Journal of Machine Learning Research 25 (2024) 1-34 Submitted 11/21; Revised 1/24; Published 1/24

Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization

Ö. Deniz Akyildiz DENIZ.AKYILDIZ@IMPERIAL.AC.UK
Department of Mathematics, Imperial College London

Sotirios Sabanis S.SABANIS@ED.AC.UK
School of Mathematics, University of Edinburgh
The Alan Turing Institute
National Technical University of Athens

Editor: Ohad Shamir

Abstract

We provide a nonasymptotic analysis of the convergence of the stochastic gradient Hamiltonian Monte Carlo (SGHMC) to a target measure in Wasserstein-2 distance without assuming log-concavity. Our analysis quantifies key theoretical properties of the SGHMC as a sampler under local conditions which significantly improves the findings of previous results. In particular, we prove that the Wasserstein-2 distance between the target and the law of the SGHMC is uniformly controlled by the step-size of the algorithm, therefore demonstrate that the SGHMC can provide high-precision results uniformly in the number of iterations. The analysis also allows us to obtain nonasymptotic bounds for nonconvex optimization problems under local conditions and implies that the SGHMC, when viewed as a nonconvex optimizer, converges to a global minimum with the best known rates. We apply our results to obtain nonasymptotic bounds for scalable Bayesian inference and nonasymptotic generalization bounds.

Keywords: Non-convex optimization, underdamped Langevin Monte Carlo, non-log-concave sampling, nonasmyptotic bounds, global optimization.

1. Introduction

We are interested in nonasymptotic estimates for the sampling problem from the probability measures of the form

$$\pi_\beta(\mathrm{d}\theta) \propto \exp(-\beta U(\theta))\mathrm{d}\theta. \tag{1}$$

when only the noisy estimate of ∇U is available. This problem arises in many cases in machine learning, most notably in large-scale (mini-batch) Bayesian inference (Welling and Teh, 2011; Ahn et al., 2012) and nonconvex stochastic optimization as for large values of β, a sample from the target measure (1) is an approximate minimizer of the potential U (Gelfand and Mitter, 1991). Consequently, nonasymptotic error bounds for the sampling schemes can be used to obtain guarantees for Bayesian inference or nonconvex optimization.

©2024 Ömer Deniz Akyildiz and Sotirios Sabanis.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/21-1423.html.

An efficient method for obtaining a sample from (1) is simulating the overdamped Langevin stochastic differential equation (SDE) which is given by

$$\mathrm{d}L_t = -h(L_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t, \tag{2}$$

with a random initial condition $L_0 := \theta_0$ where $h := \nabla U$ and $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. The Langevin SDE (2) admits π_β as the unique invariant measure, therefore simulating this process will lead to samples from π_β and can be used as a Markov chain Monte Carlo (MCMC) algorithm (Roberts et al., 1996; Roberts and Stramer, 2002). Moreover, the fact that the limiting probability measure π_β concentrates around the global minimum of U for sufficiently large values of β makes the diffusion (2) also an attractive candidate as a global optimizer (see, e.g., Hwang (1980)). However, since the continuous-time process (2) cannot be simulated, its first-order Euler discretization with the step-size $\eta > 0$ is used in practice, termed *the Unadjusted Langevin Algorithm* (ULA) (Roberts et al., 1996). The ULA scheme has become popular in recent years due to its advantages in high-dimensional settings and ease of implementation. Nonasymptotic properties of the ULA were recently established under strong convexity and smoothness assumptions by Dalalyan (2017); Durmus et al. (2017, 2019) while some extensions about relaxing smoothness assumptions or inaccurate gradients were also considered by Dalalyan and Karagulyan (2019); Brosse et al. (2019). The similar attractive properties hold for the ULA when the potential U is nonconvex (Gelfand and Mitter, 1991; Raginsky et al., 2017; Xu et al., 2018; Erdogdu et al., 2018; Sabanis and Zhang, 2019). In recent works, ULA has been extended for nonconvex cases in several different directions, e.g., under log-Sobolev inequality (Vempala and Wibisono, 2019), under Hölder continuity and specific tail growth conditions (Erdogdu and Hosseinzadeh, 2021), under Poincaré inequality (Chewi et al., 2022). Further work has extended these results, see, e.g., Balasubramanian et al. (2022) for averaged Langevin schemes, Erdogdu et al. (2022) for the analysis under dissipativity in Chi-squared and Renyi divergences, Mou et al. (2022) for results under dissipativity with smooth initalisation, and finally under weak Poincaré inequalities (Mousavi-Hosseini et al., 2023).

While the ULA performs well when the computation of the gradient $h(\cdot)$ is straightforward, this is not the case in most interesting applications. Usually, a stochastic, unbiased estimate of $h(\cdot)$ is available, either because the cost function is defined as an expectation or as a finite sum. Using stochastic gradients in the ULA leads to another scheme called stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011). The SGLD has been particularly popular in the fields of (i) *large-scale Bayesian inference* since it allows one to construct Markov chains Monte Carlo (MCMC) algorithms using only subsets of the dataset (Welling and Teh, 2011), (ii) *nonconvex optimization* since it enables one to estimate global minima using only stochastic (often cheap-to-compute) gradients (Raginsky et al., 2017). As a result, attempts for theoretical understanding of the SGLD have been recently made in several works, both for the strongly convex potentials (i.e. log-concave targets), see, e.g., Barkhagen et al. (2021); Brosse et al. (2018) and nonconvex potentials, see, e.g. Raginsky et al. (2017); Majka et al. (2018); Zhang et al. (2023b). Our particular interest is in nonasymptotic bounds for the nonconvex case, as it is relevant to our work. In their seminal paper, Raginsky et al. (2017) obtain a nonasymptotic bound between the law of the SGLD and the target measure in Wasserstein-2 distance with a rate $\eta^{5/4}n$ where η is

the step-size and n is the number of iterations. While this work is first of its kind, the error rate grows with the number of iterations. In a related contribution, Xu et al. (2018) have obtained improved rates, albeit still growing with the number of iterations n. In a more recent work, Chau et al. (2021) have obtained a uniform rate of order $\eta^{1/2}$ in Wasserstein-1 distance. Majka et al. (2018) achieved error rates of $\eta^{1/2}$ and $\eta^{1/4}$ for Wasserstein-1 and Wasserstein-2 distances, respectively, under the assumption of convexity outside a ball. Finally, Zhang et al. (2023b) achieved the same rates under only local conditions which can be verified for a class of practical problems.

An alternative to methods based on the *overdamped* Langevin SDE (2) is the class of algorithms based on the *underdamped* Langevin SDE. To be precise, the underdamped Langevin SDE is given as

$$\mathrm{d}V_t = -\gamma V_t \mathrm{d}t - h(\theta_t)\mathrm{d}t + \sqrt{\frac{2\gamma}{\beta}}\mathrm{d}B_t, \tag{3}$$

$$\mathrm{d}\theta_t = V_t \mathrm{d}t, \tag{4}$$

where $(\theta_t, V_t)_{t\geq 0}$ are called position and momentum process, respectively, and $h := \nabla U$. Similar to eq. (2), this diffusion can be used as both an MCMC sampler and nonconvex optimizer, since under appropriate conditions, the Markov process $(\theta_t, V_t)_{t\geq 0}$ has a unique invariant measure given by

$$\overline{\pi}_\beta(\mathrm{d}\theta, \mathrm{d}v) \propto \exp\left(-\beta\left(\frac{1}{2}\|v\|^2 + U(\theta)\right)\right)\mathrm{d}\theta\mathrm{d}v. \tag{5}$$

Consequently, the marginal distribution of (5) in θ is precisely the target measure defined in (1). This means that sampling from (5) in the extended space and then keeping the samples in the θ-space would define a valid sampler for the sampling problem of (1).

Due to its attractive properties, methods based on the underdamped Langevin SDE have attracted significant attention. In particular, the first order discretization of (3)–(4), which is termed *underdamped Langevin MCMC* (i.e. the *underdamped* counterpart of the ULA), has been a focus of attention, see, e.g., Duncan et al. (2017); Dalalyan and Riou-Durand (2018); Cheng et al. (2018b). Particularly, the underdamped Langevin MCMC has displayed improved convergence rates in the setting where U is convex, see, e.g., Dalalyan and Riou-Durand (2018); Cheng et al. (2018b). Similar results have been extended to the nonconvex case. In particular, Cheng et al. (2018a) have shown that the underdamped Langevin MCMC converges in Wasserstein-2 with a better dimension and step-size dependence under the assumptions smoothness and convexity outside a ball. It has been also shown that the underdamped Langevin MCMC can be seen as an accelerated optimization method in the space of measures in Kullback-Leibler divergence (Ma et al., 2019).

Similar to the case in the ULA, oftentimes $\nabla U(\cdot)$ is expensive or impossible to compute exactly, but rather an unbiased estimate of it can be obtained efficiently. The underdamped Langevin MCMC with stochastic gradients is dubbed as Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) and given as (Chen et al., 2014; Ma et al., 2015)

$$V_{n+1}^\eta = V_n^\eta - \eta[\gamma V_n^\eta + H(\theta_n^\eta, X_{n+1})] + \sqrt{\frac{2\gamma\eta}{\beta}}\xi_{n+1}, \tag{6}$$

$$\theta_{n+1}^\eta = \theta_n^\eta + \eta V_n^\eta, \tag{7}$$

where $\eta > 0$ is a step-size, $V_0^\eta = v_0$, $\theta_0^\eta = \theta_0$, and $\mathbb{E}[H(\theta, X_0)] = h(\theta)$ for every $\theta \in \mathbb{R}^d$. We note that SGHMC relies on the Euler-Maruyama discretisation which is of the possible discretisation methods that can be used. Alternative discretisation methods can be considered and lead to different algorithms, e.g., see Cheng et al. (2018b); Li et al. (2019); Zhang et al. (2023a).

In this paper, we analyze recursions (6)–(7). We achieve convergence bounds and improve the ones proved in Gao et al. (2022) and Chau and Rásonyi (2022) (see Section 2.1 for a direct comparison).

Notation. For an integer $d \geq 1$, the Borel sigma-algebra of $\mathbb{R}^d$ is denoted by $\mathcal{B}(\mathbb{R}^d)$. We denote the dot product with $\langle \cdot, \cdot \rangle$ while $|\cdot|$ denotes the associated norm. The set of probability measures defined on a measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is denoted as $\mathcal{P}(\mathbb{R}^d)$. For an $\mathbb{R}^d$-valued random variable, $\mathcal{L}(X)$ and $\mathbb{E}[X]$ are used to denote its law and its expectation respectively. Note that we also write $\mathbb{E}[X]$ as $\mathbb{E}X$. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures Γ on $\mathcal{B}(\mathbb{R}^{2d})$ so that its marginals are μ, ν. Finally, $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the Wasserstein distance of order $p \geq 1$ is defined as

$$W_p(\mu, \nu) := \inf_{\Gamma \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta - \theta'|^p \Gamma(\mathrm{d}\theta, \mathrm{d}\theta') \right)^{1/p}. \tag{8}$$

2. Main results and overview

Let $(X_n)_{n\in\mathbb{N}}$ be an $\mathbb{R}^m$-valued stochastic process adapted to $(\mathcal{G}_n)_{n\in\mathbb{N}}$ where $\mathcal{G}_n := \sigma(X_k, k \leq n)$ for $n \in \mathbb{N}$. It is assumed henceforth that $\theta_0, v_0, \mathcal{G}_\infty$, and $(\xi_n)_{n\in\mathbb{N}}$ are independent. The main assumptions about U and other quantities follow.

Assumption 2.1 *The cost function U takes nonnegative values, i.e., $U(\theta) \geq 0$.*

Next, the *local smoothness* assumptions on the stochastic gradient H are given.

Assumption 2.2 *There exist positive constants L_1, L_2 and ρ such that, for all $x, x' \in \mathbb{R}^m$ and $\theta, \theta' \in \mathbb{R}^d$,*

$$|H(\theta, x) - H(\theta', x)| \leq L_1(1 + |x|)^\rho |\theta - \theta'|,$$
$$|H(\theta, x) - H(\theta, x')| \leq L_2(1 + |x| + |x'|)^\rho (1 + |\theta|)|x - x'|.$$

The following assumption states that the stochastic gradient is assumed to be unbiased.

Assumption 2.3 *The process $(X_n)_{n\in\mathbb{N}}$ is i.i.d. with $|X_0| \in L^{4(\rho+1)}$ and $|\theta_0|, |v_0| \in L^4$. It satisfies*

$$\mathbb{E}[H(\theta, X_0)] = h(\theta).$$

It is important to note that Assumption 2.2 is a significant relaxation in comparison with the corresponding assumptions provided in the literature, see, e.g., Raginsky et al. (2017); Gao et al. (2022); Chau and Rásonyi (2022). To the best of the authors' knowledge, all relevant works in this area have focused on uniform Lipschitz assumptions with the exception of Zhang et al. (2023b), which provides a nonasymptotic analysis of the SGLD under similar assumptions to ours.

Remark 2.1 *Assumption 2.2 implies, for all $\theta, \theta' \in \mathbb{R}^d$,*

$$|h(\theta) - h(\theta')| \leq L_1 \mathbb{E}[(1 + |X_0|)^\rho]|\theta - \theta'|, \tag{9}$$

which consequently implies

$$|h(\theta)| \leq L_1 \mathbb{E}[(1 + |X_0|)^\rho]|\theta| + h_0, \tag{10}$$

where $h_0 := |h(0)|$. Let $H_0 := |H(0,0)|$, then Assumption 2.2 implies

$$|H(\theta, x)| \leq L_1(1 + |x|)^\rho|\theta| + L_2(1 + |x|)^{\rho+1} + H_0.$$

We denote $C_\rho := \mathbb{E}\left[(1 + |X_0|)^{4(\rho+1)}\right]$. Note that $C_\rho < \infty$ by Assumption 2.3.

Assumption 2.4 *There exist a measurable (symmetric matrix-valued) function $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$, $b : \mathbb{R}^m \to \mathbb{R}$ such that for any $x, y \in \mathbb{R}^d$, $\langle y, A(x)y\rangle \geq 0$, and for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,*

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

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

Note that Assumption 2.4 is a *local* dissipativity condition. This assumption implies the usual dissipativity property on the corresponding deterministic (full) gradient. In the next remark, we motivate these assumptions for the case of linear regression loss.

Remark 2.2 *We note that Assumptions 2.2 and 2.4 can be motivated even using the simplest linear regression loss. Consider the problem*

$$\min_{\theta\in\mathbb{R}^d} \mathbb{E}\left[|Z - \langle Y, \theta\rangle|^2\right],$$

where $(Z, Y) \sim P(\mathrm{d}z, \mathrm{d}y)$ on $\mathbb{R} \times \mathbb{R}^d$. The stochastic gradient here is given by, for a sample $(z_n, y_n) \sim P(\mathrm{d}z, \mathrm{d}y)$,

$$H(\theta, x_n) = -2y_n z_n - y_n\langle y_n, \theta\rangle,$$

where $x_n = (z_n, y_n)$. This loss is not globally Lipschitz however we can prove that it is locally Lipschitz. For example, it is easy to show that Assumption 2.2 is satisfied with $L_1 = 1$, $L_2 = 2$, and $\rho = 2$. Moreover, it is easy to see that this stochastic gradient is not dissipative uniformly in x_n as

$$\langle \theta, H(\theta, x_n)\rangle \geq |\langle y_n, \theta\rangle|^2 - z_n^2.$$

However, our Assumption 2.4 is satisfied with $A(x_n) = y_n y_n^\top$ and $b(x_n) = z_n^2$. We would also like to note that we do not need to assume that stochastic gradient moments are bounded as it is a result of our assumptions.

Remark 2.3 *By Assumption 2.4, we obtain $\langle h(\theta), \theta\rangle \geq a|\theta|^2 - b$, for $\theta \in \mathbb{R}^d$ and $a, b > 0$.*

Remark 2.4 *Due to Assumption 2.1, Remark 2.1 and Remark 2.3, one observes that Assumption 2.1 in Eberle et al. (2019) is satisfied and thus, due to Corollary 2.6 in Eberle et al. (2019), the underdamped Langevin SDE* (3)–(4) *has the unique invariant measure given by* (5)*.*

Below, we state our main result about the convergence of the law $\mathcal{L}(\theta_k^\eta, V_k^\eta)$, which is generated by the SGHMC recursions (6)–(7), to the extended target measure $\overline{\pi}_\beta$ in Wasserstein-2 (W_2) distance. We first define

$$\eta_{\max} = \min\left\{1, \frac{1}{\gamma}, \frac{\gamma\lambda}{2K_1}, \frac{K_3}{K_2}, \frac{\gamma\lambda}{2\bar{K}}\right\},$$

where K_1, K_2, K_3 are constants explicitly given in the proof of Lemma 3.2 and $\bar{K}$ is a constant explicitly given in the proof of Lemma 3.4. Then, the following result is obtained.

Theorem 2.1 *Let Assumptions 2.1–2.4 hold. Then, there exist constants $C_1^\star, C_2^\star, C_3^\star, C_4^\star > 0$ such that, for every $0 < \eta \leq \eta_{\max}$,*

$$W_2(\mathcal{L}(\theta_n^\eta, V_n^\eta), \overline{\pi}_\beta) \leq C_1^\star \eta^{1/2} + C_2^\star \eta^{1/4} + C_3^\star e^{-C_4^\star \eta n}. \tag{11}$$

where the constants $C_1^\star, C_2^\star, C_3^\star, C_4^\star$ are explicitly provided in the Appendix.

Proof See Section 4. ■

Remark 2.5 *We remark that $C_1^\star = \mathcal{O}(d^{1/2})$, $C_2^\star = \mathcal{O}(e^d)$, $C_3^\star = \mathcal{O}(e^d)$, and $C_4^\star = \mathcal{O}(e^{-d})$. We note that although the dependence of $C_1^\star$ on dimension is $\mathcal{O}(d^{1/2})$, and comes from our main result, the dependence of $C_2^\star$, $C_3^\star$, and $C_4^\star$ to dimension may be exponential as it is an immediate consequence of the contraction result of the underdamped Langevin SDE in Eberle et al. (2019).*

The result in Theorem 2.1 demonstrates that the error scales like $\mathcal{O}(\eta^{1/4})$ and is uniformly bounded over n which can be made arbitrarily small by choosing $\eta > 0$ small enough. This result is thus a significant improvement over the findings in Gao et al. (2022), where error bounds grow with the number of iterations, and in Chau and Rásonyi (2022), where the corresponding error bounds contain an additional term that is independent of η and relates to the variance of the unbiased estimator.

Remark 2.6 *Our proof techniques can be adapted easily when $H(\theta, x) = h(\theta)$. Hence Theorem 2.1 provides a novel convergence bound for the underdamped Langevin MCMC as well. Our result is in W_2 distance under dissipativity for the Euler-Maruyama discretisation of the underdamped Langevin dynamics, which can be contrasted with existing work which focus on alternative discretisations. Beyond the log-concave case, Ma et al. (2019) analyses a second-order discretisation of underdamped Langevin diffusion under the assumption that the target satisfies a log-Sobolev inequality with a Lipschitz Hessian condition for the log-target. More recently, Zhang et al. (2023a) removed the Lipschitz Hessian condition and established the convergence of underdamped Langevin MCMC under log-Sobolev and Poincaré inequalities. To the best of our knowledge, our result for the Euler-Maruyama discretisation of the underdamped Langevin dynamics is the first nonasymptotic result under dissipativity that is uniformly controlled by the step-size.*

Let $(\theta_k^\eta)_{k\in\mathbb{N}}$ be generated by the SGHMC algorithm. Convergence of the $\mathcal{L}(\theta_k^\eta)$ to π_β in W_2 also implies that one can prove a global convergence result (Raginsky et al., 2017). More precisely, assume that we aim at solving the problem $\theta_\star \in \arg\min_{\theta\in\mathbb{R}^d} U(\theta)$ which is a nonconvex optimization problem. We denote $U_\star := \inf_{\theta\in\mathbb{R}^d} U(\theta)$. Then we can bound the error $\mathbb{E}[U(\theta_k^\eta)] - U_\star$ which would give us a guarantee on the nonconvex optimization problem. We state it as a formal result as follows.

Theorem 2.2 *Under the assumptions of Theorem 2.1, we obtain*

$$\mathbb{E}[U(\theta_n^\eta)] - U_\star \leq \overline{C}_1^\star \eta^{1/2} + \overline{C}_2^\star \eta^{1/4} + \overline{C}_3^\star e^{-C_4^\star \eta n} + \frac{d}{2\beta}\log\left(\frac{e\overline{L}_1}{a}\left(\frac{b\beta}{d}+1\right)\right),$$

where $\overline{C}_1^\star, \overline{C}_2^\star, \overline{C}_3^\star, C_4^\star, \overline{L}_1$ are finite constants which are explicitly given in the proofs.

Proof See Section 5. ■

This result bounds the error in terms of the function value for convergence to the global minima. Note that $\overline{C}_1^\star, \overline{C}_2^\star$ anf $\overline{C}_3^\star$ have the same dependence to dimension as $C_1^\star, C_2^\star$ and $C_3^\star$ (see Remark 2.5).

2.1 Related work and contributions

Our work is related to two available analyses of the SGHMC, i.e., Gao et al. (2022) and Chau and Rásonyi (2022). We compare the bounds provided in Theorem 2.1 and 2.2 to these two works. Finally, we also briefly consider the relations to Zhang et al. (2023b) at the end of this section.

The scheme (6)–(7) is analyzed by Gao et al. (2022). In particular, Gao et al. (2022) provided a convergence rate of the SGHMC (6)–(7) to the underdamped Langevin SDE (3)–(4) which is of order $\mathcal{O}(\delta^{1/4} + \eta^{1/4})\sqrt{\eta n}\sqrt{\log(\eta n)}$. This rate grows with n, hence worsens over the number of iterations. Moreover, it is achieved under a uniform assumption on the stochastic gradient, i.e., $H(\theta, x)$ is assumed to be Lipschitz in θ uniformly in x (as opposed to our Assumption 2.2). Moreover, the mean-squared error of the gradient is assumed to be bounded whereas we do not place such an assumption in our work. Similar analyses appeared in the literature, e.g., for variance-reduced SGHMC (Zou et al., 2019) which also has growing rates with the number of iterations.

Another related work was provided by Chau and Rásonyi (2022) who also analyzed the SGHMC recursions essentially under the same assumptions as in Gao et al. (2022). However, Chau and Rásonyi (2022) improved the convergence rate of the SGHMC recursions to the underdamped Langevin SDE significantly, i.e., provided a convergence rate of order $\mathcal{O}(\delta^{1/4} + \eta^{1/4})$ where $\delta > 0$ is a constant. While this rate significantly improves the rate of Gao et al. (2022), it cannot be made to vanish by choosing $\eta > 0$ small enough, as $\delta > 0$ is (a priori assumed to be) independent of η.

In contrast, we prove that the SGHMC recursions track the underdamped Langevin SDE with a rate of order $\mathcal{O}(\eta^{1/4})$ which can be made arbitrarily small as with small $\eta > 0$. Moreover, our assumptions are significantly relaxed compared to Gao et al. (2022) and Chau and Rásonyi (2022). In particular, we relax the assumptions on stochastic gradients significantly by allowing growth in both variables (θ, x) which makes our theory hold for practical settings (Zhang et al., 2023b).

Note that we analyze the SGHMC under similar assumptions to Zhang et al. (2023b) who analyzed SGLD, since these assumptions allow for broader applicability of the results. However, the SGHMC requires the analysis of the underdamped Langevin diffusion and the discrete-time scheme on $\mathbb{R}^{2d}$ and requires significantly different techniques compared to the ones used in Zhang et al. (2023b). We show that the SGHMC can achieve a similar performance with the SGLD as showed in Zhang et al. (2023b). However, due to the nonconvexity, we do not observe any improved dimension dependence as in the convex case and this constitutes a significant direction of future work.

3. Preliminaries

In this section, preliminary results which are essential for proving the main results are provided. A central idea behind the proof of Theorem 2.1 is the introduction of continuous-time auxiliary processes whose marginals at chosen discrete times coincide with the marginals of the (joint) law $\mathcal{L}(\theta_k^\eta, V_k^\eta)$. Hence, these auxiliary stochastic processes can be used to analyze the recursions (6)–(7). We first introduce the necessary lemmas about the moments of the auxiliary processes defined in Section 3 and contraction rates of the underlying underdamped Langevin diffusion. We give the lemmas and some explanatory remarks and defer the proofs to Appendix C. We then provide the proofs of main theorems in the following section.

Consider the scaled process $(\zeta_t^\eta, Z_t^\eta) := (\theta_{\eta t}, V_{\eta t})$ given $(\theta_t, V_t)_{t\in\mathbb{R}_+}$ as in (3)–(4). We next define

$$\mathrm{d}Z_t^\eta = -\eta(\gamma Z_t^\eta + h(\zeta_t^\eta))\mathrm{d}t + \sqrt{2\gamma\eta\beta^{-1}}\mathrm{d}B_t^\eta, \tag{12}$$

$$\mathrm{d}\zeta_t^\eta = \eta Z_t^\eta \mathrm{d}t, \tag{13}$$

where $\eta > 0$ and $B_t^\eta = \frac{1}{\sqrt{\eta}} B_{\eta t}$, where $(B_s)_{s\in\mathbb{R}_+}$ is a Brownian motion with natural filtration $\mathcal{F}_t$. We denote the natural filtration of $(B_t^\eta)_{t\in\mathbb{R}_+}$ as $\mathcal{F}_t^\eta$. We note that $\mathcal{F}_t^\eta$ is independent of $\mathcal{G}_\infty \vee \sigma(\theta_0, v_0)$. Next, we define the continuous-time interpolation of the SGHMC

$$\mathrm{d}\overline{V}_t^\eta = -\eta(\gamma \overline{V}_{\lfloor t\rfloor}^\eta + H(\overline{\theta}_{\lfloor t\rfloor}^\eta, X_{\lceil t\rceil}))\mathrm{d}t + \sqrt{2\gamma\eta\beta^{-1}}\mathrm{d}B_t^\eta, \tag{14}$$

$$\mathrm{d}\overline{\theta}_t^\eta = \eta \overline{V}_{\lfloor t\rfloor}^\eta \mathrm{d}t. \tag{15}$$

The processes (14)–(15) mimic the recursions (6)–(7) at $n \in \mathbb{N}$, i.e., $\mathcal{L}(\theta_n^\eta, V_n^\eta) = \mathcal{L}(\overline{\theta}_n^\eta, \overline{V}_n^\eta)$. Finally, we define the underdamped Langevin process $(\widehat{\zeta}_t^{s,u,v,\eta}, \widehat{Z}_t^{s,u,v,\eta})$ for $s \le t$

$$\mathrm{d}\widehat{Z}_t^{s,u,v,\eta} = -\eta(\gamma \widehat{Z}_t^{s,u,v,\eta} + h(\widehat{\zeta}_t^{s,u,v,\eta}))\mathrm{d}t + \sqrt{2\gamma\eta\beta^{-1}}\mathrm{d}B_t^\eta, \tag{16}$$

$$\mathrm{d}\widehat{\zeta}_t^{s,u,v,\eta} = \eta \widehat{Z}_t^{s,u,v,\eta}\mathrm{d}t, \tag{17}$$

with initial conditions $\widehat{\theta}_s^{s,u,v,\eta} = u$ and $\widehat{V}_s^{s,u,v,\eta} = v$.

Definition 3.1 *Fix $n \in \mathbb{N}$ and for $T := \lfloor 1/\eta \rfloor$, define*

$$\overline{\zeta}_t^{\eta,n} = \widehat{\zeta}_t^{nT,\overline{\theta}_{nT}^\eta,\overline{V}_{nT}^\eta,\eta}, \qquad \text{and} \qquad \overline{Z}_t^{\eta,n} = \widehat{Z}_t^{nT,\overline{\theta}_{nT}^\eta,\overline{V}_{nT}\eta,\eta}.$$

The process $(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})_{t\geq nT}$ is an underdamped Langevin process started at time nT with $(\overline{\theta}_{nT}^{\eta}, \overline{V}_{nT}^{\eta})$.

To achieve the convergence results, we first define a Lyapunov function, common in the literature (Mattingly et al., 2002; Eberle et al., 2019), defined as

$$\mathcal{V}(\theta, v) = \beta U(\theta) + \frac{\beta}{4}\gamma^2 \left(|\theta + \gamma^{-1}v|^2 + |\gamma^{-1}v|^2 - \lambda|\theta|^2\right), \tag{18}$$

where $\lambda \in (0, 1/4]$. This Lyapunov function plays an important role in obtaining uniform moment estimates for some of the aforementioned processes. First we recall a result from the literature about the second moments of the processes $(\theta_t, V_t)_{t\geq 0}$ in continuous-time.

Lemma 3.1 *(Lemma 12(i) in Gao et al. (2022).) Let Assumptions 2.1–2.4 hold. Then*

$$\sup_{t\geq 0} \mathbb{E}|\theta_t|^2 \leq C_\theta^c := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{d+A_c}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \tag{19}$$

$$\sup_{t\geq 0} \mathbb{E}|V_t|^2 \leq C_v^c := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{d+A_c}{\lambda}}{\frac{1}{4}(1-2\lambda)\beta} \tag{20}$$

Proof See Gao et al. (2022). ■

Next, we obtain, uniform in time, second moment estimates for the discrete-time processes $(\theta_k^\eta)_{k\geq 0}$ and $(V_k^\eta)_{k\geq 0}$.

Lemma 3.2 *Let Assumptions 2.1–2.4 hold. Then, for $0 < \eta \leq \eta_{\max}$,*

$$\sup_{k\geq 0} \mathbb{E}|\theta_k^\eta|^2 \leq C_\theta := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{4(A_c+d)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2},$$

$$\sup_{k\geq 0} \mathbb{E}|V_k^\eta|^2 \leq C_v := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{4(A_c+d)}{\lambda}}{\frac{1}{4}(1-2\lambda)\beta}.$$

Proof See Appendix C.1. ■

Using Lemma 3.2, we can get the second moment bounds of our auxiliary process $(\overline{\zeta}_t^{\eta,n})_{t\geq 0}$.

Lemma 3.3 *Under the assumptions of Lemmas 3.1 and 3.2, we obtain*

$$\sup_{n\in\mathbb{N}} \sup_{t\in(nT,(n+1)T)} \mathbb{E}|\overline{\zeta}_t^{\eta,n}|^2 \leq C_\zeta := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{8(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}. \tag{21}$$

Proof See Appendix C.2. ■

Moreover, in order to obtain the error bound distance between the laws $\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})$ and $\mathcal{L}(\zeta_t^\eta, Z_t^\eta)$, we need to obtain the contraction of $\mathbb{E}\mathcal{V}^2(\theta_k^\eta, V_k^\eta)$, which is established in the following lemma.

Lemma 3.4 *Let $0 < \eta \leq \eta_{\max}$ and Assumptions 2.1–2.4 hold. Then, we have*

$$\sup_{k\in\mathbb{N}} \mathbb{E}[\mathcal{V}^2(\theta_k^\eta, V_k^\eta)] \leq \mathbb{E}[\mathcal{V}^2(\theta_0, v_0)] + \frac{2D}{\gamma\lambda}$$

where $D = \mathcal{O}(d^2)$ constant independent of η and provided explicitly in the proof.

Proof See Appendix C.3. ■

Finally, we present a convergence result for the underdamped Langevin diffusion adapted from Eberle et al. (2019). To this end, a functional for probability measures μ, ν on $\mathbb{R}^{2d}$ is introduced below

$$\mathcal{W}_\rho(\mu,\nu) = \inf_{\Gamma \in \mathcal{C}(\mu,\nu)} \int \rho((x,v),(x',v'))\Gamma(d(x,v),d(x',v')), \tag{22}$$

where ρ is defined in eq. (2.10) in Eberle et al. (2019). Thus, in view of Remarks 2.1 and 2.3, one recovers the following result.

Theorem 3.1 *(Eberle et al., 2019, Theorem 2.3 and Corollary 2.6) Let Assumptions 2.1–2.4 hold and the laws of the underdamped Langevin SDEs (θ_t, V_t) and (θ'_t, V'_t) started at $(\theta_0, V_0) \sim \mu$ and $(\theta'_0, V'_0) \sim \nu$ respectively. Then, there exist constants $\dot{c}, \dot{C} \in (0,\infty)$ such that*

$$W_2(\mathcal{L}(\theta_t, V_t), \mathcal{L}(\theta'_t, V'_t)) \leq \sqrt{\dot{C}} e^{-\dot{c}t/2} \sqrt{\mathcal{W}_\rho(\mu,\nu)}, \tag{23}$$

where the constants $\dot{c} = \mathcal{O}(e^{-d})$ and $\dot{C} = \mathcal{O}(e^d)$ are given in Appendix A.

4. Proof of Theorem 2.1

In order to prove Theorem 2.1, we note first that W_2 is a metric on $\mathcal{P}(\mathbb{R}^{2d})$. The main strategy for this proof is to bound $W_2(\mathcal{L}(\theta^\eta_n, V^\eta_n), \overline{\pi}_\beta)$ by using appropriate estimates on the continuous time interpolation of $(\theta^\eta_n, V^\eta_n)_{n\in\mathbb{N}}$. In particular, we obtain the desired bound by decomposing first as

$$\begin{aligned} W_2(\mathcal{L}(\overline{\theta}^\eta_t, \overline{V}^\eta_t), \overline{\pi}_\beta) \leq\ & W_2(\mathcal{L}(\overline{\theta}^\eta_t, \overline{V}^\eta_t), \mathcal{L}(\overline{\zeta}^{\eta,n}_t, \overline{Z}^{\eta,n}_t)) + W_2(\mathcal{L}(\overline{\zeta}^{\eta,n}_t, \overline{Z}^{\eta,n}_t), \mathcal{L}(\zeta^\eta_t, Z^\eta_t)) \\ & + W_2(\mathcal{L}(\zeta^\eta_t, Z^\eta_t), \overline{\pi}_\beta), \end{aligned} \tag{24}$$

for $nT \leq t \leq (n+1)T$ for every $n \in \mathbb{N}$ and then by obtaining suitable (decaying in n) bounds for each of the terms on the rhs of (24). This leads to the proof of our main result, namely, Theorem 2.1. All proofs are deferred to the Appendix. First, we bound the first term of (24).

Theorem 4.1 *Let Assumptions 2.1–2.4 hold and $0 < \eta \leq \eta_{\max}$. Then, for every $t \in [nT, nT+1)$,*

$$W_2(\mathcal{L}(\overline{\theta}^\eta_t, \overline{V}^\eta_t), \mathcal{L}(\overline{\zeta}^{\eta,n}_t, \overline{Z}^{\eta,n}_t)) \leq C^\star_1 \eta^{1/2} \tag{25}$$

where $C^\star_1 = \mathcal{O}(d^{1/2})$ is a finite constant.

Proof See Appendix D.1. ■

Next, we prove the following result for bounding the second term of (24).

Theorem 4.2 *Let Assumptions 2.1–2.4 hold and $0 < \eta \leq \eta_{\max}$. Then,*

$$W_2(\mathcal{L}(\overline{\zeta}^{\eta,n}_t, \overline{Z}^{\eta,n}_t), \mathcal{L}(\zeta^\eta_t, Z^\eta_t)) \leq C^\star_2 \eta^{1/4},$$

where $C^\star_2 = \mathcal{O}(e^d)$.

Proof See Appendix D.2. ■

The constant $C_2^\star$ comes from the contraction result of Eberle et al. (2019, Corollary 2.6) which might scale exponentially in d. Finally, the convergence of the last term follows from Theorem 3.1.

Theorem 4.3 *(Eberle et al., 2019) Let Assumptions 2.1–2.4 hold. Then,*

$$W_2(\mathcal{L}(\zeta_t^\eta, Z_t^\eta), \overline{\pi}_\beta) \leq C_3^\star e^{-C_4^\star \eta t},$$

where $C_3^\star = \sqrt{\dot{C}\mathcal{W}_\rho(\mu_0, \nu_0)}$ *and* $C_4^\star = \dot{c}/2$. *In particular,* $C_3^\star = \mathcal{O}(e^d)$ *while* $C_4^\star = \mathcal{O}(e^{-d})$.

Finally, considering Theorems 4.1, 4.2, and 4.3 together by putting $t = n$ leads to the full proof of our main result, namely, Theorem 2.1.

5. Proof of Theorem 2.2

The bound provided for the convergence to the target in W_2 distance can be used to obtain guarantees for the nonconvex optimization. In order to do so, we proceed by decomposing the error as follows

$$\mathbb{E}[U(\theta_n^\eta)] - U_\star = \underbrace{\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}[U(\theta_\infty)] - U_\star}_{\mathcal{T}_2},$$

where $\theta_\infty \sim \pi_\beta$. The following proposition presents a bound for $\mathcal{T}_1$ under our assumptions.

Proposition 5.1 *Under the assumptions of Theorem 2.1, we have,*

$$\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)] \leq \overline{C}_1^\star \eta^{1/2} + \overline{C}_2^\star \eta^{1/4} + \overline{C}_3^\star e^{-C_4^\star \eta n}, \tag{26}$$

where $\overline{C}_i^\star = C_i^\star(C_m \bar{L}_1 + h_0)$ *for* $i = 1, 2, 3$ *and* $C_m^2 = \max(C_\theta^c, C_\theta)$.

Proof See Appendix D.3. ■

Next, we bound the second term $\mathcal{T}_2$ as follows. This result is fairly standard in the literature (see, e.g., Raginsky et al. (2017); Gao et al. (2022); Chau and Rásonyi (2022)).

Proposition 5.2 *(Raginsky et al., 2017) Under the assumptions of Theorem 2.1, we have*

$$\mathbb{E}[U(\theta_\infty)] - U_\star \leq \frac{d}{2\beta} \log\left(\frac{e\overline{L}_1}{a}\left(\frac{b\beta}{d} + 1\right)\right).$$

Merging Props. 5.1 and 5.2 leads to the bound given in Theorem 2.2 which completes our proof.

6. Applications

In this section, we present two applications to machine learning. First, we show that the SGHMC can be used to sample from the posterior probability measure in the context of scalable Bayesian inference. We also note that our assumptions hold in a practical setting of Bayesian logistic regression. Secondly, we provide an improved generalization bound for empirical risk minimization.

6.1 Convergence rates for scalable Bayesian inference

Consider a prior distribution $\pi_0(\theta)$ and a likelihood $\mathsf{p}(y_i|\theta)$ for a sequence of data points $\{y_i\}_{i=1}^M$ where M is the dataset size. Often, one is interested in sampling from the posterior probability distribution $p(\theta|y_{1:M})\mathrm{d}\theta \propto \pi_0(\theta)\prod_{i=1}^M \mathsf{p}(y_i|\theta)\mathrm{d}\theta$. This is a sampling problem of the form (1). The SGHMC is an MCMC method to sample from the posterior measure and, therefore, our explicit convergence rates provides a guarantee for the sampling. To see this, note that the underdamped Langevin SDE with $h(\theta) = -\nabla \log p(\theta|y_{1:M})$ converges to the extended target $\overline{\pi}$ whose θ-marginal of $\overline{\pi}$ is $p(\theta|y_{1:M})$, hence the underdamped Langevin SDE samples from the posterior distribution.

We note that, our setting specifically applies to cases where M is too large. More precisely, note that we have $h(\theta) = -\nabla \log p(\theta|y_{1:M}) = -\nabla \log \pi_0(\theta) - \sum_{i=1}^M \nabla \log \mathsf{p}(y_i|\theta)$. When M is too large, evaluating $h(\theta)$ is impractical. However, one can estimate the sum in the last term in an unbiased way. To be precise, consider random indices $i_1, \ldots, i_K \sim \{1, ..., M\}$ uniformly, then one can construct a stochastic gradient by using $\mathbf{u} = \{y_{i_1}, \ldots, y_{i_K}\}$ and obtaining $H(\theta, \mathbf{u}) = -\log \pi_0(\theta) - \frac{M}{K}\sum_{k=1}^K \nabla \log \mathsf{p}(y_{i_k}|\theta)$. Then, we have the following corollary follows from Theorem 2.1.

Corollary 6.1 *Assume that the log-posterior density* $\log p(\theta|y_{1:M})$*, its gradient, and stochastic gradient* $H(\theta, \cdot)$ *satisfy the Assumptions 2.1–2.4. Then,*

$$W_2(\mathcal{L}(\theta_n), p(\theta|y_{1:M})) \leq C_1^\star \eta^{1/2} + C_2^\star \eta^{1/4} + C_3^\star e^{-C_4^\star \eta n}.$$

where $C_1^\star, C_2^\star, C_3^\star, C_4^\star$ *are finite constants.*

This setting becomes practical under our assumptions, e.g., for the Bayesian logistic regression example. Consider the Gaussian mixture prior $\pi_0(\theta) \propto \exp(-f_0(\theta)) = e^{-|\theta-m|^2/2} + e^{-|\theta+m|^2/2}$ and the likelihood $\mathsf{p}(\mathbf{z}_i|\theta) = (1/(1+e^{-z_i^\top \theta}))^{y_i}(1-1/(1+e^{-z_i^\top \theta}))^{1-y_i}$ for $\theta \in \mathbb{R}^d$ and $\mathbf{z}_i = (z_i, y_i)$. Then, it is shown by Zhang et al. (2023b) that the stochastic gradient $H(\theta, \mathbf{u})$ for a mini-batch in this case satisfies assumptions 2.1–2.4. In particular, our theoretical guarantee in Theorem 2.1 and Corollary 6.1 apply to the Bayesian logistic regression case.

6.2 A generalization bound for machine learning

Leveraging standard results in machine learning literature, e.g., Raginsky et al. (2017), we can prove a generalization bound for the empirical risk minimization problem. Note that, many problems in machine learning can be written as a finite-sum minimization problem as

$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^d} U(\theta) := \frac{1}{M}\sum_{i=1}^M f(\theta, z_i). \tag{27}$$

Applying the result of Theorem 2.2, one can get a convergence guarantee on $\mathbb{E}[U(\theta_k^\eta)] - U_\star$. However, this does not account for the so-called *generalization error*. Note that, one can see the cost function in (27) as an empirical risk (expectation) minimization problem where the risk is given by $\mathbb{U}(\theta) := \int f(\theta, z)P(\mathrm{d}z) = \mathbb{E}[f(\theta, Z)]$, where $Z \sim P(\mathrm{d}z)$ is an unknown probability measure where the real-world data is sampled from. Therefore, in order to bound

the generalization error, one needs to bound the error $\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{U}_\star$. The generalization error can be decomposed as

$$\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{U}_\star = \underbrace{\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{E}[\mathbb{U}(\theta_\infty)]}_{\mathcal{B}_1} + \underbrace{\mathbb{E}[\mathbb{U}(\theta_\infty)] - \mathbb{E}[U(\theta_\infty)]}_{\mathcal{B}_2} + \underbrace{\mathbb{E}[U(\theta_\infty)] - \mathbb{U}_\star}_{\mathcal{B}_3}.$$

In what follows, we present a series of results bounding the terms $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$. By using the results about Gibbs distributions presented in Raginsky et al. (2017), one can bound $\mathcal{B}_1$ as follows.

Proposition 6.1 *Under the assumptions of Theorem 2.1, we obtain*

$$\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{E}[\mathbb{U}(\theta_\infty)] \leq \overline{C}_1^\star \eta^{1/2} + \overline{C}_2^\star \eta^{1/4} + \overline{C}_3^\star e^{-C_4^\star \eta n},$$

where $\overline{C}_i^\star = C_i^\star (C_m \bar{L}_1 + h_0)$ *for* $i = 1, 2, 3$ *and* $C_m^2 = \max(C_\theta^c, C_\theta)$.

The proof of Proposition 6.1 is similar to the proof of Proposition 5.1 and indeed the rates match.

Next, we seek a bound for the term $\mathcal{B}_2$. In order to prove the following result, we assume that Assumption 2.2 and Assumption 2.4 hold uniformly in x, as required by, e.g., Raginsky et al. (2017).

Proposition 6.2 *(Raginsky et al., 2017) Assume that Assumptions 2.1, 2.3 hold and Assumptions 2.2 and 2.4 hold uniformly in* x*, i.e.,* $|H(\theta, x)| \leq L_1' |\theta|^2 + B_1$*. Then,*

$$\mathbb{E}[\mathbb{U}(\theta_\infty)] - \mathbb{E}[U(\theta_\infty)] \leq \frac{4\beta c_{\mathrm{LS}}}{M} \left(\frac{L_1'}{a}(b + d/\beta) + B_1 \right),$$

where c_{LS} *is the constant of the logarithmic Sobolev inequality.*

Finally, let $\Theta^\star \in \arg\min_{\theta \in \mathbb{R}} \mathbb{U}(\theta)$. We note that $\mathcal{B}_3$ is bounded trivially as

$$\mathbb{E}[U(\theta_\infty)] - \mathbb{U}_\star = \mathbb{E}[U(\theta_\infty) - U_\star] + \mathbb{E}[U_\star - U(\Theta^\star)] \leq \mathbb{E}[U(\theta_\infty) - U_\star], \tag{28}$$

which follows from the proof of Proposition 5.2. Finally, Proposition 6.1, Proposition 6.2 and (28) leads to the following generalization bound presented as a corollary.

Corollary 6.2 *Under the setting of Proposition 6.2, we obtain the generalization bound for the SGHMC,*

$$\begin{aligned}\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{U}_\star \leq{}& \overline{C}_1^\star \eta^{1/2} + \overline{C}_2^\star \eta^{1/4} + \overline{C}_3^\star e^{-C_4^\star \eta n} + \frac{4\beta c_{\mathrm{LS}}}{M} \left(\frac{L_1'}{a}(b + d/\beta) + B_1 \right) \\ &+ \frac{d}{2\beta} \log \left(\frac{e\overline{L}_1}{a} \left(\frac{b\beta}{d} + 1 \right) \right).\end{aligned} \tag{29}$$

We note that this generalization bound improves that of Raginsky et al. (2017); Gao et al. (2022); Chau and Rásonyi (2022) due to our improved W_2 bound which is reflected in Theorem 2.2 and, consequently, Proposition 6.1. In particular, while the generalization bounds of Raginsky et al. (2017) and Gao et al. (2022) grow with the number of iterations and require careful tuning between the step-size and the number of iterations, our bound decreases with the number of iterations n. We also note that our bound improves that of Chau and Rásonyi (2022), similar to the W_2 bound.

7. Conclusions

We have analyzed the convergence of the SGHMC recursions (6)–(7) to the extended target measure $\overline{\pi}_\beta$ in Wasserstein-2 distance which implies the convergence of the law of the iterates $\mathcal{L}(\theta_n^\eta)$ to the target measure π_β in W_2. We have proved that the error bound scales like $\mathcal{O}(\eta^{1/4})$ where η is the step-size. This improves the existing bounds for the SGHMC significantly which are either growing with the number of iterations or include constants cannot be made to vanish by decreasing the step-size η. This bound on sampling from π_β enables us to prove a stochastic global optimization result when $(\theta_n^\eta)_{n\in\mathbb{N}}$ is viewed as an output of a nonconvex optimizer. We have shown that our results provide convergence rates for scalable Bayesian inference and we have particularized our results to the Bayesian logistic regression. Moreover, we have shown that our improvement of W_2 bounds are reflected in improved generalization bounds for the SGHMC.

Acknowledgments

Ö. D. A. is supported by the Lloyd's Register Foundation Data Centric Engineering Programme and EPSRC Programme Grant EP/R034710/1 (CoSInES). S.S. acknowledges support by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *29th International Conference on Machine Learning, ICML 2012*, pages 1591–1598, 2012.

Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

Mathias Barkhagen, Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

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

Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

Huy N Chau and Miklós Rásonyi. Stochastic gradient hamiltonian Monte Carlo for nonconvex learning. *Stochastic Processes and their Applications*, 149:341–368, 2022.

Huy N Chau, Chaman Kumar, Miklós Rásonyi, and Sotirios Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23: 217–244, 2019.

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. 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.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference On Learning Theory*, pages 300–323, 2018b.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of Langevin Monte Carlo from poincare to log-sobolev. In *Conference on Learning Theory*, pages 1–2. PMLR, 2022.

Arnak S Dalalyan. 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.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

AB Duncan, N Nüsken, and GA Pavliotis. Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, 169(6):1098–1131, 2017.

Alain Durmus, Eric Moulines, et al. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alain Durmus, Eric Moulines, et al. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

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

Murat A Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. Convergence of Langevin Monte Carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 8151–8175. PMLR, 2022.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian Monte Carlo for nonconvex stochastic optimization: nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5): 2931–2947, 2022.

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Chii-Ruey Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic runge-kutta accelerates Langevin Monte Carlo and beyond. *Advances in neural information processing systems*, 32, 2019.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is There an Analog of Nesterov Acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

Mateusz B Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *arXiv preprint arXiv:1808.07105*, 2018.

Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *Bernoulli*, 28(3):1577–1601, 2022.

Alireza Mousavi-Hosseini, Tyler Farghly, Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. Towards a complete analysis of Langevin Monte Carlo: Beyond poincar\'e inequality. *arXiv preprint arXiv:2303.03589*, 2023.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

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

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

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

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

Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A Erdogdu. Improved discretization analysis for underdamped Langevin Monte Carlo. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 36–71. PMLR, 2023a.

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25, 2023b.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems*, pages 3830–3841, 2019.

Appendix

Appendix A. Constants in Theorem 3.1

The constants of Theorem 3.1 are given as follows Eberle et al. (2019).

$$\dot{C} = 2e^{2+\Lambda}\frac{(1+\gamma)^2}{\min(1,\alpha)^2}\max\left(1, 4(1+2\alpha+2\alpha^2)\frac{(d+A_c)\gamma^{-1}\dot{c}^{-1}}{\min(1,R_1)\beta}\right)$$

and

$$\dot{c} = \frac{\gamma}{384}\min\left(\lambda\overline{L}_1\beta^{-1}\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}\overline{L}_1\beta^{-1}\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}\right)$$

where

$$\Lambda = \overline{L}_1 R_1^2/8 = \frac{12}{5}(1+2\alpha+2\alpha^2)(d+A_c)\overline{L}_1\beta^{-1}\gamma^{-2}\lambda^{-1}(1-2\lambda)^{-1},$$

and $\alpha \in (0,\infty)$.

Appendix B. Additional lemmata

We prove the following lemma adapted from Raginsky et al. (2017).

Lemma B.1 *For all $\theta \in \mathbb{R}^d$,*

$$\frac{a}{3}|\theta|^2 - \frac{b}{2}\log 3 \leq U(\theta) \leq u_0 + \frac{\bar{L}_1}{2}|\theta|^2 + h_0|\theta|.$$

where $u_0 = U(0)$ and $\bar{L}_1 = L_1\mathbb{E}[(1+|X_0|)^\rho]$.

Proof See Appendix C.4. ■

Lemma B.2 *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be sigma-algebras. Consider two $\mathbb{R}^d$-valued random vectors, denoted X, Y, in L^p with $p \geq 1$ such that Y is measurable w.r.t. $\mathcal{H} \vee \mathcal{G}$. Then,*

$$\mathbb{E}^{1/p}[|X - \mathbb{E}[X|\mathcal{H}\vee\mathcal{G}]|^p|\mathcal{G}] \leq 2\mathbb{E}^{1/p}[|X-Y|^p|\mathcal{G}].$$

Proof See Lemma 6.1 in Chau et al. (2019). ■

Lemma B.3 *Let Assumption 2.1, 2.3, 2.2 and 2.4 hold. For any $k = 1, \ldots, K+1$ where $K+1 \leq T$, we obtain*

$$\sup_{n\in\mathbb{N}}\sup_{t\in[nT,(n+1)T]}\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\eta,n}) - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2\right] \leq \sigma_H$$

where

$$\sigma_H := 8L_2^2\sigma_Z(1+C_\zeta) < \infty,$$

where $\sigma_Z = \mathbb{E}[(1+|X_0|+|\mathbb{E}[X_0]|)^{2\rho}|X_0 - \mathbb{E}[X_0]|^2] < \infty$.

Proof See Appendix C.5. ■

Lemma B.4 *Under Assumptions 2.1–2.4*

$$\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t\rfloor} - \overline{V}^{\eta}_{t}\right|^2\right] \leq \sigma_V \eta,$$

where

$$\sigma_V = 4\eta\gamma^2 C_v + 4\eta(\widetilde{L}_1 C_\theta + \widetilde{C}_1) + 4\gamma\beta^{-1}d.$$

Moreover,

$$\mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor t\rfloor} - \overline{\theta}^{\eta}_{t}\right|^2\right] \leq \eta^2 C_v.$$

Proof See Appendix C.6. ■

Next, it is shown that a key assumption appearing in Eberle et al. (2019) holds.

Lemma B.5 *There exist constants $A_c \in (0,\infty)$ and $\lambda \in (0, 1/4]$ such that*

$$\langle \theta, h(\theta)\rangle \geq 2\lambda\left(U(\theta) + \gamma^2|\theta|^2/4\right) - 2A_c/\beta \tag{30}$$

for all $x \in \mathbb{R}^d$.

Proof See Appendix C.7. ■

Appendix C. Proofs of the preliminary results

C.1 Proof of Lemma 3.2

For this proof, we use the Lyapunov function defined by Eberle et al. (2019) and follow a similar proof presented in Gao et al. (2022). We first define the Lyapunov function as

$$\mathcal{V}(\theta, v) = \beta U(\theta) + \frac{\beta}{4}\gamma^2\left(\|\theta + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|\theta\|^2\right).$$

Next, we will use this Lyapunov function to show that the second moments of the processes $(V^\eta_n)_{n\in\mathbb{N}}$ and $(\theta^\eta_n)_{n\in\mathbb{N}}$ are finite.

We start by defining

$$M_2(k) = \mathbb{E}\mathcal{V}(\theta^\eta_k, V^\eta_k)/\beta = \mathbb{E}\left[U(\theta^\eta_k) + \frac{\gamma^2}{4}\left(|\theta^\eta_k + \gamma^{-1}V^\eta_k|^2 + |\gamma^{-1}V^\eta_k|^2 - \lambda|\theta^\eta_k|^2\right)\right]. \tag{31}$$

Recall our discrete-time recursions (6)–(7)

$$\begin{aligned} V^\eta_{k+1} &= (1-\eta\gamma)V^\eta_k - \eta H(\theta^\eta_k, X_{k+1}) + \sqrt{2\eta\gamma\beta^{-1}}\xi_{k+1}, \\ \theta^\eta_{k+1} &= \theta^\eta_k + \eta V^\eta_k, \qquad \theta^\eta_0 = \theta_0, \qquad V^\eta_0 = v_0, \end{aligned}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Normal random variables. Consequently, we have the equality

$$\begin{aligned} \mathbb{E}\left[|V^\eta_{k+1}|^2\right] &= \mathbb{E}\left[|(1-\gamma\eta)V^\eta_k - \eta H(\theta^\eta_k, X_{k+1})|^2\right] + 2\gamma\eta\beta^{-1}d, \\ &= (1-\gamma\eta)^2\mathbb{E}\left[|V^\eta_k|^2\right] - 2\eta(1-\gamma\eta)\mathbb{E}\left[\langle V^\eta_k, h(\theta^\eta_k)\rangle\right] + \eta^2\mathbb{E}\left[|H(\theta^\eta_k, X_{k+1})|^2\right] + 2\gamma\eta\beta^{-1}d, \end{aligned}$$

which immediately leads to

$$\mathbb{E}\left[|V^\eta_{k+1}|^2\right] \leq (1-\gamma\eta)^2\mathbb{E}\left[|V^\eta_k|^2\right] - 2\eta(1-\gamma\eta)\mathbb{E}\left[\langle V^\eta_k, h(\theta^\eta_k)\rangle\right] + \eta^2\left[\widetilde{L}_1\mathbb{E}\left[|\theta^\eta_k|^2\right] + \widetilde{C}_1\right] + 2\gamma\eta\beta^{-1}d, \tag{32}$$

where

$$\widetilde{L}_1 = 2L_1^2 C_\rho \qquad \text{and} \qquad \widetilde{C}_1 = 4L_2^2 C_\rho + 4H_0^2. \tag{33}$$

Next, we note that

$$\mathbb{E}\left|\theta_{k+1}^{\eta}\right|^2 = \mathbb{E}|\theta_k^{\eta}|^2 + 2\eta\mathbb{E}\langle\theta_k^{\eta}, V_k^{\eta}\rangle + \eta^2\mathbb{E}\left|V_k^{\eta}\right|^2. \tag{34}$$

Recall $h := \nabla U$ and note also that

$$U(\theta_{k+1}^{\eta}) = U(\theta_k^{\eta} + \eta V_k^{\eta}) = U(\theta_k^{\eta}) + \int_0^1 \langle h(\theta_k^{\eta} + \tau\eta V_k^{\eta}), \eta V_k^{\eta}\rangle \mathrm{d}\tau,$$

which suggests

$$\begin{aligned}\left|U(\theta_{k+1}^{\eta}) - U(\theta_k^{\eta}) - \langle h(\theta_k^{\eta}), \eta V_k^{\eta})\rangle\right| &= \left|\int_0^1 \langle h(\theta_k^{\eta} + \tau\eta V_k^{\eta}) - h(\theta_k^{\eta}), \eta V_k^{\eta}\rangle \mathrm{d}\tau\right|, \\ &\leq \int_0^1 \left|h(\theta_k^{\eta} + \tau\eta V_k^{\eta}) - h(\theta_k^{\eta})\right| \left|\eta V_k^{\eta}\right| \mathrm{d}\tau, \\ &\leq \frac{1}{2} L_1 C_\rho \eta^2 \left|V_k^{\eta}\right|^2,\end{aligned}$$

where the second line follows from the Cauchy-Schwarz inequality and the final line follows from (9). Finally we obtain

$$\mathbb{E}U(\theta_{k+1}^{\eta}) - \mathbb{E}U(\theta_k^{\eta}) \leq \eta\mathbb{E}\langle h(\theta_k^{\eta}), V_k^{\eta}\rangle + \frac{1}{2} L_1 C_\rho \eta^2 \mathbb{E}\left|V_k^{\eta}\right|^2. \tag{35}$$

Next, we continue computing

$$\begin{aligned}\mathbb{E}\left|\theta_{k+1}^{\eta} + \gamma^{-1}V_{k+1}^{\eta}\right|^2 &= \mathbb{E}\left|\theta_k^{\eta} + \gamma^{-1}V_k^{\eta} - \eta\gamma^{-1}H(\theta_k^{\eta}, X_{k+1})\right|^2 + 2\gamma^{-1}\beta^{-1}\eta d, \\ &= \mathbb{E}\left|\theta_k^{\eta} + \gamma^{-1}V_k^{\eta}\right|^2 - 2\eta\gamma^{-1}\mathbb{E}\langle\theta_k^{\eta} + \gamma^{-1}V_k^{\eta}, h(\theta_k^{\eta})\rangle \\ &+ \eta^2\gamma^{-2}\mathbb{E}\left|H(\theta_k^{\eta}, X_{k+1})\right|^2 + 2\gamma^{-1}\eta\beta^{-1}d, \\ &\leq \mathbb{E}\left|\theta_k^{\eta} + \gamma^{-1}V_k^{\eta}\right|^2 - 2\eta\gamma^{-1}\mathbb{E}\langle\theta_k^{\eta} + \gamma^{-1}V_k^{\eta}, h(\theta_k^{\eta})\rangle \\ &+ \eta^2\gamma^{-2}(\widetilde{L}_1\mathbb{E}\left|\theta_k^{\eta}\right|^2 + \widetilde{C}_1) + 2\gamma^{-1}\eta\beta^{-1}d.\end{aligned} \tag{36}$$

where $\widetilde{L}_1$ and $\widetilde{C}_1$ is defined as in (33). Next, combining (32), (34), (35), (36),

$$\begin{aligned}&M_2(k+1) - M_2(k) \\ &\quad= \mathbb{E}\left[U(\theta_{k+1}^{\eta}) - U(\theta_k^{\eta})\right] + \frac{\gamma^2}{4}\left(\mathbb{E}\left|\theta_{k+1}^{\eta} + \gamma^{-1}V_{k+1}^{\eta}\right|^2 - \mathbb{E}\left|\theta_k^{\eta} + \gamma^{-1}V_k^{\eta}\right|^2\right) \\ &\quad+ \frac{1}{4}\left(\mathbb{E}\left|V_{k+1}^{\eta}\right|^2 - \mathbb{E}\left|V_k^{\eta}\right|^2\right) - \frac{\gamma^2\lambda}{4}\left(\mathbb{E}\left|\theta_{k+1}^{\eta}\right|^2 - \mathbb{E}\left|\theta_k^{\eta}\right|^2\right), \\ &\quad\leq \eta\mathbb{E}\langle h(\theta_k^{\eta}), V_k^{\eta}\rangle + \frac{L_1 C_\rho \eta^2}{2}\mathbb{E}\left|V_k^{\eta}\right|^2 \\ &\quad+ \frac{\gamma^2}{4}\left(-2\eta\gamma^{-1}\mathbb{E}\langle\theta_k^{\eta} + \gamma^{-1}V_k^{\eta}, h(\theta_k^{\eta})\rangle + \eta^2\gamma^{-2}\left(\widetilde{L}_1\mathbb{E}\left|\theta_k^{\eta}\right|^2 + \widetilde{C}_1\right) + 2\gamma^{-1}\beta^{-1}\eta d\right) \\ &\quad+ \frac{1}{4}\left((-2\gamma\eta + \gamma^2\eta^2)\mathbb{E}\left|V_k^{\eta}\right|^2 - 2\eta(1-\gamma\eta)\mathbb{E}\langle V_k^{\eta}, h(\theta_k^{\eta})\rangle + \eta^2\left(\widetilde{L}_1\mathbb{E}|\theta_k^{\eta}|^2 + \widetilde{C}_1\right) + 2\gamma\eta\beta^{-1}d\right) \\ &\quad- \frac{\gamma^2\lambda}{4}\left(2\eta\mathbb{E}\langle\theta_k^{\eta}, V_k^{\eta}\rangle + \eta^2\mathbb{E}|V_k^{\eta}|^2\right), \\ &\quad= -\frac{\eta\gamma}{2}\mathbb{E}\langle\theta_k^{\eta}, h(\theta_k^{\eta})\rangle + \frac{\gamma\eta^2}{2}\mathbb{E}\langle h(\theta_k^{\eta}), V_k^{\eta}\rangle + \left(\frac{L_1 C_\rho \eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4}\right)\mathbb{E}\left|V_k^{\eta}\right|^2 \\ &\quad+ \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\left|\theta_k^{\eta}\right|^2 - \frac{\gamma^2\eta\lambda}{2}\mathbb{E}\langle\theta_k^{\eta}, V_k^{\eta}\rangle + \frac{\widetilde{C_1}\eta^2}{2} + \gamma\eta\beta^{-1}d, \\ &\quad\leq -\eta\gamma\lambda\mathbb{E}U(\theta_k^{\eta}) - \frac{\lambda\gamma^3\eta}{4}\mathbb{E}|\theta_k^{\eta}|^2 + A_c\eta\gamma\beta^{-1} + \frac{\gamma\eta^2}{2}\mathbb{E}\langle h(\theta_k^{\eta}), V_k^{\eta}\rangle + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\left|\theta_k^{\eta}\right|^2 \\ &\quad+ \left(\frac{L_1 C_\rho \eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4}\right)\mathbb{E}\left|V_k^{\eta}\right|^2 - \frac{\gamma^2\eta\lambda}{2}\mathbb{E}\langle\theta_k^{\eta}, V_k^{\eta}\rangle + \frac{\widetilde{C_1}\eta^2}{2} + \gamma\eta\beta^{-1}d.\end{aligned}$$

where the last line is obtained using (30). Next, using the fact that $0 < \lambda \leq 1/4$ and the form of the Lyapunov function (31), we obtain

$$-\frac{\gamma}{2}\mathbb{E}\langle \theta_k^\eta, V_k^\eta\rangle \leq -M_2(k) + \mathbb{E}U(\theta_k^\eta) + \frac{\gamma^2}{4}\mathbb{E}|\theta_k^\eta|^2 + \frac{1}{2}\mathbb{E}|V_k^\eta|^2.$$

Using this, we can obtain

$$\begin{aligned} M_2(k+1) - M_2(k) &\leq A_c\eta\gamma\beta^{-1} + \frac{\gamma\eta^2}{2}\mathbb{E}\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\,|\theta_k^\eta|^2 + \gamma\eta\beta^{-1}d \\ &+ \left(\frac{L_1C_\rho\eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4} + \frac{\gamma\lambda\eta}{2}\right)\mathbb{E}\,|V_k^\eta|^2 + \frac{\widetilde{C_1}\eta^2}{2} - \gamma\lambda\eta M_2(k). \end{aligned}$$

Next, reorganizing and using $\langle a, b\rangle \leq (|a|^2 + |b|^2)/2$

$$\begin{aligned} M_2(k+1) &\leq (1-\gamma\lambda\eta)M_2(k) + A_c\eta\gamma\beta^{-1} + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\,|\theta_k^\eta|^2 + \frac{\widetilde{C_1}\eta^2}{2} + \gamma\eta\beta^{-1}d \\ &+ \left(\frac{L_1C_\rho\eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4} + \frac{\gamma\lambda\eta}{4} + \frac{\gamma\eta^2}{4}\right)\mathbb{E}\,|V_k^\eta|^2 + \frac{\gamma\eta^2}{4}\mathbb{E}|h(\theta_k^\eta)|^2, \\ &\leq (1-\gamma\lambda\eta)M_2(k) + A_c\eta\gamma\beta^{-1} + \eta^2\left(\frac{\widetilde{L}_1}{2} + \frac{\gamma}{2}L_1^2C_\rho^2\right)\mathbb{E}\,|\theta_k^\eta|^2 + \frac{\widetilde{C_1}\eta^2}{2} + \gamma\eta\beta^{-1}d \\ &+ \eta^2\left(\frac{L_1C_\rho}{2} + \frac{\gamma^2}{4} - \frac{\gamma^2\lambda}{4} + \frac{\gamma}{4}\right)\mathbb{E}\,|V_k^\eta|^2 + \frac{\gamma\eta^2h_0^2}{2}, \end{aligned}$$

where the last inequality follows since $\lambda \leq 1/4$ and (10). We note that

$$\mathcal{V}(\theta, v) \geq \max\left\{\frac{1}{8}(1-2\lambda)\beta\gamma^2|\theta|^2, \frac{\beta}{4}(1-2\lambda)|v|^2\right\},$$

which implies by the definition of $M_2(k)$ that

$$\begin{aligned} M_2(k) &\geq \max\left\{\frac{1}{8}(1-2\lambda)\gamma^2\mathbb{E}|\theta_k^\eta|^2, \frac{1}{4}(1-2\lambda)\mathbb{E}|V_k^\eta|^2\right\}, \\ &\geq \frac{1}{16}(1-2\lambda)\gamma^2\mathbb{E}|\theta_k^\eta|^2 + \frac{1}{8}(1-2\lambda)\mathbb{E}|V_k^\eta|^2, \end{aligned} \tag{37}$$

since $\max\{x, y\} \geq (x+y)/2$ for any $x, y > 0$. Therefore, we obtain

$$M_2(k+1) \leq (1 - \gamma\lambda\eta + K_1\eta^2)M_2(k) + K_2\eta^2 + K_3\eta$$

where

$$K_1 := \max\left\{\frac{\frac{L_1C_\rho}{2} + \frac{\gamma^2}{4} - \frac{\gamma^2\lambda}{4} + \frac{\gamma}{4}}{\frac{1}{8}(1-2\lambda)}, \frac{\frac{\widetilde{L}_1}{2} + \frac{\gamma}{2}L_1^2C_\rho^2}{\frac{1}{16}(1-2\lambda)\gamma^2}\right\}$$

and

$$K_2 = \frac{\widetilde{C_1} + \gamma h_0^2}{2} \quad \text{and} \quad K_3 = (A_c + d)\gamma\beta^{-1}.$$

For $0 < \eta \leq \min\left\{\frac{K_3}{K_2}, \frac{\gamma\lambda}{2K_1}, \frac{2}{\gamma\lambda}\right\}$, we obtain

$$M_2(k+1) \leq \left(1 - \frac{\gamma\lambda\eta}{2}\right)M_2(k) + 2K_3\eta$$

which implies

$$M_2(k) \leq M_2(0) + \frac{4}{\gamma\lambda}K_3.$$

Combining this with (37) gives the result.

C.2 Proof of Lemma 3.3

We recall that $\overline{\zeta}_t^{\eta,n}$ is the Langevin diffusion started at θ_{nT}^η and run until $t \in (nT,(n+1)T)$. First notice that Lemma 3.1 implies

$$\sup_{t\in(nT,(n+1)T)} \mathbb{E}|\zeta_t^{\eta,n}|^2 \leq \frac{\mathbb{E}\mathcal{V}(\theta_{nT}^\eta, V_{nT}^\eta) + \frac{4(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \tag{38}$$

which, noting that, $\mathbb{E}\mathcal{V}(\theta_{nT}^\eta, V_{nT}^\eta) = \beta M_2(nT)$, implies

$$\sup_{t\in(nT,(n+1)T)} \mathbb{E}|\zeta_t^{\eta,n}|^2 \leq \frac{\beta M_2(0) + \frac{8(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}.$$

Substituting $M_2(0)$ gives

$$\sup_{t\in(nT,(n+1)T)} \mathbb{E}|\zeta_t^{\eta,n}|^2 \leq C_\zeta := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta,v)\mu_0(\mathrm{d}\theta,\mathrm{d}v) + \frac{8(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}.$$

C.3 Proof of Lemma 3.4

We need to obtain the contraction of $\mathbb{E}\mathcal{V}^2(\theta_k^\eta, V_k^\eta)$. Recall again the Lyapunov function defined by Eberle et al. (2019)

$$\mathcal{V}(\theta,v) = \beta U(\theta) + \frac{\beta}{4}\gamma^2\left(\|\theta+\gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|\theta\|^2\right).$$

In this proof, we follow a similar strategy to the proof of Lemma 3.2 and approaches of Gao et al. (2022) and Chau and Rásonyi (2022). We define the notation $\mathcal{V}_k = \mathcal{V}(\theta_k^\eta, V_k^\eta)$ and note

$$\mathcal{V}_k/\beta := \mathcal{V}(\theta_k^\eta, V_k^\eta)/\beta = U(\theta_k^\eta) + \frac{\gamma^2}{4}\left(|\theta_k^\eta + \gamma^{-1}V_k^\eta|^2 + |\gamma^{-1}V_k^\eta|^2 - \lambda|\theta_k^\eta|^2\right). \tag{39}$$

Recall our discrete-time recursions (6)–(7)

$$\begin{aligned} V_{k+1}^\eta &= (1-\eta\gamma)V_k^\eta - \eta H(\theta_k^\eta, X_{k+1}) + \sqrt{2\eta\gamma\beta^{-1}}\xi_{k+1}, \\ \theta_{k+1}^\eta &= \theta_k^\eta + \eta V_k^\eta, \qquad \theta_0^\eta = \theta_0, \qquad V_0^\eta = v_0, \end{aligned}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. Normal random variables. We first define $\Delta_k^1 = (1-\eta\gamma)V_k^\eta - \eta\, H(\theta_k^\eta, X_{k+1})$ and write

$$\begin{aligned} |V_{k+1}^\eta|^2 &= |\Delta_k^1|^2 + 2\sqrt{2\eta\gamma\beta^{-1}}\langle\Delta_k^1, \xi_{k+1}\rangle + 2\gamma\eta\beta^{-1}|\xi_{k+1}|^2, \\ &= |(1-\gamma\eta)V_k^\eta - \eta H(\theta_k^\eta, X_{k+1})|^2 + 2\sqrt{2\eta\gamma\beta^{-1}}\langle\Delta_k^1, \xi_{k+1}\rangle + 2\gamma\eta\beta^{-1}|\xi_{k+1}|^2, \\ &= (1-\gamma\eta)^2\,|V_k^\eta|^2 - 2\eta(1-\gamma\eta)\langle V_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle \\ &+ \eta^2\,|H(\theta_k^\eta, X_{k+1})|^2 + 2\sqrt{2\eta\gamma\beta^{-1}}\langle\Delta_k^1, \xi_{k+1}\rangle + 2\gamma\eta\beta^{-1}|\xi_{k+1}|^2. \end{aligned} \tag{40}$$

Next, we note that

$$\left|\theta_{k+1}^\eta\right|^2 = |\theta_k^\eta|^2 + 2\eta\langle\theta_k^\eta, V_k^\eta\rangle + \eta^2\,|V_k^\eta|^2. \tag{41}$$

Recall $h := \nabla U$ and note also that

$$U(\theta_{k+1}^\eta) = U(\theta_k^\eta + \eta V_k^\eta) = U(\theta_k^\eta) + \int_0^1 \langle h(\theta_k^\eta + \tau\eta V_k^\eta), \eta V_k^\eta\rangle \mathrm{d}\tau,$$

which suggests

$$\begin{aligned} \left|U(\theta_{k+1}^\eta) - U(\theta_k^\eta) - \langle h(\theta_k^\eta), \eta V_k^\eta)\rangle\right| &= \left|\int_0^1 \langle h(\theta_k^\eta + \tau\eta V_k^\eta) - h(\theta_k^\eta), \eta V_k^\eta\rangle \mathrm{d}\tau\right|, \\ &\leq \int_0^1 |h(\theta_k^\eta + \tau\eta V_k^\eta) - h(\theta_k^\eta)|\,|\eta V_k^\eta|\,\mathrm{d}\tau, \\ &\leq \frac{1}{2}L_1 C_\rho \eta^2\,|V_k^\eta|^2, \end{aligned}$$

where, similarly as in the previous proof, the second line follows from the Cauchy-Schwarz inequality and the final line follows from (9). Finally we arrive at

$$U(\theta_{k+1}^\eta) - U(\theta_k^\eta) \leq \eta\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{1}{2}L_1 C_\rho \eta^2 \left|V_k^\eta\right|^2. \tag{42}$$

Next, we note that $\Delta_k^2 = \theta_k^\eta + \gamma^{-1}V_k^\eta - \eta\gamma^{-1}H(\theta_k^\eta, X_{k+1})$ and continue computing

$$\begin{aligned}\left|\theta_{k+1}^\eta + \gamma^{-1}V_{k+1}^\eta\right|^2 &= \left|\theta_k^\eta + \gamma^{-1}V_k^\eta - \eta\gamma^{-1}H(\theta_k^\eta, X_{k+1})\right|^2 + 2\sqrt{2\gamma^{-1}\beta^{-1}\eta}\langle\Delta_k^2, \xi_{k+1}\rangle + 2\gamma^{-1}\beta^{-1}\eta|\xi_{k+1}|^2,\\ &= \left|\theta_k^\eta + \gamma^{-1}V_k^\eta\right|^2 - 2\eta\gamma^{-1}\langle\theta_k^\eta + \gamma^{-1}V_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle\\ &\quad + 2\sqrt{2\gamma^{-1}\beta^{-1}\eta}\langle\Delta_k^2, \xi_{k+1}\rangle + \eta^2\gamma^{-2}\left|H(\theta_k^\eta, X_{k+1})\right|^2 + 2\gamma^{-1}\eta\beta^{-1}|\xi_{k+1}|^2,\end{aligned} \tag{43}$$

Next, combining (40), (41), (42), (43),

$$\begin{aligned}&\frac{\mathcal{V}_{k+1} - \mathcal{V}_k}{\beta} = \left(U(\theta_{k+1}^\eta) - U(\theta_k^\eta)\right) + \frac{\gamma^2}{4}\left(\left|\theta_{k+1}^\eta + \gamma^{-1}V_{k+1}^\eta\right|^2 - \left|\theta_k^\eta + \gamma^{-1}V_k^\eta\right|^2\right)\\ &+ \frac{1}{4}\left(\left|V_{k+1}^\eta\right|^2 - |V_k^\eta|^2\right) - \frac{\gamma^2\lambda}{4}\left(\left|\theta_{k+1}^\eta\right|^2 - |\theta_k^\eta|^2\right),\\ &\leq \eta\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{1}{2}L_1C_\rho\eta^2\left|V_k^\eta\right|^2 + \frac{\gamma^2}{4}\left(-2\eta\gamma^{-1}\langle\theta_k^\eta + \gamma^{-1}V_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle\right.\\ &\left.+\eta^2\gamma^{-2}\left|H(\theta_k^\eta, X_{k+1})\right|^2 + 2\gamma^{-1}\eta\beta^{-1}|\xi_{k+1}|^2\right) + \frac{1}{4}\left((-2\gamma\eta + \gamma^2\eta^2)\left|V_k^\eta\right|^2 - 2\eta(1-\gamma\eta)\langle V_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle\right.\\ &\left.+\eta^2\left|H(\theta_k^\eta, X_{k+1})\right|^2 + 2\gamma\eta\beta^{-1}|\xi_{k+1}|^2\right) - \frac{\gamma^2\lambda}{4}\left(2\eta\langle\theta_k^\eta, V_k^\eta\rangle + \eta^2\left|V_k^\eta\right|^2\right) + \Sigma_k,\\ &\leq -\frac{\eta\gamma}{2}\langle\theta_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle + \eta\langle h(\theta_k^\eta), V_k^\eta\rangle + \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4}\right)\left|V_k^\eta\right|^2 + \frac{\eta^2}{2}\left|H(\theta_k^\eta, X_{k+1})\right|^2\\ &+ \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \left(\frac{\gamma\eta^2}{2} - \eta\right)\langle V_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle - \frac{\gamma^2\lambda\eta}{2}\langle\theta_k^\eta, V_k^\eta\rangle + \Sigma_k\end{aligned}$$

where

$$\Sigma_k = \frac{\gamma^2}{2}\sqrt{2\gamma^{-1}\beta^{-1}\eta}\langle\Delta_k^2, \xi_{k+1}\rangle + \frac{1}{2}\sqrt{2\eta\gamma\beta^{-1}}\langle\Delta_k^1, \xi_{k+1}\rangle.$$

Next, using the fact that $0 < \lambda \leq 1/4$ and the form of the Lyapunov function (39), we obtain

$$-\frac{\gamma}{2}\langle\theta_k^\eta, V_k^\eta\rangle \leq -\frac{\mathcal{V}_k}{\beta} + U(\theta_k^\eta) + \frac{\gamma^2}{4}|\theta_k^\eta|^2 + \frac{1}{2}|V_k^\eta|^2.$$

Using this and merging some terms, we obtain

$$\begin{aligned}\frac{\mathcal{V}_{k+1} - \mathcal{V}_k}{\beta} &\leq -\frac{\eta\gamma}{2}\langle\theta_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle + \eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle\\ &+ \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4}\right)\left|V_k^\eta\right|^2 + \frac{\eta^2}{2}\left|H(\theta_k^\eta, X_{k+1})\right|^2 + \eta\gamma\beta^{-1}|\xi_{k+1}|^2\\ &+ \frac{\gamma\eta^2}{2}\langle V_k^\eta, H(\theta_k^\eta, X_{k+1})\rangle - \gamma\lambda\eta\frac{\mathcal{V}_k}{\beta} + \gamma\lambda\eta U(\theta_k^\eta) + \frac{\gamma^3\lambda\eta}{4}|\theta_k^\eta|^2 + \frac{\gamma\lambda\eta}{2}|V_k^\eta|^2 + \Sigma_k.\end{aligned} \tag{44}$$

Next, by reorganizing and using $\langle a, b\rangle \leq (|a|^2 + |b|^2)/2$, we arrive at

$$\begin{aligned}&\frac{\mathcal{V}_{k+1}}{\beta} \leq (1-\gamma\lambda\eta)\frac{\mathcal{V}_k}{\beta} - \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta)\rangle + \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})\rangle + \eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle\\ &+ \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2}\right)\left|V_k^\eta\right|^2 + \frac{\eta^2}{2}\left|H(\theta_k^\eta, X_{k+1})\right|^2 + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \gamma\lambda\eta U(\theta_k^\eta) + \frac{\gamma^3\lambda\eta}{4}|\theta_k^\eta|^2\\ &+ \Sigma_k.\end{aligned} \tag{45}$$

Using (30), and $\lambda \leq 1/4$, we obtain

$$\begin{aligned}
&\frac{\mathcal{V}_{k+1}}{\beta} \leq (1-\gamma\lambda\eta)\frac{\mathcal{V}_k}{\beta} - \eta\gamma\lambda U(\theta_k^\eta) - \frac{\gamma^3\eta\lambda}{4}|\theta_k^\eta|^2 + \eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})\rangle \\
&+ \eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle + \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2}\right)|V_k^\eta|^2 + \frac{\eta^2}{2}|H(\theta_k^\eta, X_{k+1})|^2 \\
&+ \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \gamma\lambda\eta U(\theta_k^\eta) + \frac{\gamma^3\lambda\eta}{4}|\theta_k^\eta|^2 + \Sigma_k, \\
&= (1-\gamma\lambda\eta)\frac{\mathcal{V}_k}{\beta} + \eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})\rangle + \eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle \\
&+ \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2}\right)|V_k^\eta|^2 + \frac{\eta^2}{2}|H(\theta_k^\eta, X_{k+1})|^2 + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k, \\
&\leq (1-\gamma\lambda\eta)\frac{\mathcal{V}_k}{\beta} + \eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})\rangle + \eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle \\
&+ \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2}\right)|V_k^\eta|^2 \\
&+ \frac{\eta^2}{2}\left(2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k,
\end{aligned}$$

by Remark 2.1. Let $\phi = (1-\gamma\lambda\eta)$ and let $\mathcal{H}_k = \mathcal{G}_k \vee \sigma(\xi_1, \ldots, \xi_k)$. By using $\lambda \leq 1/4$, we obtain

$$\begin{aligned}
&\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} \leq \phi^2\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta}\left(\eta\gamma A_c\beta^{-1} + \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} + \frac{\gamma^2\eta^2}{4}\right)|V_k^\eta|^2\right. \\
&\left.+\frac{\eta^2}{2}\left(2L_1^2C_\rho|\theta_k^\eta|^2 + 4L_2^2C_\rho + 4H_0^2\right) + \eta\gamma\beta^{-1}d\right) + \mathbb{E}\left[\left(\eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})\rangle\right.\right. \\
&+ \eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle + \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2}\right)|V_k^\eta|^2 \\
&\left.\left.+\frac{\eta^2}{2}\left(2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k\right)^2\Big|\mathcal{H}_k\right].
\end{aligned}$$

We first note that

$$\begin{aligned}
\frac{\mathcal{V}_k}{\beta} &\geq \max\left\{\frac{1}{8}(1-2\lambda)\gamma^2|\theta_k^\eta|^2, \frac{1}{4}(1-2\lambda)|V_k^\eta|^2\right\}, \\
&\geq \frac{1}{16}(1-2\lambda)\gamma^2|\theta_k^\eta|^2 + \frac{1}{8}(1-2\lambda)|V_k^\eta|^2,
\end{aligned} \tag{46}$$

since $\max\{x, y\} \geq (x+y)/2$ for any $x, y > 0$. Then we first obtain

$$\begin{aligned}
&\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} \leq (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta}\left(\eta\gamma A_c\beta^{-1} + 2\eta^2L_2C_\rho + 2\eta^2H_0^2 + \eta\gamma\beta^{-1}d\right) \\
&+ \mathbb{E}\left[\left(\eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{2}\langle\theta_k^\eta, h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})\rangle\right.\right. \\
&+\eta\langle h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1}), V_k^\eta\rangle + \left(\frac{L_1C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2}\right)|V_k^\eta|^2 \\
&\left.\left.+\frac{\eta^2}{2}\left(2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k\right)^2\Big|\mathcal{H}_k\right]
\end{aligned}$$

where

$$\tilde{K}_1 := \max\left\{\frac{\frac{L_1C_\rho}{2} + \frac{\gamma^2}{4} - \frac{\gamma^2\lambda}{4}}{\frac{1}{8}(1-2\lambda)}, \frac{L_1^2C_\rho}{\frac{1}{16}(1-2\lambda)\gamma^2}\right\}$$

Next using $\langle a, b\rangle \leq (|a|^2 + |b|^2)/2$, we obtain

$$
\begin{aligned}
&\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} \leq (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta}\left(\eta\gamma A_c\beta^{-1} + 2\eta^2 L_2 C_\rho + 2\eta^2 H_0^2 + \eta\gamma\beta^{-1}d\right)\\
&+\mathbb{E}\Bigg[\Bigg(\eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{4}|\theta_k^\eta|^2 + \left(\frac{\eta\gamma}{4} + \frac{\eta}{2}\right)|h(\theta_k^\eta) - H(\theta_k^\eta, X_{k+1})|^2\\
&+\left(\frac{L_1 C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2} + \frac{\eta}{2}\right)|V_k^\eta|^2\\
&+\frac{\eta^2}{2}\left(2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k\Bigg)^2\Bigg|\mathcal{H}_k\Bigg]\\
&\leq (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta}\left(\eta\gamma A_c\beta^{-1} + 2\eta^2 L_2 C_\rho + 2\eta^2 H_0^2 + \eta\gamma\beta^{-1}d\right)\\
&+\mathbb{E}\Bigg[\Bigg(\eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{4}|\theta_k^\eta|^2 + \left(\frac{\eta\gamma}{2} + \eta\right)(|h(\theta_k^\eta)|^2 + |H(\theta_k^\eta, X_{k+1})|^2)\\
&+\left(\frac{L_1 C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2} + \frac{\eta}{2}\right)|V_k^\eta|^2\\
&+\frac{\eta^2}{2}\left(2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k\Bigg)^2\Bigg|\mathcal{H}_k\Bigg]
\end{aligned}
$$

where we have used $|a+b|^2 \leq 2|a|^2 + 2|b|^2$. Now using Remark 2.1, we obtain

$$
\begin{aligned}
&\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} \leq (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta}\left(\eta\gamma A_c\beta^{-1} + 2\eta^2 L_2 C_\rho + 2\eta^2 H_0^2 + \eta\gamma\beta^{-1}d\right)\\
&+\mathbb{E}\Bigg[\Bigg(\eta\gamma A_c\beta^{-1} + \frac{\eta\gamma}{4}|\theta_k^\eta|^2 + \left(\frac{\eta\gamma}{2} + \eta\right)\left(2L_1^2C_\rho^2|\theta_k^\eta|^2 + 2h_0^2 + 2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2\right.\\
&\left.+4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \left(\frac{L_1 C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2} + \frac{\eta}{2}\right)|V_k^\eta|^2\\
&+\frac{\eta^2}{2}\left(2L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k\Bigg)^2\Bigg|\mathcal{H}_k\Bigg]\\
&= (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta}\left(\eta\gamma A_c\beta^{-1} + 2\eta^2 L_2 C_\rho + 2\eta^2 H_0^2 + \eta\gamma\beta^{-1}d\right)\\
&+\mathbb{E}\Bigg[\Bigg(\eta\gamma A_c\beta^{-1} + \left(\frac{\eta\gamma}{4} + \left(\frac{\eta\gamma}{2} + \eta\right)\left(2L_1^2C_\rho^2 + 2L_1^2(1+|X_{k+1}|)^{2\rho}\right) + \frac{\eta^2}{2}2L_1^2(1+|X_{k+1}|)^{2\rho}\right)|\theta_k^\eta|^2\\
&+\left(\frac{\eta\gamma}{2} + \eta\right)\left(2h_0^2 + 4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \left(\frac{L_1 C_\rho\eta^2}{2} - \frac{\gamma^2\lambda\eta^2}{4} - \frac{\gamma\eta}{2} + \frac{\gamma^2\eta^2}{4} + \frac{\gamma\lambda\eta}{2} + \frac{\eta}{2}\right)|V_k^\eta|^2\\
&+\frac{\eta^2}{2}\left(4L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 4H_0^2\right) + \eta\gamma\beta^{-1}|\xi_{k+1}|^2 + \Sigma_k\Bigg)^2\Bigg|\mathcal{H}_k\Bigg]
\end{aligned}
$$

Now taking the expectation, regrouping and simplifying the terms above, we finally arrive at

$$
\begin{aligned}
\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} &\leq (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + 2\phi\frac{\mathcal{V}_k}{\beta^2}\eta\tilde{c}_1 + 2\phi\frac{\mathcal{V}_k}{\beta}\eta^2\hat{c}_1 + \eta^2\frac{\tilde{c}_2}{\beta^2} + \tilde{c}_3\eta^2|\theta_k^\eta|^4 + \hat{c}_3\eta^4|\theta_k^\eta|^4\\
&+\eta^2\tilde{c}_4|V_k^\eta|^4 + \eta^4\hat{c}_4|V_k^\eta|^4 + \eta^2\tilde{c}_5 + \eta^4\hat{c}_5 + \eta^2\frac{\tilde{c}_6}{\beta^2} + \mathbb{E}[\Sigma_k^2|\mathcal{H}_k]. \qquad (47)
\end{aligned}
$$

where

$$\begin{aligned}
\tilde{c}_1 &= \gamma A_c + \gamma d,\\
\hat{c}_1 &= 2L_2 C_\rho + 2H_0^2\\
\tilde{c}_2 &= 6\gamma^2 A_c^2\\
\tilde{c}_3 &= \frac{9\gamma^2}{8} + 18(\gamma+2)^2(L_1^4 C_\rho^4 + L_1^4 C_\rho),\\
\hat{c}_3 &= 18 L_1^4 C_\rho,\\
\tilde{c}_4 &= 6\frac{(1+\lambda\gamma-\gamma)^2}{4},\\
\hat{c}_4 &= 6\left(\frac{L_1 C_\rho}{2} + \frac{\gamma}{4} + \frac{1}{2}\right)^2,\\
\tilde{c}_5 &= (\gamma+2)^2\left(30h_0^4 + 120L_2^4 + 120H_0^4 + 30L_2^4 C_\rho + 30H_0^4\right),\\
\hat{c}_5 &= 48(L_4^4 + H_0^4),\\
\tilde{c}_6 &= \gamma^2 d(d+2).
\end{aligned}$$

Next, recall

$$\Sigma_k = \frac{\gamma^2}{2}\sqrt{2\gamma^{-1}\beta^{-1}\eta}\langle \Delta_k^2, \xi_{k+1}\rangle + \frac{1}{2}\sqrt{2\eta\gamma\beta^{-1}}\langle \Delta_k^1, \xi_{k+1}\rangle,$$

where

$$\begin{aligned}
\Delta_k^1 &= (1-\eta\gamma)V_k^\eta \ - \eta\ H(\theta_k^\eta, X_{k+1})\\
\Delta_k^2 &= \theta_k^\eta + \gamma^{-1}V_k^\eta - \eta\gamma^{-1}H(\theta_k^\eta, X_{k+1}).
\end{aligned}$$

Note that

$$\begin{aligned}
\Sigma_k^2 &\le 2\gamma\beta^{-1}\eta|\Delta_k^1|^2|\xi_{k+1}|^2 + 2\gamma^3\beta^{-1}\eta|\Delta_k^2|^2|\xi_{k+1}|^2,\\
&\le 2\eta\gamma\beta^{-1}|\xi_{k+1}|^2\left(|V_k^\eta - \eta(\gamma V_k^\eta + H(\theta_k^\eta, X_{k+1})|^2 + \gamma^2|\theta_k^\eta + \gamma^{-1}V_k^\eta - \eta\gamma^{-1}H(\theta_k^\eta, X_{k+1})|^2\right),\\
&\le 2\eta\gamma\beta^{-1}|\xi_{k+1}|^2\left(2(1-\eta\gamma)^2|V_k^\eta|^2 + 2|H(\theta_k, X_{k+1})|^2 + 3\gamma^2|\theta_k^\eta|^2 + 3|V_k^\eta|^2 + 3\eta^2|H(\theta_k^\eta, X_{k+1})|^2\right)\\
&\le 2\eta\gamma\beta^{-1}|\xi_{k+1}|^2\Big(2(1-\eta\gamma)^2|V_k^\eta|^2 + 2\left(3L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 3L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 3H_0^2\right)\\
&+3\gamma^2|\theta_k^\eta|^2 + 3|V_k^\eta|^2 + 3\eta^2\left(3L_1^2(1+|X_{k+1}|)^{2\rho}|\theta_k^\eta|^2 + 3L_2^2(1+|X_{k+1}|)^{2(\rho+1)} + 3H_0^2\right)\Big).
\end{aligned}$$

Therefore, taking expectation and using (46) and $\eta \le 1/\gamma$, we obtain

$$\begin{aligned}
\mathbb{E}[\Sigma_k^2|\mathcal{H}_k] &\le 2\eta\gamma\beta^{-1}d\left(5|V_k^\eta|^2 + \left(6L_1^2C_\rho + 3\gamma^2 + 9L_1^2C_\rho\right)|\theta_k^\eta|^2 + 15L_2^2C_\rho + 15H_0^2\right).\\
&\le \eta\tilde{c}_7\frac{\mathcal{V}_k}{\beta^2} + \eta\frac{\tilde{c}_8}{\beta}
\end{aligned}$$

where

$$\begin{aligned}
\tilde{c}_7 &= \max\left\{\frac{10\gamma d}{\frac{1}{8}(1-2\lambda)}, \frac{(2\gamma d)(6L_1^2C_\rho + 3\gamma^2 + 9L_1^2C_\rho)}{\frac{1}{16}(1-2\lambda)\gamma^2}\right\}\\
\tilde{c}_8 &= 30\gamma d L_2^2 C_\rho + 30\gamma d H_0^2.
\end{aligned}$$

Now plugging this into the inequality (47), we arrive at

$$\begin{aligned}
\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} &\le (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + (2\phi\tilde{c}_1 + \tilde{c}_7)\frac{\mathcal{V}_k}{\beta^2}\eta + 2\phi\frac{\mathcal{V}_k}{\beta}\eta^2\hat{c}_1 + \eta^2\frac{\tilde{c}_2}{\beta^2} + \tilde{c}_3\eta^2|\theta_k^\eta|^4 + \hat{c}_3\eta^4|\theta_k^\eta|^4\\
&+ \eta^2\tilde{c}_4|V_k^\eta|^4 + \eta^4\hat{c}_4|V_k^\eta|^4 + \eta^2\tilde{c}_5 + \eta^4\hat{c}_5 + \eta^2\frac{\tilde{c}_6}{\beta^2} + \eta\frac{\tilde{c}_8}{\beta},
\end{aligned} \tag{48}$$

Now, recall our step-size condition for the above inequality, in particular, $\eta \leq 1$, which leads to

$$\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} \leq (\phi^2 + 2\phi\tilde{K}_1\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + (2\phi\tilde{c}_1 + \tilde{c}_7)\frac{\mathcal{V}_k}{\beta^2}\eta + 2\phi\frac{\mathcal{V}_k}{\beta}\eta^2\hat{c}_1 + \eta^2\frac{\tilde{c}_2}{\beta^2} + (\tilde{c}_3 + \hat{c}_3)\eta^2|\theta_k^\eta|^4 + \eta^2(\tilde{c}_4 + \hat{c}_4)|V_k^\eta|^4 + \eta^2(\tilde{c}_5 + \hat{c}_5) + \eta^2\frac{\tilde{c}_6}{\beta^2} + \eta\frac{\tilde{c}_8}{\beta}, \tag{49}$$

Next, recall that

$$\frac{\mathcal{V}_k}{\beta} \geq \max\left\{\frac{1}{8}(1-2\lambda)\gamma^2|\theta_k^\eta|^2, \frac{1}{4}(1-2\lambda)|V_k^\eta|^2\right\},$$

which implies

$$\begin{aligned}\frac{\mathcal{V}_k^2}{\beta^2} &\geq \max\left\{\frac{1}{64}(1-2\lambda)^2\gamma^4|\theta_k^\eta|^4, \frac{1}{16}(1-2\lambda)^2|V_k^\eta|^4\right\},\\ &\geq \frac{1}{128}(1-2\lambda)^2\gamma^4|\theta_k^\eta|^4 + \frac{1}{32}(1-2\lambda)^2|V_k^\eta|^4.\end{aligned} \tag{50}$$

Now inequalities (49) and (50) together imply

$$\frac{\mathbb{E}[\mathcal{V}_{k+1}^2|\mathcal{H}_k]}{\beta^2} \leq (\phi^2 + 2\phi\tilde{K}_2\eta^2)\frac{\mathcal{V}_k^2}{\beta^2} + (2\phi\tilde{c}_1 + \tilde{c}_7)\frac{\mathcal{V}_k}{\beta^2}\eta + 2\phi\frac{\mathcal{V}_k}{\beta}\eta^2\hat{c}_1 + \eta^2\frac{\tilde{c}_2}{\beta^2} + \eta^2(\tilde{c}_5 + \hat{c}_5) + \eta^2\frac{\tilde{c}_6}{\beta^2} + \eta\frac{\tilde{c}_8}{\beta},$$

where $\tilde{K}_2 = \tilde{K}_1 + \tilde{c}_9$ and

$$\tilde{c}_9 = \max\left\{\frac{\tilde{c}_3 + \hat{c}_3}{\frac{1}{128}(1-2\lambda)^2\gamma^4}, \frac{\hat{c}_4 + \tilde{c}_4}{\frac{1}{32}(1-2\lambda)^2}\right\}.$$

Now we take unconditional expectations, use the fact that $\phi \leq 1$, and obtain

$$\mathbb{E}[\mathcal{V}_{k+1}^2] \leq \left(1 - \lambda\gamma\eta + \bar{K}\eta^2\right)\mathbb{E}[\mathcal{V}_k^2] + \tilde{c}_{10}\mathbb{E}[\mathcal{V}_k]\eta + 2\hat{c}_1\mathbb{E}[\mathcal{V}_k]\beta\eta^2 + \eta^2\tilde{c}_2 + \eta^2(\hat{c}_5 + \tilde{c}_5)\beta^2 + \eta^2\tilde{c}_6 + \eta\tilde{c}_8\beta,$$

where $\bar{K} = 2\tilde{K}_2$, $\tilde{c}_{10} = 2\tilde{c}_1 + \tilde{c}_7$. First note that, it follows from the proof of Lemma 3.2 that

$$\mathbb{E}[\mathcal{V}_k] \leq \mathbb{E}[\mathcal{V}_0] + \frac{4(A_c + d)}{\gamma\lambda} := \tilde{c}_{11}.$$

Therefore, using $\eta \leq 1$, we obtain

$$\mathbb{E}[\mathcal{V}_{k+1}^2] \leq \left(1 - \lambda\gamma\eta + \bar{K}\eta^2\right)\mathbb{E}[\mathcal{V}_k^2] + D\eta$$

where $D = \tilde{c}_{10}\tilde{c}_{11} + 2\tilde{c}_{11}\hat{c}_1\beta + \tilde{c}_2 + (\hat{c}_5 + \tilde{c}_5)\beta^2 + \tilde{c}_6 + \tilde{c}_8\beta$. Now recall that $\eta \leq \lambda\gamma/2\bar{K}$ which leads to

$$\mathbb{E}[\mathcal{V}_{k+1}^2] \leq \left(1 - \frac{\lambda\gamma\eta}{2}\right)\mathbb{E}[\mathcal{V}_k^2] + D\eta,$$

which implies that

$$\mathbb{E}[\mathcal{V}_k^2] \leq \mathbb{E}[\mathcal{V}_0^2] + \frac{2D}{\gamma\lambda},$$

which concludes the proof. □

C.4 Proof of Lemma B.1

We start by writing that

$$\begin{aligned}U(\theta) - U(0) &= \int_0^1 \langle\theta, h(t\theta)\rangle \mathrm{d}t,\\ &\leq \int_0^1 |\theta||h(t\theta)|\mathrm{d}t,\\ &\leq \int_0^1 |\theta|(\bar{L}_1 t|\theta| + h_0)\mathrm{d}t.\end{aligned}$$

from Remark (10) where $\bar{L}_1 = L_1\mathbb{E}[(1+|X_0|)^\rho]$. This in turn leads to

$$U(\theta) \leq u_0 + \frac{\bar{L}_1}{2}|\theta|^2 + h_0|\theta|.$$

where $u_0 = U(0)$. Next, we prove the lower bound. To this end, take $c \in (0,1)$ and write

$$\begin{aligned} U(\theta) &= U(c\theta) + \int_c^1 \langle \theta, h(t\theta)\rangle \mathrm{d}t, \\ &\geq \int_c^1 \frac{1}{t}\langle t\theta, h(t\theta)\rangle \mathrm{d}t, \\ &\geq \int_c^1 \frac{1}{t}\left(a|t\theta|^2 - b\right)\mathrm{d}t, \\ &= \frac{a(1-c^2)}{2}|\theta|^2 + b\log c. \end{aligned}$$

Taking $c = 1/\sqrt{3}$ leads to the bound.

C.5 Proof of Lemma B.3

Let $\mathcal{H}_t^\infty = \mathcal{F}_\infty^\eta \vee \mathcal{G}_{\lfloor t \rfloor}$. Following Zhang et al. (2023b), we obtain

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

where the first inequality holds due to Lemma B.2 and

$$\sigma_Z = \mathbb{E}[(1+|X_0|+|\mathbb{E}[X_0]|)^{2\rho}|X_0 - \mathbb{E}[X_0]|^2].$$

Then, by using Lemma 3.3, we obtain

$$\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\eta,n}) - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2\right] \leq 8L_2^2\sigma_Z + 8L_2^2\sigma_Z C_\zeta.$$

C.6 Proof of Lemma B.4

Note that for any t, we have

$$\overline{V}_t^\eta = \overline{V}_{\lfloor t \rfloor}^\eta - \eta\gamma\int_{\lfloor t \rfloor}^t \overline{V}_{\lfloor s \rfloor}^\eta \mathrm{d}s - \eta\int_{\lfloor t \rfloor}^t H(\overline{\theta}_{\lfloor s \rfloor}, X_{\lceil s \rceil})\mathrm{d}s + \sqrt{2\gamma\eta\beta^{-1}}(B_t^\eta - B_{\lfloor t \rfloor}^\eta).$$

We therefore obtain

$$\begin{aligned} \mathbb{E}\left[\left|\overline{V}_{\lfloor t \rfloor}^\eta - \overline{V}_t^\eta\right|^2\right] &= \mathbb{E}\left[\left|-\eta\gamma\int_{\lfloor t \rfloor}^t \overline{V}_{\lfloor s \rfloor}^\eta \mathrm{d}s - \eta\int_{\lfloor t \rfloor}^t H(\overline{\theta}_{\lfloor s \rfloor}, X_{\lceil s \rceil})\mathrm{d}s + \sqrt{2\gamma\eta\beta^{-1}}(B_t^\eta - B_{\lfloor t \rfloor}^\eta)\right|^2\right] \\ &\leq 4\eta^2\gamma^2 C_v + 4\eta^2(\widetilde{L}_1 C_\theta + \widetilde{C}_1) + 4\gamma\eta\beta^{-1}d. \end{aligned}$$

Next, we write

$$\overline{\theta}_t^\eta = \overline{\theta}_{\lfloor t \rfloor}^\eta + \int_{\lfloor t \rfloor}^t \eta\overline{V}_{\lfloor \tau \rfloor}^\eta \mathrm{d}\tau,$$

which implies

$$\mathbb{E}\left[\left|\overline{\theta}_t^\eta - \overline{\theta}_{\lfloor t \rfloor}^\eta\right|^2\right] \leq \eta^2 C_v.$$

C.7 Proof of Lemma B.5

Choose $\lambda > 0$ such that

$$\lambda = \min\left\{\frac{1}{4}, \frac{a}{\bar{L}_1 + 2\bar{L}_1 b + \frac{\gamma^2}{2}}\right\}.$$

Using Assumption 2.4, we obtain

$$\begin{aligned}
\langle \theta, h(\theta)\rangle &\geq a|\theta|^2 - b, \\
&\geq 2\lambda\left(\frac{\bar{L}_1}{2} + \bar{L}_1 b + \frac{\gamma^2}{4}\right)|\theta|^2 - b, \\
&\geq 2\lambda\left(U(\theta) - u_0 - \bar{L}_1 b|\theta| + \bar{L}_1 b|\theta|^2 + \frac{\gamma^2}{4}|\theta|^2\right) - b, \\
&\geq 2\lambda\left(U(\theta) - u_0 - \bar{L}_1 b + \frac{\gamma^2}{4}|\theta|^2\right) - b,
\end{aligned}$$

where the third line follows from Lemma B.1 and the last line follows from the inequality $|x| \leq 1 + |x|^2$. Consequently, we obtain

$$\langle \theta, h(\theta)\rangle \geq 2\lambda\left(U(\theta) + \frac{\gamma^2}{4}|\theta|^2\right) - 2A_c/\beta$$

where

$$A_c = \frac{\beta}{2}(b + 2\lambda u_0 + 2\lambda\bar{L}_1 b),$$

which proves the claim.

Appendix D. Proofs of main results

D.1 Proof of Theorem 4.1

We note that

$$W_2(\mathcal{L}(\overline{\theta}_t^\eta, \overline{V}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})) \leq \mathbb{E}\left[\left|\overline{\theta}_t^\eta - \overline{\zeta}_t^{\eta,n}\right|^2\right]^{1/2} + \mathbb{E}\left[\left|\overline{V}_t^\eta - \overline{Z}_t^{\eta,n}\right|^2\right]^{1/2}. \tag{51}$$

We first bound the first term of (51). We start by employing the synchronous coupling and obtain

$$\left|\overline{\theta}_t^\eta - \overline{\zeta}_t^{\eta,n}\right| \leq \eta\int_{nT}^t \left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right| \mathrm{d}s,$$

which implies

$$\begin{aligned}
\sup_{nT\leq u\leq t}\mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] &\leq \eta\sup_{nT\leq u\leq t}\int_{nT}^u \mathbb{E}\left[\left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right]\mathrm{d}s, \\
&= \eta\int_{nT}^t \mathbb{E}\left[\left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right]\mathrm{d}s,
\end{aligned}$$

Next, we write for any $t \in [nT, (n+1)T)$

$$\begin{aligned}
\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right| &\leq \left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right| + \left|\overline{V}^{\eta}_t - \overline{Z}^{\eta,n}_t\right|, \\
&\leq \left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right| + \left|-\gamma\eta \int_{nT}^t [\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s]\mathrm{d}s - \eta \int_{nT}^t [H(\overline{\theta}^{\eta}_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|, \\
&\leq \left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right| + \gamma\eta \int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right| \mathrm{d}s + \eta \left|\int_{nT}^t [H(\overline{\theta}^{\eta}_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|, \\
&\leq \left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right| + \gamma\eta \int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right| \mathrm{d}s \\
&+ \eta \int_{nT}^t \left|H(\overline{\theta}^{\eta}_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil})\right| \mathrm{d}s + \eta \left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|, \\
&\leq \left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right| + \gamma\eta \int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right| \mathrm{d}s \\
&+ \eta \int_{nT}^t L_1(1 + X_{\lceil s \rceil})^{\rho} \left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right| \mathrm{d}s + \eta \left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|.
\end{aligned}$$

We take the squares of both sides and use $(a+b)^2 \leq 2(a^2+b^2)$ twice to obtain

$$\begin{aligned}
\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right|^2 &\leq 4\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right|^2 + 4\gamma^2\eta^2 \left(\int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right| \mathrm{d}s\right)^2 \\
&+ 4\eta^2 \left(\int_{nT}^t L_1(1 + X_{\lceil s \rceil})^{\rho} \left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right| \mathrm{d}s\right)^2 + 4\eta^2 \left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2 \\
&\leq 4\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right|^2 + 4\gamma^2\eta \int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2 \mathrm{d}s \\
&+ 4\eta L_1^2 \int_{nT}^t (1 + X_{\lceil s \rceil})^{2\rho} \left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2 \mathrm{d}s + 4\eta^2 \left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2.
\end{aligned}$$

Taking expectations of both sides, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right|^2\right] &\leq 4\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right|^2\right] + 4\gamma^2\eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2\right] \mathrm{d}s \\
&+ 4\eta L_1^2 C_\rho \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2\right] \mathrm{d}s + 4\eta^2 \mathbb{E}\left[\left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2\right], \\
&\leq 4\sigma_V\eta + 4\gamma^2\eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2\right] \mathrm{d}s \\
&+ 4\eta L_1^2 C_\rho \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2\right] \mathrm{d}s + 4\eta^2 \mathbb{E}\left[\left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2\right].
\end{aligned}$$

By applying Grönwall's lemma, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right|^2\right] &\leq 4c_1\sigma_V\eta + 4\eta c_1 L_1^2 C_\rho \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2\right] \mathrm{d}s \\
&+ 4c_1\eta^2 \mathbb{E}\left[\left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2\right],
\end{aligned}$$

where $c_1 = \exp(4\gamma^2)$ since $\eta T \leq 1$. Next, we write

$$\begin{aligned}
\sup_{nT\leq u\leq t} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] &\leq \eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right] \mathrm{d}s, \\
&\leq 4c_1\eta\sigma_V + 4c_1\eta^2 L_1^2 C_\rho \int_{nT}^t \int_{nT}^s \mathbb{E}\left[\left|\overline{\theta}_{\lfloor s'\rfloor}^\eta - \overline{\zeta}_{s'}^{\eta,n}\right|^2\right] \mathrm{d}s'\mathrm{d}s \\
&+ 4c_1\eta^3 \int_{nT}^t \mathbb{E}\left[\left|\int_{nT}^s [H(\overline{\zeta}_{s'}^{\eta,n}, X_{\lceil s'\rceil}) - h(\overline{\zeta}_{s'}^{\eta,n})]\mathrm{d}s'\right|^2\right] \mathrm{d}s, \\
&\leq 4c_1\eta\sigma_V + 4c_1\eta L_1^2 C_\rho \sup_{nT\leq s\leq t} \int_{nT}^s \mathbb{E}\left[\left|\overline{\theta}_{\lfloor s'\rfloor}^\eta - \overline{\zeta}_{s'}^{\eta,n}\right|^2\right] \mathrm{d}s' \\
&+ 4c_1\eta^3 \int_{nT}^t \mathbb{E}\left[\left|\int_{nT}^s [H(\overline{\zeta}_{s'}^{\eta,n}, X_{\lceil s'\rceil}) - h(\overline{\zeta}_{s'}^{\eta,n})]\mathrm{d}s'\right|^2\right] \mathrm{d}s. \qquad (52)
\end{aligned}$$

First, we bound the supremum term (i.e. the second term of (52)) as

$$\begin{aligned}
\sup_{nT\leq s\leq t} \int_{nT}^s \mathbb{E}\left[\left|\overline{\theta}_{\lfloor s'\rfloor}^\eta - \overline{\zeta}_{s'}^{\eta,n}\right|^2\right] \mathrm{d}s' &= \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}_{\lfloor s'\rfloor}^\eta - \overline{\zeta}_{s'}^{\eta,n}\right|^2\right] \mathrm{d}s' \\
&\leq \int_{nT}^t \sup_{nT\leq u\leq s'} \mathbb{E}\left[\left|\overline{\theta}_{\lfloor u\rfloor}^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] \mathrm{d}s' \\
&\leq \int_{nT}^t \sup_{nT\leq u\leq s'} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] \mathrm{d}s'. \qquad (53)
\end{aligned}$$

Next, we bound the last term of (52) by partitioning the integral. Assume that $nT+K \leq s \leq t \leq nT+K+1$ where $K+1 \leq T$. Thus we can write

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

where

$$I_k = \int_{nT+(k-1)}^{nT+k} [h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k})]\mathrm{d}s' \qquad \text{and} \qquad R_K = \int_{nT+K}^s [h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+K+1})]\mathrm{d}s'.$$

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, it remains to take the expectations of both sides. We begin by defining the filtration $\mathcal{H}_s^\infty = \mathcal{F}_\infty^\eta \vee \mathcal{G}_{\lfloor s\rfloor}$ and note that for any $k = 2, \ldots, K$, $j = 1, \ldots, k-1$,

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

By the same argument $\mathbb{E}\langle I_k, R_K\rangle = 0$ for all $1 \le k \le K$. Therefore,

$$\begin{aligned}
&\int_{nT}^{t} \mathbb{E}\left[\left|\int_{nT}^{s} [H(\overline{\zeta}_{s'}^{\eta,n}, X_{\lceil s' \rceil}) - h(\overline{\zeta}_{s'}^{\eta,n})]\mathrm{d}s'\right|^2\right] \mathrm{d}s \\
&= \int_{nT}^{t} \left[\sum_{k=1}^{K} \mathbb{E}\left[\left|\int_{nT+(k-1)}^{nT+k} [h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k})]\mathrm{d}s'\right|^2\right]\right] \mathrm{d}s \\
&+ \int_{nT}^{t} \mathbb{E}\left[\left|\int_{nT+K}^{s} [h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+K+1})]\mathrm{d}s'\right|^2\right] \mathrm{d}s \\
&\le \int_{nT}^{t} \left[\sum_{k=1}^{K} \int_{nT+(k-1)}^{nT+k} \mathbb{E}\left[\left|h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k})\right|^2\right] \mathrm{d}s'\right] \mathrm{d}s \\
&+ \int_{nT}^{t} \int_{nT+K}^{s} \mathbb{E}\left[\left|h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+K+1})\right|^2\right] \mathrm{d}s'\mathrm{d}s \\
&\le T^2\sigma_H + T\sigma_H. \qquad (54)
\end{aligned}$$

Using (52), (53), and (54), we eventually obtain

$$\begin{aligned}
\sup_{nT\le u\le t} \mathbb{E}\left[\left|\overline{\theta}_u^{\eta} - \overline{\zeta}_u^{\eta,n}\right|^2\right] &\le 4c_1\eta\sigma_V + 4\eta L_1^2 C_\rho \int_{nT}^{t} \sup_{nT\le u\le s'} \mathbb{E}\left[\left|\overline{\theta}_u^{\eta} - \overline{\zeta}_u^{\eta,n}\right|^2\right] \mathrm{d}s', \\
&+ 4c_1\eta^3(T^2\sigma_H + T\sigma_H), \\
&\le 4c_1\eta\sigma_V + 4c_1\eta L_1^2 C_\rho \int_{nT}^{t} \sup_{nT\le u\le s'} \mathbb{E}\left[\left|\overline{\theta}_u^{\eta} - \overline{\zeta}_u^{\eta,n}\right|^2\right] \mathrm{d}s' \\
&+ 4c_1\eta\sigma_H + 4c_1\eta^2\sigma_H, \qquad (55)
\end{aligned}$$

since $\eta T \le 1$. Finally, applying Grönwall's inequality and using again $\eta T \le 1$ provides

$$\sup_{nT\le u\le t} \mathbb{E}\left[\left|\overline{\theta}_u^{\eta} - \overline{\zeta}_u^{\eta,n}\right|^2\right] \le \exp(4c_1 L_1^2 C_\rho)(4c_1\sigma_V + 4c_1\sigma_H + 4c_1\eta\sigma_H)\eta,$$

which implies that

$$\mathbb{E}\left[\left|\overline{\theta}_t^{\eta} - \overline{\zeta}_t^{\eta,n}\right|^2\right]^{1/2} \le C_{1,1}^{\star}\sqrt{\eta} \qquad (56)$$

with $C_{1,1}^{\star} = \sqrt{\exp(4c_1 L_1^2 C_\rho)(4c_1\sigma_V + 4c_1\sigma_H + 4c_1\eta\sigma_H)}$. Note that $\sigma_V = \mathcal{O}(d)$ and $\sigma_H = \mathcal{O}(d)$ hence $C_{1,1}^{\star} = \mathcal{O}(\sqrt{d})$.

Next, we upper bound the second term of (51). To prove it, we write

$$\left|\overline{V}_t^{\eta} - \overline{Z}_t^{\eta,n}\right| \le \left|\gamma\eta \int_{nT}^{t} \left[\overline{V}_{\lfloor s \rfloor}^{\eta} - \overline{Z}_s^{\eta,n}\right] \mathrm{d}s\right| + \eta\left|\int_{nT}^{t} \left[H(\overline{\theta}_{\lfloor s \rfloor}^{\eta}) - h(\overline{\zeta}_s^{\eta,n})\right] \mathrm{d}s\right|,$$

which leads to

$$\mathbb{E}\left[\left|\overline{V}_t^{\eta} - \overline{Z}_t^{\eta,n}\right|^2\right] \le 2\gamma^2\eta \int_{nT}^{t} \mathbb{E}\left[\left|\overline{V}_{\lfloor s \rfloor}^{\eta} - \overline{Z}_s^{\eta,n}\right|^2\right] + 2\eta^2\mathbb{E}\left[\left|\int_{nT}^{t} \left[H(\overline{\theta}_{\lfloor s \rfloor}^{\eta}) - h(\overline{\zeta}_s^{\eta,n})\right] \mathrm{d}s\right|^2\right].$$

By similar arguments we have used for bounding the the first term, we obtain

$$\mathbb{E}\left[\left|\overline{V}_t^{\eta} - \overline{Z}_t^{\eta,n}\right|^2\right] \le 2\gamma^2\eta \int_{nT}^{t} \mathbb{E}\left[\left|\overline{V}_{\lfloor s \rfloor}^{\eta} - \overline{Z}_s^{\eta,n}\right|^2\right] + \exp(4c_1 L_1^2 C_\rho)(4c_1\sigma_H\eta + 4c_1\eta^2\sigma_H\eta^2).$$

Using the fact that the rhs is an increasing function of t and we obtain

$$\sup_{nT\le u\le t} \mathbb{E}\left[\left|\overline{V}_u^{\eta} - \overline{Z}_u^{\eta,n}\right|^2\right] \le 2\gamma^2\eta \int_{nT}^{t} \sup_{nT\le u\le s} \mathbb{E}\left[\left|\overline{V}_u^{\eta} - \overline{Z}_u^{\eta,n}\right|^2\right] \mathrm{d}s + \exp(4c_1 L_1^2 C_\rho)(4c_1\sigma_H\eta + 4c_1\eta^2\sigma_H\eta^2),$$

Applying Gronwall's lemma and $\eta T \leq 1$ yields

$$\sup_{nT \leq u \leq t} \mathbb{E}\left[\left|\overline{V}_u^\eta - \overline{Z}_u^{\eta,n}\right|^2\right] \leq \exp(2\gamma^2 + 4c_1 L_1^2 C_\rho) 4c_1(\sigma_H \eta + \sigma_H \eta^2),$$

which leads to

$$\mathbb{E}\left[\left|\overline{V}_t^\eta - \overline{Z}_t^{\eta,n}\right|^2\right]^{1/2} \leq C_{1,2}^\star \sqrt{\eta}, \tag{57}$$

where $C_{1,2}^\star = \sqrt{\exp(2\gamma^2 + 4c_1 L_1^2 C_\rho) 4c_1 \sigma_H (1+\eta)}$. Note again that $\sigma_H = \mathcal{O}(d)$, hence $C_{1,2}^\star = \mathcal{O}(d^{1/2})$.

Therefore, combining (51), (56), (57), we obtain

$$W_2(\mathcal{L}(\overline{\theta}_t^\eta, \overline{V}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})) \leq C_1^\star \eta^{1/2},$$

where $C_1^\star = C_{1,1}^\star + C_{1,2}^\star = \mathcal{O}(d^{1/2})$.

D.2 Proof of Theorem 4.2

Triangle inequality implies that

$$W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) \leq \sum_{k=1}^n W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,k}, \overline{Z}_t^{\eta,k}), \mathcal{L}(\overline{\zeta}_t^{\eta,k-1}, \overline{Z}_t^{\eta,k-1})),$$

since $(\overline{\zeta}_t^{\eta,0}, \overline{Z}_t^{\eta,0}) = (\widehat{\zeta}_t^{0,\overline{\theta}_0^\eta,\overline{V}_0^\eta,\eta}, \widehat{Z}_t^{0,\overline{\theta}_0^\eta,\overline{V}_0^\eta,\eta}) = (\zeta_t^\eta, Z_t^\eta)$ by definition. Next we write out the definition of the process $(\overline{\zeta}_t^{\eta,k}, \overline{Z}_t^{\eta,k})$ and obtain

$$\begin{aligned}
&W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) \\
&\quad \leq \sum_{k=1}^n W_2(\mathcal{L}(\widehat{\zeta}_t^{kT,\overline{\theta}_{kT}^\eta,\overline{V}_{kT}^\eta,\eta}, \widehat{Z}_t^{kT,\overline{\theta}_{kT}^\eta,\overline{V}_{kT}^\eta,\eta}), \mathcal{L}(\widehat{\zeta}_t^{(k-1)T,\overline{\theta}_{(k-1)T}^\eta,\overline{V}_{(k-1)T}^\eta,\eta}, \widehat{Z}_t^{(k-1)T,\overline{\theta}_{(k-1)T}^\eta,\overline{V}_{(k-1)T}^\eta,\eta})).
\end{aligned}$$

At this point, we have two processes and in order to be able to use the contraction result, we need their starting times to match. For notational simplicity, let us define

$$\widehat{B}_t^{s,\overline{\theta}_s^\eta,\overline{V}_s^\eta,\eta} = (\widehat{\zeta}_t^{s,\overline{\theta}_s^\eta,\overline{V}_s^\eta,\eta}, \widehat{Z}_t^{s,\overline{\theta}_s^\eta,\overline{V}_s^\eta,\eta})$$

Therefore, in order to be able to use a contraction result, we note

$$\mathcal{L}(\widehat{B}_t^{(k-1)T,\overline{\theta}_{(k-1)T}^\eta,\overline{V}_{(k-1)T}^\eta,\eta}) = \mathcal{L}(\widehat{B}_t^{kT,\widehat{B}_{kT}^{(k-1)T,\overline{\theta}_{(k-1)T}^\eta,\overline{V}_{(k-1)T}^\eta,\eta},\eta}).$$

This leads to

$$\begin{aligned}
&W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) \\
&\leq \sum_{k=1}^n W_2(\mathcal{L}(\widehat{\zeta}_t^{kT,\theta_{kT}^\eta,V_{kT}^\eta,\eta}, \widehat{Z}_t^{kT,\theta_{kT}^\eta,V_{kT}^\eta,\eta}), \mathcal{L}(\widehat{\zeta}_t^{kT,\widehat{B}_{kT}^{(k-1)T,\overline{\theta}_{(k-1)T}^\eta,\overline{V}_{(k-1)T}^\eta,\eta},\eta}, \widehat{Z}_t^{kT,\widehat{B}_{kT}^{(k-1)T,\overline{\theta}_{(k-1)T}^\eta,\overline{V}_{(k-1)T}^\eta,\eta},\eta})).
\end{aligned}$$

The reader should notice at this point that this quantity can be upper bounded by a *contraction result* as both processes are defined for time t and both started at time kT. By using Theorem 3.1, we obtain

$$W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) \leq \sqrt{\dot{C}} \sum_{k=1}^n e^{-\eta\dot{c}(t-kT)/2} \sqrt{\mathcal{W}_\rho(\mathcal{L}(\theta_{kT}^\eta, V_{kT}^\eta), \mathcal{L}(\overline{\zeta}_{kT}^{\eta,(k-1)T}, \overline{Z}_{kT}^{\eta,(k-1)T})}.$$

Next, using Lemma 5.4 of Chau and Rásonyi (2022) to upper bound the the last term leads

$$\begin{aligned}
&W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) \\
&\leq 3\max\{1+\alpha, \gamma^{-1}\}\sqrt{\dot{C}} \sum_{k=1}^n e^{-\eta\dot{c}(t-kT)/2} \sqrt{1 + \varepsilon_c \mathbb{E}^{1/2}[\mathcal{V}^2(\theta_{kT}^\eta, V_{kT}^\eta)] + \varepsilon_c \mathbb{E}^{1/2}[\mathcal{V}^2(\overline{\zeta}_{kT}^{\eta,(k-1)T}, \overline{Z}_{kT}^{\eta,(k-1)T})]} \\
&\times \sqrt{W_2(\mathcal{L}(\theta_{kT}^\eta, V_{kT}^\eta), \mathcal{L}(\overline{\zeta}_{kT}^{\eta,(k-1)T}, \overline{Z}_{kT}^{\eta,(k-1)T}))}, \\
&\leq C_2^\star \eta^{1/4},
\end{aligned}$$

where the last line follows from Lemma 3.4 and Theorem 4.1. The bound for the term relating to the continuous dynamics (i.e. $\mathbb{E}^{1/2}[\mathcal{V}^2(\overline{\zeta}_{kT}^{\eta,(k-1)T}, \overline{Z}_{kT}^{\eta,(k-1)T})]$) follows from Remark 2.3 (dissipativity of the gradient) and we have a constant diffusion coefficient, and uniform bound in Lemma 3.1 and, finally, the fact that the initial data (which is the iterates of the numerical scheme) has finite fourth moments (Lemma 3.4). We note that $C_2^\star = \mathcal{O}(e^d)$.

D.3 Proof of Proposition 5.1

We denote $\pi_{n,\beta}^\eta := \mathcal{L}(\theta_n^\eta, V_n^\eta)$ and write

$$\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)] = \int_{\mathbb{R}^{2d}} U(\theta)\pi_{n,\beta}^\eta(\mathrm{d}\theta, \mathrm{d}v) - \int_{\mathbb{R}^{2d}} U(\theta)\pi_\beta(\mathrm{d}\theta, \mathrm{d}v).$$

Recall from (10), we have

$$|h(\theta)| \leq \overline{L}_1|\theta| + h_0.$$

Using Raginsky et al. (2017, Lemma 6), we arrive at

$$\left|\int_{\mathbb{R}^{2d}} U(\theta)\pi_{n,\beta}^\eta(\mathrm{d}\theta, \mathrm{d}v) - \int_{\mathbb{R}^{2d}} U(\theta)\pi_\beta(\mathrm{d}\theta, \mathrm{d}v)\right| \leq (\overline{L}_1 C_m + h_0) W_2(\pi_{n,\beta}^\eta, \pi_\beta),$$

where,

$$C_m^2 := \max\left(\int_{\mathbb{R}^{2d}} \|\theta\|^2 \pi_{n,\beta}^\eta(\mathrm{d}\theta, \mathrm{d}v), \int_{\mathbb{R}^{2d}} \|\theta\|^2 \pi_\beta(\mathrm{d}\theta, \mathrm{d}v)\right) = \max(C_\theta^c, C_\theta).$$

We therefore obtain using Theorem 2.1 that

$$\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)] \leq (\overline{L}_1 C_m + h_0)\left(C_1^\star d^{1/2}\eta^{1/2} + C_2^\star \eta^{1/4} + C_3^\star e^{-C_4^\star \eta n}\right).$$