-u/8|^2}(u-8\theta)(u^\mathsf{T}/16+(4\theta^\mathsf{T}-u^\mathsf{T}/2)/8)}{16(e^{|u|^2/32}+e^{2|\theta-u/8|^2})^2}\Bigg)\Bigg)$$

$$+\frac{z_i z_i^\mathsf{T}}{32}\left(\frac{7e^{-z_i^\mathsf{T}(u/4+\theta)}}{(1+e^{-z_i^\mathsf{T}(u/4+\theta)})^2}-\frac{e^{-z_i^\mathsf{T}(u/4-\theta)}}{(1+e^{-z_i^\mathsf{T}(u/4-\theta)})^2}\right),$$

which implies Assumption 2 holds with $L_2 = 1/4$ and $\eta(u) = 9/2 + 8e^{|u|^2/32} + \sum_{i=1}^n |z_i|^2 + 4|\hat{a}|^2 + 3|u|^2/8$. □

Proof of Proposition 3.2 First, we show that the objective function U defined in (12) is not necessarily convex. We consider the case where $N = 2$, and similar arguments can be applied for $N \geq 2$. By using (13), the Hessian matrix of U, denoted by $\nabla^2 U$, is given by:

$$\nabla^2 U(\theta) = \begin{pmatrix} 2\hat{\eta} + M(\theta) & -M(\theta) \\ -M(\theta) & 2\hat{\eta} + M(\theta) \end{pmatrix}, \quad \theta \in \mathbb{R}^2$$

where $M(\theta) := 2g_1^2(\theta)g_2^2(\theta)\mathbb{E}[(X_2 - X_1)^2] + 2g_1(\theta)g_2(\theta)(1-2g_1(\theta))\mathbb{E}[(Y - g_1(\theta)X_1 - g_2(\theta)X_2)(X_2 - X_1)]$. Then, for any $v = (v_1, v_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$, it follows that

$$\langle v, \nabla^2 U(\theta)v\rangle = 2\hat{\eta}|v|^2 + M(\theta)(v_1 - v_2)^2,$$

which is not necessarily nonnegative for all $\theta \in \mathbb{R}^2$. To see this, we consider the following example. Let

$$\begin{aligned} &\mathbb{E}[X_1] = 0.03, \quad \mathbb{E}[X_2] = 0.04, \quad \mathbb{E}[Y] = 0.033, \\ &\mathrm{Var}(X_1) = 5\times 10^{-5}, \quad \mathrm{Var}(X_2) = 2\times 10^{-4}, \quad \mathrm{Var}(Y) = 5.5\times 10^{-5}, \\ &\mathrm{Cov}(X_1, X_2) = 10^{-5}, \quad \mathrm{Cov}(X_1, Y) = 5\times 10^{-6}, \quad \mathrm{Cov}(X_2, Y) = -9\times 10^{-5}, \end{aligned}$$

which implies $\mathbb{E}[YX_2] = 1.23\times 10^{-3}$, $\mathbb{E}[X_1X_2] = 1.21\times 10^{-3}$, $\mathbb{E}[X_1Y] = 9.95\times 10^{-4}$, $\mathbb{E}[X_1^2] = 9.5\times 10^{-4}$. Then, set $\hat{\eta} = 10^{-6}$, $v = (1,0)$. For $\theta = (1, \ln 2)$, i.e., $g_1(\theta) = 1/3$, one obtains,

$$\begin{aligned} M(\theta) &= 2g_1(\theta)g_2^2(\theta)(3g_1(\theta)-1)\mathbb{E}[(X_2 - X_1)^2] \\ &\quad + 2g_1(\theta)g_2(\theta)(1-2g_1(\theta))\mathbb{E}[(Y - X_1)(X_2 - X_1)] \\ &= 2g_1(\theta)g_2(\theta)(1-2g_1(\theta))\mathbb{E}[(YX_2 - X_1X_2 - YX_1 + X_1^2)] = -\frac{1}{270000}, \end{aligned}$$

which indicates $\langle v, \nabla^2 U(\theta)v\rangle = -\frac{23}{13500000} < 0$. In addition, for $\theta = (1,1)$, i.e. $g_1(\theta) = 1/2$, one obtains

$$\begin{aligned} M(\theta) &= 2g_1(\theta)g_2^2(\theta)(3g_1(\theta)-1)\mathbb{E}[(X_2 - X_1)^2] \\ &\quad + 2g_1(\theta)g_2(\theta)(1-2g_1(\theta))\mathbb{E}[(Y - X_1)(X_2 - X_1)] \\ &= 2g_1(\theta)g_2^2(\theta)(3g_1(\theta)-1)\mathbb{E}[(X_2 - X_1)^2] \geq 0. \end{aligned}$$

which implies $\langle v, \nabla^2 U(\theta)v\rangle \geq 0$. Thus, one concludes that U is in general nonconvex.

Next, we prove that the stochastic gradient H given in (14) satisfies Assumptions 1, 2, and 3. Recall the explicit expressions for H_m, $m = 1, \ldots, N$ are given as follows:

$$H_m(\theta, z) = 2\hat{\eta}\theta_m + 2\left(y - \sum_{i=1}^N g_i(\theta)x_i\right) g_m(\theta) \sum_{i\neq m}^N g_i(\theta)(x_i - x_m).$$

It is easily checkable that Assumption 1 holds. To see Assumption 3 is satisfied, one calculates the following: for any $\theta \in \mathbb{R}^N, z \in \mathbb{R}^{N+1}$,

$$\langle \theta, H(\theta,z)\rangle = \sum_{m=1}^N \theta_m H_m(\theta,z) \geq \hat{\eta}|\theta|^2 - \hat{\eta}^{-1}N\left(\left(|y| + \sum_{i=1}^N |x_i|\right)\sum_{i\neq m}^N (|x_i|+|x_m|)\right)^2 .$$

Assumption 2 is also satisfied. Indeed, for any $m = 1,\dots,N$, $\theta,\theta' \in \mathbb{R}^N, z \in \mathbb{R}^{N+1}$, it follows that

$$\begin{aligned}|H_m(\theta,z) - H_m(\theta',z)| &\leq 2\hat{\eta}|\theta_m - \theta'_m| \\ &\quad + 2\left|\left(y - \sum_{i=1}^N g_i(\theta)x_i\right) g_m(\theta)\sum_{i\neq m}^N g_i(\theta)(x_i - x_m)\right. \\ &\quad \left. - \left(y - \sum_{i=1}^N g_i(\theta')x_i\right) g_m(\theta')\sum_{i\neq m}^N g_i(\theta')(x_i - x_m)\right| \\ &\leq 6\sqrt{N}\left(\hat{\eta} + \left(|y| + \sum_{i=1}^N |x_i|\right)\sum_{i\neq m}^N (|x_i|+|x_m|)\right)|\theta-\theta'|,\end{aligned}$$

where the last inequality holds due to the following: for any $m = 1,\dots,N$, $\theta,\theta' \in \mathbb{R}^N$.

$$|g_m(\theta) - g_m(\theta')| \leq \sqrt{N}|\theta-\theta'|.$$

Then, one obtains $|H(\theta,z) - H(\theta',z)| \leq 6N\eta(x)|\theta-\theta'|$, where $\eta(z) = \hat{\eta} + \left(1 + |y| + \sum_{i=1}^N |x_i|\right)\left(1 + \sum_{i\neq m}^N (|x_i|+|x_m|)\right)$. Similarly, for any $m = 1,\dots,N$, $\theta \in \mathbb{R}^N$, $z, z' \in \mathbb{R}^{N+1}$, one obtains

$$\begin{aligned}|H_m(\theta,z) - H_m(\theta,z')| &\leq 2\left|\left(y - \sum_{i=1}^N g_i(\theta)x_i\right) g_m(\theta)\sum_{i\neq m}^N g_i(\theta)(x_i - x_m)\right. \\ &\quad \left. - \left(y' - \sum_{i=1}^N g_i(\theta)x'_i\right) g_m(\theta)\sum_{i\neq m}^N g_i(\theta)(x'_i - x'_m)\right| \\ &\leq 4(N+1)\left(|y'| + \sum_{i=1}^N |x'_i| + \sum_{i\neq m}^N (|x_i|+|x_m|)\right)|z-z'| \\ &\leq 4(N+1)(\eta(z)+\eta(z'))|z-z'|,\end{aligned}$$

which further implies $|H(\theta,z) - H(\theta,z')| \leq 4\sqrt{N}(N+1)(\eta(z)+\eta(z'))|z-z'|$. □

Appendix C Proofs of the results in Section 2 and 4

Proof of Remark 2.1 To prove (5), one notices that by using Assumptions 1 and 2

$$|h(\theta) - h(\theta')| \leq \mathbb{E}[|H(\theta,X_0) - H(\theta',X_0)|] \leq L_1\mathbb{E}[\eta(X_0)]|\theta-\theta'|.$$

Then, to prove (6), one calculates by using Assumption 2

$$|H(\theta,x)| \leq |H(\theta,x) - H(0,x)| + |H(0,x) - H(0,0)| + |H(0,0)|$$

$$\begin{aligned}&\leq L_1\eta(x)|\theta| + L_2(\eta(x)+\eta(0))|x| + |H(0,0)|\\ &\leq L_1\eta(x)|\theta| + L_2\bar{\eta}(x) + H_\star,\end{aligned}$$

where $\bar{\eta}(x) := (\eta(x)+\eta(0))|x|$, and $H_\star := |H(0,0)|$. □

Proof of Lemma 4.2 For any $n \in \mathbb{N}$ and $t \in (n, n+1]$, define $\Delta_{n,t} := \bar{\theta}^\lambda_n - \lambda H(\bar{\theta}^\lambda_n, X_{n+1})(t-n)$. By using (18), it is easily seen that for $t \in (n, n+1]$

$$\mathbb{E}\left[|\bar{\theta}^\lambda_t|^2 \,\middle|\, \bar{\theta}^\lambda_n\right] = \mathbb{E}\left[|\Delta_{n,t}|^2 \,\middle|\, \bar{\theta}^\lambda_n\right] + (2\lambda/\beta)d(t-n).$$

Then, by using Assumptions 1, 2, 3 and Remark 2.1, one obtains

$$\begin{aligned}\mathbb{E}\left[|\Delta_{n,t}|^2 \,\middle|\, \bar{\theta}^\lambda_n\right] &= |\bar{\theta}^\lambda_n|^2 - 2\lambda(t-n)\mathbb{E}\left[\left\langle \bar{\theta}^\lambda_n, H(\bar{\theta}^\lambda_n, X_{n+1})\right\rangle \,\middle|\, \bar{\theta}^\lambda_n\right]\\ &\quad + \lambda^2(t-n)^2\mathbb{E}\left[|H(\bar{\theta}^\lambda_n, X_{n+1})|^2 \,\middle|\, \bar{\theta}^\lambda_n\right]\\ &\leq |\bar{\theta}^\lambda_n|^2 - 2\lambda(t-n)\left\langle \bar{\theta}^\lambda_n, \mathbb{E}\left[A(X_0)\right]\bar{\theta}^\lambda_n\right\rangle + 2\lambda(t-n)b\\ &\quad + \lambda^2(t-n)^2\mathbb{E}\left[|L_1\eta(X_{n+1})|\bar{\theta}^\lambda_n| + L_2\bar{\eta}(X_{n+1}) + H_\star|^2 \,\middle|\, \bar{\theta}^\lambda_n\right]\\ &\leq (1-2a\lambda(t-n))|\bar{\theta}^\lambda_n|^2 + 2\lambda^2(t-n)^2L_1^2\mathbb{E}\left[\eta^2(X_0)\right]|\bar{\theta}^\lambda_n|^2\\ &\quad + 4\lambda^2(t-n)^2L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + 4\lambda^2(t-n)^2H_\star^2 + 2\lambda(t-n)b,\end{aligned}$$

where the last inequality is obtained by using $(a+b)^2 \leq 2a^2+2b^2$, for $a,b \geq 0$ twice. For $\lambda < \lambda_{\max} < a/(2L_1^2\mathbb{E}\left[\eta^2(X_0)\right])$,

$$\mathbb{E}\left[|\Delta_{n,t}|^2 \,\middle|\, \bar{\theta}^\lambda_n\right] \leq (1-a\lambda(t-n))|\bar{\theta}^\lambda_n|^2 + \lambda(t-n)c_0,$$

where $c_0 := 4\lambda_{\max}L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + 4\lambda_{\max}H_\star^2 + 2b$. Therefore, one obtains

$$\mathbb{E}\left[|\bar{\theta}^\lambda_t|^2 \,\middle|\, \bar{\theta}^\lambda_n\right] \leq (1-a\lambda(t-n))|\bar{\theta}^\lambda_n|^2 + \lambda(t-n)c_1,$$

where $c_1 := c_0 + 2d/\beta$ and the result follows by induction. To calculate a higher moment, denote by $U^\lambda_{n,t} := \{2\lambda\beta^{-1}\}^{1/2}(\tilde{B}^\lambda_t - \tilde{B}^\lambda_n)$, for $t \in (n, n+1]$, one calculates

$$\begin{aligned}\mathbb{E}\left[|\bar{\theta}^\lambda_t|^4 \,\middle|\, \bar{\theta}^\lambda_n\right] &= \mathbb{E}\left[\left(|\Delta_{n,t}|^2 + |U^\lambda_{n,t}|^2 + 2\left\langle \Delta_{n,t}, U^\lambda_{n,t}\right\rangle\right)^2 \,\middle|\, \bar{\theta}^\lambda_n\right]\\ &= \mathbb{E}\left[|\Delta_{n,t}|^4 + |U^\lambda_{n,t}|^4 + 2|\Delta_{n,t}|^2|U^\lambda_{n,t}|^2 + 4|\Delta_{n,t}|^2\left\langle \Delta_{n,t}, U^\lambda_{n,t}\right\rangle\right.\\ &\quad \left. +4|U^\lambda_{n,t}|^2\left\langle \Delta_{n,t}, U^\lambda_{n,t}\right\rangle + 4\left(\left\langle \Delta_{n,t}, U^\lambda_{n,t}\right\rangle\right)^2 \,\middle|\, \bar{\theta}^\lambda_n\right]\\ &\leq \mathbb{E}\left[|\Delta_{n,t}|^4 + |U^\lambda_{n,t}|^4 + 6|\Delta_{n,t}|^2|U^\lambda_{n,t}|^2 \,\middle|\, \bar{\theta}^\lambda_n\right]\\ &\leq (1+a\lambda(t-n))\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}^\lambda_n\right] + (1+9/(a\lambda(t-n)))\mathbb{E}\left[|U^\lambda_{n,t}|^4\right].\end{aligned} \tag{C3}$$

where the last inequality holds due to $2uv \leq \dot{\epsilon}u^2 + \dot{\epsilon}^{-1}v^2$, for $u, v \geq 0$ and $\dot{\epsilon} > 0$ with $u = |\Delta_{n,t}|^2, v = 3|U^\lambda_{n,t}|^2, \dot{\epsilon} = a\lambda(t-n)$, and due to the independence of $U^\lambda_{n,t}$ and $\bar{\theta}^\lambda_n$. Then, by using the Cauchy-Schwarz inequality, one obtains

$$\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}^\lambda_n\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(|\bar{\theta}_n^\lambda|^2 - 2\lambda(t-n)\left\langle \bar{\theta}_n^\lambda, H(\bar{\theta}_n^\lambda, X_{n+1})\right\rangle + \lambda^2(t-n)^2|H(\bar{\theta}_n^\lambda, X_{n+1})|^2\right)^2 \middle| \bar{\theta}_n^\lambda\right] \\
&= \mathbb{E}\left[|\bar{\theta}_n^\lambda|^4 + 4\lambda^2(t-n)^2\left(\left\langle \bar{\theta}_n^\lambda, H(\bar{\theta}_n^\lambda, X_{n+1})\right\rangle\right)^2\right. \\
&\quad + \lambda^4(t-n)^4|H(\bar{\theta}_n^\lambda, X_{n+1})|^4 - 4\lambda(t-n)\left\langle \bar{\theta}_n^\lambda, H(\bar{\theta}_n^\lambda, X_{n+1})\right\rangle|\bar{\theta}_n^\lambda|^2 \\
&\quad + 2\lambda^2(t-n)^2|\bar{\theta}_n^\lambda|^2|H(\bar{\theta}_n^\lambda, X_{n+1})|^2 \\
&\quad \left. -4\lambda^3(t-n)^3\left\langle \bar{\theta}_n^\lambda, H(\bar{\theta}_n^\lambda, X_{n+1})\right\rangle|H(\bar{\theta}_n^\lambda, X_{n+1})|^2 \middle| \bar{\theta}_n^\lambda\right] \\
&\leq |\bar{\theta}_n^\lambda|^4 + \mathbb{E}\left[6\lambda^2(t-n)^2|\bar{\theta}_n^\lambda|^2|H(\bar{\theta}_n^\lambda, X_{n+1})|^2 - 4\lambda(t-n)\left\langle \bar{\theta}_n^\lambda, H(\bar{\theta}_n^\lambda, X_{n+1})\right\rangle|\bar{\theta}_n^\lambda|^2\right. \\
&\quad \left. -4\lambda^3(t-n)^3|H(\bar{\theta}_n^\lambda, X_{n+1})|^2\left\langle \bar{\theta}_n^\lambda, H(\bar{\theta}_n^\lambda, X_{n+1})\right\rangle + \lambda^4(t-n)^4|H(\bar{\theta}_n^\lambda, X_{n+1})|^4 \middle| \bar{\theta}_n^\lambda\right].
\end{aligned}
$$

By Remark 2.1, the independence of X_{n+1} and $\bar{\theta}_n^\lambda$, and by using $(u+v+\nu)^s \leq 2^{s-1}(u+v)^s + 2^{s-1}\nu^s \leq 2^{2s-2}(u^s+v^s) + 2^{s-1}\nu^s$ for $u, v, \nu \geq 0, s \geq 1$, it follows that, for $q \geq 1$,

$$
\mathbb{E}\left[|H(\bar{\theta}_n^\lambda, X_{n+1})|^q \middle| \bar{\theta}_n^\lambda\right] \leq 2^{q-1}L_1^q\mathbb{E}\left[\eta^q(X_0)\right]|\bar{\theta}_n^\lambda|^q + 2^{2q-2}L_2^q\mathbb{E}\left[|\bar{\eta}^q(X_0)\right] + 2^{2q-2}H_\star^q. \tag{C4}
$$

By using Assumption 3 and by taking $q = 2, 3, 4$ in (C4), one obtains

$$
\begin{aligned}
\mathbb{E}\left[|\Delta_{n,t}|^4 \middle| \bar{\theta}_n^\lambda\right] \leq\ & (1 - 4a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + 4b\lambda(t-n)|\bar{\theta}_n^\lambda|^2 \\
& + 12\lambda^2(t-n)^2L_1^2\mathbb{E}\left[\eta^2(X_0)\right]|\bar{\theta}_n^\lambda|^4 + 16\lambda^3(t-n)^3L_1^3\mathbb{E}\left[\eta^3(X_0)\right]|\bar{\theta}_n^\lambda|^4 \\
& + 8\lambda^4(t-n)^4L_1^4\mathbb{E}\left[\eta^4(X_0)\right]|\bar{\theta}_n^\lambda|^4 \\
& + 24\lambda^2(t-n)^2\left(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2\right)|\bar{\theta}_n^\lambda|^2 \\
& + 64\lambda^3(t-n)^3\left(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3\right)|\bar{\theta}_n^\lambda| \\
& + 64\lambda^4(t-n)^4\left(L_2^4\mathbb{E}\left[\bar{\eta}^4(X_0)\right] + H_\star^4\right)
\end{aligned}
$$

which implies, by using $\lambda < \lambda_{\max}$

$$
\begin{aligned}
\mathbb{E}\left[|\Delta_{n,t}|^4 \middle| \bar{\theta}_n^\lambda\right] \leq\ & (1 - 3a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + 4b\lambda(t-n)|\bar{\theta}_n^\lambda|^2 \\
& + 24\lambda^2(t-n)^2\left(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2\right)|\bar{\theta}_n^\lambda|^2 \\
& + 64\lambda^3(t-n)^3\left(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3\right)|\bar{\theta}_n^\lambda| \\
& + 64\lambda^4(t-n)^4\left(L_2^4\mathbb{E}\left[\bar{\eta}^4(X_0)\right] + H_\star^4\right).
\end{aligned} \tag{C5}
$$

For $|\bar{\theta}_n^\lambda| > (8ba^{-1} + 48a^{-1}\lambda_{\max}(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2))^{1/2}$, we have

$$
-\frac{a\lambda(t-n)}{2}|\bar{\theta}_n^\lambda|^4 + 4b\lambda(t-n)|\bar{\theta}_n^\lambda|^2 + 24\lambda^2(t-n)^2\left(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2\right)|\bar{\theta}_n^\lambda|^2 < 0. \tag{C6}
$$

Similarly, for $|\bar{\theta}_n^\lambda| > (128a^{-1}\lambda_{\max}^2(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3))^{1/3}$

$$
-\frac{a\lambda(t-n)}{2}|\bar{\theta}_n^\lambda|^4 + 64\lambda^3(t-n)^3\left(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3\right)|\bar{\theta}_n^\lambda| < 0. \tag{C7}
$$

Denote by

$$\begin{aligned} M := \max\{&(8ba^{-1} + 48a^{-1}\lambda_{\max}(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2))^{1/2}, \\ &(128a^{-1}\lambda_{\max}^2(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3))^{1/3}\}. \end{aligned} \tag{C8}$$

Moreover, denote by $\mathsf{A}_{n,M} := \{\omega \in \Omega : |\bar{\theta}_n^\lambda(\omega)| > M\}$. Then, by substituting (C6), (C7) into (C5), one obtains,

$$\begin{aligned} \mathbb{E}\left[|\Delta_{n,t}|^4 \mathbb{1}_{\mathsf{A}_{n,M}} \middle| \bar{\theta}_n^\lambda\right] \leq{}& (1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 \mathbb{1}_{\mathsf{A}_{n,M}} \\ &+ 64\lambda^4(t-n)^4\left(L_2^4\mathbb{E}\left[\bar{\eta}^4(X_0)\right] + H_\star^4\right)\mathbb{1}_{\mathsf{A}_{n,M}}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \mathbb{E}\left[|\Delta_{n,t}|^4 \mathbb{1}_{\mathsf{A}^{\mathsf{c}}_{n,M}} \middle| \bar{\theta}_n^\lambda\right] \leq{}& (1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 \mathbb{1}_{\mathsf{A}^{\mathsf{c}}_{n,M}} + 4b\lambda(t-n)M^2\mathbb{1}_{\mathsf{A}^{\mathsf{c}}_{n,M}} \\ &+ 24\lambda^2(t-n)^2\left(L_2^2\mathbb{E}\left[\bar{\eta}^2(X_0)\right] + H_\star^2\right)M^2\mathbb{1}_{\mathsf{A}^{\mathsf{c}}_{n,M}} \\ &+ 64\lambda^3(t-n)^3\left(L_2^3\mathbb{E}\left[\bar{\eta}^3(X_0)\right] + H_\star^3\right)M\mathbb{1}_{\mathsf{A}^{\mathsf{c}}_{n,M}} \\ &+ 64\lambda^4(t-n)^4\left(L_2^4\mathbb{E}\left[\bar{\eta}^4(X_0)\right] + H_\star^4\right)\mathbb{1}_{\mathsf{A}^{\mathsf{c}}_{n,M}}. \end{aligned}$$

Combining the two cases yields

$$\mathbb{E}\left[|\Delta_{n,t}|^4 \middle| \bar{\theta}_n^\lambda\right] \leq (1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + \lambda(t-n)c_2, \tag{C9}$$

where $c_2 := 4bM^2 + 152(1+\lambda_{\max})^3\left((1+L_2)^4\mathbb{E}\left[(1+\bar{\eta}(X_0))^4\right] + (1+H_\star)^4\right)(1+M)^2$with M given in (C8). Substituting (C9) into (C3), one obtains

$$\begin{aligned} \mathbb{E}\left[|\bar{\theta}_t^\lambda|^4 \middle| \bar{\theta}_n^\lambda\right] \leq{}& (1 + a\lambda(t-n))(1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 \\ &+ (1 + a\lambda(t-n))\lambda(t-n)c_2 + 12d^2\lambda^2\beta^{-2}(t-n)^2(1 + 9/(a\lambda(t-n))) \\ \leq{}& (1 - a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + \lambda(t-n)c_3, \end{aligned}$$

where $c_3 := (1 + a\lambda_{\max})c_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9a^{-1})$. Finally, for any $n \in \mathbb{N}, t \in (n, n+1]$, $0 < \lambda \leq \lambda_{\max}$, one obtains,

$$\begin{aligned} \mathbb{E}\left[|\bar{\theta}_t^\lambda|^4\right] &\leq (1 - a\lambda(t-n))\mathbb{E}\left[|\bar{\theta}_n^\lambda|^4\right] + \lambda(t-n)c_3 \\ &\leq (1 - a\lambda(t-n))(1 - a\lambda)\mathbb{E}\left[|\bar{\theta}_{n-1}^\lambda|^4\right] + \lambda_{\max}c_3 + \lambda c_3 \\ &\leq (1 - a\lambda(t-n))(1 - a\lambda)^2\mathbb{E}\left[|\bar{\theta}_{n-2}^\lambda|^4\right] \\ &\quad + \lambda_{\max}c_3 + \lambda c_3(1 + (1 - a\lambda)) \\ &\leq \ldots \\ &\leq (1 - a\lambda(t-n))(1 - a\lambda)^n\mathbb{E}\left[|\theta_0|^4\right] + c_3(\lambda_{\max} + 1/a), \end{aligned}$$

which completes the proof. □

Proof of Lemma 4.5 For any $p \geq 1$, application of Ito's lemma and taking expectation yields

$$\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})] = \mathbb{E}[V_p(\bar{\theta}_{nT}^\lambda)] + \int_{nT}^t \mathbb{E}\left[\lambda\frac{\Delta V_p(\bar{\zeta}_s^{\lambda,n})}{\beta} - \lambda\langle h(\bar{\zeta}_s^{\lambda,n}), \nabla V_p(\bar{\zeta}_s^{\lambda,n})\rangle\right] \mathrm{d}s.$$

Differentiating both sides and using Lemma 4.4, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})] = \mathbb{E}\left[\lambda\frac{\Delta V_p(\bar{\zeta}_t^{\lambda,n})}{\beta} - \lambda\langle h(\bar{\zeta}_t^{\lambda,n}), \nabla V_p(\bar{\zeta}_t^{\lambda,n})\rangle\right] \leq -\lambda\bar{c}(p)\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})] + \lambda\tilde{c}(p),$$

which yields

$$\begin{aligned}\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})] &\leq e^{-\lambda(t-nT)\bar{c}(p)}\mathbb{E}[V_p(\bar{\theta}_{nT}^{\lambda})] + \tilde{c}(p)/\bar{c}(p)\left(1 - e^{-\lambda\bar{c}(p)(t-nT)}\right)\\ &\leq e^{-\lambda(t-nT)\bar{c}(p)}\mathbb{E}[V_p(\bar{\theta}_{nT}^{\lambda})] + \tilde{c}(p)/\bar{c}(p).\end{aligned}$$

Now for $p=2$, by using Lemma 4.2, Corollary 4.3 and Lemma 4.4, we obtain

$$\begin{aligned}\mathbb{E}[V_2(\bar{\zeta}_t^{\lambda,n})] &\leq e^{-\lambda(t-nT)\bar{c}(2)}\mathbb{E}[V_2(\bar{\theta}_{nT}^{\lambda})] + \tilde{c}(2)/\bar{c}(2)\\ &\leq (1-a\lambda)^{nT}e^{-\lambda(t-nT)\bar{c}(2)}\mathbb{E}[V_2(\theta_0)] + \tilde{c}(2)/\bar{c}(2) + c_1(\lambda_{\max} + a^{-1}) + 1\\ &\leq e^{-a\lambda t/2}\mathbb{E}[V_2(\theta_0)] + 3\mathrm{v}_2(\overline{M}_2) + c_1(\lambda_{\max} + a^{-1}) + 1,\end{aligned}$$

where the last inequality holds due to $0 \leq 1-z \leq e^{-z}$ for $z \geq 0$ and $\bar{c}(2) = a/2$. Similarly, for $p=4$, one obtains

$$\begin{aligned}\mathbb{E}[V_4(\bar{\zeta}_t^{\lambda,n})] &\leq e^{-\lambda(t-nT)\bar{c}(4)}\mathbb{E}[V_4(\bar{\theta}_{nT}^{\lambda})] + \tilde{c}(4)/\bar{c}(4)\\ &\leq 2(1-a\lambda)^{nT}e^{-\lambda(t-nT)\bar{c}(4)}\mathbb{E}[V_4(\theta_0)] + \tilde{c}(4)/\bar{c}(4) + 2c_3(\lambda_{\max} + a^{-1}) + 2\\ &\leq 2e^{-a\lambda t}\mathbb{E}[V_4(\theta_0)] + 3\mathrm{v}_4(\overline{M}_4) + 2c_3(\lambda_{\max} + a^{-1}) + 2,\end{aligned}$$

where the last inequality holds due to $0 \leq 1-z \leq e^{-z}$ for $z \geq 0$ and $\bar{c}(4) = a$. □

Proof of Lemma 4.7 To handle the first term in (22), we start by establishing an upper bound in Wasserstein-2 distance and the statment follows by noticing $W_1 \leq W_2$. By employing synchronous coupling, using (18) and the definition of $\bar{\zeta}_t^{\lambda,n}$ in Definition 4.1, one obtains, for any $t \in (nT, (n+1)T]$,

$$\begin{aligned}\left|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^{\lambda}\right| &\leq \lambda\left|\int_{nT}^{t}\left[H(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}, X_{\lceil s\rceil}) - h(\bar{\zeta}_s^{\lambda,n})\right]\mathrm{d}s\right|\\ &\leq \lambda\left|\int_{nT}^{t}\left[H(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}, X_{\lceil s\rceil}) - H(\bar{\zeta}_s^{\lambda,n}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|\\ &\quad + \lambda\left|\int_{nT}^{t}\left[h(\bar{\zeta}_s^{\lambda,n}) - H(\bar{\zeta}_s^{\lambda,n}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|\\ &\leq \lambda L_1\int_{nT}^{t}\eta(X_{\lceil s\rceil})\left|\bar{\theta}_{\lfloor s\rfloor}^{\lambda} - \bar{\zeta}_s^{\lambda,n}\right|\mathrm{d}s + \lambda\left|\int_{nT}^{t}\left[h(\bar{\zeta}_s^{\lambda,n}) - H(\bar{\zeta}_s^{\lambda,n}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|,\end{aligned}$$

where the last inequality holds due to Assumption 2. Now taking squares of both sides, using $(a+b)^2 \leq 2a^2 + 2b^2$ for $a,b>0$, and then taking expectations lead to

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^{\lambda}\right|^2\right] &\leq 2\lambda L_1^2\int_{nT}^{t}\mathbb{E}\left[\eta^2(X_0)\right]\mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\lambda} - \bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s\\ &\quad + 2\lambda^2\mathbb{E}\left[\left|\int_{nT}^{t}\left[h(\bar{\zeta}_s^{\lambda,n}) - H(\bar{\zeta}_s^{\lambda,n}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right].\end{aligned}$$

where the expectation splits over terms in the first integral due to the independence of $X_{\lceil s\rceil}$ from the rest of the random variables. Using $\lambda T \leq 1$, Lemma A.2 and $(a+b)^2 \leq 2a^2 + 2b^2$ once again, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_t^{\lambda}\right|^2\right] \leq{}& 4\lambda L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\int_{nT}^t \mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\lambda}-\bar{\theta}_s^{\lambda}\right|^2\right]\mathrm{d}s \\
&+4\lambda L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\int_{nT}^t \mathbb{E}\left[\left|\bar{\theta}_s^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s \\
&+2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right] \\
\leq{}& 4\lambda L_1^2\mathbb{E}\left[\eta^2(X_0)\right](e^{-a\lambda nT}\bar{\sigma}_Y\mathbb{E}[V_2(\theta_0)]+\tilde{\sigma}_Y) \\
&+4\lambda L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\int_{nT}^t \mathbb{E}\left[\left|\bar{\theta}_s^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s \\
&+2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right]. \qquad \text{(C10)}
\end{aligned}
$$

where $\bar{\sigma}_Y$ and $\tilde{\sigma}_Y$ are provided in (A2). Next, we bound the last term in (C10) by partitioning the last integral. Assume that $nT+K<t\leq nT+K+1$ where $K+1\leq T, K\in\mathbb{N}$. Thus we can write

$$
\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right| = \left|\sum_{k=1}^K I_k + R_K\right|
$$

where $I_k := \int_{nT+(k-1)}^{nT+k}[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{nT+k})]\mathrm{d}s$, and $R_K := \int_{nT+K}^t[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{nT+K+1})]\mathrm{d}s$. Taking squares of both sides

$$
\left|\sum_{k=1}^K I_k+R_K\right|^2 = \sum_{k=1}^K|I_k|^2+2\sum_{k=2}^K\sum_{j=1}^{k-1}\langle I_k,I_j\rangle+2\sum_{k=1}^K\langle I_k,R_K\rangle+|R_K|^2,
$$

Finally, we take expectations of both sides. Define the filtration $\mathcal{H}_t=\mathcal{F}_\infty^\lambda\vee\mathcal{G}_{\lfloor t\rfloor}$. We first note that for any $k=2,\ldots,K$, $j=1,\ldots,k-1$,

$$
\begin{aligned}
\mathbb{E}\left[\langle I_k,I_j\rangle\right] &= \mathbb{E}\left[\mathbb{E}[\langle I_k,I_j\rangle|\mathcal{H}_{nT+j}]\right], \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle\int_{nT+(k-1)}^{nT+k}[H(\bar{\zeta}_s^{\lambda,n},X_{nT+k})-h(\bar{\zeta}_s^{\lambda,n})]\mathrm{d}s,\right.\right.\right. \\
&\qquad\left.\left.\left.\left.\int_{nT+(j-1)}^{nT+j}[H(\bar{\zeta}_s^{\lambda,n},X_{nT+j})-h(\bar{\zeta}_s^{\lambda,n})]\mathrm{d}s\right\rangle\right|\mathcal{H}_{nT+j}\right]\right], \\
&= \mathbb{E}\left[\left\langle\int_{nT+(k-1)}^{nT+k}\mathbb{E}\left[H(\bar{\zeta}_s^{\lambda,n},X_{nT+k})-h(\bar{\zeta}_s^{\lambda,n})\middle|\mathcal{H}_{nT+j}\right]\mathrm{d}s,\right.\right. \\
&\qquad\left.\left.\int_{nT+(j-1)}^{nT+j}[H(\bar{\zeta}_s^{\lambda,n},X_{nT+j})-h(\bar{\zeta}_s^{\lambda,n})]\mathrm{d}s\right\rangle\right]=0.
\end{aligned}
$$

By the same argument $\mathbb{E}\langle I_k,R_K\rangle=0$ for all $1\leq k\leq K$. Therefore, the last term of (C10) is bounded as

$$
2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right] = 2\lambda^2\sum_{k=1}^K\mathbb{E}\left[|I_k|^2\right]+2\lambda^2\mathbb{E}\left[|R_K|^2\right]
$$

$$\leq 4e^{-a\lambda nT/2}\lambda(\bar{\sigma}_Z\mathbb{E}[V_2(\theta_0)]+\tilde{\sigma}_Z),$$

where the last inequality holds due to Lemma A.1 and $\bar{\sigma}_Z$ and $\tilde{\sigma}_Z$ are provided in (A1). Therefore, the bound (C10) becomes

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_t^{\lambda}\right|^2\right] \leq\ & 4\lambda L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\int_{nT}^{t}\mathbb{E}\left[\left|\bar{\theta}_s^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s\\ &+4e^{-a\lambda nT/2}\lambda(L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\bar{\sigma}_Y+\bar{\sigma}_Z)\mathbb{E}[V_2(\theta_0)]\\ &+4\lambda(L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\tilde{\sigma}_Y+\tilde{\sigma}_Z).\end{aligned}$$

Using Grönwall's inequality leads

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_t^{\lambda}\right|^2\right] \leq \lambda e^{4L_1^2\mathbb{E}[\eta^2(X_0)]}\Big[&4e^{-a\lambda nT/2}(L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\bar{\sigma}_Y+\bar{\sigma}_Z)\mathbb{E}[V_2(\theta_0)]\\ &+4(L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\tilde{\sigma}_Y+\tilde{\sigma}_Z)\Big].\end{aligned}$$

which implies by $\lambda T \geq 1/2$,

$$W_2^2(\mathcal{L}(\bar{\theta}_t^{\lambda}),\mathcal{L}(\bar{\zeta}_t^{\lambda,n})) \leq \mathbb{E}\left|\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_t^{\lambda}\right|^2 \leq \lambda(e^{-an/4}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)]+\bar{C}_{2,2}), \tag{C11}$$

where

$$\begin{aligned}\bar{C}_{2,1} &:= 4e^{4L_1^2\mathbb{E}[\eta^2(X_0)]}(L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\bar{\sigma}_Y+\bar{\sigma}_Z),\\ \bar{C}_{2,2} &:= 4e^{4L_1^2\mathbb{E}[\eta^2(X_0)]}(L_1^2\mathbb{E}\left[\eta^2(X_0)\right]\tilde{\sigma}_Y+\tilde{\sigma}_Z)\end{aligned} \tag{C12}$$

with $\bar{\sigma}_Y$, $\tilde{\sigma}_Y$ provided in (A2) and $\bar{\sigma}_Z$, $\tilde{\sigma}_Z$ given in (A1). □

Proof of Lemma 4.8 To upper bound the second term $W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}),\mathcal{L}(Z_t^{\lambda}))$ in (22), we adapt the proof from Lemma 3.18 in [14]. Recall the definition of $w_{1,2}$ given in (21), and the fact that $W_1(\mu,\nu)\leq w_{1,2}(\mu,\nu)$ for any $\mu,\nu\in\mathcal{P}_{V_2}$. By Proposition 4.6, one obtains, for any $t\in(nT,(n+1)T]$,

$$\begin{aligned}W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}),\mathcal{L}(Z_t^{\lambda})) &\leq \sum_{k=1}^{n}W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,k}),\mathcal{L}(\bar{\zeta}_t^{\lambda,k-1}))\\ &\leq \sum_{k=1}^{n}w_{1,2}(\mathcal{L}(\zeta_t^{kT,\bar{\theta}_{kT}^{\lambda},\lambda}),\mathcal{L}(\zeta_t^{kT,\bar{\zeta}_{kT}^{\lambda,k-1},\lambda}))\\ &\leq \hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))w_{1,2}(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}),\mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})),\end{aligned}$$

which implies, by using Cauchy-Schwarz inequality, Young's inequality, Lemma 4.7, Corollary 4.3 and Lemma 4.5,

$$\begin{aligned}W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}),\mathcal{L}(Z_t^{\lambda})) \leq \hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))W_2(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}),\mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1}))\Big[&1+\left\{\mathbb{E}[V_4(\bar{\theta}_{kT}^{\lambda})]\right\}^{1/2}\\ &+\left\{\mathbb{E}[V_4(\bar{\zeta}_{kT}^{\lambda,k-1})]\right\}^{1/2}\Big]\\ \leq (\sqrt{\lambda})^{-1}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))W_2^2(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}),\mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1}))&\end{aligned}$$

$$
\begin{aligned}
&\quad + 3\sqrt{\lambda}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))\left[1+\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{kT})]+\mathbb{E}[V_4(\bar{\zeta}^{\lambda,k-1}_{kT})]\right]\\
&\le \sqrt{\lambda}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))(e^{-a(k-1)/4}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)]+\bar{C}_{2,2})\\
&\quad + 3\sqrt{\lambda}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))\left[1+\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{kT})]+\mathbb{E}[V_4(\bar{\zeta}^{\lambda,k-1}_{kT})]\right]\\
&\le \sqrt{\lambda}e^{-\min\{\dot{c},a/4\}n}n\hat{c}(e^{\min\{\dot{c},a/4\}}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)]+12\mathbb{E}[V_4(\theta_0)])\\
&\quad + \sqrt{\lambda}\frac{\hat{c}}{1-\exp(-\dot{c})}(\bar{C}_{2,2}+12c_3(\lambda_{\max}+a^{-1})+9\mathrm{v}_4(\overline{M}_4)+15)\\
&\le \sqrt{\lambda}(e^{-\min\{\dot{c},a/4\}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)]+\bar{C}_{2,4})\\
&= \sqrt{\lambda}(e^{-\dot{c}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)]+\bar{C}_{2,4}),
\end{aligned}
$$

where the last inequality holds by applying the inequality $e^{-\alpha n}(n+1)\le 1+\alpha^{-1}$, for $\alpha>0$ with $\alpha=\min\{\dot{c},a/4\}/2$, and the last equality holds by noticing $\min\{\dot{c},a/4\}=\dot{c}$ with $\dot{c}$ given in (23). The explicit expressions for the constants $\bar{C}_{2,3},\bar{C}_{2,4}$ are given below:

$$
\begin{aligned}
\bar{C}_{2,3} &:= \hat{c}\left(1+\frac{2}{\dot{c}}\right)(e^{a/4}\bar{C}_{2,1}+12)\\
\bar{C}_{2,4} &:= \frac{\hat{c}}{1-\exp(-\dot{c})}(\bar{C}_{2,2}+12c_3(\lambda_{\max}+a^{-1})+9\mathrm{v}_4(\overline{M}_4)+15)
\end{aligned}
\tag{C13}
$$

with $\bar{C}_{2,1}$, $\bar{C}_{2,2}$ given in (C12), $\hat{c}$, $\dot{c}$ given in Lemma 4.11, c_3 given in (20), and $\overline{M}_4$ given in Lemma 4.4. □

Proof of Corollary 4.9 One notices that $W_2\le\sqrt{2w_{1,2}}$, then, by using similar arguments as in the proof of Lemma 4.8, one obtains

$$
\begin{aligned}
&W_2(\mathcal{L}(\bar{\zeta}^{\lambda,n}_t),\mathcal{L}(Z^{\lambda}_t))\\
&\le \sum_{k=1}^{n}W_2(\mathcal{L}(\bar{\zeta}^{\lambda,k}_t),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_t))\\
&\le \sum_{k=1}^{n}\sqrt{2}w_{1,2}^{1/2}(\mathcal{L}(\zeta_t^{kT,\bar{\theta}^{\lambda}_{kT},\lambda}),\mathcal{L}(\zeta_t^{kT,\bar{\zeta}^{\lambda,k-1}_{kT},\lambda}))\\
&\le \sqrt{2\hat{c}}\sum_{k=1}^{n}\exp(-\dot{c}(n-k)/2)W_2^{1/2}(\mathcal{L}(\bar{\theta}^{\lambda}_{kT}),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_{kT}))\left[1+\left\{\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{kT})]\right\}^{1/2}\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\left.+\left\{\mathbb{E}[V_4(\bar{\zeta}^{\lambda,k-1}_{kT})]\right\}^{1/2}\right]^{1/2}\\
&\le \lambda^{-1/4}\sqrt{2\hat{c}}\sum_{k=1}^{n}\exp(-\dot{c}(n-k)/2)W_2(\mathcal{L}(\bar{\theta}^{\lambda}_{kT}),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_{kT}))\\
&\quad + \lambda^{1/4}\sqrt{2\hat{c}}\sum_{k=1}^{n}\exp(-\dot{c}(n-k)/2)\left[1+\left\{\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{kT})]\right\}^{1/2}+\left\{\mathbb{E}[V_4(\bar{\zeta}^{\lambda,k-1}_{kT})]\right\}^{1/2}\right]\\
&\le \sqrt{2\hat{c}}\lambda^{1/4}e^{-\min\{\dot{c},a/4\}n/2}n(e^{\min\{\dot{c},a/4\}/2}\bar{C}_{2,1}^{1/2}\mathbb{E}^{1/2}[V_2(\theta_0)]+2\sqrt{2}\mathbb{E}^{1/2}[V_4(\theta_0)])
\end{aligned}
$$

$$
\begin{aligned}
&+\sqrt{2\hat{c}}\lambda^{1/4}\frac{1}{1-\exp(-\dot{c}/2)}(\bar{C}_{2,2}^{1/2}+2\sqrt{2c_3}(\lambda_{\max}+a^{-1})^{1/2}+\sqrt{3}\mathrm{v}_4^{1/2}(\overline{M}_4)+\sqrt{15})\\
&\le \lambda^{1/4}(e^{-\min\{\dot{c},a/4\}n/4}\bar{C}^*_{2,3}\mathbb{E}^{1/2}[V_4(\theta_0)]+\bar{C}^*_{2,4})\\
&= \lambda^{1/4}(e^{-\dot{c}n/4}\bar{C}^*_{2,3}\mathbb{E}^{1/2}[V_4(\theta_0)]+\bar{C}^*_{2,4}),
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{C}^*_{2,3} &:= \sqrt{2\hat{c}}\,(1+4/\dot{c})\,(e^{a/8}\bar{C}_{2,1}^{1/2}+2\sqrt{2})\\
\bar{C}^*_{2,4} &:= \frac{\sqrt{2\hat{c}}}{1-\exp(-\dot{c}/2)}(\bar{C}_{2,2}^{1/2}+2\sqrt{2c_3}(\lambda_{\max}+a^{-1})^{1/2}+\sqrt{3}\mathrm{v}_4^{1/2}(\overline{M}_4)+\sqrt{15}),
\end{aligned}
\tag{C14}
$$

with $\bar{C}_{2,1}$, $\bar{C}_{2,2}$ given in (C12), $\hat{c}$, $\dot{c}$ given in Lemma 4.11, c_3 given in (20), and $\overline{M}_4$ given in Lemma 4.4. This completes the proof. □

Proof of Lemma 4.10 By using Lemma 4.7 and 4.8, one obtains

$$
\begin{aligned}
W_1(\mathcal{L}(\bar{\theta}^\lambda_t),\mathcal{L}(Z^\lambda_t)) &\le W_1(\mathcal{L}(\bar{\theta}^\lambda_t),\mathcal{L}(\bar{\zeta}^{\lambda,n}_t))+W_1(\mathcal{L}(\bar{\zeta}^{\lambda,n}_t),\mathcal{L}(Z^\lambda_t))\\
&\le \sqrt{\lambda}(e^{-an/8}\bar{C}_{2,1}^{1/2}\mathbb{E}^{1/2}[V_2(\theta_0)]+\bar{C}_{2,2}^{1/2})\\
&\quad+\sqrt{\lambda}(e^{-\dot{c}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)]+\bar{C}_{2,4})\\
&\le (\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2}+\bar{C}_{2,3}+\bar{C}_{2,4})\sqrt{\lambda}(e^{-\dot{c}n/2}\mathbb{E}[V_4(\theta_0)]+1).
\end{aligned}
$$

□

Proof of Lemma 4.11 To obtain the contraction constant $\dot{c}$, we apply the arguments in the proof of [15, Theorem 2.2] to SDE (17). More precisely, we replace $h(r)$ in [15, Eqn. (5.14)] by

$$
h(r) := \frac{\beta}{4}\int_0^r s\kappa\,\mathrm{d}s + 2Q(\epsilon)r, \tag{C15}
$$

where $\kappa = L_1\mathbb{E}[\eta(X_0)]$ and $Q(\epsilon)$ are given in [15, Eqn. (2.24)], and replace [15, Eqn. (2.25)] by $(4\tilde{c}(2)\epsilon)^{-1} \ge \frac{\beta}{2}\int_0^{R_1}\int_0^s \exp\left(\frac{\beta}{4}\int_r^s u\kappa\,\mathrm{d}u + 2Q(\epsilon)(s-r)\right)\,\mathrm{d}r\,\mathrm{d}s$. Then, following the proof of [15, Theorem 2.2], one can derive the expressions for $\dot{c}$: $\dot{c} := \min\{\phi, \bar{c}(2), 4\tilde{c}(2)\epsilon\bar{c}(2)\}/2$, where $\bar{c}(2) = a/2$, $\tilde{c}(2) = (3/2)a\mathrm{v}_2(\overline{M}_2)$ with $\overline{M}_2$ given in Lemma 4.4, ϕ is given by $\phi^{-1} := \int_0^{R_2}\int_0^s \exp\left(\frac{\beta}{4}\int_r^s u\kappa\,\mathrm{d}u + 2Q(\epsilon)(s-r)\right)\,\mathrm{d}r\,\mathrm{d}s$ with R_2 given in [15, Eqn. (2.29)], and $\epsilon \in (0,1]$ is required to satisfy $\epsilon^{-1} \ge 2\beta\tilde{c}(2)\int_0^{R_1}\int_0^s \exp\left(\frac{\beta}{4}\int_r^s u\kappa\,\mathrm{d}u + 2Q(\epsilon)(s-r)\right)\,\mathrm{d}r\,\mathrm{d}s$ with R_1 given in [15, Eqn. (2.29)]. To simply the expressions for ϕ and ϵ, we follow the proof of [14, Lemma 3.24], and thus (23), (24), (25) can be obtained.

To obtain an explicit expression for $\hat{c}$, one first notes that, by using (C15), [15, Eqn. (5.4)] becomes: for any $r \in [0, R_2]$, $r\exp(-\beta\kappa R_2^2/8 - 2Q(\epsilon)R_2) \le \Phi(r) \le 2f(r) \le 2\Phi(r) \le 2r$. Then, in view of [14, Eqn. (60)], and by applying the same arguments as in the proof of [14, Lemma 3.24], one obtains $C_9 := C_{11}/C_{10} \le \hat{c} := 2(1+\overline{R}_2)\exp(\beta K_1\overline{R}_2^2/8 + 2\overline{R}_2)/\epsilon$, where $\overline{R}_2 = \bar{b} := 2\sqrt{4\tilde{c}(2)(1+\bar{c}(2))/\bar{c}(2) - 1}$, $K_1 := L_1\mathbb{E}[\eta(X_0)]$, and ϵ is given in (25). □

Appendix D Table of constant

Table D1 Analytic expressions of constants

Constant		Full expression
Lemma 4.4	$\overline{M}_p$	$\sqrt{1/3 + 4b/(3a) + 4d/(3a\beta) + 4(p-2)/(3a\beta)}$
	$\bar{c}(p)$	$ap/4$
	$\tilde{c}(p)$	$(3/4)ap\mathrm{v}_p(\overline{M}_p)$
Lemma 4.7	$\bar{C}_{2,1}$	$4e^{4L_1^2\mathbb{E}\left[\eta^2(X_0)\right]}\left(2\lambda_{\max}L_1^4(\mathbb{E}\left[\eta^2(X_0)\right])^2 + 8L_2^2\mathbb{E}[(\eta(X_0)+\eta(\mathbb{E}[X_0]))^2\|X_0-\mathbb{E}[X_0]\|^2]\right)$
	$\tilde{C}_{2,2}$	$4e^{4L_1^2\mathbb{E}\left[\eta^2(X_0)\right]}\Big(2\lambda_{\max}L_1^4(\mathbb{E}[\eta^2(X_0)])^2c_1(\lambda_{\max}+a^{-1})$ $+4\lambda_{\max}L_1^2L_2^2\mathbb{E}[\eta^2(X_0)]\mathbb{E}[\bar{\eta}^2(X_0)] + 4\lambda_{\max}H_\star^2L_1^2\mathbb{E}\left[\eta^2(X_0)\right] + 2d\beta^{-1}L_1^2\mathbb{E}\left[\eta^2(X_0)\right]$ $+8L_2^2\mathbb{E}[(\eta(X_0)+\eta(\mathbb{E}[X_0]))^2\|X_0-\mathbb{E}[X_0]\|^2](3\mathrm{v}_2(\overline{M}_2)+c_1(\lambda_{\max}+a^{-1})+1)\Big)$
Lemma 4.8	$\bar{C}_{2,3}$	$\hat{c}\left(1+\frac{2}{\dot{c}}\right)(e^{a/4}\bar{C}_{2,1}+12)$
	$\bar{C}_{2,4}$	$\frac{\hat{c}}{1-\exp(-\dot{c})}(\bar{C}_{2,2}+12c_3(\lambda_{\max}+a^{-1})+9\mathrm{v}_4(\overline{M}_4)+15)$
Corollary 4.9	$\bar{C}^*_{2,3}$	$\sqrt{2\hat{c}}\left(1+\frac{4}{\dot{c}}\right)(e^{a/8}\bar{C}_{2,1}^{1/2}+2\sqrt{2})$
	$\bar{C}^*_{2,4}$	$\frac{\sqrt{2\hat{c}}}{1-\exp(-\dot{c}/2)}(\bar{C}_{2,2}^{1/2}+2\sqrt{2c_3}(\lambda_{\max}+a^{-1})^{1/2}+\sqrt{3}\mathrm{v}_4^{1/2}(\overline{M}_4)+\sqrt{15})$
Theorem 2.4	C_1	$2e^{\dot{c}/2}\left[(\lambda_{\max}^{1/2}(\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2}+\bar{C}_{2,3}+\bar{C}_{2,4})+\hat{c})+\hat{c}\left(1+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right)\right]$
	C_2	$\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2}$
	C_3	$\bar{C}_{2,3}+\bar{C}_{2,4}$
Corollary 2.5	C_4	$2\left(\lambda_{\max}^{1/2}(\bar{C}_{2,1}^{1/2}+\bar{C}_{2,2}^{1/2})+\lambda_{\max}^{1/4}(\bar{C}^*_{2,3}+\bar{C}^*_{2,4})+\sqrt{2}\hat{c}^{1/2}\right)$
	C_5	$\lambda_{\max}^{1/4}\bar{C}_{2,1}^{1/2}+\lambda_{\max}^{1/4}\bar{C}_{2,2}^{1/2}$
	C_6	$\bar{C}^*_{2,3}+\bar{C}^*_{2,4}$
Corollary 2.8	$C_1^\sharp$	$C_4\left(L_1\mathbb{E}[\eta(X_0)](\mathbb{E}[\|\theta_0\|^2]+c_1(\lambda_{\max}+a^{-1}))+L_2\mathbb{E}[\bar{\eta}(X_0)]+H_\star\right)\mathbb{E}[\|\theta_0\|^4+1]$
	$C_2^\sharp$	$(C_5+C_6)\left(L_1\mathbb{E}[\eta(X_0)](\mathbb{E}[\|\theta_0\|^2]+c_1(\lambda_{\max}+a^{-1}))+L_2\mathbb{E}[\bar{\eta}(X_0)]+H_\star\right)$
	$C_3^\sharp$	$\frac{d}{2\beta}\log\left(\frac{eL_1\mathbb{E}[\eta(X_0)]}{a}\left(\frac{b\beta}{d}+1\right)\right)$

Table D2 Constants in Lemma 4.2 and Lemma 4.11, and their dependency on key parameters

Constant	Key parameters		
	d	β	Moments of X_0
c_1	$O(1+d/\beta)$	$O(1+d/\beta)$	$O(\mathbb{E}[(1+\|X_0\|)\eta(X_0)])$
c_3	$O(1+(d/\beta)^2)$	$O(1+(d/\beta)^2)$	$O(\mathbb{E}^{3/2}[(1+\|X_0\|)^4\eta^4(X_0)])$
$\dot{c}$	$\left(\frac{32\sqrt{\pi}(1+a^2)(1+\beta)}{a^2\sqrt{\beta}}\left(1+\frac{1}{\sqrt{L_1\mathbb{E}[\eta(X_0)]}}\right)e^{\left(8C^\star(a,b)(1+\beta L_1\mathbb{E}[\eta(X_0)])(1+\frac{d}{\beta})+\frac{16}{\beta L_1\mathbb{E}[\eta(X_0)]}\right)}\right)^{-1}$ [1]		
$\hat{c}$	$O\left(\sqrt{\frac{\beta}{L_1\mathbb{E}[\eta(X_0)]}}(1+\frac{d}{\beta})^2e^{\left(12C^\star(a,b)(1+\beta L_1\mathbb{E}[\eta(X_0)])(1+\frac{d}{\beta})+\frac{16}{\beta L_1\mathbb{E}[\eta(X_0)]}\right)}\right)$ [1]		

[1] $C^\star(a,b) := (1+2/a)(1+a+b)$.

References

[1] Hwang, C.-R.: Laplace's method revisited: weak convergence of probability measures. The Annals of Probability **8**(6), 1177–1182 (1980)

[2] Dalalyan, A.S.: Theoretical guarantees for approximate sampling from

smooth and log-concave densities. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(3), 651–676 (2017)

[3] Durmus, A., Moulines, E.: Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. The Annals of Applied Probability **27**(3), 1551–1587 (2017)

[4] Durmus, A., Moulines, E.: High-dimensional Bayesian inference via the unadjusted Langevin algorithm. Bernoulli **25**(4A), 2854–2882 (2019)

[5] Brosse, N., Durmus, A., Moulines, É., Sabanis, S.: The tamed unadjusted Langevin algorithm. Stochastic Processes and their Applications **129**(10), 3638–3663 (2019)

[6] Dalalyan, A.S., Karagulyan, A.: User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stochastic Processes and their Applications (2019)

[7] Sabanis, S., Zhang, Y.: Higher order Langevin Monte Carlo algorithm. Electronic Journal of Statistics **13**(2), 3805–3850 (2019)

[8] Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 681–688 (2011)

[9] Barkhagen, M., Chau, N.H., Moulines, É., Rásonyi, M., Sabanis, S., Zhang, Y.: On stochastic gradient langevin dynamics with dependent data streams in the logconcave case. Bernoulli **27**(1), 1–33 (2021)

[10] Brosse, N., Durmus, A., Moulines, E.: The promises and pitfalls of stochastic gradient Langevin dynamics. In: Advances in Neural Information Processing Systems, pp. 8268–8278 (2018)

[11] Dalalyan, A.: Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In: Kale, S., Shamir, O. (eds.) Proceedings of the 2017 Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 65, pp. 678–689. PMLR, ??? (2017). https://proceedings.mlr.press/v65/dalalyan17a.html

[12] Raginsky, M., Rakhlin, A., Telgarsky, M.: Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In: Conference on Learning Theory, pp. 1674–1703 (2017)

[13] Xu, P., Chen, J., Zou, D., Gu, Q.: Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In: Advances in Neural Information Processing Systems, pp. 3122–3133 (2018)

[14] Chau, N.H., Moulines, É., Rásonyi, M., Sabanis, S., Zhang, Y.: On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. SIAM Journal on Mathematics of Data Science **3**(3), 959–986 (2021)

[15] Eberle, A., Guillin, A., Zimmer, R.: Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. Transactions of the American Mathematical Society **371**(10), 7135–7173 (2019)

[16] Cheng, X., Chatterji, N.S., Abbasi-Yadkori, Y., Bartlett, P.L., Jordan, M.I.: Sharp convergence rates for Langevin dynamics in the nonconvex setting. arXiv preprint arXiv:1805.01648 (2018)

[17] Majka, M.B., Mijatović, A., Szpruch, L.: Non-asymptotic bounds for sampling algorithms without log-concavity. arXiv preprint arXiv:1808.07105v3 (2019)

[18] Eberle, A.: Reflection couplings and contraction rates for diffusions. Probability theory and related fields **166**(3-4), 851–886 (2016)

[19] Erdogdu, M.A., Mackey, L., Shamir, O.: Global non-convex optimization with discretized diffusions. In: Advances in Neural Information Processing Systems, pp. 9671–9680 (2018)

[20] Zheng, Y., Chen, B., Hospedales, T.M., Yang, Y.: Index tracking with cardinality constraints: A stochastic neural networks approach. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 1242–1249 (2020)

[21] Gaivoronski, A.A., Krylov, S., Van der Wijst, N.: Optimal portfolio selection and dynamic benchmark tracking. European Journal of operational research **163**(1), 115–131 (2005)

[22] Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning **1**(1–2), 1–305 (2008)

[23] Price, R.: A useful theorem for nonlinear devices having gaussian inputs. IRE Transactions on Information Theory **4**(2), 69–72 (1958)

[24] Salimans, T., Knowles, D.A.: Fixed-form variational posterior approximation through stochastic linear regression. Bayesian Analysis **8**(4), 837–882 (2013)

[25] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[26] Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. ICML'14, pp. 1278–1286. JMLR.org, ??? (2014). http://dl.acm.org/citation.cfm?id=3044805.3045035

[27] Mattingly, J., Stuart, A., Higham, D.: Ergodicity for sdes and approximations: Locally lipschitz vector fields and degenerate noise. Stochastic Processes and their Applications **101**, 185–232 (2002)

[28] Cox, S., Hutzenthaler, M., Jentzen, A.: Local lipschitz continuity in the initial value and strong completeness for nonlinear stochastic differential equations. arXiv preprint arXiv:1309.5595 (2013)

[29] Chau, H.N., Kumar, C., Rásonyi, M., Sabanis, S.: On fixed gain recursive estimators with discontinuity in the parameters. ESAIM: Probability and Statistics **23**, 217–244 (2019)

Statements and Declarations

- **Funding** This work was supported by The Alan Turing Institute for Data Science and AI under EPSRC grant EP/N510129/1. Y. Z. was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh. Ö. D. A. is supported by the Lloyd's Register Foundation Data Centric Engineering Programme and EPSRC Programme Grant EP/R034710/1. T. D. acknowledges support from EPSRC EP/T004134/1, UKRI Turing AI Fellowship EP/V02678X/1, and Lloyd's Register Foundation programme on Data Centric Engineering through the London Air Quality project.
- **Competing interests** The authors have no relevant financial or non-financial interests to disclose.
- **Authors' contributions** All authors contributed to the study conception and design. The first draft of the manuscript was written by Ying Zhang and Ömer Deniz Akyildiz, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.