Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization: Non-Asymptotic Performance Bounds and Momentum-Based Acceleration

Xuefeng Gao [*†], Mert Gürbüzbalaban [*‡], Lingjiong Zhu [*§]

November 19, 2020

Abstract

Stochastic gradient Hamiltonian Monte Carlo (SGHMC) is a variant of stochastic gradient with momentum where a controlled and properly scaled Gaussian noise is added to the stochastic gradients to steer the iterates towards a global minimum. Many works reported its empirical success in practice for solving stochastic non-convex optimization problems, in particular it has been observed to outperform overdamped Langevin Monte Carlo-based methods such as stochastic gradient Langevin dynamics (SGLD) in many applications. Although asymptotic global convergence properties of SGHMC are well known, its finite-time performance is not well-understood. In this work, we study two variants of SGHMC based on two alternative discretizations of the underdamped Langevin diffusion. We provide finite-time performance bounds for the global convergence of both SGHMC variants for solving stochastic non-convex optimization problems with explicit constants. Our results lead to non-asymptotic guarantees for both population and empirical risk minimization problems. For a fixed target accuracy level, on a class of non-convex problems, we obtain complexity bounds for SGHMC that can be tighter than those available for SGLD.

[*]The authors are in alphabetical order.

[†]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong; xfgao@se.cuhk.edu.hk.

[‡]Department of Management Science and Information Systems and the DIMACS Institute, Rutgers University, Piscataway, NJ-08854, United States of America; mg1366@rutgers.edu.

[§]Department of Mathematics, Florida State University, 1017 Academic Way, Tallahassee, FL-32306, United States of America; zhu@math.fsu.edu.

1 Introduction

We consider the stochastic non-convex optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) := \mathbb{E}_{Z\sim\mathcal{D}}[f(x,Z)]\,, \tag{1.1}$$

where Z is a random variable whose probability distribution $\mathcal{D}$ is unknown, supported on some unknown set $\mathcal{Z}$, the objective F is the expectation of a random function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ where the functions $x \mapsto f(x,z)$ are continuous and non-convex. Having access to independent and identically distributed (i.i.d.) samples $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ where each Z_i is a random variable distributed with the population distribution $\mathcal{D}$, the goal is to compute an approximate minimizer $\hat{x}$ (possibly with a randomized algorithm) of the *population risk*, i.e. we want to compute $\hat{x}$ such that $\mathbb{E}F(\hat{x}) - F^* \leq \hat{\varepsilon}$ for a given target accuracy $\hat{\varepsilon} > 0$, where $F^* = \min_{x\in\mathbb{R}^d} F(x)$ is the minimum value and the expectation is taken with respect to both $\mathbf{Z}$ and the randomness encountered (if any) during the iterations of the algorithm to compute $\hat{x}$. This formulation arises frequently in several contexts including machine learning. A prominent example is deep learning where x denotes the set of trainable weights for a deep learning model and $f(x,z_i)$ is the penalty (loss) of prediction using weight x with the individual sample value $Z_i = z_i \in \mathcal{Z}$.

Because the population distribution $\mathcal{D}$ is unknown, a common popular approach is to consider the *empirical risk minimization* problem

$$\min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,z_i)\,, \tag{1.2}$$

based on the dataset $\mathbf{z} := (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$ as a proxy to the problem (1.1) and minimize the *empirical risk*

$$\mathbb{E}F_{\mathbf{z}}(x) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \tag{1.3}$$

instead, where the expectation is taken with respect to any randomness encountered during the algorithm to generate x.[1] Many algorithms have been proposed to solve the problem (1.1) and its finite-sum version (1.2). Among these, gradient descent, stochastic gradient and their variance-reduced or momentum-based variants come with guarantees for finding a local minimizer or a stationary point for non-convex problems. In some applications, convergence to a local minimum can be satisfactory ([GLM17, DLT+18]). However, in general, methods with global convergence guarantees are also desirable and preferable in many settings ([HLSS16, ŞimşekliYN+18]).

It has been well known that sampling from a distribution which concentrates around a global minimizer of F is a similar goal to computing an approximate global minimizer

[1] We note that in our notation $\mathbf{Z}$ is a random vector, whereas $\mathbf{z}$ is deterministic vector associated to a dataset that corresponds to a realization of the random vector $\mathbf{Z}$.

of F. For example such connections arise in the study of simulated annealing algorithms in optimization which admit several asymptotic convergence guarantees (see e.g. [Gid85, Haj85, GM91, KGV83, BT93, BLNR15, BM99]). Recent studies made such connections between the fields of statistics and optimization stronger, justifying and popularizing the use of Langevin Monte Carlo-based methods in stochastic non-convex optimization and large-scale data analysis further (see e.g. [CCS+17, Dal17, RRT17, CCG+16, ŞimşekliBCR16, ŞimşekliYN+18, WT11, Wib18]).

Stochastic gradient algorithms based on Langevin Monte Carlo are popular variants of stochastic gradient which admit asymptotic global convergence guarantees where a properly scaled Gaussian noise is added to the gradient estimate. Two popular Langevin-based algorithms that have demonstrated empirical success are stochastic gradient Langevin dynamics (SGLD) ([WT11, CDC15]) and stochastic gradient Hamiltonian Monte Carlo (SGHMC) ([CFG14, CDC15, Nea10, DKPR87]) and their variants to improve their efficiency and accuracy ([AKW12, MCF15, PT13, DFB+14, Wib18]). In particular, SGLD can be viewed as the analogue of stochastic gradient in the Markov Chain Monte Carlo (MCMC) literature whereas SGHMC is the analogue of stochastic gradient with momentum (see e.g. [CFG14]). SGLD iterations consist of

$$X_{k+1} = X_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k\,,$$

where $\eta > 0$ is the stepsize parameter, $\beta > 0$ is the inverse temperature, g_k is a conditionally unbiased estimate of the gradient of $F_{\mathbf{z}}$ and $\xi_k \in \mathbb{R}^d$ is a sequence of i.i.d. centered Gaussian random vector with unit covariance matrix. When the gradient variance is zero, SGLD dynamics corresponds to (explicit) Euler discretization of the first-order (a.k.a. overdamped) Langevin stochastic differential equation (SDE)

$$dX(t) = -\nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\beta^{-1}}dB(t)\,, \quad t \geq 0\,, \tag{1.4}$$

where $\{B(t) : t \geq 0\}$ is the standard Brownian motion in $\mathbb{R}^d$. The process X admits a unique stationary distribution $\pi_{\mathbf{z}}(dx) \propto \exp(-\beta F_{\mathbf{z}}(x))dx$, also known as the *Gibbs measure*, under some assumptions on $F_{\mathbf{z}}$ (see e.g. [CHS87, HKS89]). For β chosen properly (large enough), it is easy to see that this distribution will concentrate around approximate global minimizers of $F_{\mathbf{z}}$. Recently, [Dal17] established novel theoretical guarantees for the convergence of the overdamped Langevin MCMC and the SGLD algorithm for sampling from a smooth and log-concave density and these results have direct implications to stochastic convex optimization; see also [DK19]. In a seminal work, [RRT17] showed that SGLD iterates track the overdamped Langevin SDE closely and obtained finite-time performance bounds for SGLD. Their results show that SGLD converges to ε-approximate global minimizers after $\mathcal{O}(\text{poly}(\frac{1}{\lambda_*}, \beta, d, \frac{1}{\varepsilon}))$ iterations where λ_* is the uniform spectral gap that controls the convergence rate of the overdamped Langevin diffusion which is in general exponentially small in both β and the dimension d ([RRT17, TLR18]). A related result of [ZLC17] shows that a modified version of the SGLD algorithm will find an ε-approximate local minimum

after polynomial time (with respect to all parameters). Recently, [XCZG18] improved the ε dependency of the upper bounds of [RRT17] further in the mini-batch setting, and obtained several guarantees for the gradient Langevin dynamics and variance-reduced SGLD algorithms.

On the other hand, the SGHMC algorithm is based on the underdamped (a.k.a. second-order or kinetic) Langevin diffusion

$$dV(t) = -\gamma V(t)dt - \nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\gamma\beta^{-1}}dB(t), \tag{1.5}$$

$$dX(t) = V(t)dt, \tag{1.6}$$

where $\gamma > 0$ is the friction coefficient, $X(t), V(t) \in \mathbb{R}^d$ models the position and the momentum of a particle moving in a field of force (described by the gradient of $F_{\mathbf{z}}$) plus a random (thermal) force described by Brownian noise, first derived by [Kra40]. It is known that under some assumptions on $F_{\mathbf{z}}$, the Markov process $(X(t), V(t))_{t\geq 0}$ is ergodic and admits a unique stationary distribution

$$\pi_{\mathbf{z}}(dx, dv) = \frac{1}{\Gamma_{\mathbf{z}}} \exp\left(-\beta\left(\frac{1}{2}\|v\|^2 + F_{\mathbf{z}}(x)\right)\right) dxdv, \tag{1.7}$$

(see e.g. [HN04, Pav14]) where $\Gamma_{\mathbf{z}}$ is the normalizing constant:

$$\Gamma_{\mathbf{z}} = \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(-\beta\left(\frac{1}{2}\|v\|^2 + F_{\mathbf{z}}(x)\right)\right) dxdv = \left(\frac{2\pi}{\beta}\right)^{d/2} \int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(x)}dx.$$

Hence, the x-marginal distribution of stationary distribution $\pi_{\mathbf{z}}(dx, dv)$ is exactly the invariant distribution of the overdamped Langevin diffusion.[2] SGHMC dynamics correspond to the discretization of the underdamped Langevin SDE where the gradients are replaced with their unbiased estimates. Although various discretizations of the underdamped Langevin SDE has also been considered and studied ([CDC15, LMS15]), the following first-order Euler scheme is the simplest approach that is easy to implement, and a common scheme among the practitioners ([TTV16, CCG+16, CDC15]):

$$V_{k+1} = V_k - \eta[\gamma V_k + g(X_k, U_{\mathbf{z},k})] + \sqrt{2\gamma\beta^{-1}\eta}\xi_k, \tag{1.8}$$

$$X_{k+1} = X_k + \eta V_k, \tag{1.9}$$

where $(\xi_k)_{k=0}^{\infty}$ is a sequence of i.i.d standard Gaussian random vectors in $\mathbb{R}^d$, $\{U_{\mathbf{z},k} : k = 0, 1, \ldots\}$ is a sequence of i.i.d random elements such that

$$\mathbb{E}g(x, U_{\mathbf{z},k}) = \nabla F_{\mathbf{z}}(x) \quad \text{for any } x \in \mathbb{R}^d.$$

[2] With slight abuse of notation, we use $\pi_{\mathbf{z}}(dx)$ to denote the x-marginal of the equilibrium distribution $\pi_{\mathbf{z}}(dx, dv)$.

In this paper, we focus on the unadjusted dynamics (without Metropolis-Hastings type of correction) that works well in many applications ([CFG14, CDC15]), as Metropolis-Hastings correction is typically computationally expensive for applications in machine learning and large-scale optimization when the size of the dataset n is large and low to medium accuracy is enough in practice (see e.g. [WT11, CFG14]).

There is also an alternative discretization to (1.8)-(1.9), recently proposed by [CCBJ18] which leads to state-of-the-art estimates in the special case that improves upon the Euler discretization when the objective is strongly convex ([CCBJ18]). To introduce this alternative discretization by [CCBJ18], we first define a sequence of functions ψ_k by $\psi_0(t) = e^{-\gamma t}$ and $\psi_{k+1}(t) = \int_0^t \psi_k(s)ds$, $k \geq 0$. The iterates $(\hat{X}_k, \hat{V}_k)$ are then defined by the following recursion:

$$\hat{V}_{k+1} = \psi_0(\eta)\hat{V}_k - \psi_1(\eta)g(\hat{X}_k, U_{\mathbf{z},k}) + \sqrt{2\gamma\beta^{-1}}\xi_{k+1}, \tag{1.10}$$

$$\hat{X}_{k+1} = \hat{X}_k + \psi_1(\eta)\hat{V}_k - \psi_2(\eta)g(\hat{X}_k, U_{\mathbf{z},k}) + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1}, \tag{1.11}$$

where (ξ_{k+1}, ξ'_{k+1}) is a $2d$-dimensional centered Gaussian vector so that (ξ_j, ξ'_j)'s are independent and identically distributed (i.i.d.) and independent of the initial condition, and for any fixed j, the random vectors $((\xi_j)_1, (\xi'_j)_1)$, $((\xi_j)_2, (\xi'_j)_2)$, $\ldots$ $((\xi_j)_d, (\xi'_j)_d)$ are i.i.d. with the covariance matrix:

$$C(\eta) = \int_0^\eta [\psi_0(t), \psi_1(t)]^T [\psi_0(t), \psi_1(t)]\, dt. \tag{1.12}$$

In the rest of the paper, we refer to Euler discretization (1.8)-(1.9) as SGHMC1 whereas the alternative discretization (1.10)-(1.11) as SGHMC2.

Recently, [EGZ19] show that the underdamped SDE converges to its stationary distribution faster than that of the best known convergence rate of overdamped SDE in the 2-Wasserstein metric under some assumptions, where $F_{\mathbf{z}}$ can be non-convex. Their result is for the continuous-time underdamped dynamics. This raises the natural question whether the discretized underdamped dynamics (SGHMC), can lead to better guarantees than the SGLD method for solving stochastic non-convex optimization problems. Indeed, experimental results show that SGHMC can outperform SGLD dynamics in many applications (see e.g. [EGZ19, CDC15, CFG14]). Although asymptotic convergence guarantees for SGHMC exist (see e.g. [CFG14] [MSH02, Section 3], [LMS15]), there is a lack of finite-time explicit performance bounds for solving non-convex stochastic optimization problems with SGHMC in the literature including risk minimization problems.

1.1 Contributions

Our main contributions can be summarized as follows:

- We provide for the first time the non-asymptotic provable guarantees for SGHMC to find approximate minimizers of both empirical and population risks with explicit

constants. We establish the results under some regularity and growth assumptions for the component functions $f(x, z)$ and the noise in the gradients, but we do not assume f is strongly convex in any region.

- We show that for a class of non-convex problems, SGHMC2 can improve upon the (vanilla) SGLD algorithm in terms of the *gradient complexity*, i.e. the total number of stochastic gradients required to achieve a global minimum. Here, "improvement" means the best available bounds for SGHMC2, which we prove in our paper, are better than the best available bounds for SGLD for some class of problems; see Section 5 for details. As a consequence, our analysis gives further theoretical justification to the success of momentum-based methods for solving non-convex machine learning problems, empirically observed in practice (see e.g. [SMDH13]).

- We illustrate the applications of our theoretical results using two examples including binary linear classification and robust ridge regression.

- On the technical side, we adapt the proof techniques of [RRT17] developed for the overdamped dynamics to the underdamped dynamics and combine it with the analysis of [EGZ19] which quantifies the convergence rate of the underdamped Langevin SDE to its equilibrium. The main new technical results we derive in this paper, relative to these studies, include controlling the discretization errors between SGHMC and the continuous-time underdamped Langevin SDE, and bounding the moments of underdamped dynamics.

1.2 Related Work and Comparison to Existing Literature

In a recent work, [ŞimşekliYN+18] obtained a finite-time performance bound for the ergodic average of the SGHMC iterates in the presence of delays in gradient computations. Their analysis highlights the dependency of the optimization error on the delay in the gradient computations and the stepsize explicitly, however it hides some implicit constants which can be exponential both in β and d in the worst case. A comparison with the SGLD algorithm is also not given. On the contrary, in our paper, we make all the constants explicit. This allows us to make gradient complexity comparisons with respect to overdamped MCMC approaches such as SGLD.

[CCA+18] considered the problem of sampling from a target distribution $p(x) \propto \exp(-F(x))$ where $F : \mathbb{R}^d \to \mathbb{R}$ is L-smooth everywhere and m-strongly convex outside a ball of finite radius $\mathcal{R}$. They proved upper bounds for the time required to sample from a distribution that is within ε of the target distribution with respect to the 1-Wasserstein distance for both underdamped and overdamped methods that scales polynomially in ε and d. They also show that underdamped MCMC has a better dependency with respect to ε and d by a square root factor. Compared to this paper, in our analysis, we consider a larger class of non-convex functions $F(x)$ that satisfy the dissipativity condition, a weaker

condition that does not require strong convexity outside a region. Under our assumptions, the best known bounds are such that the distance to the invariant distribution scales exponentially with dimension d in the worst-case but not polynomially in d (see e.g. [RRT17, XCZG18]). When F is globally strongly convex (or equivalently when the target distribution $p(x) \propto \exp(-F(x))$ is strongly log-concave), there is also a growing interesting literature that establish performance bounds for both overdamped MCMC (see e.g. [Dal17]) and underdamped MCMC methods (see e.g. [CCBJ18, MS17]). In this particular setting, the fact that underdamped Langevin MCMC (also known as Hamiltonian MCMC) can improve upon the best available bounds for overdamped Langevin MCMC algorithms has also been proven ([CCBJ18, MS17, DRD20, CFM$^+$18]. Similar results have also been established when $F(x)$ is convex but not strongly convex ([DKRD19]). Compared to these papers where $F(x)$ is convex, our assumptions are weaker as we allow $F(x)$ to be non-convex as long as it is dissipative.

A related paper [XCZG18] applies variance reduction techniques to overdamped MCMC to improve performance when the empirical risk can be non-convex satisfying the same dissipativity assumption considered in our paper. However, these results do not give guarantees for the risk minimization problem (1.1). Furthermore, such variance reduction techniques require objectives in the form of a finite sum and do not apply to the *streaming data setting* when each data point is used only once. In this work, we obtain guarantees for both the risk minimization problem and the empirical risk minimization and our results apply to the streaming data setting. Also, the convergence guarantees provided in [XCZG18] depends on a spectral gap-related parameter that is not provided explicitly; whereas all our results are explicit and this allows us to have explicit performance comparisons between the upper bounds of SGLD and SGHMC algorithms.

We also note that underdamped Langevin MCMC (also known as Hamiltonian MCMC) and its practical applications have also been analyzed further in a number of recent works (see e.g. [LV18, BBLG17, Bet17, BBG14, MPS18]). In particular, [MPS18] provide a characterization of the conductance of Hamiltonian Monte Carlo (HMC) in continuous time using Liouville's theorem and invoking the Cheeger's inequality, they obtain upper and lower bounds on the spectral gap of HMC in continuous-time. Although the formula provided in [MPS18] for the conductance of HMC is elegant, it is not an explicit formula. In our analysis, our focus is to obtain performance bounds with explicit constants and therefore we build on the coupling techniques of [EGZ19] which leads to explicit constants for the class of problems we consider.

We also note that [MPS18] consider sampling from the target distribution $\frac{1}{2}\mathcal{N}(-1,\sigma^2)+\frac{1}{2}\mathcal{N}(1,\sigma^2)$ in dimension one and estimate the spectral gap of HMC in the regime as $\sigma \to 0$. This is a mixture of two Gaussians with the same variance σ^2 centered at -1 and 1 respectively where they argue that for this specific example HMC does not lead to much improvement over the Random Walk approach for sampling. In our paper, our results apply to more general targets that are not necessarily mixture of Gaussians. However, if we consider sampling from the distribution $\frac{1}{2}\mathcal{N}(-a,\sigma^2)+\frac{1}{2}\mathcal{N}(a,\sigma^2)$ as $a \to \infty$ for σ^2

fixed, Proposition 11 is applicable and it implies that HMC will be more efficient than overdamped Langevin dynamics in terms of dependency to a (which measures the distance between the modes) in the sense that the mixing time will be $\mathcal{O}(a)$ in HMC whereas it will be $\mathcal{O}(a^2)$ in Random Walk. This does not contradict results of [MPS18] because we consider different scaling regimes: We fix $\sigma > 0$ and let $a \to \infty$ whereas [MPS18] fix $a = 1$ and let $\sigma \to 0$.

There are also some connections of our work to existing momentum-based optimization algorithms. More specifically, if the term with $dB(t)$ involving the Brownian noise is removed in the underdamped SDE (1.5)–(1.6), this results in a second-order ODE in $X(t)$. Momentum-based algorithms for strongly convex objectives such as Polyak's heavy ball method as well as Nesterov's accelerated gradient method can be both viewed as (alternative) discretizations of this ODE (see e.g. [Pol87, SBC14, SDJS18, WRJ16]). It is known ([SBC14, SDJS18, WRJ16]) that Nesterov's accelerated gradient method tracks this second-order ODE (also referred to as the Nesterov's ODE in the literature), whereas the first-order non-accelerated methods such as the classical gradient descent are known to track a first-order ODE in $X(t)$ called the *gradient flow* dynamics. Furthermore, existing analysis shows that Nesterov's ODE converges to its equilibrium faster (in time) than the first-order gradient flow ODE in terms of upper bounds and this speed-up is also inherited by the discretized dynamics. Roughly speaking, our results can be interpreted as the analogue of these results in the non-convex optimization setting where we deal with SDEs instead of ODEs building on the theory of Markov processes and show that SGHMC tracks the second-order (underdamped) Langevin SDE closely and inherits its favorable convergence guarantees (in terms of upper bounds on the expected suboptimality) compared to that of overdamped Langevin SDE.

Acceleration of first-order gradient or stochastic gradient methods and their variance-reduced versions for finding a local stationary point (a point with a gradient less than ε in norm) are also studied in the literature (see e.g. [CDHS18, Nes83, GL16, JT19, AZH16]). It has also been shown that under some assumptions momentum-based accelerated methods can escape saddle points faster (see e.g. [OW19, LCZZ18]). In contrast, in this work, our focus is obtaining performance guarantees for convergence to global minimizers instead.

2 Preliminaries and Assumptions

In our analysis, we will use the following 2-Wasserstein distance: For any two probability measures ν_1, ν_2 on $\mathbb{R}^{2d}$, we define

$$\mathcal{W}_2(\nu_1, \nu_2) = \left(\inf_{Y_1 \sim \nu_1, Y_2 \sim \nu_2} \mathbb{E}\left[\|Y_1 - Y_2\|^2 \right] \right)^{1/2},$$

where $\|\cdot\|$ is the usual Euclidean norm, ν_1, ν_2 are two Borel probability measures on $\mathbb{R}^{2d}$ with finite second moments, and the infimum is taken over all random couples (Y_1, Y_2)

taking values in $\mathbb{R}^{2d} \times \mathbb{R}^{2d}$ with marginals $Y_1 \sim \nu_1, Y_2 \sim \nu_2$ (see e.g. [Vil08]). We let $C^1(\mathbb{R}^d)$ denote the set of continuously differentiable functions on $\mathbb{R}^d$ and $L^2(\pi_{\mathbf{z}})$ denote the space of square-integrable functions on $\mathbb{R}^d$ with respect to the measure $\pi_{\mathbf{z}}$.

We first state the assumptions used in this paper below in Assumption 1. Note that we do not assume the component functions $f(x, z)$ to be convex; they can be non-convex. The first assumption of non-negativity of f can be assumed without loss of generality by subtracting a constant and shifting the coordinate system as long as f is bounded below. The second assumption of Lipschitz gradients is in general unavoidable for discretized Langevin algorithms to be convergent (see e.g. [MSH02]), and the third assumption is known as the *dissipativity condition* (see e.g. [Hal88]) and is standard in the literature to ensure the convergence of Langevin diffusions to the stationary distribution (see e.g. [RRT17, EGZ19, MSH02]). The fourth assumption is regarding the amount of noise present in the gradient estimates and allows not only constant variance noise but allows the noise variance to grow with the norm of the iterates (which is the typical situation in mini-batch methods in stochastic gradient methods, see e.g. [RRT17]). Finally, the fifth assumption is a mild assumption saying that the initial distribution μ_0 for the SGHMC dynamics should have a reasonable decay rate of the tails to ensure convergence to the stationary distribution. For instance, if the algorithm is started from any arbitrary point $(x_0, v_0) \in \mathbb{R}^{2d}$, then the Dirac measure $\mu_0(dx, dv) = \delta_{(x_0,v_0)}(dx, dv)$ would work. If the initial distribution $\mu_0(dx, dv)$ is supported on a Euclidean ball with radius being some universal constant, it would also work. Similar assumptions on the initial distribution μ_0 is also necessary to achieve convergence to a stationary measure in continuous-time underdamped dynamics as well (see e.g. [HN04]).

Assumption 1. *We impose the following assumptions.*

(*i*) *The function f is continuously differentiable, takes non-negative real values, and there exist constants $A_0, B \geq 0$ so that*

$$|f(0, z)| \leq A_0, \qquad \|\nabla f(0, z)\| \leq B,$$

for any $z \in \mathcal{Z}$.

(*ii*) *For each $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is M-smooth:*

$$\|\nabla f(w, z) - \nabla f(v, z)\| \leq M\|w - v\|.$$

(*iii*) *For each $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is (m, b)-dissipative:*

$$\langle x, \nabla f(x, z)\rangle \geq m\|x\|^2 - b\,.$$

(*iv*) *There exists a constant $\delta \in [0, 1)$ such that for every $\mathbf{z}$:*

$$\mathbb{E}[\|g(x, U_{\mathbf{z}}) - \nabla F_{\mathbf{z}}(x)\|^2] \leq 2\delta(M^2\|x\|^2 + B^2)\,.$$

(*v*) *The probability law* μ_0 *of the initial state* (X_0, V_0) *satisfies:*

$$\int_{\mathbb{R}^{2d}} e^{\alpha \mathcal{V}(x,v)} \mu_0(dx, dv) < \infty \,,$$

where $\mathcal{V}$ *is a Lyapunov function to be used repeatedly for the rest of the paper:*

$$\mathcal{V}(x, v) := \beta F_{\mathbf{z}}(x) + \frac{\beta}{4}\gamma^2(\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2) \,, \tag{2.1}$$

and γ *is the friction coefficient as in* (1.5)*,* λ *is a positive constant less than* $\min(1/4, m/(M + \gamma^2/2))$*, and* $\alpha = \lambda(1 - 2\lambda)/12$*.*

We note that the Lyapunov function $\mathcal{V}$ is used in [EGZ19] to study the rate of convergence to equilibrium for underdamped Langevin diffusion, which itself is motivated by e.g. [MSH02]. It follows from the above assumptions (applying Lemma 25) that there exists a constant $A \in (0, \infty)$ so that

$$x \cdot \nabla F_{\mathbf{z}}(x) \geq m\|x\|^2 - b \geq 2\lambda(F_{\mathbf{z}}(x) + \gamma^2\|x\|^2/4) - 2A/\beta \,. \tag{2.2}$$

This drift condition, which will be used later, guarantees the stability and the existence of Lyapunov function $\mathcal{V}$ for the underdamped Langevin diffusion in (1.5)–(1.6), see [EGZ19].

3 Main Results for SGHMC1 Algorithm

Our first result shows SGHMC1 iterates (X_k, V_k) in (1.8)–(1.9) track the underdamped Langevin SDE in the sense that the expectation of the empirical risk $F_{\mathbf{z}}$ with respect to the probability law of (X_k, V_k) conditional on the sample $\mathbf{z}$, denoted by $\mu_{k,\mathbf{z}}$, and the stationary distribution $\pi_{\mathbf{z}}$ of the underdamped SDE is small when k is large enough. The difference in expectations decomposes as a sum of two terms $\mathcal{J}_0(\mathbf{z}, \varepsilon)$ and $\mathcal{J}_1(\varepsilon)$ while the former term quantifies the dependency on the initialization and the dataset $\mathbf{z}$ whereas the latter term is controlled by the discretization error and the amount of noise in the gradients which depends on the parameter δ. We also note that the parameter μ_* (see Table 1) in our bounds governs the speed of convergence to the equilibrium of the underdamped Langevin diffusion.

Theorem 2. *Consider the SGHMC1 iterates* (X_k, V_k) *defined by the recursion* (1.8)–(1.9) *from the initial state* (X_0, V_0) *which has the law* μ_0*. If Assumption 1 is satisfied, then for* $\beta, \varepsilon > 0$*, we have*

$$\begin{aligned} \left|\mathbb{E}F_{\mathbf{z}}(X_k) - \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X))\right| &= \left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\mu_{k,\mathbf{z}}(dx, dv) - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx, dv)\right| \\ &\leq \mathcal{J}_0(\mathbf{z}, \varepsilon) + \mathcal{J}_1(\varepsilon) \,, \end{aligned}$$

where

$$\mathcal{J}_0(\mathbf{z},\varepsilon) := (M\sigma + B)\cdot C\sqrt{\mathcal{H}_\rho(\mu_0,\pi_{\mathbf{z}})}\cdot\varepsilon, \tag{3.1}$$

$$\mathcal{J}_1(\varepsilon) := (M\sigma + B)\cdot\left(\left(\frac{C_0}{\mu_*^{3/2}}(\log(1/\varepsilon))^{3/2}\delta^{1/4} + \frac{C_1}{\mu_*^{3/2}}\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))} + \frac{C_2}{\mu_*}\frac{\varepsilon^2}{(\log(1/\varepsilon))^2}\right), \tag{3.2}$$

with σ defined by (A.20) provided that

$$\eta \le \min\left\{\left(\frac{\varepsilon}{(\log(1/\varepsilon))^{3/2}}\right)^4, 1, \frac{\gamma}{K_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2K_1}, \frac{2}{\gamma\lambda}\right\}, \tag{3.3}$$

and

$$k\eta = \frac{1}{\mu_*}\log\left(\frac{1}{\varepsilon}\right) \ge e. \tag{3.4}$$

Here $\mathcal{H}_\rho$ is a semi-metric for probability distributions defined by (A.12). All the constants are made explicit and are summarized in Table 1.

The proof of Theorem 2 will be presented in details in Section A in the Appendix. In the following subsections, we discuss that this theorem combined with some basic properties of the equilibrium distribution $\pi_{\mathbf{z}}$ leads to a number of results which provide performance guarantees for both the empirical risk and population risk minimization.

3.1 Performance bound for the empirical risk minimization

In order to obtain guarantees for the empirical risk given in (1.3), in light of Theorem 2, one has to control the quantity

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x)\,,$$

which is a measure of how much the $x-$marginal of the equilibrium distribution $\pi_{\mathbf{z}}$ concentrates around a global minimizer of the empirical risk. As β goes to infinity, it can be verified that this quantity goes to zero. For finite β, [RRT17] (see Proposition 11) derives an explicit bound of the form

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \le \mathcal{J}_2 := \frac{d}{2\beta}\log\left(\frac{eM}{m}\left(\frac{b\beta}{d}+1\right)\right), \tag{3.5}$$

(which is also provided in the Appendix for the sake of completeness, see Lemma 28). This combined with Theorem 2 immediately leads to the following performance bound for the empirical risk minimization. The proof is omitted.

Corollary 3 (Empirical risk minimization with SGHMC1)**.** *Under the setting of Theorem 2, the empirical risk minimization problem admits the performance bounds:*

$$\mathbb{E}F_{\mathbf{z}}(X_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \leq \mathcal{J}_0(\varepsilon, \mathbf{z}) + \mathcal{J}_1(\varepsilon) + \mathcal{J}_2\,,$$

provided that conditions (3.3) and (3.4) hold where the terms $\mathcal{J}_0(\mathbf{z},\varepsilon)$, $\mathcal{J}_1(\varepsilon)$ and $\mathcal{J}_2$ are defined by (3.1), (3.2) and (3.5) respectively.

3.2 Performance bound for the population risk minimization

By exploiting the fact that the $x-$marginal of the invariant distribution for the underdamped dynamics is the same as it is in the overdamped case, it can be shown that the generalization error $F(X_k) - F_{\mathbf{Z}}(X_k)$ is no worse than that of the available bounds for SGLD given in [RRT17], and therefore, we have the following corollary. A more detailed proof will be given in Section A in the Appendix.

Corollary 4 (Population risk minimization with SGHMC1)**.** *Under the setting of Theorem 2, the expected population risk of X_k (the iterates in (1.9)) is bounded by*

$$\mathbb{E}F(X_k) - F^* \leq \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon) + \mathcal{J}_2 + \mathcal{J}_3(n)\,,$$

with

$$\overline{\mathcal{J}}_0(\varepsilon) := (M\sigma + B)\cdot C\cdot\sqrt{\overline{\mathcal{H}}_\rho(\mu_0)}\cdot\varepsilon, \tag{3.6}$$

$$\mathcal{J}_3(n) := \frac{4\beta c_{LS}}{n}\left(\frac{M^2}{m}(b + d/\beta) + B^2\right), \tag{3.7}$$

where σ is defined by (A.20), $\overline{\mathcal{H}}_\rho(\mu_0)$ is defined by (A.18), $\mathcal{J}_1(\varepsilon)$ and $\mathcal{J}_2$ are defined by (3.2) and (3.5) respectively and c_{LS} is a constant satisfying

$$c_{LS} \leq \frac{2m^2 + 8M^2}{m^2 M\beta} + \frac{1}{\lambda_*}\left(\frac{6M(d+\beta)}{m} + 2\right),$$

and λ_* is the uniform spectral gap for overdamped Langevin dynamics [3]:

$$\lambda_* := \inf_{\mathbf{z}\in\mathcal{Z}^n}\inf\left\{\frac{\beta^{-1}\int_{\mathbb{R}^d}\|\nabla g\|^2 d\pi_{\mathbf{z}}}{\int_{\mathbb{R}^d} g^2 d\pi_{\mathbf{z}}} : g\in C^1(\mathbb{R}^d)\cap L^2(\pi_{\mathbf{z}}), g\neq 0, \int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0\right\}. \tag{3.8}$$

[3] In [RRT17], their formula for λ_* missed β^{-1} factor.

Constants	*Source*
$C_x^c = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \quad C_v^c = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{(d+A)}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}$	(A.1), (A.2)
$K_1 = \max\left\{ \frac{32M^2\left(\frac{1}{2}+\gamma+\delta\right)}{(1-2\lambda)\beta\gamma^2}, \frac{8\left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)}{\beta(1-2\lambda)} \right\}$	(A.3)
$K_2 = B^2\left(1+2\gamma+2\delta\right)$	(A.4)
$C_x^d = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \quad C_v^d = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}$	(A.5), (A.6)
$\sigma = \max\left\{\sqrt{C_x^c}, \sqrt{C_x^d}\right\} = \sqrt{C_x^d}$	(A.20)
$C_0 = \hat{\gamma}\cdot\left(\left(M^2C_x^d + B^2\right)\beta/\gamma + \sqrt{\left(M^2C_x^d + B^2\right)\beta/\gamma}\right)^{1/2}$	(A.8)
$C_1 = \hat{\gamma}\cdot\left(\beta M^2(C_2)^2/(2\gamma) + \sqrt{\beta M^2(C_2)^2/(2\gamma)}\right)^{1/2}$	(A.9)
$C_2 = \left(2\gamma^2 C_v^d + (4+2\delta)\left(M^2C_x^d + B^2\right) + 2\gamma\beta^{-1}\right)^{1/2}$	(A.10)
$\hat{\gamma} = \frac{2\sqrt{2}}{\sqrt{\alpha_0}}\left(\frac{5}{2} + \log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv) + \frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)\right)\right)^{1/2}$	(A.11)
$\alpha_0 = \frac{\alpha(1-2\lambda)\beta\gamma^2}{64+32\gamma^2}, \qquad \alpha = \frac{\lambda(1-2\lambda)}{12}$	(A.7)
$\mu_* = \frac{\gamma}{768}\min\{\lambda M\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}M\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}\}$	(A.13)
$C = \frac{(1+\gamma)\sqrt{2}e^{1+\frac{\Lambda}{2}}}{\min\{1,\alpha_1\}}\sqrt{\max\{1, 4(1+2\alpha_1+2\alpha_1^2)(d+A)\beta^{-1}\gamma^{-1}\mu_*^{-1}/\min\{1,R_1\}\}}$	(A.14)
$\Lambda = \frac{12}{5}(1+2\alpha_1+2\alpha_1^2)(d+A)M\gamma^{-2}\lambda^{-1}(1-2\lambda)^{-1}, \qquad \alpha_1 = (1+\Lambda^{-1})M\gamma^{-2}$	(A.15)
$\varepsilon_1 = 4\gamma^{-1}\mu_*/(d+A)$	(A.16)
$R_1 = 4\cdot(6/5)^{1/2}(1+2\alpha_1+2\alpha_1^2)^{1/2}(d+A)^{1/2}\beta^{-1/2}\gamma^{-1}(\lambda-2\lambda^2)^{-1/2}$	(A.17)
$\overline{\mathcal{H}}_\rho(\mu_0) = R_1 + R_1\varepsilon_1\max\left\{M+\frac{1}{2}\beta\gamma^2, \frac{3}{4}\beta\right\}\|(x,v)\|^2_{L^2(\mu_0)} + R_1\varepsilon_1\left(M+\frac{1}{2}\beta\gamma^2\right)\frac{b+d/\beta}{m} + R_1\varepsilon_1\frac{3}{4}d + 2R_1\varepsilon_1\left(\beta A_0 + \frac{\beta B^2}{2M}\right)$	(A.18)

Table 1: Summary of the constants and where they are defined in the text.

3.3 Generalization error of SGHMC1 in the one pass regime

Since the predictor X_k is random, it is natural to consider the expected generalization error $\mathbb{E}F(X_k) - \mathbb{E}F_{\mathbf{Z}}(X_k)$ (see e.g. [HRS16]) which admits the decomposition

$$\begin{aligned}\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k) = (\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)) + (\mathbb{E}F_{\mathbf{Z}}(X^\pi) - \mathbb{E}F(X^\pi)) \\ + (\mathbb{E}F(X^\pi) - \mathbb{E}F(X_k))\ ,\end{aligned} \tag{3.9}$$

where X^π is the Gibbs output, i.e. its distribution conditional on $\mathbf{Z} = \mathbf{z}$ is given by $\pi_{\mathbf{z}}$. If every sample is used once, i.e. if only one pass is made over the dataset, then the second term in (3.9) disappears. As a consequence, the generalization error is controlled by the bound

$$|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k)| \le |\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)| + |\mathbb{E}F(X^\pi) - \mathbb{E}F(X_k)|\,. \tag{3.10}$$

The following result provides a bound on this quantity. The proof is similar to the proof of Theorem 2 and its corollaries, and hence omitted.

Theorem 5 (Generalization error of SGHMC1)**.** *Under the setting of Theorem 2, we have*

$$\begin{aligned}|\mathbb{E}F(X_k) - \mathbb{E}F(X^\pi)| &\le \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon)\,, \\ |\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)| &\le \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon),\end{aligned}$$

provided that (3.3) and (3.4) hold where X^π is the output of the underdamped Langevin dynamics, i.e. its distribution conditional on $\mathbf{Z} = \mathbf{z}$ is given by $\pi_{\mathbf{z}}$ and $\overline{\mathcal{J}}_0(\varepsilon)$ is defined by (3.6). Then, it follows from (3.10) that if each data point is used once, the expected generalization error satisfies

$$|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k)| \le 2\overline{\mathcal{J}}_0(\varepsilon) + 2\mathcal{J}_1(\varepsilon).$$

4 Main Results for SGHMC2 Algorithm

Recall the SGHMC2 algorithm $(\hat{X}_k, \hat{V}_k)$ defined in (1.10)-(1.11), and denote the probability law of $(\hat{X}_k, \hat{V}_k)$ conditional on the sample $\mathbf{z}$ by $\hat{\mu}_{k,\mathbf{z}}(dx, dv)$. Similar to our analysis for SGHMC1, we can derive similar performance guarantees for SGHMC2 in terms of empirical risk, population risk and the generalization error. The main difference is that the term $\mathcal{J}_1(\varepsilon)$ is controlled by the accuracy of the discretization and has to be replaced by another term $\hat{\mathcal{J}}_1(\varepsilon)$, as SGHMC2 algorithm is based on an alternative discretization. In particular, the performance bounds we get for SGHMC2 are tighter than SGHMC1, as will be elaborated further in the Section 5.

Theorem 6. *Consider the SGHMC2 iterates* $(\hat{X}_k, \hat{V}_k)$ *defined by the recursion* (1.10)–(1.11) *from the initial state* (X_0, V_0) *which has the law* μ_0*. If Assumption 1 is satisfied, then for* $\beta, \varepsilon > 0$*, we have*

$$\left|\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X))\right| = \left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\hat{\mu}_{k,\mathbf{z}}(dx,dv) - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv)\right| \leq \mathcal{J}_0(\mathbf{z},\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon)\,,$$

where $\mathcal{J}_0(\mathbf{z},\varepsilon)$ *is defined in* (3.1) *and*

$$\hat{\mathcal{J}}_1(\varepsilon) := (M\sigma + B)\cdot\left(\frac{C_0}{\sqrt{\mu_*}}\sqrt{\log(1/\varepsilon)}\delta^{1/4} + \frac{\hat{C}_1}{\sqrt{\mu_*}}\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}, \tag{4.1}$$

with σ *defined by* (A.20) *provided that*

$$\eta \leq \min\left\{\left(\frac{\varepsilon}{\sqrt{\log(1/\varepsilon)}}\right)^2, 1, \frac{\gamma}{\hat{K}_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2\hat{K}_1}, \frac{2}{\gamma\lambda}\right\}, \tag{4.2}$$

and

$$k\eta = \frac{1}{\mu_*}\log\left(\frac{1}{\varepsilon}\right) \geq e. \tag{4.3}$$

Here $\mathcal{H}_\rho$ *is a semi-metric for probability distributions defined by* (A.12)*. All the constants are made explicit and are summarized in Table 1 and Table 2.*

The proof of Theorem 6 is given in Section B in the Appendix. Relying on Theorem 6, one can readily derive the following result on the performance bound for the empirical risk minimization with the SGHMC2 algorithm. The proof follows a similar argument as discussed in Section 3.1, and is omitted.

Corollary 7 (Empirical risk minimization with SGHMC2)**.** *Under the setting of Theorem 6, the empirical risk minimization problem admits the performance bounds:*

$$\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \leq \mathcal{J}_0(\mathbf{z},\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon) + \mathcal{J}_2\,,$$

provided that conditions (4.2) *and* (4.3) *hold where the terms* $\mathcal{J}_0(\mathbf{z},\varepsilon)$*,* $\hat{\mathcal{J}}_1(\varepsilon)$ *and* $\mathcal{J}_2$ *are defined by* (3.1)*,* (4.1) *and* (3.5) *respectively.*

Next, we present the performance bound for the population risk minimization with the SGHMC2 algorithm. Similar as in Section 3.2, to control the population risk during SGHMC2 iterations, one needs to control the difference between the finite sample size problem (1.2) and the original problem (1.1) in addition to the empirical risk. This leads to the following result. The details of the proof are given in Section B in the Appendix.

Constants	*Source*
$\hat{K}_1 = K_1 + Q_1\frac{4}{1-2\lambda} + Q_2\frac{8}{(1-2\lambda)\gamma^2}$	(B.1)
$\hat{K}_2 = K_2 + Q_3$	(B.2)
$Q_1 = \frac{1}{2}c_0\left((5M+4-2\gamma+(c_0+\gamma^2)) + (1+\gamma)\left(\frac{5}{2}+c_0(1+\gamma)\right) + 2\gamma^2\lambda\right)$	(B.3)
$Q_2 = \frac{1}{2}c_0\left[\left((1+\gamma)\left(c_0(1+\gamma)+\frac{5}{2}\right)+c_0+2+\lambda\gamma^2+2(Mc_0+M+1)\right)\left(2(1+\delta)M^2\right) + \left(2M^2+\gamma^2\lambda+\frac{3}{2}\gamma^2(1+\gamma)\right)\right]$	(B.4)
$Q_3 = c_0\left((1+\gamma)\left(c_0(1+\gamma)+\frac{5}{2}\right)+c_0+2+\lambda\gamma^2+2(Mc_0+M+1)\right)(1+\delta)B^2 + c_0B^2 + \frac{1}{2}\gamma^3\beta^{-1}c_{22} + \gamma^2\beta^{-1}c_{12} + M\gamma\beta^{-1}c_{22}$	(B.5)
$c_0 = 1+\gamma^2, \qquad c_{12} = \frac{d}{2}, \qquad c_{22} = \frac{d}{3}$	(B.6)
$\hat{C}_1 = \hat{\gamma}\cdot\left(\frac{3\beta M^2}{2\gamma}\left(C_v^d + \left(2(1+\delta)M^2C_x^d + 2(1+\delta)B^2\right) + \frac{2d\gamma\beta^{-1}}{3}\right) + \sqrt{\frac{3\beta M^2}{2\gamma}\left(C_v^d + (2(1+\delta)M^2C_x^d + 2(1+\delta)B^2) + \frac{2d\gamma\beta^{-1}}{3}\right)}\right)^{1/2}$	(B.8)

Table 2: Summary of the constants and where they are defined in the text.

Corollary 8 (Population risk minimization with SGHMC2)**.** *Under the setting of Theorem 6, the expected population risk of $\hat{X}_k$ (the iterates in (1.11)) is bounded by*

$$\mathbb{E}F(\hat{X}_k) - F^* \le \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon) + \mathcal{J}_2 + \mathcal{J}_3(n)\,,$$

where $\overline{\mathcal{J}}_0(\varepsilon)$, $\hat{\mathcal{J}}_1(\varepsilon)$, $\mathcal{J}_2$, $\mathcal{J}_3(n)$ are defined in (3.6), (4.1), (3.5) and (3.7).

Finally, we present a result on the generalization error of the SGHMC2 algorithm in the one pass regime. The proof follows from Theorem 6 and the discussion for SGHMC1 algorithm in Section 3.3, and hence is omitted.

Theorem 9 (Generalization error of SGHMC2)**.** *Under the setting of Theorem 6, we have*

$$\begin{aligned}\left|\mathbb{E}F(\hat{X}_k) - \mathbb{E}F(X^\pi)\right| &\le \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon)\,,\\ \left|\mathbb{E}F_{\mathbf{Z}}(\hat{X}_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)\right| &\le \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon),\end{aligned}$$

provided that (4.2) and (4.3) hold where X^π is the output of the underdamped Langevin dynamics, i.e. its distribution conditional on $\mathbf{Z} = \mathbf{z}$ is given by $\pi_{\mathbf{z}}$ and $\overline{\mathcal{J}}_0(\varepsilon)$ is defined by (3.6). Then, it follows from (3.10) that if each data point is used once, the expected generalization error satisfies

$$|\mathbb{E}F_{\mathbf{Z}}(\hat{X}_k) - \mathbb{E}F(\hat{X}_k)| \le 2\overline{\mathcal{J}}_0(\varepsilon) + 2\hat{\mathcal{J}}_1(\varepsilon).$$

5 Performance comparison with respect to SGLD algorithm

In this section, we compare our performance bounds for SGHMC1 and SGHMC2 to SGLD. We use the notations $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$ to give explicit dependence on the parameters d, β, μ_* but it hides factors that depend (at worst polynomially) on other parameters $m, M, B, \lambda, \gamma, b$ and A_0. Without loss of generality, we assume here the initial distribution $\mu_0(dx, dv)$ is supported on a Euclidean ball with radius being some universal constant for the simplicity of performance comparison.

Generalization error in the one-pass setting. A consequence of Theorem 5 is that the generalization error of the SGHMC1 iterates $|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k)|$ in the one-pass setting satisfy

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon + \frac{(d+\beta)^{3/2}}{\beta(\mu_*)^{3/2}}\left((\log(1/\varepsilon))^{3/2}\delta^{1/4} + \varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))} + \frac{d+\beta}{\beta}\frac{\varepsilon^2}{\mu_*(\log(1/\varepsilon))^2}\right) \quad (5.1)$$

for $k = K_{SGHMC1} := \tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^4}\log^7(1/\varepsilon)\right)$ iterations, and similarly, Theorem 9 implies the generalization error of the SGHMC2 iterates $|\mathbb{E}F_{\mathbf{Z}}(\hat{X}_k) - \mathbb{E}F(\hat{X}_k)|$ in the one-pass setting

satisfy

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon + \frac{(d+\beta)^{3/2}}{\beta\sqrt{\mu_*}}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}\right), \tag{5.2}$$

for $k = K_{SGHMC2} := \tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^2}\log^2(1/\varepsilon)\right)$ iterations (see the discussion in Section G in the Appendix for details). On the other hand, the results of Theorem 1 in [RRT17] imply that the generalization error for the SGLD algorithm after K_{SGLD} iterations in the one-pass setting scales as

$$\tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*}\left(\delta^{1/4}\log(1/\varepsilon)+\varepsilon\right)\right) \quad \text{for} \quad K_{SGLD} = \tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_*\varepsilon^4}\log^5(1/\varepsilon)\right). \tag{5.3}$$

The constants λ_* (see (3.8)) and μ_* (see Table 1) are exponentially small in both β and d in the worst case, but under some extra assumptions the dependency on d can be polynomial (see e.g. [CCBJ18]) although the exponential dependence to β is unavoidable in the presence of multiple minima in general (see [BGK05]). One can readily see that K_{SGHMC2} has better dependency on ϵ than K_{SGHMC1}, and infer from (5.1)–(5.2) that the performance of SGHMC2 is better than SGHMC1. Hence, in the rest of the section, we will only focus on the comparison between SGHMC2 and SGLD.

We see that the generalization error for SGHMC2 (5.2) is bounded by

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\sqrt{\log\log(1/\varepsilon)}\right), \tag{5.4}$$

as μ_* is small, and if we ignore the $\sqrt{\log\log(1/\varepsilon)}$ factor [4], then, we get

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\right) \quad \text{for} \quad K_{SGHMC2} = \tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^2}\log^2(1/\varepsilon)\right), \tag{5.5}$$

iterations of the SGHMC2 algorithm whereas the corresponding bound for SGLD from [RRT17, Theorem 1] is

$$\tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*}\left(\log(1/\varepsilon)\delta^{1/4}+\varepsilon\right)\right) \quad \text{for} \quad K_{SGLD} = \tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_*\varepsilon^4}\log^5(1/\varepsilon)\right) \tag{5.6}$$

iterations of the SGLD algorithm. Note that K_{SGHMC2} and K_{SGLD} do not have the same dependency to ε up to log factors (the former scales with ε as $\log^2(1/\varepsilon)\varepsilon^{-2}$ and the latter

[4] We emphasize that the effect of the last term $\sqrt{\log\log(1/\varepsilon)}$ appearing in (5.4) is typically negligible compared to other parameters. For instance even if $\varepsilon = 2^{-2^{16}}$ is double-exponentially small, we have $\sqrt{\log\log(1/\varepsilon)} \le 4$.

$\log^5(1/\varepsilon)\varepsilon^{-4}$), and this improvement on ε dependency is due to better diffusion approximation of SGHMC2 (see Lemma 22) compared to SGLD and the exponential integrability estimate we have in Lemma 17 which improves the estimate in [RRT17] and using the same argument, one can improve the $\log^5(1/\varepsilon)/\varepsilon^4$ term in (5.6) to $\log^3(1/\varepsilon)/\varepsilon^4$.

To make the comparison to SGLD simpler, we notice that in both expressions (5.5) and (5.6), we see a term scaling with $\delta^{1/4}$ due to the gradient noise level δ (δ is fixed in the one-pass setting), and we fix the error in (5.5) and (5.6) without the δ term to be the same order, and then compare the number of iterations K_{SGHMC2} and K_{SGLD}. More precisely, given $\hat{\varepsilon} > 0$ and we choose $\varepsilon > 0$ such that $\frac{(d+\beta)^{3/2}}{\beta\mu_*}\varepsilon = \hat{\varepsilon}$ in (5.5) so that the generalization error for SGHMC2 is

$$\tilde{\mathcal{O}}\left(\hat{\varepsilon} + \frac{(d+\beta)^{3/2}}{\beta\mu_*}\sqrt{\log\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*\hat{\varepsilon}}\right)}\delta^{1/4}\right) \quad \text{for} \quad K_{SGHMC2} = \tilde{\Omega}\left(\frac{(d+\beta)^3}{\beta^2\mu_*^3\hat{\varepsilon}^2}\log^2\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*\hat{\varepsilon}}\right)\right). \tag{5.7}$$

Similarly, the generalization error for SGLD is

$$\tilde{\mathcal{O}}\left(\hat{\varepsilon} + \frac{\beta(\beta+d)^2}{\lambda_*}\log\left(\frac{\beta(\beta+d)^2}{\lambda_*\hat{\varepsilon}}\right)\delta^{1/4}\right) \quad \text{for} \quad K_{SGLD} = \tilde{\Omega}\left(\frac{\beta^5(d+\beta)^9}{\lambda_*^5\hat{\varepsilon}^4}\log^5\left(\frac{\beta(\beta+d)^2}{\lambda_*\hat{\varepsilon}}\right)\right). \tag{5.8}$$

When λ_* and μ_* are on the same order or μ_* is larger, since typically $\beta \geq 1$, the term involving δ in the generalization error for SGHMC2 above is (smaller) better than the counterpart for SGLD, and this is guaranteed to be achieved in a less number of iterations ignoring the log factors and universal constants for K_{SGHMC2} in (5.7) and K_{SGLD} in (5.8).

Comparing λ_* and μ_* on arbitrary non-convex functions seems not trivial, however we give a class of non-convex functions (see Proposition 11 and Example 10) where $\frac{1}{\mu_*} = \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{\lambda_*}}\right)$. For this class, we can infer from (5.7) that K_{SGHMC2} has a dependency $1/\mu_*^3 = \tilde{\mathcal{O}}(1/\lambda_*^{3/2})$ which is much smaller in contrast to $1/\lambda_*^5$ for K_{SGLD} in (5.8).

Empirical risk minimization. The empirical risk minimization bound given in Corollary 7 has an additional term $\mathcal{J}_2$ compared to the $\overline{\mathcal{J}}_0(\varepsilon)$ and $\hat{\mathcal{J}}_1(\varepsilon)$ terms appearing in the one-pass generalization bounds. Note also that $\mathcal{J}_0(\mathbf{z},\varepsilon) \leq \overline{\mathcal{J}}_0(\varepsilon)$. As a consequence, SGHMC2 algorithm has expected empirical risk

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon + \frac{(d+\beta)^{3/2}}{\beta\sqrt{\mu_*}}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4} + \varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))} + d\cdot\frac{\log(1+\beta)}{\beta}\right), \tag{5.9}$$

after $K_{SGHMC2} = \tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^2}\log^2(1/\varepsilon)\right)$ iterations as opposed to

$$\tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*}\left(\delta^{1/4}\log(1/\varepsilon) + \varepsilon\right) + d\cdot\frac{\log(1+\beta)}{\beta}\right), \tag{5.10}$$

after $K_{SGLD} = \tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_* \varepsilon^4} \log^5(1/\varepsilon)\right)$ iterations required in [RRT17]. Comparing (5.9) and (5.10), we see that the last terms are the same. If this term is the dominant term, then the empirical risk upper bounds for SGLD and SGHMC2 will be similar except that K_{SGHMC2} can be smaller than K_{SGLD} for instance when $\frac{1}{\mu_*} = \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{\lambda_*}}\right)$. Otherwise, if the last term is not the dominant one and can be ignored with respect to other terms, then, the performance comparison will be similar to the discussion about the generalization bounds (5.4) and (5.6) discussed above.

We next briefly discuss the comparisons of SGHMC2 and SGLD based on the total number of stochastic gradient evaluations (gradient complexity), and we compare with a recent work [XCZG18] which established a faster convergence result and improved the gradient complexity for SGLD in the mini-batch setting compared with [RRT17]. Here, the total number of stochastic gradient evaluations of an algorithm is defined as the number of stochastic gradients calculated per iteration (which is equal to the batch size in the mini-batch setting) times the total number of iterations. [XCZG18] showed that it suffices to take

$$\hat{K}_{SGLD} = \tilde{\Omega}\left(\frac{d^7}{\hat{\lambda}^5 \hat{\varepsilon}^5}\right) \tag{5.11}$$

stochastic gradient evaluations to converge to an $\hat{\varepsilon}$ neighborhood of an almost ERM where $\tilde{\Omega}(\cdot)$ hides some factors in β and $\hat{\lambda}$ is the spectral gap of the discrete overdamped Langevin dynamics, i.e. SGLD with zero gradient noise. This improves upon the result in [RRT17] which showed that the same task requires $\tilde{\Omega}\left(\frac{d^{17}}{\lambda_*^9 \hat{\varepsilon}^8}\right)$ stochastic gradient evaluations. Our results show that (see e.g. (5.9)) for SGHMC2, it suffices to have

$$\hat{K}_{SGHMC2} = \tilde{\Omega}\left(\frac{d^9}{\mu_*^4 \hat{\varepsilon}^6}\right) \tag{5.12}$$

stochastic gradient evaluations, ignoring the log factors in the parameters $\hat{\varepsilon}, \mu_*, d$ and hiding factors in β that can be made explicit. To see (5.12), we infer from (5.9) that for fixed precision $\hat{\varepsilon} > 0$ and dimension d, by ignoring the log factors and β, we can choose ε so that $d^{3/2}\varepsilon/\mu_* = \hat{\varepsilon}$ and choose the gradient noise level δ so that $d^{3/2}\delta^{1/4}/\sqrt{\mu_*} = \hat{\varepsilon}$. So the number of SGHMC2 iterations is

$$K_{SGHMC2} = \tilde{\Omega}\left(\frac{1}{\mu_* \varepsilon^2}\right) = \tilde{\Omega}\left(\frac{d^3}{\mu_*^2 \hat{\varepsilon}^2}\right).$$

On the other hand, the mini-batch size to achieve gradient noise level δ is given by $1/\delta$ (see [RRT17]), which is equal to $d^6/(\mu_*^2 \hat{\varepsilon}^4)$. Hence, we obtain (5.12) which is the product of the mini-batch size and number of iterations.

It is hard to compare $\hat{\lambda}$ in (5.11) and μ_* in (5.12) in general since $\hat{\lambda}$ is the spectral gap of the discrete overdamped Langevin dynamics (i.e. SGLD with zero gradient noise) without a simple closed-form formula. However, when the stepsize is small enough, we

expect $\hat{\lambda}$ will be similar to λ_*, which is the spectral gap of the continuous-time overdamped Langevin diffusion. As a consequence, when the stepsize η is small enough (which is the case for instance, when target accuracy $\hat{\varepsilon}$ is small enough), we will have $\hat{\lambda} \approx \lambda_*$ and $\frac{1}{\mu_*} = \mathcal{O}\left(\sqrt{\frac{1}{\lambda_*}}\right) = \mathcal{O}\left(\sqrt{\frac{1}{\hat{\lambda}}}\right)$ for the class of non-convex functions we discuss in Proposition 11 and Example 10. For this class of problems, comparing (5.11) and (5.12), we see that we obtain an improvement in the spectral gap parameter (μ_*^4 vs. $\hat{\lambda}^5$), however $\hat{\varepsilon}$ and d dependency of the bound (5.11) is better than (5.12).

Population risk minimization. If samples are recycled and multiple passes over the dataset is made, then one can see from Corollary 4 that there is an extra term $\mathcal{J}_3$ that needs to be added to the bounds given in (5.9) and (5.10). This term satisfies

$$\mathcal{J}_3 = \tilde{\mathcal{O}}\left(\frac{(\beta+d)^2}{\lambda_* n}\right).$$

If this term is dominant compared to other terms $\overline{\mathcal{J}}_0, \mathcal{J}_1$ and $\mathcal{J}_2$, for instance this may happen if the number of samples n is not large enough, then the performance guarantees for population risk minimization via SGLD and SGHMC2 will be similar. Otherwise, if n is large and β is chosen in a way to keep the $\mathcal{J}_2$ term on the order $\overline{\mathcal{J}_0}$, then similar improvement can be achieved.

Comparison of λ_* and μ_*. The parameters λ_* (see (3.8)) and μ_* (see Table 1) govern the convergence rate to the equilibrium of the overdamped and underdamped Langevin SDE, they can be both exponentially small in dimension d and in β. They appear naturally in the complexity estimates of SGHMC2 and SGLD method as these algorithms can be viewed as discretizations of Langevin SDEs (when the discretization step is small and the gradient noise $\delta = 0$, the discrete dynamics will behave similarly as the continuous dynamics). Next, to get further intuition, first we discuss some toy examples of non-convex functions below where $\frac{1}{\mu_*} = \mathcal{O}\left(\sqrt{\frac{1}{\lambda_*}}\right)$. For these examples if the other parameters (β, d, δ) are fixed, then SGHMC2 can lead to an improvement upon the SGLD performance. We will then show in Proposition 11 that these examples generalize to a more general class of non-convex functions.

Example 10. *Consider the following symmetric double-well potential in $\mathbb{R}^d$ studied previously in the context of Langevin diffusions ([EGZ19]):*

$$f_a(x) = U(x/a) \quad \textit{with} \quad U(x) := \begin{cases} \frac{1}{2}(\|x\|-1)^2 & \textit{for} \quad \|x\| \geq \frac{1}{2}, \\ \frac{1}{4} - \frac{\|x\|^2}{2} & \textit{for} \quad \|x\| \leq \frac{1}{2}, \end{cases}$$

where $a > 0$ is a scaling parameter which is illustrated in the left panel of Figure 1. For this example, there are two minima that are apart at a distance $\mathcal{R} = \mathcal{O}(a)$. For simplicity,

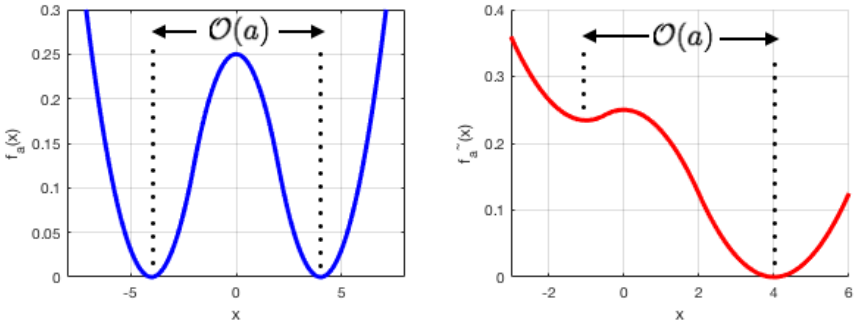

Figure 1: The illustration of the functions $f_a(x)$ (left) and $\tilde{f}_a(x)$ (right) for $a = 4$.

we assume there is only one sample, i.e. $\mathbf{z} = (z_1)$ *and* $F_{\mathbf{z}}(x) = f(x, z_1) = f_a(x)$. *We consider the non-convex optimization problem* (1.2) *with both the SGHMC2 algorithm and the SGLD algorithm. [EGZ19] showed that* $\mu_* \geq \Theta(\frac{1}{a})$ *for this example whereas* $\lambda_* \leq \Theta(\frac{1}{a^2})$ *making the constants hidden by the* Θ *explicit. This shows that the contraction rate of the underdamped diffusion* μ_* *is (faster) larger than that of the overdamped diffusion* λ_* *by a square root factor when* a *is large where all the constants can be made explicit. Such results extend to a more general class of non-convex functions with multiple-wells and higher dimensions as long as the gradient of the objective satisfies a growth condition (see Example 1.1, Example 1.13 in [EGZ19] for a further discussion).*

For computing an ε*-approximate global minimizer of* $f_a = f(x, z_1)$ *(or more generally for a non-convex problem satisfying Assumption 1),* β *is chosen large enough so that the stationary measure concentrates around the global minimizer. Using the tight characterization of* λ_* *from Theorem 1.2 in [BGK05] for* β *large, further comparisons with similar conclusions between the rate of convergence to the equilibrium distribution between the underdamped and overdamped dynamics can also be made. For example, consider the non-convex objective* $F_{\mathbf{z}}(x) = \tilde{f}_a(x) = \tilde{U}(x/a)$ *instead, illustrated in the right panel of Figure 1 for* $a = 4$ *where*

$$\tilde{U}(x) = \begin{cases} \frac{1}{2}(x-1)^2 & \text{for} \quad x \geq \frac{1}{2}, \\ \frac{1}{4} - \frac{x^2}{2} & \text{for} \quad -\frac{1}{8} \leq x \leq \frac{1}{2}, \\ \frac{1}{2}(x+\frac{1}{4})^2 + \frac{15}{64} & \text{for} \quad x \leq -\frac{1}{8}, \end{cases}$$

is the asymmetric double well potential in dimension one. It follows from Theorem 19 (see also [EGZ19]) that the contraction rate satisfies $\mu_* = \Theta\left(a^{-1}\right)$, *whereas it follows from Theorem 1.2 in [BGK05] that* $\lambda_* = \Theta(1/a^2)$. *This shows that when the separation between minima, or alternatively the scaling factor* a *is large enough,* μ_* *is larger than* λ_* *by a square root factor up to constants.*

The behavior in these toy examples can be generalized to more general non-convex objectives with a finite-sum structure satisfying Assumption 1. Proposition 11 below gives a class of functions where μ_* is on the order of the square root of λ_*. The proof will be presented in details in Section F.

Proposition 11. *Suppose that the functions $f_a(x,z)$ indexed by a satisfies Assumption 1 (i)-(iii) with $m = m_1 a^{-2}$, $M = M_1 a^{-2}$ and $B = B_1 a^{-1}$ for some fixed constants m_1, M_1, and B_1. Then, we have as $a \to \infty$,*

$$\lambda_* = \mathcal{O}(a^{-2}), \qquad \mu_* = \Theta(a^{-1}). \tag{5.13}$$

This result is more general than the previous example. In particular, if $f(x,z)$ satisfies Assumption 1 (i)-(iii) with m, M, B replaced by m_1, M_1, B_1, then $f_a(x,z) := f(x/a, z)$ satisfies Assumption 1 (i)-(iii) with $m = m_1 a^{-2}$, $M = M_1 a^{-2}$ and $B = B_1 a^{-1}$. Proposition 11 essentially says that if we consider the normalized empirical risk objective $F_{\mathbf{z}}(x/a) = \frac{1}{n}\sum_{i=1}^n f(x/a, z_i)$ where a is a (normalization) scaling parameter and $f(x,z)$ satisfies Assumption 1, then for large enough values of a, the empirical risk surface will be relatively flat and the convergence rate of momentum variant SGHMC2 to an ε-neighborhood of the global minimum will be governed by the parameter μ_* which will be larger than that of the parameter λ_* of SGLD when a is sufficiently large. This will lead to improved performance bounds for SGHMC2 compared to known performance bounds for SGLD.

6 Applications

We note that several non-convex stochastic optimization problems of interest can satisfy Assumption 1 under appropriate noise assumptions for the underlying dataset. For example, Lasso problems with non-convex regularizers (see e.g. [HLM+17]), non-convex formulations of the phase retrieval problem around global minimum (see e.g. [ZZLC17]) or non-convex stochastic optimization problems defined on a compact set including but not limited to dictionary learning over the sphere (see e.g. [SQW16]), training deep learning models subject to norm constraints in the model parameters (see e.g. [ALG19]). In this section, we discuss some applications of our results where we provide two specific examples.

6.1 Binary linear classification

In linear binary classification, the aim is to learn a predictive model of the form $\mathbb{P}(Y = 1 | A_{in} = a) = \sigma_c(\langle \tilde{x}, a \rangle)$, where $\tilde{x} \in \mathbb{R}^d$ is a parameter vector to be learned, A_{in} is the input variable (feature vector), Y is the binary output and $\sigma_c : \mathbb{R} \to [0,1]$ is a threshold function. Binary classification arises in many data-driven applications in operations research from diagnosing patients in healthcare [WL07] to predicting directions in the stock market [JWHT13].

A number of empirical studies have demonstrated that non-convex choices of the σ_c function can often lead to superior classification accuracy and robustness properties compared to convex choices of σ_c such as the hinge loss [CDT+09, CSWB06, WL07, NS13]. Given access to a dataset of input-output pairs $z_i = (a_i, y_i)$, a standard way of estimating $\tilde{x}$ is based on minimizing the *regularized squared loss* over the dataset, i.e.

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \sigma_c(\langle x, a_i \rangle))^2 + \frac{\lambda_r}{2} \|x\|^2, \tag{6.1}$$

where $\lambda_r > 0$ is a regularization parameter that may depend on the number of samples n. By Lagrangian duality, this problem is equivalent to the constrained optimization problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \sigma_c(\langle x, a_i \rangle))^2 \quad \text{subject to} \quad \|x\| \leq R,$$

for some R, which has also been considered in the literature (see e.g. [MBM18, FSS18, WCX19]). For non-convex $\sigma_c(\cdot)$, this problem is also non-convex in general. We consider minimizing the objective (6.1) in the mini-batch setting where the gradients in SGHMC iterations are estimated from n_b data points sampled with replacement, i.e. the gradient is estimated as

$$g(x, U_{\mathbf{z}}) = \frac{1}{n_b} \sum_{j=1}^{n_b} \nabla f(x, z_j)\,, \tag{6.2}$$

where z_j are i.i.d. with a uniform distribution over the set of indices $\{1, 2, \ldots, n\}$. We also consider the following assumption for the threshold function σ_c which are satisfied by many choices of σ_c in practice. A prominent example is the logistic (or sigmoid) function in which case $\sigma_c(z) = 1/(1 + e^{-z})$ which is also used in deep learning. Another possible choice is the *probit function* which corresponds to $\sigma_c(t) = \Phi(t)$ where Φ is the cumulative distribution function of the standard normal distribution.

Assumption 12. *The threshold function σ_c is twice continuously differentiable on $\mathbb{R}$. It is bounded and has bounded first and second derivatives, i.e. there exists a constant $L_{\sigma_c} > 0$ such that $\max\left\{\|\sigma_c\|_\infty, \|\sigma_c'\|_\infty, \|\sigma_c''\|_\infty\right\} \leq L_{\sigma_c}$. The distribution of the input data A_{in} has compact support, i.e. $\|A_{in}\| \leq D$ for some $D > 0$.*

We show in the next lemma that if Assumption 12 holds, then Assumption 1 holds with explicit constants A_0, B, M, m, b and σ_c that we can precise. The proof can be found in the Appendix.

Lemma 13. *In the setting of binary linear classification, consider the SGHMC method applied to the objective (6.1) where gradients are estimated according to (6.2) where the*

probability law μ_0 of the initial state has compact support. If Assumption 12 holds; then Assumption 1 hold for any $\delta \in [\frac{1}{4n_b}, 1)$ with the following constants:

$$A_0 = (1 + \|\sigma_c(0)\|)^2, \quad B = 2D\,(1 + \|\sigma_c\|_\infty)\,\|\sigma_c'\|_\infty, \tag{6.3}$$

$$M = 2D^2\|\sigma_c'\|_\infty^2 + 2D^2(1 + \|\sigma_c\|_\infty)\|\sigma_c''\|_\infty + 5\lambda_r, \tag{6.4}$$

$$m = \lambda/2, \quad b = 8(1 + \|\sigma_c\|_\infty)^2\|\sigma_c'\|_\infty^2 D^2/\lambda_r. \tag{6.5}$$

We conclude from Lemma 13 that the objective is dissipative and our main results for SGHMC1 and SGHMC2 algorithms described in Sections 3–5 apply to binary linear classification under Assumption 12 with the constants given in Lemma 13 and where μ_* is given by the formula in Table 1. For example, if $D = \mathcal{O}(1)$, then we have $\frac{1}{\mu_*} = \tilde{\Theta}(\sqrt{d+\beta}e^{\tilde{\Theta}(d+\beta)})$ (see (G.1)) and we conclude from (5.12) that it suffices to have

$$\hat{K}_{SGHMC2} = \tilde{\Omega}\left(\frac{d^9}{\mu_*^4\hat{\varepsilon}^6}\right) = \tilde{\Omega}\left(\frac{d^9 e^{\tilde{\Theta}(d+\beta)}}{(d^2+\beta^2)\hat{\varepsilon}^6}\right)$$

stochastic gradient evaluations to converge to an $\hat{\varepsilon}$ neighborhood of an almost ERM ignoring the log factors in the parameters $\hat{\varepsilon}, \mu_*, d$ and hiding other constants that can be made explicit based on Lemma 13.[5]

6.2 Robust Ridge Regression

Given an input (feature) vector $A_{in} \in \mathbb{R}^d$, the aim is to predict the output $Y \in \mathbb{R}$. Given access to a dataset of input-output pairs $z_i = (a_i, y_i)$, we assume a linear model $y_i = a_i^T\tilde{x} + \varepsilon_i$ where the errors ε_i are i.i.d. with mean zero. The standard *ridge regression* estimate of $\tilde{x}$ minimizes a penalized residual sum of squares [HK70], i.e. minimizes $\sum_{i=1}^n \|y_i - \langle x, a_i\rangle\|^2 + \lambda_r\|x\|^2$ where $\lambda_r > 0$ is a regularization parameter.[6] However, this formulation can be sensitive to outliers. Robust formulations of the ridge regression [ROK12] can be obtained if one solves instead the following problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f(x, z_i), \quad f(x, z_i) = \rho\,(y_i - \langle x, a_i\rangle) + \frac{\lambda_r}{2}\|x\|^2, \tag{6.6}$$

where $\lambda_r > 0$ is a regularization parameter and $\rho : \mathbb{R} \to \mathbb{R}$ is a suitably chosen loss function. In particular, for achieving robustness to outliers, the non-convex choices of the function ρ that are either bounded or slowly growing near infinity has been considered in the literature

[5] We also note that under further assumptions on the statistical nature of the input and if the number of data points is large enough, it can be shown that the objective (6.1) admits a unique minimizer and the objective is strongly convex in some regions [MBM18]. However, our assumptions here are weaker, therefore such arguments are not directly applicable.

[6] See [KO01] for details regarding the choice of the parameter λ_r.

(as opposed to the standard ridge regression setting which corresponds to $\rho(t) = \|t\|^2$). For example, popular choices of the function $t \mapsto \rho(t)$ include *Tukey's bisquare loss* defined as

$$\rho_{Tukey}(t) = \begin{cases} 1 - (1 - (t/t_0)^2)^3 & \text{for} \quad \|t\| \le t_0, \\ 1 & \text{for} \quad |t| \ge t_0, \end{cases}$$

(see e.g. [MBM18]) and exponential squared loss [WJHZ13]: $\rho_{exp}(t) = 1 - e^{-\|t\|^2/t_0}$, where $t_0 > 0$ is a tuning parameter. In the following, similar to [WCX19], we assume that the data A_{in} is bounded and the threshold function and its derivatives up to order two are bounded, similar to [MBM18]. This assumption for ρ is satisfied in several cases, including Tukey's bisquare loss and exponential squares loss mentioned above.

Assumption 14. *The function ρ is twice continuously differentiable on $\mathbb{R}$. The function ρ is bounded and has bounded first and second derivatives; i.e. there exists a constant L_ρ such that $\max(\|\rho\|_\infty, \|\rho'\|_\infty, \|\rho''\|_\infty) \le L_\rho$. Furthermore, the distribution of the input data A_{in} has compact support, i.e. there exists D such that $\|A_{in}\| \le D$.*

The following lemma shows that under Assumption 14, our assumptions (Assumption 1) for analyzing SGHMC methods hold with proper initialization.

Lemma 15. *In the setting of robust regression, consider the objective (6.1) where gradients are estimated according to (6.2) where the probability law μ_0 of the initial state has compact support. If Assumption 12 holds; then Assumption 1 hold for both SGHMC1 and SGHMC2 methods for any choice of $\delta \in [\frac{1}{4n_b}, 1)$ with the following constants:*

$$A_0 = \|\rho\|_\infty, \qquad B = 4\|\rho'\|_\infty D, \tag{6.7}$$

$$M = \|\rho''\|_\infty D^2 + \lambda_r, \qquad m = \lambda_r/2, \qquad b = \frac{2\|\rho'\|_\infty^2 D^2}{\lambda_r}. \tag{6.8}$$

Similarly, we conclude from Lemma 15 that our main results for SGHMC1 and SGHMC2 algorithms described in Sections 3–5 apply to the problem of robust regression under Assumption 14.

7 Outline of the Proof

To obtain the main results in this paper, we adapt the proof techniques of [RRT17] developed for the overdamped dynamics to the underdamped dynamics and combine it with the analysis of [EGZ19] which quantifies the convergence rate of the underdamped Langevin SDE to its equilibrium. In an analogy to the fact that momentum-based first-order optimization methods require a different Lyapunov function and a quite different set of analysis tools (compared to their non-accelerated variants) to achieve fast rates (see e.g. [LFM18, SBC14, Nes83]), our analysis of the momentum-based SGHMC1 and SGHMC2

algorithms requires studying a different Lyapunov function $\mathcal{V}$ defined in (2.1) that also depends on the objective f as opposed to the classic Lyapunov function $\mathcal{H}(x) = \|x\|^2$ arising in the study of the SGLD algorithm (see e.g. [MSH02, RRT17]). This fact introduces some challenges for the adaptation of the existing analysis techniques for SGLD to SGHMC. For this purpose, we take the following steps:

First, we show that SGHMC1 and SGHMC2 iterates track the underdamped Langevin diffusion closely in the 2-Wasserstein metric. As this metric requires finiteness of second moments, we first establish uniform (in time) L^2 bounds for both the underdamped Langevin SDE and SGHMC1 and SGHMC2 iterates (see Lemma 16 and Lemma 21 in Appendix), exploiting the structure of the Lyapunov function $\mathcal{V}$. Second, we obtain a bound for the Kullback-Leibler divergence between the discrete and continuous underdamped dynamics making use of the Girsanov theorem, which is then converted to bounds in the 2-Wasserstein metric by an application of an optimal transportation inequality of [BV05]. This step requires proving a certain exponential integrability property of the underdamped Langevin diffusion (Lemma 17 in Appendix). We show in Lemma 17 that the exponential moments grow at most linearly in time, which strictly improves the exponential growth in time in Lemma 4 in [RRT17]. [7] As a result, the method improves upon the ε dependence of the number of iterates (see equations (5.5) and (5.6)).

Second, we apply the seminal result of [EGZ19] which showed that the continuous-time underdamped Langevin SDE is geometrically ergodic with an explicit rate μ_* in the 2-Wasserstein metric. In order to get explicit performance guarantees, we derive new bounds that make the dependence of the constants to the initialization in [EGZ19] explicit (see Lemma 20 in Appendix).

As the x-marginal of the equilibrium distribution $\pi_{\mathbf{z}}(dx, dv)$ of the underdamped Langevin SDE concentrates around the global minimizers of $F_{\mathbf{z}}$ for β appropriately chosen, and we can control the error between the discrete-time SGHMC1 and SGHMC2 dynamics and the underdamped SDE by choosing the step size accordingly; this leads to performance bounds for the empirical risk minimizations for SGHMC1 and SGHMC2 algorithms in Corollary 3 and Corollary 7. For controlling the population risk during SGHMC iterations, in addition to the empirical risk, one has to control the *generalization error* $F(X_k) - F_{\mathbf{Z}}(X_k)$ that accounts for the differences between the finite sample size problem (1.2) and the original problem (1.1). By exploiting the fact that the $x-$marginal of the invariant distribution for the underdamped dynamics is the same as it is in the overdamped case, we control the generalization error in Corollary 4 and Corollary 8 which is no worse than that of the available bounds for SGLD given in [RRT17].

[7] The method that is used in the proof of Lemma 17 in Appendix can indeed be adapted to improve the exponential integrability and hence the overall estimates in [RRT17] for SGLD as well.

8 Conclusion

SGHMC is a momentum-based popular variant of stochastic gradient where a controlled amount of isotropic Gaussian noise is added to the gradient estimates for optimizing a non-convex function. We obtained first-time finite-time guarantees for the convergence of SGHMC1 and SGHMC2 algorithms to the ε-global minimizers under some regularity assumption on the non-convex objective f. We also show that on a class of non-convex problems, SGHMC2 can be faster than overdamped Langevin MCMC approaches such as SGLD in the sense that the best available bounds for SGHMC2, which we prove in our paper, are better than the best available bounds for SGLD. This effect is due to the momentum term in the underdamped SDE. Furthermore, our results show that momentum-based acceleration is possible on a class of non-convex problems under some conditions if we compare known upper bounds between SGLD and SGHMC. Finally, we mention a few limitations in our work that may lead to some future research directions. In our paper, the performance dependence on dimension is exponential in general. In the future, we will investigate for what class of (non-convex) target functions f we can obtain performance bound independent of dimension d or has polynomial dependence on d. In addition, our results suggest that momentum-based SGHMC methods will work particularly well when the (non-convex) target functions have relatively flat landscapes. In the future, we will investigate whether we can obtain theoretical results for SGHMC on a wider class of non-convex problems.

Acknowledgements

We thank Agostino Capponi, Xiuli Chao, Wenbin Chen, Jim Dai, Murat A. Erdogdu, Fuqing Gao, Jianqiang Hu, Jin Ma, Sanjoy Mitter, Asuman Ozdaglar, Pablo Parrilo, Umut Şimşekli, and S. R. S. Varadhan for helpful discussions. Xuefeng Gao acknowledges support from Hong Kong RGC Grants 24207015 and 14201117. Mert Gürbüzbalaban's research is supported in part by the grants NSF DMS-1723085 and NSF CCF-1814888. Lingjiong Zhu is grateful to the support from the grant NSF DMS-1613164.

References

[AKW12] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning*, pages 1771–1778, 2012.

[ALG19] Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301, 2019.

[AZH16] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

[BBG14] MJ Betancourt, Simon Byrne, and Mark Girolami. Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*, 2014.

[BBLG17] Michael Betancourt, Simon Byrne, Sam Livingstone, and Mark Girolami. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, 2017.

[Bet17] Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[BGK05] Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

[BLNR15] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265, 2015.

[BM99] Vivek S Borkar and Sanjoy K Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of Optimization Theory and Applications*, 100(3):499–513, 1999.

[BT93] Dimitris Bertsimas and John Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.

[BV05] François Bolley and Cédric Villani. Weighted Csiszár-Kullback-pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 14(3):331–352, 2005.

[CCA+18] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting. *arXiv preprint arXiv: 1805.01648*, 2018.

[CCBJ18] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Annual Conference on Learning Theory*, 2018.

[CCG+16] Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1051–1060, 2016.

[CCS+17] P Chaudhari, Anna Choromanska, S Soatto, Yann LeCun, C Baldassi, C Borgs, J Chayes, Levent Sagun, and R Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*, 2017.

[CDC15] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2278–2286, 2015.

[CDHS18] Y. Carmon, J. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[CDT+09] Olivier Chapelle, Chuong B Do, Choon H Teo, Quoc V Le, and Alex J Smola. Tighter bounds for structured estimation. In *Advances in Neural Information Processing Systems*, pages 281–288, 2009.

[CFG14] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

[CFM+18] Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773, 2018.

[CHJ13] Sonja Cox, Martin Hutzenthaler, and Arnulf Jentzen. Local Lipschitz continuity in the initial value and strong completeness for nonlinear stochastic differential equations. *arXiv preprint arXiv:1309.5595*, 2013.

[CHS87] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

[CSWB06] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 201–208, 2006.

[Dal17] 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.

[DFB+14] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3203–3211, 2014.

[DK19] Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

[DKPR87] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[DKRD19] Arnak S Dalalyan, Avetik Karagulyan, and Lionel Riou-Durand. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *arXiv:1906.08530*, 2019.

[DLT+18] Simon S Du, Jason D Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1339–1348, 2018.

[DRD20] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[DS19] Susanne Ditlevsen and Adeline Samson. Hypoelliptic diffusions: filtering and inference from complete and partial observations. *Journal of the Royal Statistical Society Series B*, 81:361–384, 2019.

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, 47(4):1982–2010, 2019.

[FSS18] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8745–8756, 2018.

[GGZ20] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. In *Advances in Neural Information Processing Systems*, 2020.

[Gid85] Basilis Gidas. Nonstationary Markov chains and convergence of the annealing algorithm. *Journal of Statistical Physics*, 39(1-2):73–131, 1985.

[GL16] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[GLM17] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[GM91] 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.

[Haj85] Bruce Hajek. A tutorial survey of theory and applications of simulated annealing. In *1985 24th IEEE Conference on Decision and Control*, pages 755–760. IEEE, 1985.

[Hal88] JK Hale. *Asymptotic Behavior of Dissipative Systems*, volume 25. American Mathematical Society, 1988.

[HK70] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HKS89] Richard A Holley, Shigeo Kusuoka, and Daniel W Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *Journal of Functional Analysis*, 83(2):333–347, 1989.

[HLM$^+$17] Yaohua Hu, Chong Li, Kaiwen Meng, Jing Qin, and Xiaoqi Yang. Group sparse optimization via $\ell_{p,q}$ regularization. *The Journal of Machine Learning Research*, 18(1):960–1011, 2017.

[HLSS16] Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *International Conference on Machine Learning*, pages 1833–1841, 2016.

[HN04] Frédéric Hérau and Francis Nier. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Archive for Rational Mechanics and Analysis*, 171(2):151–218, 2004.

[HRS16] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[JT19] A. Jofré and P. Thompson. On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Mathematical Programming*, 174:253–292, 2019.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[KGV83] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[KO01] Misha E Kilmer and Dianne P O'Leary. Choosing regularization parameters in iterative methods for ill-posed problems. *SIAM Journal on matrix analysis and applications*, 22(4):1204–1221, 2001.

[Kra40] Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

[LCZZ18] Tianyi Liu, Zhehui Chen, Enlu Zhou, and Tuo Zhao. Toward deeper understanding of nonconvex stochastic optimization with momentum using diffusion approximations. *arXiv preprint arXiv:1802.05155*, 2018.

[LFM18] Haihao Lu, Robert M Freund, and Vahab Mirrokni. Accelerating greedy coordinate descent methods. In *International Conference on Machine Learning*, pages 3257–3266, 2018.

[LMS15] Benedict Leimkuhler, Charles Matthews, and Gabriel Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA Journal of Numerical Analysis*, 36(1):13–79, 2015.

[LS13] Robert S Liptser and Albert N Shiryaev. *Statistics of Random Processes: I. General Theory*, volume 5. Springer Science & Business Media, 2013.

[LV18] Yin Tat Lee and Santosh S Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121. ACM, 2018.

[MBM18] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[MCF15] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2917–2925, 2015.

[MPS18] Oren Mangoubi, S. Pillai, Natesh, and Aaron Smith. Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? *arXiv preprint arXiv:1808.03230*, 2018.

[MS17] O. Mangoubi and A. Smith. Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions. *arXiv preprint arXiv: 1708.07114*, 2017.

[MSH02] 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.

[Nea10] RM Neal. MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds.), 2010.

[Nes83] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

[NS13] Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093, 2013.

[Øks03] B. K. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 2003.

[OW19] Michael O'Neill and Stephen J Wright. Behavior of accelerated gradient methods near critical points of nonconvex problems. *Mathematical Programming*, 176:403–427, 2019.

[Pav14] Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.

[Pol87] Boris T Polyak. Introduction to Optimization. *Optimization Software. New York*, 1987.

[PT13] Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3102–3110, 2013.

[PW16] Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

[ROK12] S Alireza Razavi, Esa Ollila, and Visa Koivunen. Robust greedy algorithms for compressed sensing. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 969–973. IEEE, 2012.

[RRT17] 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.

[SBC14] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[SDJS18] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[ŞimşekliBCR16] Umut Şimşekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. Stochastic quasi-Newton Langevin Monte Carlo. In *International Conference on Machine Learning*, pages 642–651, 2016.

[ŞimşekliYN+18] U. Şimşekli, Ç. Yıldız, T. H. Nguyen, G. Richard, and A. Taylan Cemgil. Asynchronous Stochastic Quasi-Newton MCMC for Non-Convex Optimization. In *International Conference on Machine Learning*, pages 4674–4683, 2018.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

[SQW16] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

[TLR18] Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. In *Conference on Learning Theory*, pages 857–875, 2018.

[TTV16] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

[Vil08] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[WCX19] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535, 2019.

[Wib18] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Annual Conference on Learning Theory*, 2018.

[WJHZ13] Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502):632–643, 2013.

[WL07] Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

[WRJ16] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

[XCZG18] 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.

[ZLC17] Yuchen Zhang, Percy. Liang, and Moses. Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.

[ZZLC17] Huishuai Zhang, Yi Zhou, Yingbin Liang, and Yuejie Chi. A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *The Journal of Machine Learning Research*, 18(1):5164–5198, 2017.

A Proof of Theorem 2 and Corollary 4

We first present several technical lemmas that will be used in our analysis and review existing results for the underdamped Langevin SDE. The proof of these lemmas will be deferred to Section C.

Our analysis for analyzing the convergence speed of the SGHMC1 algorithm and its comparison to the underdamped Langevin SDE is based on the 2-Wasserstein distance and this requires the L^2 norm of the iterates to be finite. In the next lemma, we show that L^2 norm of the both discrete and continuous dynamics are uniformly bounded over time with explicit constants. The main idea is to make use of the properties of the Lyapunov function $\mathcal{V}$ which is designed originally for the continuous-time process and show that the discrete dynamics can also be controlled by it.

Lemma 16 (Uniform L^2 bounds)**.** *(i) It holds that*

$$\sup_{t\geq 0} \mathbb{E}_{\mathbf{z}} \|X(t)\|^2 \leq C_x^c := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) d\mu_0(x,v) + \frac{d+A}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2} < \infty, \tag{A.1}$$

$$\sup_{t\geq 0} \mathbb{E}_{\mathbf{z}} \|V(t)\|^2 \leq C_v^c := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) d\mu_0(x,v) + \frac{d+A}{\lambda}}{\frac{\beta}{4}(1-2\lambda)} < \infty. \tag{A.2}$$

(ii) For $0 < \eta \leq \min\left\{\frac{\gamma}{K_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2K_1}, \frac{2}{\gamma\lambda}\right\}$*, where*

$$K_1 := \max\left\{\frac{32M^2\left(\frac{1}{2}+\gamma+\delta\right)}{(1-2\lambda)\beta\gamma^2}, \frac{8\left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)}{\beta(1-2\lambda)}\right\}, \tag{A.3}$$

$$K_2 := 2B^2\left(\frac{1}{2}+\gamma+\delta\right), \tag{A.4}$$

we have

$$\sup_{j\geq 0} \mathbb{E}_{\mathbf{z}} \|X_j\|^2 \leq C_x^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) \mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2} < \infty, \tag{A.5}$$

$$\sup_{j\geq 0} \mathbb{E}_{\mathbf{z}} \|V_j\|^2 \leq C_v^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) \mu_0(dx,dv) + \frac{4(d+A)}{\lambda}}{\frac{\beta}{4}(1-2\lambda)} < \infty. \tag{A.6}$$

Since SGHMC1 is a discretization of the underdamped SDE (except that noise is also added to the gradients), we expect SGHMC1 to follow the underdamped SDE dynamics. It is natural to seek for bounds between the probability law $\mu_{\mathbf{z},k}$ of the SGHMC1 algorithm at step k with time step η and that of the underdamped SDE at time $t = k\eta$ which we denote by $\nu_{\mathbf{z},k\eta}$. In our analysis, we first control the Kullback-Leibler (KL) divergence between

these two, and then convert these bounds into bounds in terms of the 2-Wasserstein metric, applying an optimal transportation inequality by [BV05]. Note that Bolley and Villani theorem has also been successfully applied to analyzing the SGLD dynamics in [RRT17]. However, the analysis in [RRT17] does not directly apply to our setting as underdamped dynamics require a different Lyapunov function. This step requires an exponential integrability property of the underdamped SDE process which we establish next, before stating our result in Lemma 18 about the diffusion approximation of the SGHMC1 iterates.

Lemma 17 (Exponential integrability)**.** *For every t,*

$$\mathbb{E}_{\mathbf{z}}\left[e^{\alpha_0\|(X(t),V(t))\|^2}\right] \leq \int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv) + \frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)t,$$

where

$$\alpha_0 := \frac{\alpha}{\frac{64}{(1-2\lambda)\beta\gamma^2} + \frac{32}{\beta(1-2\lambda)}}, \qquad \alpha := \frac{\lambda(1-2\lambda)}{12}. \tag{A.7}$$

We showed in the above Lemma 17 that the exponential moments grow at most linearly in time t, which is a strict improvement from the exponential growth in time t in [RRT17]. As a result, in the following Lemma 18 for the diffusion approximation, our upper bound is of the order $(k\eta)^{3/2}\sqrt{\log(k\eta)}(\delta^{1/4}+\eta^{1/4}) + k\eta\sqrt{\eta}$ compared to $k\eta(\delta^{1/4}+\eta^{1/4})$ in [RRT17]. The method that is used in the proof of Lemma 17 for the underdamped dynamics can indeed be adapted to the case of the overdamped dynamics to improve the results in [RRT17].

Lemma 18 (Diffusion approximation)**.** *For any $k \in \mathbb{N}$ and any $\eta \leq 1$, so that $k\eta \geq e$ and η satisfies the condition in Part (ii) of Lemma 16. Then, we have*

$$\mathcal{W}_2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) \leq (C_0\delta^{1/4} + C_1\eta^{1/4}) \cdot (k\eta)^{3/2} \cdot \sqrt{\log(k\eta)} + C_2(k\eta)\sqrt{\eta},$$

where C_0, C_1 and C_2 are given by:

$$C_0 = \hat{\gamma} \cdot \left(\left(M^2C_x^d + B^2\right)\frac{\beta}{\gamma} + \sqrt{\left(M^2C_x^d + B^2\right)\frac{\beta}{\gamma}}\right)^{1/2}, \tag{A.8}$$

$$C_1 = \hat{\gamma} \cdot \left(\left(\frac{M^2\beta\eta}{\gamma} + \frac{\beta\eta\gamma}{2}\right)(C_2)^2 + \sqrt{\left(\frac{M^2\beta\eta}{\gamma} + \frac{\beta\eta\gamma}{2}\right)(C_2)^2}\right)^{1/2}, \tag{A.9}$$

$$C_2 = \left(2\gamma^2C_v^d + (4+2\delta)\left(M^2C_x^d + B^2\right) + 2\gamma\beta^{-1}\right)^{1/2}, \tag{A.10}$$

$$\hat{\gamma} = \frac{2\sqrt{2}}{\sqrt{\alpha_0}}\left(\frac{5}{2} + \log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv) + \frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)\right)\right)^{1/2}. \tag{A.11}$$

A.1 Convergence rate to the equilibrium of the underdamped SDE

We consider the underdamped SDE and bound the 2-Wasserstein distance $\mathcal{W}_2(\nu_{z,t}, \pi_{\mathbf{z}})$ to the equilibrium for a fix arbitrary time $t \geq 0$. Crucial to the analysis is [EGZ19], which quantifies the convergence to equilibrium for underdamped Langevin diffusions. We first review the results from [EGZ19]. Let us recall from (2.1) the definition of the Lyapunov function $\mathcal{V}(x, v)$:

$$\mathcal{V}(x, v) = \beta F_{\mathbf{z}}(x) + \frac{\beta}{4}\gamma^2(\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2).$$

For any $(x, v), (x', v') \in \mathbb{R}^{2d}$, we set:

$$\begin{aligned} r((x, v), (x', v')) &= \alpha_1\|x - x'\| + \|x - x' + \gamma^{-1}(v - v')\|, \\ \rho((x, v), (x', v')) &= h(r((x, v), (x', v'))) \cdot (1 + \varepsilon_1\mathcal{V}(x, v) + \varepsilon_1\mathcal{V}(x', v')), \end{aligned}$$

where $\alpha_1, \varepsilon_1 > 0$ are appropriately chosen constants, and $h : [0, \infty) \to [0, \infty)$ is continuous, non-decreasing concave function such that $h(0) = 0$, h is C^2 on $(0, R_1)$ for some constant $R_1 > 0$ with right-sided derivative $h'_+(0) = 1$ and left-sided derivative $h'_-(R_1) > 0$ and h is constant on $[R_1, \infty)$. For any two probability measures μ, ν on $\mathbb{R}^{2d}$, we define

$$\mathcal{H}_\rho(\mu, \nu) := \inf_{(X,V)\sim\mu,(X',V')\sim\nu} \mathbb{E}[\rho((X, V), (X', V'))]. \tag{A.12}$$

Note that $\mathcal{H}_\rho$ is a semi-metric, but not necessarily a metric. A simplified version of the main result from [EGZ19] which will be used in our setting is given below.

Theorem 19 (Theorem 2.3. and Corollary 2.6. in [EGZ19])**.** *There exist constants $\alpha_1, \epsilon_1 \in (0, \infty)$ and a continuous non-decreasing function $h : [0, \infty) \to [0, \infty)$ with $h(0) = 0$ such that we have*

$$\mathcal{W}_2(\nu_{\mathbf{z},k\eta}, \pi_{\mathbf{z}}) \leq C\sqrt{\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}})}e^{-\mu_* k\eta},$$

where

$$\mu_* = \frac{\gamma}{768}\min\{\lambda M\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}M\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}\}, \tag{A.13}$$

$$C = \sqrt{2}e^{1+\frac{\Lambda}{2}}\frac{1+\gamma}{\min\{1, \alpha_1\}}\sqrt{\max\{1, 4(1 + 2\alpha_1 + 2\alpha_1^2)(d + A)\beta^{-1}\gamma^{-1}\mu_*^{-1}/\min\{1, R_1\}\}}, \tag{A.14}$$

$$\Lambda = \frac{12}{5}(1 + 2\alpha_1 + 2\alpha_1^2)(d + A)M\gamma^{-2}\lambda^{-1}(1 - 2\lambda)^{-1}, \qquad \alpha_1 = (1 + \Lambda^{-1})M\gamma^{-2}, \tag{A.15}$$

$$\varepsilon_1 = 4\gamma^{-1}\mu_*/(d + A), \tag{A.16}$$

$$R_1 = 4 \cdot (6/5)^{1/2}(1 + 2\alpha_1 + 2\alpha_1^2)^{1/2}(d + A)^{1/2}\beta^{-1/2}\gamma^{-1}(\lambda - 2\lambda^2)^{-1/2}. \tag{A.17}$$

We remark that the definitions of Λ, α_1 in (A.15) are coupled and there exists $\alpha_1 \in (0, \infty)$ so that Λ, α_1 in (A.15) are well defined; see Theorem 2.3. in [EGZ19]. In order to get explicit performance bounds, we also derive an upper bound for $\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}})$ in the next lemma. It is based on the (integrability properties) structure of the stationary distribution $\pi_{\mathbf{z}}$ and the Lyapunov function $\mathcal{V}$ that controls the L^2 norm of the initial distribution μ_0.

Lemma 20 (Bounding initialization error)**.** *If parts (i), (ii), (iii) and (iv) of Assumption 1 hold, then we have*

$$\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}}) \le \overline{\mathcal{H}}_\rho(\mu_0) := R_1 + R_1\varepsilon_1 \max\left\{M + \frac{1}{2}\beta\gamma^2, \frac{3}{4}\beta\right\} \|(x,v)\|^2_{L^2(\mu_0)} + R_1\varepsilon_1\left(M + \frac{1}{2}\beta\gamma^2\right)\frac{b + d/\beta}{m} + R_1\varepsilon_1\frac{3}{4}d + 2R_1\varepsilon_1\left(\beta A_0 + \frac{\beta B^2}{2M}\right), \quad \text{(A.18)}$$

where $\|(x,v)\|^2_{L^2(\mu_0)} := \int_{\mathbb{R}^{2d}} \|(x,v)\|^2 \mu_0(dx, dv)$.

A.2 Proof of Theorem 2

As the function $F_{\mathbf{z}}$ satisfies the conditions in Lemma 26 in Section E with $c_1 = M$ and $c_2 = B$ (Lemma 25 in Section E), and the probability measures $\mu_{k,\mathbf{z}}, \pi_{\mathbf{z}}$ have finite second moments (Lemma 16), we can apply Lemma 26 and deduce that

$$\left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\mu_{k,\mathbf{z}}(dx, dv) - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx, dv)\right| \le (M\sigma + B)\cdot \mathcal{W}_2(\mu_{\mathbf{z},k}, \pi_{\mathbf{z}}). \quad \text{(A.19)}$$

Here, one can obtain from Lemma 16 and Theorem 19 (convergence in 2-Wasserstein distance implies convergence of second moments) that

$$\sigma^2 = \max\left\{C_x^c, C_x^d\right\} = C_x^d. \quad \text{(A.20)}$$

Now, by Lemma 18 and Theorem 19, we have

$$\begin{aligned}\mathcal{W}_2(\mu_{\mathbf{z},k}, \pi_{\mathbf{z}}) &\le \mathcal{W}_2(\mu_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) + \mathcal{W}_2(\nu_{\mathbf{z},k\eta}, \pi_{\mathbf{z}}) \\ &\le (C_0\delta^{1/4} + C_1\eta^{1/4})\cdot(k\eta)^{3/2}\cdot\sqrt{\log(k\eta)} + C_2(k\eta)\sqrt{\eta} + C\sqrt{\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}})}e^{-\mu_* k\eta}.\end{aligned}$$

It then follows from (A.19) that

$$\begin{aligned}&\left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\mu_{k,\mathbf{z}}(dx, dv) - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx, dv)\right| \\ &\le (M\sigma + B)\cdot\left(C\sqrt{\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}})}e^{-\mu_* k\eta} + (C_0\delta^{1/4} + C_1\eta^{1/4})\cdot(k\eta)^{3/2}\cdot\sqrt{\log(k\eta)} + C_2(k\eta)\sqrt{\eta}\right).\end{aligned}$$

Let $k\eta \geq e$, and

$$k\eta = \frac{1}{\mu_*} \log\left(\frac{1}{\varepsilon}\right).$$

Then for any η satisfying the condition in Lemma 16 and $\eta \leq \left(\frac{\varepsilon}{(\log(1/\varepsilon))^{3/2}}\right)^4$, we have

$$\begin{aligned}
&\left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\mu_{k,\mathbf{z}}(dx,dv) - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv)\right| \\
&\leq (M\sigma + B)\cdot\left(C\sqrt{\mathcal{H}_\rho(\mu_0,\pi_{\mathbf{z}})}\varepsilon + \left(\frac{C_0}{\mu_*^{3/2}}(\log(1/\varepsilon))^{3/2}\delta^{1/4} + \frac{C_1}{\mu_*^{3/2}}\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}\right. \\
&\qquad\left. + \frac{C_2}{\mu_*}\frac{\epsilon^2}{(\log(1/\varepsilon))^2}\right).
\end{aligned}$$

The proof is therefore complete.

A.3 Proof of Corollary 4

With a slight abuse of notations, consider the random elements $(\hat{X},\hat{V})$ and $(\hat{X}^*,\hat{V}^*)$ with $\mathrm{Law}((\hat{X},\hat{V})|\mathbf{Z}=\mathbf{z}) = \mu_{\mathbf{z},k}$ and $\mathrm{Law}((\hat{X}^*,\hat{V}^*)|\mathbf{Z}=\mathbf{z}) = \pi_{\mathbf{z}}$. Then we can decompose the expected population risk of $\hat{X}$ (which has the same distribution as X_k) as follows:

$$\mathbb{E}F(\hat{X}) - F^* = \Big(\mathbb{E}F(\hat{X}) - \mathbb{E}F(\hat{X}^*)\Big) + \Big(\mathbb{E}F(\hat{X}^*) - \mathbb{E}F_{\mathbf{Z}}(\hat{X}^*)\Big) + \Big(\mathbb{E}F_{\mathbf{Z}}(\hat{X}^*) - F^*\Big). \quad \text{(A.21)}$$

The first term in (A.21) can be written as:

$$\mathbb{E}F(\hat{X}) - \mathbb{E}F(\hat{X}^*) = \int_{\mathcal{Z}^n} P^n(d\mathbf{z})\left(\int_{\mathbb{R}^{2d}} F_{\mathbf{z}}(x)\mu_{k,\mathbf{z}}(dx,dv) - \int_{\mathbb{R}^{2d}} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv)\right),$$

where P^n is the product measure of independent random variables $Z_1,\ldots,Z_n$. Then it follows from Theorem 2 and Lemma 20 that

$$\mathbb{E}F(\hat{X}) - \mathbb{E}F(\hat{X}^*) \leq \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon).$$

Next, we bound the second and third terms in (A.21). Note that

$$\int_{\mathbb{R}^{2d}} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) = \int_{\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx),$$

where $\pi_{\mathbf{z}}(dx) = \Lambda_{\mathbf{z}} e^{-\beta F_{\mathbf{z}}(x)}dx$ and $\Lambda_{\mathbf{z}} = \int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(x)}dx$. The distribution $\pi_{\mathbf{z}}(dx)$, i.e., the $x-$marginal of $\pi_{\mathbf{z}}(dx,dv)$, is the same as the stationary distribution of the overdamped Langevin SDE in (1.4). Therefore the second term and the third term in (A.21) can be bounded the same as in [RRT17] for the overdamped dynamics.

Specifically, the second term in (A.21) can be bounded as

$$\mathbb{E}F(\hat{X}^*) - \mathbb{E}F_{\mathbf{Z}}(\hat{X}^*) \leq \frac{4\beta c_{LS}}{n}\left(\frac{M^2}{m}(b + d/\beta) + B^2\right) = \mathcal{J}_3(n),$$

by applying Lemma 27, and the last term in (A.21) can be bounded as

$$\begin{aligned}\mathbb{E}F_{\mathbf{Z}}(\hat{X}^*) - F^* &= \mathbb{E}\left[F_{\mathbf{Z}}(\hat{X}^*) - \min_{x\in\mathbb{R}^d} F_{\mathbf{Z}}(x)\right] + \mathbb{E}\left[\min_{x\in\mathbb{R}^d} F_{\mathbf{Z}}(x) - F_{\mathbf{Z}}(x^*)\right] \\ &\leq \mathbb{E}\left[F_{\mathbf{Z}}(\hat{X}^*) - \min_{x\in\mathbb{R}^d} F_{\mathbf{Z}}(x)\right] \leq \mathcal{J}_2,\end{aligned}$$

where x^* is any minimizer of $F(x)$, i.e., $F(x^*) = F^*$, and the last step is due to Lemma 28. The proof is complete.

B Proof of Theorem 6 and Corollary 8

The proof of Theorem 6 (Corollary 8) is similar to the proof of Theorem 2 (Corollary 4). There are two key new results that we need to establish: a uniform (in time) L^2 bound for the SGHMC2 iterates $(\hat{X}_k, \hat{V}_k)$, and the diffusion approximation that characterizes the 2-Wasserstein distance between the SGHMC2 iterates and the continuous-time underdampled Langevin diffusion. We summarize these two results in the following two lemmas and defer their proofs to Section D. With these two lemmas, Theorem 6 and Corollary 8 readily follow and we omit the proof details.

Lemma 21 (Uniform L^2 bounds for SGHMC2 iterates)**.**
For $0 < \eta \leq \min\left\{1, \frac{\gamma}{\hat{K}_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2\hat{K}_1}, \frac{2}{\gamma\lambda}\right\}$, *where*

$$\hat{K}_1 := K_1 + Q_1\frac{4}{1-2\lambda} + Q_2\frac{8}{(1-2\lambda)\gamma^2}, \tag{B.1}$$

$$\hat{K}_2 := K_2 + Q_3, \tag{B.2}$$

where K_1, K_2 are defined in (A.3) and (A.4), and

$$Q_1 := \frac{1}{2}c_0\left((5M+4-2\gamma+(c_0+\gamma^2))+(1+\gamma)\left(\frac{5}{2}+c_0(1+\gamma)\right)+2\gamma^2\lambda\right), \tag{B.3}$$

$$Q_2 := \frac{1}{2}c_0\left[\left((1+\gamma)\left(c_0(1+\gamma)+\frac{5}{2}\right)+c_0+2+\lambda\gamma^2+2(Mc_0+M+1)\right)\left(2(1+\delta)M^2\right)\right.$$
$$\left.+\left(2M^2+\gamma^2\lambda+\frac{3}{2}\gamma^2(1+\gamma)\right)\right], \tag{B.4}$$

$$Q_3 := c_0\left((1+\gamma)\left(c_0(1+\gamma)+\frac{5}{2}\right)+c_0+2+\lambda\gamma^2+2(Mc_0+M+1)\right)(1+\delta)B^2+c_0B^2$$
$$+\frac{1}{2}\gamma^3\beta^{-1}c_{22}+\gamma^2\beta^{-1}c_{12}+M\gamma\beta^{-1}c_{22}, \tag{B.5}$$

where

$$c_0 := 1+\gamma^2, \qquad c_{12} := \frac{d}{2}, \qquad c_{22} := \frac{d}{3}, \tag{B.6}$$

we have

$$\sup_{j\geq 0}\mathbb{E}_{\mathbf{z}}\|\hat{X}_j\|^2 \leq C_x^d, \qquad \sup_{j\geq 0}\mathbb{E}_{\mathbf{z}}\|\hat{V}_j\|^2 \leq C_v^d, \tag{B.7}$$

where C_x^d and C_v^d are defined in (A.5) and (A.6).

Next, let us provide a diffusion approximation between the SGHMC2 algorithm $(\hat{X}_k, \hat{V}_k)$ and the continuous time underdamped diffusion process $(X(k\eta), V(k\eta))$, and we use $\hat{\mu}_{\mathbf{z},k}$ to denote the law of $(\hat{X}_k, \hat{V}_k)$ and $\nu_{\mathbf{z},k}$ to denote the law of $(X(k\eta), V(k\eta))$.

Lemma 22 (Diffusion approximation)**.** *For any $k \in \mathbb{N}$ and any η, so that $k\eta \geq e$ and η satisfies the condition in Lemma 21, we have*

$$\mathcal{W}_2(\hat{\mu}_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) \leq (C_0\delta^{1/4} + \hat{C}_1\eta^{1/2}) \cdot \sqrt{k\eta} \cdot \sqrt{\log(k\eta)},$$

where C_0 is defined in (A.8) and $\hat{C}_1$ is given by:

$$\hat{C}_1 := \hat{\gamma}\cdot\left(\frac{3\beta M^2}{2\gamma}\left(C_v^d+\left(2(1+\delta)M^2C_x^d+2(1+\delta)B^2\right)+\frac{2d\gamma\beta^{-1}}{3}\right)\right.$$
$$\left.+\sqrt{\frac{3\beta M^2}{2\gamma}\left(C_v^d+(2(1+\delta)M^2C_x^d+2(1+\delta)B^2)+\frac{2d\gamma\beta^{-1}}{3}\right)}\right)^{1/2}, \tag{B.8}$$

where $\hat{\gamma}$ is defined in (A.11).

C Proofs of Lemmas in Section A

C.1 Proof of Lemma 16

(i) We first prove the continuous–time case. The main idea is to use the following Lyapunov function (see (2.1)) introduced in [EGZ19] for the underdamped Langevin diffusion:

$$\mathcal{V}(x,v) = \beta F_{\mathbf{z}}(x) + \frac{\beta}{4}\gamma^2(\|x+\gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2). \tag{C.1}$$

Lemma 1.3 in [EGZ19] showed that if the drift condition in (2.2) holds, then

$$\mathcal{L}\mathcal{V} \le \gamma(d + A - \lambda\mathcal{V}), \tag{C.2}$$

where $\mathcal{L}$ is the infinitesimal generator of the underdamped Langevin diffusion (X,V) defined in (1.5)–(1.6):

$$\mathcal{L}\mathcal{V} = -(\gamma v + \nabla F_{\mathbf{z}}(x))\nabla_v\mathcal{V} + \gamma\beta^{-1}\Delta_v\mathcal{V} + v\nabla_x\mathcal{V}. \tag{C.3}$$

To show part (i), we first note that for $\lambda \le \frac{1}{4}$,

$$\begin{aligned}\mathcal{V}(x,v) &\ge \beta F_{\mathbf{z}}(x) + \frac{\beta}{4}(1-2\lambda)\gamma^2(\|x+\gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2)\\ &\ge \max\left\{\frac{1}{8}(1-2\lambda)\beta\gamma^2\|x\|^2, \frac{\beta}{4}(1-2\lambda)\|v\|^2\right\}.\end{aligned} \tag{C.4}$$

Now let us set for each $t \ge 0$,

$$L(t) := \mathbb{E}_{\mathbf{z}}[\mathcal{V}(X(t),V(t))], \tag{C.5}$$

and we will provide an upper bound for $L(t)$.

First, we can compute that

$$\nabla_v\mathcal{V} = \beta v + \frac{\beta\gamma}{2}x, \tag{C.6}$$

By Itô's formula and (C.6),

$$\begin{aligned}d(e^{\gamma\lambda t}\mathcal{V}(X(t),V(t))) &= \gamma\lambda e^{\gamma\lambda t}\mathcal{V}(X(t),V(t))dt + e^{\gamma\lambda t}\mathcal{L}\mathcal{V}(X(t),V(t))dt\\ &\quad + e^{\gamma\lambda t}\left(\beta V(t) + \frac{\beta\gamma}{2}X(t)\right)\cdot\sqrt{2\gamma\beta^{-1}}dB(t),\end{aligned}$$

which together with (C.2) implies that

$$\begin{aligned}e^{\gamma\lambda t}\mathcal{V}(X(t),V(t)) &\le \mathcal{V}(X(0),V(0)) + \gamma(d+A)\int_0^t e^{\lambda\gamma s}ds\\ &\quad - \int_0^t e^{\gamma\lambda s}\left(\beta V(s) + \frac{\beta\gamma}{2}X(s)\right)\cdot\sqrt{2\gamma\beta^{-1}}dB(s).\end{aligned} \tag{C.7}$$

Note that $\nabla F_{\mathbf{z}}(x)$ is Lipschitz continuous by part (ii) of Assumption 1, and hence $(X(t), V(t))$ is the unique strong solution of the SDE (1.5)-(1.6), and thus $\mathbb{E}[\int_0^T \|V(t)\|^2 + \|X(t)\|^2 dt] < \infty$ for every $T > 0$ (See e.g. [Øks03]). Therefore, for every $T > 0$, we have

$$\int_0^T e^{2\gamma\lambda s} \left\| \beta V(s) + \frac{\beta\gamma}{2} X(s) \right\|^2 (2\gamma\beta^{-1}) ds < \infty,$$

and hence $\int_0^t e^{\gamma\lambda s} \left(\beta V(s) + \frac{\beta\gamma}{2} X(s) \right) \cdot \sqrt{2\gamma\beta^{-1}} B(s)$ is a martingale. Then we can infer from (C.7) and (C.5) that for any $t \geq 0$,

$$L(t) = \mathbb{E}_{\mathbf{z}}[\mathcal{V}(X(t), V(t))] \leq L(0) e^{-\gamma\lambda t} + \frac{d + A}{\lambda} (1 - e^{-\gamma\lambda t}).$$

In combination with (C.4), we obtain that (X, V) are uniformly (in time) L^2 bounded. Indeed, we have

$$\begin{aligned} \frac{1}{8}(1 - 2\lambda)\beta\gamma^2 \mathbb{E}_{\mathbf{z}} \|X(t)\|^2 &\leq \mathbb{E}_{\mathbf{z}}[\mathcal{V}(X_0, V_0)] + \frac{d + A}{\lambda}, \\ \frac{\beta}{4}(1 - 2\lambda)\mathbb{E}_{\mathbf{z}} \|V(t)\|^2 &\leq \mathbb{E}_{\mathbf{z}}[\mathcal{V}(X_0, V_0)] + \frac{d + A}{\lambda}. \end{aligned}$$

The proof of part (i) is complete by noting that $\mathbb{E}_{\mathbf{z}}[\mathcal{V}(X_0, V_0)]$ is finite from part (v) of Assumption 1.

(ii) Next, we prove the uniform (in time) L^2 bounds for (X_k, V_k). Let us recall the dynamics:

$$V_{k+1} = V_k - \eta[\gamma V_k + g(X_k, U_{\mathbf{z},k})] + \sqrt{2\gamma\beta^{-1}\eta}\xi_k, \tag{C.8}$$

$$X_{k+1} = X_k + \eta V_k, \tag{C.9}$$

where $\mathbb{E}g(x, U_{\mathbf{z},k}) = \nabla F_{\mathbf{z}}(x)$ for any x. We again use the Lyapunov function $\mathcal{V}(x, v)$ in (C.1), and set for each $k = 0, 1, \ldots,$

$$L_2(k) = \mathbb{E}_{\mathbf{z}} \mathcal{V}(X_k, V_k)/\beta = \mathbb{E}_{\mathbf{z}} \left[F_{\mathbf{z}}(X_k) + \frac{1}{4}\gamma^2 \left(\|X_k + \gamma^{-1} V_k\|^2 + \|\gamma^{-1} V_k\|^2 - \lambda \|X_k\|^2 \right) \right]. \tag{C.10}$$

We show below that one can find explicit constants $K_1, K_2 > 0$, such that

$$(L_2(k+1) - L_2(k))/\eta \leq \gamma(d/\beta + A/\beta - \lambda L_2(k)) + (K_1 L_2(k) + K_2) \cdot \eta.$$

We proceed in several steps in upper bounding $L_2(k+1)$.

First, by using the independence of $V_k - \eta[\gamma V_k + g_k(X_k, U_{z,k})]$ and ξ_k, we can obtain from (C.8) that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\|V_{k+1}\|^2 \\
&= \mathbb{E}_{\mathbf{z}}\|V_k - \eta[\gamma V_k + g_k(X_k, U_{\mathbf{z},k})]\|^2 + 2\gamma\beta^{-1}\eta\mathbb{E}_{\mathbf{z}}\|\xi_k\|^2 \\
&= \mathbb{E}_{\mathbf{z}}\|V_k - \eta[\gamma V_k + g_k(X_k, U_{\mathbf{z},k})]\|^2 + 2\gamma\beta^{-1}\eta d \\
&= \mathbb{E}_{\mathbf{z}}\|V_k - \eta[\gamma V_k + \nabla F_{\mathbf{z}}(X_k)]\|^2 + 2\gamma\beta^{-1}\eta d + \eta^2\mathbb{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(X_k) - g_k(X_k, U_{\mathbf{z},k})\|^2 \\
&\leq (1-\eta\gamma)^2\mathbb{E}_{\mathbf{z}}\|V_k\|^2 - 2\eta(1-\eta\gamma)\mathbb{E}_{\mathbf{z}}[\langle V_k, \nabla F_{\mathbf{z}}(X_k)\rangle] + \eta^2\mathbb{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(X_k)\|^2 + 2\gamma\beta^{-1}\eta d \\
&\qquad + 2\delta\eta^2 M^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + 2\delta\eta^2 B^2 \\
&\leq (1-\eta\gamma)^2\mathbb{E}_{\mathbf{z}}\|V_k\|^2 - 2\eta(1-\eta\gamma)\mathbb{E}_{\mathbf{z}}[\langle V_k, \nabla F_{\mathbf{z}}(X_k)\rangle] \\
&\quad + \eta^2(M^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + B^2 + 2MB\mathbb{E}_{\mathbf{z}}\|X_k\|) + 2\gamma\beta^{-1}\eta d \\
&\qquad + 2\delta\eta^2 M^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + 2\delta\eta^2 B^2,
\end{aligned}$$

where we have used part (iv) of Assumption 1 and Lemma 25 in Section E in the Appendix. By using $|x| \leq \frac{|x|^2+1}{2}$, we immediately get

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}}\|V_{k+1}\|^2 \leq (1-\eta\gamma)^2\mathbb{E}_{\mathbf{z}}\|V_k\|^2 - 2\eta\mathbb{E}_{\mathbf{z}}[\langle V_k, \nabla F_{\mathbf{z}}(X_k)\rangle] + 2\eta^2\gamma\mathbb{E}_{\mathbf{z}}[\langle V_k, \nabla F_{\mathbf{z}}(X_k)\rangle] \\
+ \left(\eta^2 M^2 + \eta^2 MB + \delta\eta^2 M^2\right)\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + (\eta^2 MB + 2\gamma\beta^{-1}\eta d + 2\delta\eta^2 B^2).
\end{aligned} \tag{C.11}$$

Second, we can compute from (C.9) that

$$\mathbb{E}_{\mathbf{z}}\|X_{k+1}\|^2 = \mathbb{E}_{\mathbf{z}}\|X_k\|^2 + 2\eta\mathbb{E}_{\mathbf{z}}[\langle X_k, V_k\rangle] + \eta^2\mathbb{E}_{\mathbf{z}}\|V_k\|^2. \tag{C.12}$$

Third, note that

$$F_{\mathbf{z}}(X_{k+1}) = F_{\mathbf{z}}(X_k + \eta V_k) = F_{\mathbf{z}}(X_k) + \int_0^1 \langle \nabla F_{\mathbf{z}}(X_k + \tau\eta V_k), \eta V_k\rangle d\tau,$$

which immediately suggests that

$$\begin{aligned}
|F_{\mathbf{z}}(X_{k+1}) - F_{\mathbf{z}}(X_k) - \langle \nabla F_{\mathbf{z}}(X_k), \eta V_k\rangle| &= \left|\int_0^1 \langle \nabla F_{\mathbf{z}}(X_k + \tau\eta V_k) - \nabla F_{\mathbf{z}}(X_k), \eta V_k\rangle d\tau\right| \\
&\leq \int_0^1 \|\nabla F_{\mathbf{z}}(X_k + \tau\eta V_k) - \nabla F_{\mathbf{z}}(X_k)\| \cdot \|\eta V_k\| \, d\tau \\
&\leq \frac{1}{2}M\eta^2\|V_k\|^2,
\end{aligned}$$

where the last inequality is due to the $M-$smoothness of $F_{\mathbf{z}}$. This implies

$$\mathbb{E}_{\mathbf{z}}F_{\mathbf{z}}(X_{k+1}) - \mathbb{E}_{\mathbf{z}}F_{\mathbf{z}}(X_k) \leq \eta\mathbb{E}_{\mathbf{z}}\langle \nabla F_{\mathbf{z}}(X_k), V_k\rangle + \frac{1}{2}M\eta^2\mathbb{E}_{\mathbf{z}}\|V_k\|^2. \tag{C.13}$$

Finally, we can compute that

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}} \left\| X_{k+1} + \gamma^{-1} V_{k+1} \right\|^2 \\
&= \mathbb{E}_{\mathbf{z}} \left\| X_k + \gamma^{-1} V_k - \eta\gamma^{-1} g(X_k, U_{\mathbf{z},k}) + \sqrt{2\gamma^{-1}\beta^{-1}\eta}\xi_k \right\|^2 \\
&= \mathbb{E}_{\mathbf{z}} \left\| X_k + \gamma^{-1} V_k - \eta\gamma^{-1} g(X_k, U_{\mathbf{z},k}) \right\|^2 + 2\gamma^{-1}\beta^{-1}\eta d \\
&= \mathbb{E}_{\mathbf{z}} \left\| X_k + \gamma^{-1} V_k - \eta\gamma^{-1} \nabla F_{\mathbf{z}}(X_k) \right\|^2 + 2\gamma^{-1}\beta^{-1}\eta d \\
&\qquad\qquad + \mathbb{E}_{\mathbf{z}} \left\| \eta\gamma^{-1} g(X_k, U_{\mathbf{z},k}) - \eta\gamma^{-1} \nabla F_{\mathbf{z}}(X_k) \right\|^2 \\
&\leq \mathbb{E}_{\mathbf{z}} \left\| X_k + \gamma^{-1} V_k - \eta\gamma^{-1} \nabla F_{\mathbf{z}}(X_k) \right\|^2 + 2\gamma^{-1}\beta^{-1}\eta d + 2\eta^2\gamma^{-2}\delta(M^2 \mathbb{E}_{\mathbf{z}} \left\| X_k \right\|^2 + B^2) \\
&= \mathbb{E}_{\mathbf{z}} \left\| X_k + \gamma^{-1} V_k \right\|^2 - 2\eta\gamma^{-1} \mathbb{E}_{\mathbf{z}} \langle X_k + \gamma^{-1} V_k, \nabla F_{\mathbf{z}}(X_k) \rangle \\
&\qquad + \eta^2\gamma^{-2} \mathbb{E}_{\mathbf{z}} \left\| \nabla F_{\mathbf{z}}(X_k) \right\|^2 + 2\gamma^{-1}\beta^{-1}\eta d + 2\eta^2\gamma^{-2}\delta(M^2 \mathbb{E}_{\mathbf{z}} \| X_k \|^2 + B^2),
\end{aligned} \tag{C.14}
$$

where we have used part (iv) of Assumption 1 in the inequality above.

Combining the equations (C.11), (C.12), (C.13) and (C.14), we get

$$
\begin{aligned}
&(L_2(k+1) - L_2(k))/\eta \\
&= \Big(\mathbb{E}_{\mathbf{z}}[F_{\mathbf{z}}(X_{k+1})] - \mathbb{E}_{\mathbf{z}}[F_{\mathbf{z}}(X_k)] + \frac{1}{4}\gamma^2 \left(\mathbb{E}_{\mathbf{z}} \| X_{k+1} + \gamma^{-1} V_{k+1} \|^2 - \mathbb{E}_{\mathbf{z}} \| X_k + \gamma^{-1} V_k \|^2 \right) \\
&\qquad + \frac{1}{4} \left(\mathbb{E}_{\mathbf{z}} \| V_{k+1} \|^2 - \mathbb{E}_{\mathbf{z}} \| V_k \|^2 \right) - \frac{1}{4}\gamma^2\lambda \left(\mathbb{E}_{\mathbf{z}} \| X_{k+1} \|^2 - \mathbb{E}_{\mathbf{z}} \| X_k \|^2 \right) \Big) \Big/ \eta \\
&\leq \mathbb{E}_{\mathbf{z}} \langle \nabla F_{\mathbf{z}}(X_k), V_k \rangle + \frac{1}{2} M\eta \mathbb{E}_{\mathbf{z}} \| V_k \|^2 - \frac{1}{2}\gamma \mathbb{E} \left\langle X_k + \gamma^{-1} V_k, \nabla F_{\mathbf{z}}(X_k) \right\rangle \\
&\qquad + \frac{1}{4}\eta \mathbb{E} \| \nabla F_{\mathbf{z}}(X_k) \|^2 + \frac{1}{2}\gamma\beta^{-1} d + \frac{1}{2}\eta\delta(M^2 \mathbb{E} \| X_k \|^2 + B^2) \\
&\qquad\qquad + \frac{1}{4}(-2\gamma + \eta\gamma^2) \mathbb{E}_{\mathbf{z}} \| V_k \|^2 - \frac{1}{2}(1 - \eta\gamma) \mathbb{E}_{\mathbf{z}}[\langle V_k, \nabla F_{\mathbf{z}}(X_k) \rangle] \\
&\qquad + \frac{1}{4}\eta(M^2 \mathbb{E}_{\mathbf{z}} \| X_k \|^2 + B^2 + 2MB \mathbb{E}_{\mathbf{z}} \| X_k \|) + \frac{1}{2}\gamma\beta^{-1} d \\
&\qquad\qquad + \frac{1}{2}\delta\eta M^2 \mathbb{E}_{\mathbf{z}} \| X_k \|^2 + \frac{1}{2}\delta\eta B^2 - \frac{1}{2}\gamma^2\lambda \mathbb{E}_{\mathbf{z}} \langle X_k, V_k \rangle - \frac{1}{4}\gamma^2\lambda\eta \mathbb{E}_{\mathbf{z}} \| V_k \|^2 \\
&= -\frac{\gamma}{2} \mathbb{E}_{\mathbf{z}} \langle \nabla F_{\mathbf{z}}(X_k), X_k \rangle - \frac{\gamma}{2} \mathbb{E}_{\mathbf{z}} \| V_k \|^2 - \frac{\gamma^2\lambda}{2} \mathbb{E}_{\mathbf{z}} \langle X_k, V_k \rangle + \gamma\beta^{-1} d + \mathcal{E}_k \eta \\
&\leq -\gamma\lambda \mathbb{E}_{\mathbf{z}}[F_{\mathbf{z}}(X_k)] - \frac{1}{4}\lambda\gamma^3 \mathbb{E}_{\mathbf{z}} \| X_k \|^2 + \gamma A/\beta - \frac{\gamma}{2} \mathbb{E}_{\mathbf{z}} \| V_k \|^2 - \frac{\gamma^2\lambda}{2} \mathbb{E}_{\mathbf{z}} \langle X_k, V_k \rangle + \gamma\beta^{-1} d + \mathcal{E}_k \eta,
\end{aligned} \tag{C.15}
$$

where we used the drift condition (2.2) in the last inequality, and

$$\mathcal{E}_k := \left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda\right)\mathbb{E}_{\mathbf{z}}\|V_k\|^2 + \frac{1}{4}\mathbb{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(X_k)\|^2 + \delta(M^2\mathbb{E}\|X_k\|^2 + B^2)$$
$$+ \frac{1}{2}\gamma\mathbb{E}_{\mathbf{z}}[\langle V_k, \nabla F_{\mathbf{z}}(X_k)\rangle] + \frac{1}{4}(M^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + B^2 + 2MB\mathbb{E}_{\mathbf{z}}\|X_k\|).$$

We can upper bound $\mathcal{E}_k$ as follows:

$$\begin{aligned}
\mathcal{E}_k &\le \left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda\right)\mathbb{E}_{\mathbf{z}}\|V_k\|^2 + \frac{1}{4}\mathbb{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(X_k)\|^2 + \delta(M^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + B^2) \\
&\quad + \gamma\mathbb{E}_{\mathbf{z}}\|V_k\|^2 + \gamma\mathbb{E}_{\mathbf{z}}\|\nabla F_{\mathbf{z}}(X_k)\|^2 + \frac{1}{4}\mathbb{E}_{\mathbf{z}}(M\|X_k\| + B)^2 \\
&\le \left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)\mathbb{E}_{\mathbf{z}}\|V_k\|^2 + \delta(M^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + B^2) \\
&\quad + \left(\frac{1}{4} + \gamma\right)\mathbb{E}_{\mathbf{z}}(M\|X_k\| + B)^2 + \frac{1}{4}\mathbb{E}_{\mathbf{z}}(M\|X_k\| + B)^2 \\
&\le \left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)\mathbb{E}_{\mathbf{z}}\|V_k\|^2 \\
&\qquad + 2M^2\left(\frac{1}{2} + \gamma + \delta\right)\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + 2B^2\left(\frac{1}{2} + \gamma + \delta\right).
\end{aligned}$$

Since $\lambda \le \frac{1}{4}$, we obtain from (C.4) and (C.10) that

$$\begin{aligned}
L_2(k) &\ge \max\left\{\frac{1}{8}(1-2\lambda)\gamma^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2, \frac{1}{4}(1-2\lambda)\mathbb{E}_{\mathbf{z}}\|V_k\|^2\right\} \\
&\ge \frac{1}{16}(1-2\lambda)\gamma^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + \frac{1}{8}(1-2\lambda)\mathbb{E}_{\mathbf{z}}\|V_k\|^2.
\end{aligned} \tag{C.16}$$

Therefore,

$$\mathcal{E}_k \le K_1 L_2(k) + K_2, \tag{C.17}$$

where we recall from (A.3) and (A.4) that

$$K_1 = \max\left\{\frac{2M^2\left(\frac{1}{2} + \gamma + \delta\right)}{\frac{1}{16}(1-2\lambda)\gamma^2}, \frac{\left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)}{\frac{1}{8}(1-2\lambda)}\right\},$$

and

$$K_2 = 2B^2\left(\frac{1}{2} + \gamma + \delta\right).$$

Moreover, since $\lambda \le \frac{1}{4}$, we infer from the definition of $L_2(k)$ in (C.10) that

$$\begin{aligned}
L_2(k) &= \mathbb{E}_{\mathbf{z}}[F_{\mathbf{z}}(X_k)] + \frac{1}{4}\gamma^2(1-\lambda)\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + \frac{1}{2}\gamma\mathbb{E}_{\mathbf{z}}[\langle X_k, V_k\rangle] + \frac{1}{2}\mathbb{E}_{\mathbf{z}}\|V_k\|^2 \\
&\le \mathbb{E}_{\mathbf{z}}[F_{\mathbf{z}}(X_k)] + \frac{1}{4}\gamma^2\mathbb{E}_{\mathbf{z}}\|X_k\|^2 + \frac{1}{2}\gamma\mathbb{E}_{\mathbf{z}}[\langle X_k, V_k\rangle] + \frac{1}{2\lambda}\mathbb{E}_{\mathbf{z}}\|V_k\|^2.
\end{aligned}$$

Together with (C.15) and (C.17), we deduce that

$$(L_2(k+1) - L_2(k))/\eta \le \gamma(d/\beta + A/\beta - \lambda L_2(k)) + (K_1 L_2(k) + K_2)\eta.$$

For $0 < \eta \le \min\left\{\frac{\gamma}{K_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2K_1}\right\}$, we get

$$(L_2(k+1) - L_2(k))/\eta \le 2\gamma(d/\beta + A/\beta) - \frac{1}{2}\gamma\lambda L_2(k),$$

which implies

$$L_2(k+1) \le \rho L_2(k) + K,$$

where

$$\rho := 1 - \eta\gamma\lambda/2, \qquad K := 2\eta\gamma(d/\beta + A/\beta),$$

and we have $\rho \in [0, 1)$, where we used the assumption that $\eta \le \frac{2}{\gamma\lambda}$. It follows that

$$L_2(k) \le L_2(0) + \frac{K}{1-\rho} = \mathbb{E}_{\mathbf{z}}\left[\mathcal{V}(X_0, V_0)/\beta\right] + \frac{4(d/\beta + A/\beta)}{\lambda}.$$

The result then follows from the inequality above and (C.16).

C.2 Proof of Lemma 17

From (C.1)–(C.3), we can directly obtain that

$$\begin{aligned}
\mathcal{L}e^{\alpha\mathcal{V}} &= \left[-(\gamma v + \nabla F_{\mathbf{z}}(x))\alpha\nabla_v\mathcal{V} + \gamma\beta^{-1}\alpha\Delta_v\mathcal{V} + \gamma\beta^{-1}\alpha^2\|\nabla_v\mathcal{V}\|^2 + v\alpha\nabla_x\mathcal{V}\right] e^{\alpha\mathcal{V}} \\
&= \left[\alpha\mathcal{L}\mathcal{V} + \gamma\beta^{-1}\alpha^2\|\nabla_v\mathcal{V}\|^2\right] e^{\alpha\mathcal{V}} \\
&\le \left[\alpha\gamma d + \alpha\gamma A - \alpha\gamma\lambda\mathcal{V} + \alpha^2\gamma\beta^{-1}\|\nabla_v\mathcal{V}\|^2\right] e^{\alpha\mathcal{V}}. \qquad \text{(C.18)}
\end{aligned}$$

Moreover, we recall from (C.6) that

$$\nabla_v\mathcal{V} = \beta v + \frac{\beta\gamma}{2}x,$$

and thus

$$\|\nabla_v\mathcal{V}\|^2 \le 2\beta^2\|v\|^2 + \frac{\beta^2\gamma^2}{2}\|x\|^2.$$

We recall from (C.4) that

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

Therefore, we have

$$\|\nabla_v\mathcal{V}\|^2 \le \left[\frac{8\beta^2}{\beta(1-2\lambda)} + \frac{4\beta^2\gamma^2}{(1-2\lambda)\beta\gamma^2}\right]\mathcal{V} = \frac{12\beta}{1-2\lambda}\mathcal{V}. \qquad \text{(C.19)}$$

By choosing:

$$\alpha = \frac{\lambda\beta}{\frac{12\beta}{1-2\lambda}} = \frac{\lambda(1-2\lambda)}{12}, \tag{C.20}$$

we get

$$\mathcal{L}e^{\alpha\mathcal{V}} \leq \alpha\gamma(d+A)e^{\alpha\mathcal{V}}. \tag{C.21}$$

Since $\mathcal{L}e^{\alpha\mathcal{V}} = \left[\mathcal{L}\alpha\mathcal{V} + \gamma\beta^{-1}\|\nabla_v\alpha\mathcal{V}\|^2\right]e^{\alpha\mathcal{V}}$, we have showed that

$$\mathcal{L}\alpha\mathcal{V} + \gamma\beta^{-1}\|\nabla_v\alpha\mathcal{V}\|^2 \leq \alpha\gamma(d+A).$$

Applying an exponential integrability result, e.g. Corollary 2.4. in [CHJ13], we get

$$\mathbb{E}\left[e^{\alpha\mathcal{V}(X(t),V(t))}\right] \leq \mathbb{E}\left[e^{\alpha\mathcal{V}(X(0),V(0))}\right]e^{\alpha\gamma(d+A)t}.$$

That is,

$$\mathbb{E}_{\mathbf{z}}\left[e^{\alpha\mathcal{V}(X(t),V(t))}\right] \leq \int_{\mathbb{R}^{2d}} e^{\alpha\mathcal{V}(x,v)+\alpha\gamma(d+A)t}\mu_0(dx,dv) < \infty. \tag{C.22}$$

Next, applying Itô's formula to $e^{\frac{1}{4}\alpha\mathcal{V}(X(t),V(t))}$, we obtain

$$\begin{aligned} e^{\frac{1}{4}\alpha\mathcal{V}(X(t),V(t))} = e^{\frac{1}{4}\alpha\mathcal{V}(X(0),V(0))} &+ \int_0^t \mathcal{L}e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}ds \\ &+ \int_0^t \frac{1}{2}\left(\beta V(s) + \frac{\beta\gamma}{2}X(s)\right)e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}\cdot dB(s). \end{aligned} \tag{C.23}$$

For every $T > 0$,

$$\begin{aligned} &\int_0^T \mathbb{E}\left\|\frac{1}{2}\left(\beta V(s) + \frac{\beta\gamma}{2}X(s)\right)e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}\right\|^2 ds \\ &\leq \frac{\beta^2}{2}\int_0^T \mathbb{E}\left[\left(\|V(s)\|^2 + \gamma^2\|X(s)\|^2\right)e^{\frac{1}{2}\alpha\mathcal{V}(X(s),V(s))}\right]ds \\ &\leq \frac{6\beta}{1-2\lambda}\int_0^T \mathbb{E}\left[\mathcal{V}(X(s),V(s))e^{\frac{1}{2}\alpha\mathcal{V}(X(s),V(s))}\right]ds \\ &\leq \frac{12\beta}{1-2\lambda}\int_0^T \mathbb{E}\left[e^{\alpha\mathcal{V}(X(s),V(s))}\right]ds < \infty, \end{aligned}$$

where we used (C.4) and (C.22). Thus, $\int_0^t \frac{1}{2}\left(\beta V(s) + \frac{\beta\gamma}{2}X(s)\right)e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}\cdot dB(s)$ is a martingale. By taking expectations on both hand sides of (C.23), we get

$$\mathbb{E}\left[e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}\right] = \mathbb{E}\left[e^{\frac{1}{4}\alpha\mathcal{V}(X(0),V(0))}\right] + \int_0^t \mathbb{E}\left[\mathcal{L}e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}\right]ds. \tag{C.24}$$

From (C.18), (C.19) and (C.20), we can infer that

$$\begin{aligned}\mathcal{L}e^{\frac{1}{4}\alpha\mathcal{V}} &\leq \left(\frac{1}{4}\alpha\gamma(d+A) - \frac{1}{4}\alpha\gamma\lambda\mathcal{V} + \gamma\beta^{-1}\frac{\alpha^2}{16}\|\nabla_v\mathcal{V}\|^2\right)e^{\frac{1}{4}\alpha\mathcal{V}} \\ &\leq \left(\frac{1}{4}\alpha\gamma(d+A) - \frac{3}{16}\alpha\gamma\lambda\mathcal{V}\right)e^{\frac{1}{4}\alpha\mathcal{V}} \\ &\leq \frac{1}{4}\alpha\gamma(d+A)e^{\frac{\alpha(d+A)}{3\lambda}},\end{aligned}$$

where in the last inequality we used the facts that $\mathcal{V} \geq 0$ and $\frac{1}{4}\alpha\gamma(d+A) - \frac{3}{16}\alpha\gamma\lambda\mathcal{V} \geq 0$ if and only if $\mathcal{V} \leq \frac{4(d+A)}{3\lambda}$. Therefore, it follows from (C.24) that

$$\mathbb{E}\left[e^{\frac{1}{4}\alpha\mathcal{V}(X(s),V(s))}\right] \leq \mathbb{E}\left[e^{\frac{1}{4}\alpha\mathcal{V}(X(0),V(0))}\right] + \frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)t.$$

Finally, by (C.4) again,

$$\|(x,v)\|^2 \leq 2\|x\|^2 + 2\|v\|^2 \leq \left[\frac{16}{(1-2\lambda)\beta\gamma^2} + \frac{8}{\beta(1-2\lambda)}\right]\mathcal{V}(x,v).$$

Hence, the conclusion follows.

C.3 Proof of Lemma 18

The proof is inspired by the proof of Lemma 7 in [RRT17] although more delicate in our setting. Note that the main technical difficulty here is that the underdamped Langevin diffusion is a hypoelliptic diffusion, i.e. the diffusion matrix of the stochastic differential equation defining the multidimensional diffusion process is not of full rank, but its solutions admit a smooth density, see [DS19]. In our case, there is no Brownian noise in $dX(t)$ term in (1.6) and the underdamped Langevin diffusion (1.5)-(1.6) is hypoelliptic. Consider the following continuous-time interpolation of (X_k, V_k):

$$\overline{V}(t) = V_0 - \int_0^t \gamma\overline{V}(\lfloor s/\eta\rfloor\eta)ds - \int_0^t g(\overline{X}(\lfloor s/\eta\rfloor\eta), \overline{U}_{\mathbf{z}}(s))ds + \sqrt{2\gamma\beta^{-1}}\int_0^t dB(s), \quad \text{(C.25)}$$

$$\overline{X}(t) = X_0 + \int_0^t \overline{V}(\lfloor s/\eta\rfloor\eta)ds, \quad \text{(C.26)}$$

where $\overline{U}_{\mathbf{z}}(t) := U_{\mathbf{z},k}$ for $k\eta \leq t < (k+1)\eta$. Then $(\overline{X}(k\eta), \overline{V}(k\eta))$ and (X_k, V_k) have the same distribution $\mu_{\mathbf{z},k}$ for each $k \geq 0$. Since there is no Brownian noise in $dX(t)$ term in (1.6) and $d\overline{X}(t)$ term in (C.26), and their dynamics are different, there does not exist a solution to equation (7.115) in Theorem 7.18 in [LS13], and one can not apply Girsanov theorem to compute the relative entropy between $(X(t), V(t))$ and $(\overline{V}(t), \overline{X}(t))$, which

is the main technical difficulty here. To overcome this challenge, we define an auxiliary diffusion process $(\tilde{X}(t), \tilde{V}(t))$:

$$\tilde{V}(t) = V_0 - \int_0^t \gamma \tilde{V}(\lfloor s/\eta \rfloor \eta) ds - \int_0^t g\left(X_0 + \int_0^{\lfloor s/\eta \rfloor \eta} \tilde{V}(\lfloor u/\eta \rfloor \eta) du, \overline{U}_{\mathbf{z}}(s) \right) ds + \sqrt{2\gamma\beta^{-1}} \int_0^t dB(s), \tag{C.27}$$

$$\tilde{X}(t) = X_0 + \int_0^t \tilde{V}(s) ds, \tag{C.28}$$

which serves as a bridge between the underdamped Langevin diffusion $(X(k\eta), V(k\eta))$ and the discrete time SGHMC1 iterates (X_k, V_k). Then, it is easy to see that $\tilde{V}(k\eta)$ has the same distribution as V_k, though $\tilde{X}(k\eta)$ is not distributed the same as X_k. Since the drift term in (C.28) in the auxiliary diffusion process $(\tilde{X}(t), \tilde{V}(t))$ has the same dynamics as the $dX(t)$ term in (1.6), Girsanov theorem is applicable according to Theorem 7.18 in [LS13].

Let $\mathbb{P}$ be the probability measure associated with the underdamped Langevin diffusion $(X(t), V(t))$ in (1.5)–(1.6) and $\tilde{\mathbb{P}}$ be the probability measure associated with the $(\tilde{X}(t), \tilde{V}(t))$ process in (C.27)–(C.28). Let $\mathcal{F}_t$ be the natural filtration up to time t. Then, the Radon-Nikodym derivative of $\mathbb{P}$ w.r.t. $\tilde{\mathbb{P}}$ is given by the Girsanov theorem (see e.g. Section 7.6 in [LS13]):

$$\left.\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}}\right|_{\mathcal{F}_t} = e^{-\sqrt{\frac{\beta}{2\gamma}} \int_0^t \left(\gamma \tilde{V}(s) - \gamma \tilde{V}(\lfloor s/\eta \rfloor \eta) + \nabla F_{\mathbf{z}}(\tilde{X}(s)) - g\left(X_0 + \int_0^{\lfloor s/\eta \rfloor \eta} \tilde{V}(\lfloor u/\eta \rfloor \eta) du, \overline{U}_{\mathbf{z}}(s) \right) \right) \cdot dB(s)} \cdot e^{-\frac{\beta}{4\gamma} \int_0^t \left\| \gamma \tilde{V}(s) - \gamma \tilde{V}(\lfloor s/\eta \rfloor \eta) + \nabla F_{\mathbf{z}}(\tilde{X}(s)) - g\left(X_0 + \int_0^{\lfloor s/\eta \rfloor \eta} \tilde{V}(\lfloor u/\eta \rfloor \eta) du, \overline{U}_{\mathbf{z}}(s) \right) \right\|^2 ds}.$$

Then by writing $\mathbb{P}_t$ and $\tilde{\mathbb{P}}_t$ as the probability measures $\mathbb{P}$ and $\tilde{\mathbb{P}}$ conditional on the filtration

$\mathcal{F}_t$,

$$
\begin{aligned}
&D(\tilde{\mathbb{P}}_t\|\mathbb{P}_t)\\
&:= -\int d\tilde{\mathbb{P}}_t \log \frac{d\mathbb{P}_t}{d\tilde{\mathbb{P}}_t}\\
&= \frac{\beta}{4\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\gamma\tilde{V}(s) - \gamma\tilde{V}(\lfloor s/\eta\rfloor\eta) + \nabla F_{\mathbf{z}}(\tilde{X}(s)) - g\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du, \overline{U}_{\mathbf{z}}(s)\right)\right\|^2 ds\\
&\le \frac{\beta}{2\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(u)du\right) - g\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du, \overline{U}_{\mathbf{z}}(s)\right)\right\|^2 ds\\
&\qquad + \frac{\beta}{2\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\gamma\tilde{V}(s) - \gamma\tilde{V}(\lfloor s/\eta\rfloor\eta)\right\|^2 ds\\
&\le \frac{\beta}{\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(u)du\right) - \nabla F_{\mathbf{z}}\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right)\right\|^2 ds\\
&\quad + \frac{\beta}{\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right) - g\left(X_0 + \int_0^{\lfloor s/\eta\rfloor\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du, \overline{U}_{\mathbf{z}}(s)\right)\right\|^2 ds\\
&\qquad + \frac{\beta}{2\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\gamma\tilde{V}(s) - \gamma\tilde{V}(\lfloor s/\eta\rfloor\eta)\right\|^2 ds,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})\\
&\le \frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0 + \int_0^{j\eta}\tilde{V}(u)du\right) - \nabla F_{\mathbf{z}}\left(X_0 + \int_0^{j\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right)\right\|^2\\
&\quad + \frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0 + \int_0^{j\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right) - g\left(X_0 + \int_0^{j\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du, U_{\mathbf{z},j}\right)\right\|^2\\
&\qquad + \frac{\beta}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\gamma\tilde{V}(s) - \gamma\tilde{V}(\lfloor s/\eta\rfloor\eta)\right\|^2 ds. \qquad \text{(C.29)}
\end{aligned}
$$

We first bound the first term in (C.29):

$$\begin{aligned}
&\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0+\int_0^{j\eta}\tilde{V}(u)du\right)-\nabla F_{\mathbf{z}}\left(X_0+\int_0^{j\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right)\right\|^2\\
&\leq M^2\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}\mathbb{E}_{\mathbf{z}}\left\|\int_0^{j\eta}\left(\tilde{V}(u)-\tilde{V}(\lfloor u/\eta\rfloor\eta)\right)du\right\|^2\\
&\leq M^2\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}j\eta\int_0^{j\eta}\mathbb{E}_{\mathbf{z}}\left\|\tilde{V}(u)-\tilde{V}(\lfloor u/\eta\rfloor\eta)\right\|^2du\\
&= M^2\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}j\eta\sum_{i=0}^{j-1}\int_{i\eta}^{(i+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\tilde{V}(u)-\tilde{V}(\lfloor u/\eta\rfloor\eta)\right\|^2du\,,
\end{aligned}$$

where we used part (ii) of Assumption 1 Cauchy-Schwarz inequality.

For $i\eta < u \leq (i+1)\eta$, we have

$$\tilde{V}(u)-\tilde{V}(\lfloor u/\eta\rfloor\eta) = -(u-i\eta)\gamma V_i-(u-i\eta)g\left(X_i,U_{\mathbf{z},i}\right)+\sqrt{2\gamma\beta^{-1}}(B(u)-B(i\eta)), \quad \text{(C.30)}$$

in distribution. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left\|\tilde{V}(u)-\tilde{V}(\lfloor u/\eta\rfloor\eta)\right\|^2\\
&=(u-i\eta)^2\mathbb{E}_{\mathbf{z}}\left\|\gamma V_i+g\left(X_i,U_{\mathbf{z},i}\right)\right\|^2+2\gamma\beta^{-1}(u-i\eta)\\
&=(u-i\eta)^2\mathbb{E}_{\mathbf{z}}\left\|\gamma V_i+\nabla F_{\mathbf{z}}(X_i)\right\|^2+(u-i\eta)^2\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(X_i)-g\left(X_i,U_{\mathbf{z},i}\right)\right\|^2+2\gamma\beta^{-1}(u-i\eta)\\
&\leq 2\eta^2\mathbb{E}_{\mathbf{z}}\left\|\gamma V_i\right\|^2+2\eta^2\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(X_i)\right\|^2+\eta^2 2\delta(M^2\mathbb{E}_{\mathbf{z}}\left\|X_i\right\|^2+B^2)+2\gamma\beta^{-1}\eta\\
&\leq 2\gamma^2\eta^2\mathbb{E}_{\mathbf{z}}\left\|V_i\right\|^2+4\eta^2\left(M^2\mathbb{E}_{\mathbf{z}}\left\|X_i\right\|^2+B^2\right)+\eta^2 2\delta(M^2\mathbb{E}_{\mathbf{z}}\left\|X_i\right\|^2+B^2)+2\gamma\beta^{-1}\eta.
\end{aligned} \quad \text{(C.31)}$$

This implies that

$$\begin{aligned}
&M^2\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}j\eta\sum_{i=0}^{j-1}\int_{i\eta}^{(i+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\tilde{V}(u)-\tilde{V}(\lfloor u/\eta\rfloor\eta)\right\|^2du\\
&\leq M^2\frac{\beta}{\gamma}(k\eta)^3\left(2\gamma^2\eta^2\sup_{j\geq 0}\mathbb{E}_{\mathbf{z}}\left\|V_j\right\|^2+(4+2\delta)\eta^2\left(M^2\sup_{j\geq 0}\mathbb{E}_{\mathbf{z}}\left\|X_j\right\|^2+B^2\right)+2\gamma\beta^{-1}\eta\right).
\end{aligned}$$

We can also bound the second term in (C.29):

$$\begin{aligned}
&\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}\left(X_0+\int_0^{j\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right)-g\left(X_0+\int_0^{j\eta}\tilde{V}(\lfloor u/\eta\rfloor\eta)du,U_{\mathbf{z},j}\right)\right\|^2\\
&\leq\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}2\delta\left(M^2\mathbb{E}_{\mathbf{z}}\left\|X_0+\int_0^{s}\tilde{V}(\lfloor u/\eta\rfloor\eta)du\right\|^2+B^2\right)\\
&=\frac{\beta\eta}{\gamma}\sum_{j=0}^{k-1}2\delta\left(M^2\mathbb{E}_{\mathbf{z}}\left\|X_j\right\|^2+B^2\right)\\
&\leq\frac{2\beta\delta}{\gamma}k\eta\left(M^2\sup_{j\geq 0}\mathbb{E}_{\mathbf{z}}\left\|X_j\right\|^2+B^2\right),
\end{aligned}$$

where the first inequality follows from part (iv) of Assumption 1.

Finally, let us bound the third term in (C.29) as follows:

$$\begin{aligned}
&\frac{\beta}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\gamma\tilde{V}(s)-\gamma\tilde{V}(\lfloor s/\eta\rfloor\eta)\right\|^2ds\\
&\leq\frac{\beta}{2\gamma}(k\eta)\gamma^2\left(2\gamma^2\eta^2C_v^d+(4+2\delta)\eta^2\left(M^2C_x^d+B^2\right)+2\gamma\beta^{-1}\eta\right),
\end{aligned}$$

where we used the estimate in (C.31).

Hence, together with Lemma 16, we conclude that that

$$\begin{aligned}
D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})\leq M^2\frac{\beta}{\gamma}(k\eta)^3&\left(2\gamma^2\eta^2C_v^d+(4+2\delta)\eta^2\left(M^2C_x^d+B^2\right)+2\gamma\beta^{-1}\eta\right)\\
&+\frac{2\beta\delta}{\gamma}k\eta\left(M^2C_x^d+B^2\right)\\
&+\frac{\beta}{2\gamma}(k\eta)\gamma^2\left(2\gamma^2\eta^2C_v^d+(4+2\delta)\eta^2\left(M^2C_x^d+B^2\right)+2\gamma\beta^{-1}\eta\right).
\end{aligned}$$

We can then apply the following result of [BV05], that is, for any two Borel probability measures μ,ν on $\mathbb{R}^{2d}$ with finite second moments,

$$\mathcal{W}_2(\mu,\nu)\leq C_\nu\left[\sqrt{D(\mu\|\nu)}+\left(\frac{D(\mu\|\nu)}{2}\right)^{1/4}\right],$$

where

$$C_\nu=2\inf_{\lambda>0}\left(\frac{1}{\lambda}\left(\frac{3}{2}+\log\int_{\mathbb{R}^{2d}}e^{\lambda\|w\|^2}\nu(dw)\right)\right)^{1/2}.$$

From the exponential integrability of the measure $\nu_{\mathbf{z},k\eta}$ in Lemma 17, we have

$$C_{\nu_{\mathbf{z},k\eta}} \le 2\left(\frac{1}{\alpha_0}\left(\frac{3}{2}+\log\int_{\mathbb{R}^{2d}} e^{\alpha_0\|(x,v)\|^2}\nu_{\mathbf{z},k\eta}(dx,dv)\right)\right)^{1/2}$$
$$\le 2\left(\frac{1}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)k\eta\right)\right)\right)^{1/2}.$$

Hence

$$\mathcal{W}_2^2(\tilde{\mathbb{P}}_{k\eta},\nu_{\mathbf{z},k\eta}) \le \frac{4}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)k\eta\right)\right)$$
$$\cdot\left[\sqrt{D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})}+\left(\frac{D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})}{2}\right)^{1/4}\right]^2, \tag{C.32}$$

where

$$D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta}) \le (k\eta)^3\left[\left(\frac{M^2\beta\eta}{\gamma}+\frac{\beta\eta\gamma}{2}\right)\left(2\gamma^2\eta C_v^d+(4+2\delta)\eta\left(M^2C_x^d+B^2\right)+2\gamma\beta^{-1}\right)\right.$$
$$\left.+\frac{\beta\delta}{\gamma}\left(M^2C_x^d+B^2\right)\right].$$

Note that $\eta\le 1$ so that $2\gamma^2\eta C_v^d+(4+2\delta)\eta\left(M^2C_x^d+B^2\right)+2\gamma\beta^{-1}\le (C_2)^2$, where C_2 is defined in (A.10). Then, we have

$$D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta}) \le (k\eta)^3\left[\left(\frac{M^2\beta\eta}{\gamma}+\frac{\beta\eta\gamma}{2}\right)(C_2)^2+\frac{\beta\delta}{\gamma}\left(M^2C_x^d+B^2\right)\right].$$

By using $(x+y)^2\le 2(x^2+y^2)$, we get

$$\mathcal{W}_2^2(\tilde{\mathbb{P}}_{k\eta},\nu_{\mathbf{z},k\eta}) \le \frac{8}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)k\eta\right)\right)$$
$$\cdot\left[D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})+\sqrt{D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})}\right]. \tag{C.33}$$

Since $k\eta\ge e>1$, we get

$$\frac{8}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)k\eta\right)\right)$$
$$\le\frac{8}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)\right)+\log(k\eta)\right)$$
$$\le\frac{8}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)\right)+1\right)\log(k\eta), \tag{C.34}$$

and

$$
\begin{aligned}
&D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta}) + \sqrt{D(\tilde{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta})} \\
&\leq \left(\left(\frac{M^2\beta\eta}{\gamma} + \frac{\beta\eta\gamma}{2}\right)(C_2)^2 + \sqrt{\left(\frac{M^2\beta\eta}{\gamma} + \frac{\beta\eta\gamma}{2}\right)(C_2)^2}\right)(k\eta)^3\eta^{1/2} \\
&\qquad + \left(\left(M^2C_x^d + B^2\right)\frac{\beta}{\gamma} + \sqrt{\left(M^2C_x^d + B^2\right)\frac{\beta}{\gamma}}\right)(k\eta)^3\sqrt{\delta},
\end{aligned}
$$

which implies that

$$
\mathcal{W}_2^2(\tilde{\mathbb{P}}_{k\eta}, \nu_{\mathbf{z},k\eta}) \leq (C_0^2\sqrt{\delta} + C_1^2\sqrt{\eta})(k\eta)^3\log(k\eta),
$$

where C_0 and C_1 are defined in (A.8) and (A.9). The result then follows from the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for non-negative real numbers x and y.

Finally, let us provide a bound on $\mathcal{W}_2(\mu_{\mathbf{z},k}, \tilde{\mathbb{P}}_{k\eta})$. Note that by the definition of $\tilde{V}$, we have that $\left(X_0 + \int_0^{k\eta}\tilde{V}(\lfloor s/\eta\rfloor\eta)ds, \tilde{V}(k\eta)\right)$ has the same law as $\mu_{\mathbf{z},k}$, and we can compute that

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left\|\tilde{X}(k\eta) - X_0 - \int_0^{k\eta}\tilde{V}(\lfloor s/\eta\rfloor\eta)ds\right\|^2 \\
&= \mathbb{E}_{\mathbf{z}}\left\|\int_0^{k\eta}\tilde{V}(s) - \tilde{V}(\lfloor s/\eta\rfloor\eta)ds\right\|^2 \\
&\leq k\eta\int_0^{k\eta}\mathbb{E}_{\mathbf{z}}\left\|\tilde{V}(s) - \tilde{V}(\lfloor s/\eta\rfloor\eta)\right\|^2 ds \\
&\leq (k\eta)^2\eta\left(2\gamma^2\eta C_v^d + (4+2\delta)\eta\left(M^2C_x^d + B^2\right) + 2\gamma\beta^{-1}\right) \leq (k\eta)^2\eta(C_2)^2.
\end{aligned}
$$

where we used the assumption $\eta \leq 1$ so that $2\gamma^2\eta C_v^d + (4+2\delta)\eta\left(M^2C_x^d + B^2\right) + 2\gamma\beta^{-1} \leq (C_2)^2$ in the last inequality above, where C_2 is defined in (A.10). Therefore,

$$
\mathcal{W}_2(\mu_{\mathbf{z},k}, \tilde{\mathbb{P}}_{k\eta}) \leq C_2k\eta\sqrt{\eta}.
$$

The proof is complete.

C.4 Proof of Lemma 20

We recall first from (C.4) that

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

Since $\int_{\mathbb{R}^{2d}} e^{\alpha \mathcal{V}(x,v)} \mu_0(dx,dv) < \infty$ with $\alpha > 0$, we have $\|(x,v)\|_{L^2(\mu_0)} < \infty$.

Next, let us notice that by the concavity of the function h, we have (see [EGZ19])

$$h(r) \le \min\{r, h(R_1)\} \le \min\{r, R_1\}, \quad \text{for any } r \ge 0.$$

It follows that

$$\begin{aligned}\rho((x,v),(x',v')) &\le \min\{r((x,v),(x',v')), R_1\}(1 + \varepsilon_1 \mathcal{V}(x,v) + \varepsilon_1 \mathcal{V}(x',v')) \\ &\le R_1(1 + \varepsilon_1 \mathcal{V}(x,v) + \varepsilon_1 \mathcal{V}(x',v')).\end{aligned}$$

Moreover, by the definition of $\mathcal{V}$ in (2.1) and Lemma 25, we deduce that

$$\begin{aligned}\mathcal{V}(x,v) &\le \beta\left(\frac{M}{2}\|x\|^2 + B\|x\| + A_0\right) + \frac{1}{4}\beta\gamma^2(\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2) \\ &\le \beta\left(\frac{M}{2}\|x\|^2 + B\|x\| + A_0\right) + \frac{1}{4}\beta\gamma^2(2\|x\|^2 + 2\gamma^{-2}\|v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2) \\ &\le \beta\left(M\|x\|^2 + A_0 + \frac{B^2}{2M}\right) + \frac{1}{4}\beta\gamma^2(2\|x\|^2 + 2\gamma^{-2}\|v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2) \\ &\le \left(\beta M + \frac{1}{2}\beta\gamma^2\right)\|x\|^2 + \frac{3}{4}\beta\|v\|^2 + \beta A_0 + \frac{\beta B^2}{2M}.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}&\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}}) \\ &\le R_1 + R_1\varepsilon_1\left(\left(M + \frac{1}{2}\beta\gamma^2\right)\int_{\mathbb{R}^{2d}}\|x\|^2\mu_0(dx,dv) + \frac{3}{4}\beta\int_{\mathbb{R}^{2d}}\|v\|^2\mu_0(dx,dv) + \beta A_0 + \frac{\beta B^2}{2M}\right) \\ &\quad + R_1\varepsilon_1\left(\left(M + \frac{1}{2}\beta\gamma^2\right)\int_{\mathbb{R}^{2d}}\|x\|^2\pi_{\mathbf{z}}(dx,dv) + \frac{3}{4}\beta\int_{\mathbb{R}^{2d}}\|v\|^2\pi_{\mathbf{z}}(dx,dv) + \beta A_0 + \frac{\beta B^2}{2M}\right).\end{aligned} \tag{C.35}$$

It has been shown in [RRT17, Section 3.5] that

$$\int_{\mathbb{R}^{2d}}\|x\|^2\pi_{\mathbf{z}}(dx,dv) \le \frac{b + d/\beta}{m}.$$

In addition, from the explicit expression of $\pi_{\mathbf{z}}(dx,dv)$ in (1.7), we have

$$\int_{\mathbb{R}^{2d}}\|v\|^2\pi_{\mathbf{z}}(dx,dv) = (2\pi\beta^{-1})^{-d/2}\int_{\mathbb{R}^d}\|v\|^2 e^{-\|v\|^2/(2\beta^{-1})}dv = d/\beta.$$

Hence, the conclusion follows from (C.35).

D Proofs of Lemmas in Section B

D.1 Proof of Lemma 21

Before we proceed to the proof of Lemma 21, let us state two technical lemmas, which will be used in the proof of Lemma 21. Recall $\psi_0(t) = e^{-\gamma t}$ and $\psi_{k+1}(t) = \int_0^t \psi_k(s)ds$, and (ξ_{k+1}, ξ'_{k+1}) is a $2d$-dimensional centered Gaussian vector from the SGHMC2 iterates $(\hat{X}_k, \hat{V}_k)$ given in (1.10)–(1.11). Using the definitions, it is straightforward to establish these two lemmas, so we omit the details of their proofs.

Lemma 23. *For any $\eta \geq 0$,*

$$\max\{|\psi_0(\eta) - 1 + \gamma\eta|, |\eta - \psi_1(\eta)|, |\psi_2(\eta)|\} \leq c_0\eta^2, \tag{D.1}$$

where $c_0 := 1 + \gamma^2$.

Lemma 24. *For any $\eta \geq 0$,*

$$C_{11}(\eta) := \mathbb{E}\|\xi_k\|^2 \leq c_{11}\eta := d\eta, \tag{D.2}$$

$$C_{22}(\eta) := \mathbb{E}\|\xi'_k\|^2 \leq c_{22}\eta^3 := \frac{d}{3}\eta^3, \tag{D.3}$$

$$C_{12}(\eta) := \mathbb{E}\langle\xi_k, \xi'_k\rangle \leq c_{12}\eta^2 := \frac{d}{2}\eta^2. \tag{D.4}$$

Now, we are ready to prove Lemma 21, i.e. the uniform (in time) L^2 bounds for $(\hat{X}_k, \hat{V}_k)$ defined in (1.10)–(1.11). We can rewrite the dynamics of the SGHMC2 iterates as follows:

$$\hat{V}_{k+1} = (1 - \gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}) + \hat{E}_k + \sqrt{2\gamma\beta^{-1}}\xi_{k+1}, \tag{D.5}$$

$$\hat{X}_{k+1} = \hat{X}_k + \eta\hat{V}_k + \hat{E}'_k + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1}, \tag{D.6}$$

where

$$\hat{E}_k := (\psi_0(\eta) - 1 + \gamma\eta)\hat{V}_k + (\eta - \psi_1(\eta))g(\hat{X}_k, U_{\mathbf{z},k}), \tag{D.7}$$

$$\hat{E}'_k := (\psi_1(\eta) - \eta)\hat{V}_k - \psi_2(\eta)g(\hat{X}_k, U_{\mathbf{z},k}), \tag{D.8}$$

where $\mathbb{E}g(x, U_{\mathbf{z},k}) = \nabla F_{\mathbf{z}}(x)$ for any x. We again use the Lyapunov function $\mathcal{V}(x, v)$ defined as before, and set for each $k = 0, 1, \ldots,$

$$\hat{L}_2(k) = \mathbb{E}_{\mathbf{z}}\mathcal{V}(\hat{X}_k, \hat{V}_k)/\beta = \mathbb{E}_{\mathbf{z}}\left[F_{\mathbf{z}}(\hat{X}_k) + \frac{1}{4}\gamma^2\left(\|\hat{X}_k + \gamma^{-1}\hat{V}_k\|^2 + \|\gamma^{-1}\hat{V}_k\|^2 - \lambda\|\hat{X}_k\|^2\right)\right]. \tag{D.9}$$

We can compute that

$$\begin{aligned}\mathbb{E}F_{\mathbf{z}}(\hat{X}_{k+1}) &= \mathbb{E}F_{\mathbf{z}}\left(\hat{X}_k + \eta\hat{V}_k + \hat{E}'_k + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right)\\ &\leq \mathbb{E}F_{\mathbf{z}}\left(\hat{X}_k\right) + \mathbb{E}\left\langle \nabla F_{\mathbf{z}}(\hat{X}_k), \eta\hat{V}_k + \hat{E}'_k\right\rangle + \frac{M}{2}\mathbb{E}\left\|\eta\hat{V}_k + \hat{E}'_k + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right\|^2\\ &= \mathbb{E}F_{\mathbf{z}}\left(\hat{X}_k\right) + \mathbb{E}\left\langle \nabla F_{\mathbf{z}}(\hat{X}_k), \eta\hat{V}_k\right\rangle + \frac{M}{2}\eta^2\mathbb{E}\left\|\hat{V}_k\right\|^2 + \delta_1(k),\end{aligned}$$

where

$$\delta_1(k) := \mathbb{E}\left\langle \nabla F_{\mathbf{z}}(\hat{X}_k), \hat{E}'_k\right\rangle + \frac{M}{2}\mathbb{E}\left\|\hat{E}'_k + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right\|^2 + M\mathbb{E}\left\langle \eta\hat{V}_k, \hat{E}'_k + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right\rangle \tag{D.10}$$

$$= \mathbb{E}\left\langle \nabla F_{\mathbf{z}}(\hat{X}_k), \hat{E}'_k\right\rangle + \frac{M}{2}\mathbb{E}\left\|\hat{E}'_k\right\|^2 + M\gamma\beta^{-1}C_{22}(\eta) + M\mathbb{E}\left\langle \eta\hat{V}_k, \hat{E}'_k\right\rangle. \tag{D.11}$$

We can also compute that

$$\begin{aligned}&\frac{1}{4}\gamma^2\mathbb{E}\left\|\hat{X}_{k+1} + \gamma^{-1}\hat{V}_{k+1}\right\|^2\\ &= \frac{1}{4}\mathbb{E}\left\|\gamma\hat{X}_{k+1} + \hat{V}_{k+1}\right\|^2\\ &= \frac{1}{4}\mathbb{E}\left\|\left(\gamma\hat{X}_k + \hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}) + \sqrt{2\gamma\beta^{-1}}\xi_{k+1}\right) + \gamma\hat{E}'_k + \gamma\sqrt{2\gamma\beta^{-1}}\xi'_{k+1} + \hat{E}_k\right\|^2\\ &= \frac{1}{4}\mathbb{E}\left\|\gamma\hat{X}_k + \hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k})\right\|^2 + \delta_2(k),\end{aligned}$$

where

$$\begin{aligned}\delta_2(k) := \frac{1}{2}\gamma\beta^{-1}C_{11}(\eta) + \frac{1}{2}\gamma^3\beta^{-1}C_{22}(\eta) + \gamma^2\beta^{-1}C_{12}(\eta)\\ + \frac{1}{4}\mathbb{E}\left\|\gamma\hat{E}'_k + \hat{E}_k\right\|^2 + \frac{1}{2}\mathbb{E}\left\langle \gamma\hat{X}_k + \hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}), \gamma\hat{E}'_k + \hat{E}_k\right\rangle.\end{aligned}$$

We can also compute that

$$\begin{aligned}\frac{1}{4}\gamma^2\mathbb{E}\left\|\gamma^{-1}\hat{V}_{k+1}\right\|^2 &= \frac{1}{4}\mathbb{E}\left\|\hat{V}_{k+1}\right\|^2\\ &= \frac{1}{4}\mathbb{E}\left\|(1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}) + \hat{E}_k + \sqrt{2\gamma\beta^{-1}}\xi_{k+1}\right\|^2\\ &= \frac{1}{4}\mathbb{E}\left\|(1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}) + \hat{E}_k\right\|^2 + \frac{1}{2}\gamma\beta^{-1}C_{11}(\eta)\\ &= \frac{1}{4}\mathbb{E}\left\|(1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k})\right\|^2 + \delta_3(k),\end{aligned}$$

where

$$\delta_3(k) := \frac{1}{4}\mathbb{E}\left\|\hat{E}_k\right\|^2 + \frac{1}{2}\mathbb{E}\left\langle (1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}), \hat{E}_k\right\rangle + \frac{1}{2}\gamma\beta^{-1}C_{11}(\eta). \tag{D.12}$$

Finally, we can compute that

$$\begin{aligned}
-\frac{1}{4}\gamma^2\lambda\mathbb{E}\left\|\hat{X}_{k+1}\right\|^2 &= -\frac{1}{4}\gamma^2\lambda\mathbb{E}\left\|\hat{X}_k+\eta\hat{V}_k+\hat{E}'_k+\sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right\|^2 \\
&= -\frac{1}{4}\gamma^2\lambda\mathbb{E}\left\|\hat{X}_k+\eta\hat{V}_k\right\|^2-\frac{1}{4}\gamma^2\lambda\mathbb{E}\left\|\hat{E}'_k+\sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right\|^2 \\
&\qquad -\frac{1}{2}\gamma^2\lambda\mathbb{E}\left\langle\hat{X}_k+\eta\hat{V}_k,\hat{E}'_k+\sqrt{2\gamma\beta^{-1}}\xi'_{k+1}\right\rangle \\
&\le -\frac{1}{4}\gamma^2\lambda\mathbb{E}\left\|\hat{X}_k+\eta\hat{V}_k\right\|^2+\delta_4(k),
\end{aligned}$$

where

$$\delta_4(k) := -\frac{1}{2}\gamma^2\lambda\mathbb{E}\left\langle\hat{X}_k+\eta\hat{V}_k,\hat{E}'_k\right\rangle. \tag{D.13}$$

By following the proofs of the L_2 uniform bound for SGHMC1 iterates, we get

$$\frac{\hat{L}_2(k+1)-\hat{L}_2(k)}{\eta} \le \gamma(A/\beta-\lambda\hat{L}_2(k))+(K_1\hat{L}_2(k)+K_2)\cdot\eta+\frac{\delta_1(k)+\delta_2(k)+\delta_3(k)+\delta_4(k)}{\eta},$$

where K_1 and K_2 are given in (A.3) and (A.4).

Next, we can estimate that

$$\begin{aligned}
\delta_1(k) &= \mathbb{E}\left\langle\nabla F_{\mathbf{z}}(\hat{X}_k),\hat{E}'_k\right\rangle+\frac{M}{2}\mathbb{E}\left\|\hat{E}'_k\right\|^2+M\mathbb{E}\left\langle\eta\hat{V}_k,\hat{E}'_k\right\rangle+M\gamma\beta^{-1}C_{22}(\eta) \\
&\le c_0\eta^2\mathbb{E}\left[\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|\cdot\left(\|\hat{V}_k\|+\|g(\hat{X}_k,U_{\mathbf{z},k})\|\right)\right]+\frac{M}{2}c_0^2\eta^4\mathbb{E}\left(\|\hat{V}_k\|+\|g(\hat{X}_k,U_{\mathbf{z},k})\|\right)^2 \\
&\qquad +Mc_0\eta^2\mathbb{E}\left[\|\hat{V}_k\|\cdot\left(\|\hat{V}_k\|+\|g(\hat{X}_k,U_{\mathbf{z},k})\|\right)\right]+M\gamma\beta^{-1}c_{22}\eta^3 \\
&\le \frac{1}{2}c_0\eta^2\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2+Mc_0^2\eta^4\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})\|^2+M\gamma\beta^{-1}c_{22}\eta^3 \\
&\qquad +\frac{1}{2}Mc_0\eta^2(1+2\eta^2)\mathbb{E}\|\hat{V}_k\|^2+\frac{1}{2}(M+1)c_0\eta^2\mathbb{E}\left(\|\hat{V}_k\|+\|g(\hat{X}_k,U_{\mathbf{z},k})\|\right)^2 \\
&\le \frac{1}{2}c_0\eta^2\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2+Mc_0^2\eta^4\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})\|^2+M\gamma\beta^{-1}c_{22}\eta^3 \\
&\quad +\frac{1}{2}Mc_0\eta^2(1+2\eta^2)\mathbb{E}\|\hat{V}_k\|^2+(M+1)c_0\eta^2\mathbb{E}\|\hat{V}_k\|^2+(M+1)c_0\eta^2\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})\|^2 \\
&= \frac{1}{2}c_0\eta^2\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2+c_0\eta^2(Mc_0\eta^2+M+1)\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})\|^2+M\gamma\beta^{-1}c_{22}\eta^3 \\
&\quad +\frac{1}{2}c_0\eta^2(M(1+2\eta^2)+2M+2)\mathbb{E}\|\hat{V}_k\|^2,
\end{aligned}$$

and

$$
\begin{aligned}
\delta_2(k) &= \frac{1}{2}\gamma\beta^{-1}C_{11}(\eta) + \frac{1}{2}\gamma^3\beta^{-1}C_{22}(\eta) + \gamma^2\beta^{-1}C_{12}(\eta) \\
&\qquad + \frac{1}{4}\mathbb{E}\|\gamma\hat{E}'_k + \hat{E}_k\|^2 + \frac{1}{2}\mathbb{E}\left\langle \gamma\hat{X}_k + \hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}), \gamma\hat{E}'_k + \hat{E}_k \right\rangle \\
&\leq \frac{1}{2}\gamma\beta^{-1}c_{11}\eta + \frac{1}{2}\gamma^3\beta^{-1}c_{22}\eta^3 + \gamma^2\beta^{-1}c_{12}\eta^2 \\
&\qquad + \frac{1}{4}c_0^2\eta^4(1+\gamma)^2\mathbb{E}\left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)^2 \\
&\qquad\quad + \frac{1}{2}c_0\eta^2(1+\gamma)\mathbb{E}\left[\left\|\gamma\hat{X}_k + \hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k})\right\| \cdot \left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)\right] \\
&\leq \frac{1}{2}\gamma\beta^{-1}c_{11}\eta + \frac{1}{2}\gamma^3\beta^{-1}c_{22}\eta^3 + \gamma^2\beta^{-1}c_{12}\eta^2 \\
&\qquad + \frac{1}{4}c_0\eta^2(1+\gamma)(1+c_0\eta^2(1+\gamma))\mathbb{E}\left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)^2 \\
&\qquad\quad + \frac{1}{4}c_0\eta^2(1+\gamma)\mathbb{E}\left\|\gamma\hat{X}_k + \hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k})\right\|^2 \\
&\leq \frac{1}{2}\gamma\beta^{-1}c_{11}\eta + \frac{1}{2}\gamma^3\beta^{-1}c_{22}\eta^3 + \gamma^2\beta^{-1}c_{12}\eta^2 \\
&\qquad + \frac{1}{2}c_0\eta^2(1+\gamma)(1+c_0\eta^2(1+\gamma))\mathbb{E}\|\hat{V}_k\|^2 \\
&\qquad\quad + \frac{1}{2}c_0\eta^2(1+\gamma)(1+c_0\eta^2(1+\gamma))\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2 \\
&\qquad\quad + \frac{3}{4}c_0\eta^2(1+\gamma)\gamma^2\mathbb{E}\|\hat{X}_k\|^2 + \frac{3}{4}c_0\eta^2(1+\gamma)\mathbb{E}\|\hat{V}_k\|^2 \\
&\qquad + \frac{3}{4}c_0\eta^2(1+\gamma)\eta^2\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2 \\
&= \frac{1}{2}\gamma\beta^{-1}c_{11}\eta + \frac{1}{2}\gamma^3\beta^{-1}c_{22}\eta^3 + \gamma^2\beta^{-1}c_{12}\eta^2 \\
&\qquad + \frac{1}{2}c_0\eta^2(1+\gamma)\left(\frac{5}{2} + c_0\eta^2(1+\gamma)\right)\mathbb{E}\|\hat{V}_k\|^2 \\
&\qquad\quad + \frac{1}{2}c_0\eta^2(1+\gamma)\left(1 + c_0\eta^2(1+\gamma) + \frac{3}{2}\eta^4\right)\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2 \\
&\qquad\quad + \frac{3}{4}c_0\eta^2(1+\gamma)\gamma^2\mathbb{E}\|\hat{X}_k\|^2,
\end{aligned}
$$

and we can compute that

$$\begin{aligned}
\delta_3(k) &= \frac{1}{4}\mathbb{E}\|\hat{E}_k\|^2 + \frac{1}{2}\mathbb{E}\left\langle (1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k}), \hat{E}_k \right\rangle + \frac{1}{2}\gamma\beta^{-1}C_{11}(\eta) \\
&\leq \frac{1}{4}c_0^2\eta^4\mathbb{E}\left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)^2 \\
&\quad + \frac{1}{2}c_0\eta^2\mathbb{E}\left[\left\|(1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k})\right\| \cdot \left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)\right] + \frac{1}{2}\gamma\beta^{-1}c_{11}\eta \\
&\leq \frac{1}{4}c_0\eta^2(1+c_0\eta^2)\mathbb{E}\left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)^2 \\
&\quad + \frac{1}{4}c_0\eta^2\mathbb{E}\left\|(1-\gamma\eta)\hat{V}_k - \eta g(\hat{X}_k, U_{\mathbf{z},k})\right\|^2 + \frac{1}{2}\gamma\beta^{-1}c_{11}\eta \\
&\leq \frac{1}{2}c_0\eta^2(1+c_0\eta^2)\mathbb{E}\|\hat{V}_k\|^2 + \frac{1}{2}c_0\eta^2(1+c_0\eta^2)\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2 \\
&\quad + \frac{1}{2}c_0\eta^2(1-\gamma\eta)^2\mathbb{E}\|\hat{V}_k\|^2 + \frac{1}{2}c_0\eta^4\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2 + \frac{1}{2}\gamma\beta^{-1}c_{11}\eta \\
&= \frac{1}{2}c_0\eta^2(2-2\gamma\eta + (c_0+\gamma^2)\eta^2)\mathbb{E}\|\hat{V}_k\|^2 \\
&\qquad + \frac{1}{2}c_0\eta^2(1+(c_0+1)\eta^2)\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2 + \frac{1}{2}\gamma\beta^{-1}c_{11}\eta,
\end{aligned}$$

and finally we can compute that

$$\begin{aligned}
\delta_4(k) &= -\frac{1}{2}\gamma^2\lambda\mathbb{E}\left\langle \hat{X}_k + \eta\hat{V}_k, \hat{E}'_k \right\rangle \\
&\leq \frac{1}{2}\gamma^2\lambda c_0\eta^2\mathbb{E}\left[\left\|\hat{X}_k + \eta\hat{V}_k\right\| \cdot \left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)\right] \\
&\leq \frac{1}{4}\gamma^2\lambda c_0\eta^2\mathbb{E}\|\hat{X}_k + \eta\hat{V}_k\|^2 + \frac{1}{4}\gamma^2\lambda c_0\eta^2\mathbb{E}\left(\|\hat{V}_k\| + \|g(\hat{X}_k, U_{\mathbf{z},k})\|\right)^2 \\
&\leq \frac{1}{2}\gamma^2\lambda c_0\eta^2\mathbb{E}\|\hat{X}_k\|^2 + \frac{1}{2}\gamma^2\lambda c_0\eta^2(1+\eta^2)\mathbb{E}\|\hat{V}_k\|^2 + \frac{1}{2}\gamma^2\lambda c_0\eta^2\mathbb{E}\|g(\hat{X}_k, U_{\mathbf{z},k})\|^2.
\end{aligned}$$

Putting everything together, we have

$$
\begin{aligned}
&\frac{\hat{L}_2(k+1)-\hat{L}_2(k)}{\eta} \\
&\le \gamma(A/\beta-\lambda\hat{L}_2(k))+(K_1\hat{L}_2(k)+K_2)\cdot\eta+\frac{\delta_1(k)+\delta_2(k)+\delta_3(k)+\delta_4(k)}{\eta} \\
&\le \gamma((d+A)/\beta-\lambda\hat{L}_2(k))+(K_1\hat{L}_2(k)+K_2)\cdot\eta \\
&\quad+\frac{1}{2}c_0\eta\Bigg((M(1+2\eta^2)+2M+4-2\gamma\eta+(c_0+\gamma^2)\eta^2) \\
&\qquad+(1+\gamma)\left(\frac{5}{2}+c_0\eta^2(1+\gamma)\right)+\gamma^2\lambda(1+\eta^2)\Bigg)\mathbb{E}\|\hat{V}_k\|^2 \\
&\quad+\frac{1}{2}c_0\eta\Bigg((1+\gamma)\left(1+c_0\eta^2(1+\gamma)+\frac{3}{2}\eta^4\right)+1+(c_0+1)\eta^2 \\
&\qquad+\lambda\gamma^2+2(Mc_0\eta^2+M+1)\Bigg)\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})\|^2 \\
&\qquad+\frac{1}{2}c_0\eta\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2+\frac{1}{2}\gamma^2c_0\eta\left(\lambda+\frac{3}{2}(1+\gamma)\right)\mathbb{E}\|\hat{X}_k\|^2 \\
&\quad+\frac{1}{2}\gamma^3\beta^{-1}c_{22}\eta^2+\gamma^2\beta^{-1}c_{12}\eta+M\gamma\beta^{-1}c_{22}\eta^2,
\end{aligned}
$$

where we used the fact that $c_{11}=d$. Moreover,

$$
\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2\le\mathbb{E}(M\|\hat{X}_k\|+B)^2\le 2M^2\mathbb{E}\|\hat{X}_k\|^2+2B^2,
$$

and

$$
\begin{aligned}
\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})\|^2&=\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2+\mathbb{E}\|g(\hat{X}_k,U_{\mathbf{z},k})-\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2 \\
&\le\mathbb{E}\|\nabla F_{\mathbf{z}}(\hat{X}_k)\|^2+2\delta M^2\mathbb{E}\|\hat{X}_k\|^2+2\delta B^2 \\
&\le 2(1+\delta)M^2\mathbb{E}\|\hat{X}_k\|^2+2(1+\delta)B^2. \qquad \text{(D.14)}
\end{aligned}
$$

Therefore, we have

$$\begin{aligned}\frac{\hat{L}_2(k+1)-\hat{L}_2(k)}{\eta} &\le \gamma((d+A)/\beta-\lambda\hat{L}_2(k))+(K_1\hat{L}_2(k)+K_2)\cdot\eta\\ &\quad+\frac{1}{2}c_0\eta\Bigg((M(1+2\eta^2)+2M+4-2\gamma\eta+(c_0+\gamma^2)\eta^2)\\ &\qquad+(1+\gamma)\left(\frac{5}{2}+c_0\eta^2(1+\gamma)\right)+\gamma^2\lambda(1+\eta^2)\Bigg)\mathbb{E}\|\hat{V}_k\|^2\\ &\quad+\frac{1}{2}c_0\eta\Bigg[\Bigg((1+\gamma)\left(1+c_0\eta^2(1+\gamma)+\frac{3}{2}\eta^4\right)+1+(c_0+1)\eta^2\\ &\qquad+\lambda\gamma^2+2(Mc_0\eta^2+M+1)\Bigg)\left(2(1+\delta)M^2\right)\\ &\qquad+\left(2M^2+\gamma^2\lambda+\frac{3}{2}\gamma^2(1+\gamma)\right)\Bigg]\mathbb{E}\|\hat{X}_k\|^2\\ &\quad+c_0\eta\Bigg((1+\gamma)\left(1+c_0\eta^2(1+\gamma)+\frac{3}{2}\eta^4\right)+1+(c_0+1)\eta^2\\ &\qquad+\lambda\gamma^2+2(Mc_0\eta^2+M+1)\Bigg)(1+\delta)B^2+c_0B^2\eta\\ &\quad+\frac{1}{2}\gamma^3\beta^{-1}c_{22}\eta^2+\gamma^2\beta^{-1}c_{12}\eta+M\gamma\beta^{-1}c_{22}\eta^2,\end{aligned}$$

By applying the assumption $\eta \le 1$, we have

$$\begin{aligned}\frac{\hat{L}_2(k+1)-\hat{L}_2(k)}{\eta} &\le \gamma((d+A)/\beta-\lambda\hat{L}_2(k))+(K_1\hat{L}_2(k)+K_2)\cdot\eta\\ &\quad+\eta Q_1\mathbb{E}\|\hat{V}_k\|^2+\eta Q_2\mathbb{E}\|\hat{X}_k\|^2+\eta Q_3,\end{aligned}$$

where the constants Q_1, Q_2, Q_3 are given in (B.3)–(B.5). Let us recall that for $\lambda \le \frac{1}{4}$,

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

Thus, we have

$$\begin{aligned}\frac{\hat{L}_2(k+1)-\hat{L}_2(k)}{\eta} &\le \gamma((d+A)/\beta-\lambda\hat{L}_2(k))+(K_1\hat{L}_2(k)+K_2)\cdot\eta\\ &\quad+\eta\left(Q_1\frac{4}{1-2\lambda}+Q_2\frac{8}{(1-2\lambda)\gamma^2}\right)\hat{L}_2(k)+\eta Q_3,\end{aligned}$$

Therefore, for

$$0 < \eta \le \min\left\{\frac{\gamma}{\hat{K}_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2\hat{K}_1}\right\}, \tag{D.15}$$

where $\hat{K}_1 := K_1 + \frac{4Q_1}{1-2\lambda} + \frac{8Q_2}{(1-2\lambda)\gamma^2}$, and $\hat{K}_2 := K_2 + Q_3$, we get

$$(\hat{L}_2(k+1) - \hat{L}_2(k))/\eta \le 2\gamma(d/\beta + A/\beta) - \frac{1}{2}\gamma\lambda\hat{L}_2(k).$$

This implies $\hat{L}_2(k+1) \le \rho\hat{L}_2(k) + K$, where $\rho := 1 - \eta\gamma\lambda/2 \in [0,1)$, where we used the assumption $\eta \le \frac{2}{\gamma\lambda}$, and $K := 2\eta\gamma(d/\beta + A/\beta)$. It follows that

$$\hat{L}_2(k) \le \hat{L}_2(0) + \frac{K}{1-\rho} = \mathbb{E}_{\mathbf{z}}\left[\mathcal{V}(\hat{X}_0, \hat{V}_0)/\beta\right] + \frac{4(d/\beta + A/\beta)}{\lambda}.$$

The uniform L^2 bounds then readily follow.

D.2 Proof of Lemma 22

We follow similar steps as in the proof of Lemma 7 in [RRT17]. Recall that with the same initialization, the SGHMC2 iterates $(\hat{X}_k, \hat{V}_k)$ has the same distribution as $(\hat{X}(k\eta), \hat{V}(k\eta))$ where $(\hat{X}(\cdot), \hat{V}(\cdot))$ is a continuous-time process satisfying

$$d\hat{V}(t) = -\gamma\hat{V}(t)dt - g(\hat{X}(\lfloor t/\eta\rfloor\eta), U_{\mathbf{z}}(t))dt + \sqrt{2\gamma\beta^{-1}}dB(t), \tag{D.16}$$

$$d\hat{X}(t) = \hat{V}(t)dt, \tag{D.17}$$

Let $\mathbb{P}$ be the probability measure associated with the underdamped Langevin diffusion $(X(t), V(t))$ in (1.5)–(1.6) and $\hat{\mathbb{P}}$ be the probability measure associated with the $(\hat{X}(t), \hat{V}(t))$ process. Let $\mathcal{F}_t$ be the natural filtration up to time t. Then, the Radon-Nikodym derivative of $\mathbb{P}$ w.r.t. $\hat{\mathbb{P}}$ is given by the Girsanov theorem (see e.g. Section 7.6 in [LS13]):

$$\left.\frac{d\mathbb{P}}{d\hat{\mathbb{P}}}\right|_{\mathcal{F}_t} = e^{-\sqrt{\frac{\beta}{2\gamma}}\int_0^t (\nabla F_{\mathbf{z}}(\hat{X}(s)) - g(\hat{X}(\lfloor s/\eta\rfloor\eta), U_{\mathbf{z}}(s)))\cdot dB(s) - \frac{\beta}{4\gamma}\int_0^t \|\nabla F_{\mathbf{z}}(\hat{X}(s)) - g(\hat{X}(\lfloor s/\eta\rfloor\eta), U_{\mathbf{z}}(s))\|^2 ds}.$$

Then by writing $\mathbb{P}_t$ and $\hat{\mathbb{P}}_t$ as the probability measures $\mathbb{P}$ and $\hat{\mathbb{P}}$ conditional on the filtration $\mathcal{F}_t$,

$$\begin{aligned} D(\hat{\mathbb{P}}_t \| \mathbb{P}_t) &:= -\int d\hat{\mathbb{P}}_t \log \frac{d\mathbb{P}_t}{d\hat{\mathbb{P}}_t} \\ &= \frac{\beta}{4\gamma}\int_0^t \mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}(s)) - g(\hat{X}(\lfloor s/\eta\rfloor\eta), U_{\mathbf{z}}(s))\right\|^2 ds. \end{aligned}$$

Then, we get

$$D(\hat{\mathbb{P}}_{k\eta}\|\mathbb{P}_{k\eta}) = \frac{\beta}{4\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}(s)) - g(\hat{X}(\lfloor s/\eta\rfloor\eta), U_{\mathbf{z}}(s))\right\|^2 ds$$
$$\leq \frac{\beta}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}(s)) - \nabla F_{\mathbf{z}}(\hat{X}(\lfloor s/\eta\rfloor\eta))\right\|^2 ds$$
$$+ \frac{\beta}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}(\lfloor s/\eta\rfloor\eta)) - g(\hat{X}(\lfloor s/\eta\rfloor\eta), U_{\mathbf{z}}(s))\right\|^2 ds. \tag{D.18}$$

We first bound the first term in (D.18). Before we proceed, let us notice that for any $k\eta \leq s < (k+1)\eta$,

$$\hat{X}(s) = \hat{X}_k + \psi_1(s-k\eta)\hat{V}_k - \psi_2(s-k\eta)g(\hat{X}_k, U_{\mathbf{z},k}) + \sqrt{2\gamma\beta^{-1}}\xi'_{k+1,s-k\eta}, \tag{D.19}$$

in distribution, where $\xi'_{k+1,s-k\eta}$ is centered Gaussian independent of $\mathcal{F}_k$ and $\mathbb{E}\|\xi'_{k+1,s-k\eta}\|^2 = C_{22}(s-k\eta) \leq \frac{d}{3}(s-k\eta)^3 \leq \frac{d}{3}\eta^3$. Moreover, $\psi_1(s-k\eta) = \int_0^{s-k\eta} e^{-\gamma t}dt \leq (s-k\eta) \leq \eta$, and $\psi_2(s-k\eta) = \int_0^{s-k\eta}\psi_1(t)dt \leq \int_0^{s-k\eta} tdt \leq \eta^2$. Therefore, we can compute that

$$\frac{\beta}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}(s)) - \nabla F_{\mathbf{z}}(\hat{X}(\lfloor s/\eta\rfloor\eta))\right\|^2 ds$$
$$\leq \frac{\beta M^2}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\hat{X}(s) - \hat{X}(\lfloor s/\eta\rfloor\eta)\right\|^2 ds$$
$$= \frac{\beta M^2}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\psi_1(s-j\eta)\hat{V}_j - \psi_2(s-j\eta)g(\hat{X}_j, U_{\mathbf{z},j}) + \sqrt{2\gamma\beta^{-1}}\xi'_{j+1,s-j\eta}\right\|^2 ds$$
$$\leq \frac{3\beta M^2}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\left(\mathbb{E}_{\mathbf{z}}\left\|\psi_1(s-j\eta)\hat{V}_j\right\|^2 + \mathbb{E}_{\mathbf{z}}\left\|\psi_2(s-j\eta)g(\hat{X}_j, U_{\mathbf{z},j})\right\|^2\right.$$
$$\left. + \mathbb{E}_{\mathbf{z}}\left\|\sqrt{2\gamma\beta^{-1}}\xi'_{j+1,s-j\eta}\right\|^2\right)ds$$
$$\leq \frac{3\beta M^2}{2\gamma}(k\eta)\left(\eta^2\sup_{j\geq 0}\mathbb{E}_{\mathbf{z}}\|\hat{V}_j\|^2 + \eta^4\left(2(1+\delta)M^2\sup_{j\geq 0}\mathbb{E}\|\hat{X}_j\|^2 + 2(1+\delta)B^2\right) + \frac{d\eta^3}{3}2\gamma\beta^{-1}\right)$$
$$\leq \frac{3\beta M^2}{2\gamma}(k\eta)\eta^2\left(C_v^d + \left(2(1+\delta)M^2C_x^d + 2(1+\delta)B^2\right) + \frac{2d\gamma\beta^{-1}}{3}\right),$$

where we used (D.14), the assumption $\eta \leq 1$ and Lemma 21.

We can also bound the second term in (D.18):

$$\begin{aligned}
&\frac{\beta}{2\gamma}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}(\lfloor s/\eta\rfloor\eta))-g(\hat{X}(\lfloor s/\eta\rfloor\eta),U_{\mathbf{z}}(s))\right\|^2 ds\\
&=\frac{\beta}{2\gamma}\eta\sum_{j=0}^{k-1}\mathbb{E}_{\mathbf{z}}\left\|\nabla F_{\mathbf{z}}(\hat{X}_j)-g(\hat{X}_j,U_{\mathbf{z},j})\right\|^2\\
&\leq\frac{\beta}{2\gamma}\eta\delta\sum_{j=0}^{k-1}2(M^2\mathbb{E}_{\mathbf{z}}\|\hat{X}_j\|^2+B^2)\\
&\leq\left(M^2C_x^d+B^2\right)\frac{\beta}{\gamma}k\eta\delta,
\end{aligned}$$

where the first inequality follows from part (iv) of Assumption 1, and we also used Lemma 21. Hence, we conclude that

$$\begin{aligned}
D(\hat{\mu}_{\mathbf{z},k}\|\nu_{\mathbf{z},k\eta})\leq\frac{3\beta M^2}{2\gamma}(k\eta)\eta^2\left(C_v^d+\left(2(1+\delta)M^2C_x^d+2(1+\delta)B^2\right)+\frac{2d\gamma\beta^{-1}}{3}\right)\\
+\left(M^2C_x^d+B^2\right)\frac{\beta}{\gamma}k\eta\delta. \qquad \text{(D.20)}
\end{aligned}$$

To complete the proof, we can follow similar steps as in the proof of Lemma 18. By using the estimate in (D.20), the result from [BV05], and the exponential integrability of the measure $\nu_{\mathbf{z},k\eta}$ in Lemma 17, we can infer that

$$\begin{aligned}
&D(\hat{\mu}_{\mathbf{z},k}\|\nu_{\mathbf{z},k\eta})+\sqrt{D(\hat{\mu}_{\mathbf{z},k}\|\nu_{\mathbf{z},k\eta})}\\
&\leq\left(\frac{3\beta M^2}{2\gamma}\left(C_v^d+\left(2(1+\delta)M^2C_x^d+2(1+\delta)B^2\right)+\frac{2d\gamma\beta^{-1}}{3}\right)\right.\\
&\qquad\left.+\sqrt{\frac{3\beta M^2}{2\gamma}\left(C_v^d+(2(1+\delta)M^2C_x^d+2(1+\delta)B^2)+\frac{2d\gamma\beta^{-1}}{3}\right)}\right)k\eta^2\\
&\qquad+\left(\left(M^2C_x^d+B^2\right)\frac{\beta}{\gamma}+\sqrt{(M^2C_x^d+B^2)\frac{\beta}{\gamma}}\right)k\eta\sqrt{\delta},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{W}_2^2(\hat{\mu}_{\mathbf{z},k},\nu_{\mathbf{z},k\eta})\leq\frac{8}{\alpha_0}\left(\frac{3}{2}+\log\left(\int_{\mathbb{R}^{2d}}e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv)+\frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)k\eta\right)\right)\\
\cdot\left[D(\hat{\mu}_{\mathbf{z},k}\|\nu_{\mathbf{z},k\eta})+\sqrt{D(\hat{\mu}_{\mathbf{z},k}\|\nu_{\mathbf{z},k\eta})}\right],
\end{aligned}$$

which together implies that

$$\mathcal{W}_2^2(\hat{\mu}_{\mathbf{z},k}, \nu_{\mathbf{z},k\eta}) \leq (C_0^2\sqrt{\delta} + \hat{C}_1^2\eta)(k\eta)\log(k\eta),$$

where C_0 and $\hat{C}_1$ are defined in (A.8) and (B.8). The result then follows from the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for non-negative real numbers x and y.

E Supporting Lemmas

In this section, we present several supporting lemmas from the existing literature. These lemmas are used in our proofs, so we include them here for the sake of completeness. The first lemma shows that f admits lower and upper bounds that are quadratic functions.

Lemma 25 (See [RRT17, Lemma 2]). *If parts (i) and (ii) of Assumption 1 hold, then for all $x \in \mathbb{R}^d$ and z,*

$$\|\nabla f(x,z)\| \leq M\|x\| + B,$$

and

$$\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3 \leq f(x,z) \leq \frac{M}{2}\|x\|^2 + B\|x\| + A_0.$$

The next lemma shows a 2-Wasserstein continuity result for functions of quadratic growth. This lemma was also used in [RRT17] to study the SGLD dynamics.

Lemma 26 (See [PW16]). *Let μ, ν be two probability measures on $\mathbb{R}^{2d}$ with finite second moments, and let $G : \mathbb{R}^{2d} \to \mathbb{R}$ be a C^1 function obeying*

$$\|\nabla G(w)\| \leq c_1\|w\| + c_2,$$

for some constants $c_1 > 0$ and $c_2 \geq 0$. Then,

$$\left|\int_{\mathbb{R}^{2d}} G d\mu - \int_{\mathbb{R}^{2d}} G d\nu\right| \leq (c_1\sigma + c_2)\mathcal{W}_2(\mu,\nu),$$

where

$$\sigma^2 = \max\left\{\int_{\mathbb{R}^{2d}} \|w\|^2\mu(dw), \int_{\mathbb{R}^{2d}} \|w\|^2\nu(dw)\right\}.$$

The next lemma shows a uniform stability of $\pi_{\mathbf{z}}$. Note that the $x-$marginal of $\pi_{\mathbf{z}}(dx,dv)$ for the underdamped diffusion is the same as the stationary distribution for the overdamped diffusion studied in [RRT17]. For two $n-$tuples $\mathbf{z} = (z_1,\ldots,z_n), \overline{\mathbf{z}} = (\overline{z}_1,\ldots,\overline{z}_n) \in \mathcal{Z}^n$, we say $\mathbf{z}$ and $\overline{\mathbf{z}}$ differ only in a single coordinate if $\text{card}|\{i : z_i \neq \overline{z}_i\}| = 1$.

Lemma 27 (Proposition 12, [RRT17]). *For any two $\mathbf{z}, \overline{\mathbf{z}} \in \mathcal{Z}^n$ that differ only in a single coordinate,*

$$\sup_{z\in\mathcal{Z}} \left| \int_{\mathbb{R}^{2d}} f(x,z)\pi_{\mathbf{z}}(dx,dv) - \int_{\mathbb{R}^{2d}} f(x,z)\pi_{\overline{\mathbf{z}}}(dx,dv) \right| \leq \frac{4\beta c_{LS}}{n} \left(\frac{M^2}{m}(b+d/\beta) + B^2 \right),$$

where

$$c_{LS} \leq \frac{2m^2+8M^2}{m^2 M\beta} + \frac{1}{\lambda_*}\left(\frac{6M(d+\beta)}{m} + 2 \right),$$

where λ_* is the uniform spectral gap for overdamped Langevin dynamics:

$$\lambda_* = \inf_{\mathbf{z}\in\mathcal{Z}^n} \inf \left\{ \frac{\beta^{-1}\int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi_{\mathbf{z}}}{\int_{\mathbb{R}^d} g^2 d\pi_{\mathbf{z}}} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_{\mathbf{z}}), g \neq 0, \int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0 \right\}.$$

The next lemma show that for large values of β, the $x-$marginal of the stationary distribution $\pi_{\mathbf{z}}(dx,dv)$ is concentrated at the minimizer of $F_{\mathbf{z}}$. Note in Proposition 11 of [RRT17], they have the assumption $\beta \geq 2/m$, which seems to be only used to derive their Lemma 4, but not used in deriving their Proposition 11.

Lemma 28 (Proposition 11, [RRT17]). *It holds that*

$$\int_{\mathbb{R}^{2d}} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \leq \frac{d}{2\beta} \log\left(\frac{eM}{m}\left(\frac{b\beta}{d} + 1 \right) \right).$$

F Proof of Proposition 11

Let us first prove that $\lambda_* = \mathcal{O}(a^{-2})$. We first recall that λ_* is the uniform spectral gap for overdamped Langevin dynamics:

$$\lambda_* := \inf_{\mathbf{z}\in\mathcal{Z}^n} \inf \left\{ \frac{\beta^{-1}\int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi_{\mathbf{z}}}{\int_{\mathbb{R}^d} g^2 d\pi_{\mathbf{z}}} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_{\mathbf{z}}), g \neq 0, \int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0 \right\}.$$

In particular, fix any $\mathbf{z} \in \mathcal{Z}^n$ so that for every $g \in C^1(\mathbb{R}^d) \cap L^2(\pi_{\mathbf{z}})$, such that $g \neq 0$, and $\int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0$, we have

$$\lambda_* \leq \frac{\beta^{-1}\int_{\mathbb{R}^d} \|\nabla g\|^2 e^{-\beta F_{\mathbf{z}}(x)} dx}{\int_{\mathbb{R}^d} g^2 e^{-\beta F_{\mathbf{z}}(x)} dx}.$$

It follows from Lemma 25 that

$$\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3 \leq F_{\mathbf{z}}(x) \leq \frac{M^2}{2}\|x\|^2 + B\|x\| + A_0, \tag{F.1}$$

with $m = m_1 a^{-2}$, $M = M_1 a^{-2}$, and $B = B_1 a^{-1}$.

Next, let us take the test function $g_1(x) := \|x\|^2$. And we further define

$$c_1 := \int_{\mathbb{R}^d} g_1 d\pi_{\mathbf{z}} = \frac{\int_{\mathbb{R}^d} g_1(x) e^{-\beta F_{\mathbf{z}}(x)} dx}{\int_{\mathbb{R}^d} e^{-\beta F_{\mathbf{z}}(x)} dx}, \tag{F.2}$$

and we also define

$$g(x) := g_1(x) - c_1,$$

so that $g \in C^1(\mathbb{R}^d) \cap L^2(\pi_{\mathbf{z}})$, $g \neq 0$, and $\int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0$. Moreover, we have

$$\|\nabla g(x)\| = \|\nabla g_1(x)\| = 2\|x\|, \qquad \text{and} \qquad g_1(ax) = a^2 g_1(x) = a^2 \|x\|^2.$$

Next, by the definition of c_1 in (F.2) and the bounds in (F.1), we get

$$c_1 \geq \frac{\int_{\mathbb{R}^d} \|x\|^2 e^{-\beta(\frac{M^2}{2}\|x\|^2 + B\|x\| + A_0)} dx}{\int_{\mathbb{R}^d} e^{-\beta(\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx} = \frac{\int_{\mathbb{R}^d} \|ax\|^2 e^{-\beta(\frac{M^2}{2}\|ax\|^2 + B\|ax\| + A_0)} dx}{\int_{\mathbb{R}^d} e^{-\beta(\frac{m}{3}\|ax\|^2 - \frac{b}{2}\log 3)} dx} = a^2 \underline{c}_1,$$

$$c_1 \leq \frac{\int_{\mathbb{R}^d} \|x\|^2 e^{-\beta(\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx}{\int_{\mathbb{R}^d} e^{-\beta(\frac{M^2}{2}\|x\|^2 + B\|x\| + A_0)} dx} = \frac{\int_{\mathbb{R}^d} \|ax\|^2 e^{-\beta(\frac{m}{3}\|ax\|^2 - \frac{b}{2}\log 3)} dx}{\int_{\mathbb{R}^d} e^{-\beta(\frac{M^2}{2}\|ax\|^2 + B\|ax\| + A_0)} dx} = a^2 \bar{c}_1,$$

where

$$\underline{c}_1 := \frac{\int_{\mathbb{R}^d} \|x\|^2 e^{-\beta(\frac{M_1^2}{2}\|x\|^2 + B_1\|x\| + A_0)} dx}{\int_{\mathbb{R}^d} e^{-\beta(\frac{m_1}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx},$$

$$\bar{c}_1 := \frac{\int_{\mathbb{R}^d} \|x\|^2 e^{-\beta(\frac{m_1}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx}{\int_{\mathbb{R}^d} e^{-\beta(\frac{M_1^2}{2}\|x\|^2 + B_1\|x\| + A_0)} dx}.$$

Hence, we have

$$\begin{aligned}
\lambda_* &\leq \frac{\beta^{-1} \int_{\mathbb{R}^d} \|\nabla g(x)\|^2 e^{-\beta(\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx}{\int_{\mathbb{R}^d} g(x)^2 e^{-\beta(\frac{M^2}{2}\|x\|^2 + B\|x\| + A_0)} dx} \\
&= \frac{\beta^{-1} \int_{\mathbb{R}^d} 4\|x\|^2 e^{-\beta(\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx}{\int_{\mathbb{R}^d} (g_1(x) - c_1)^2 e^{-\beta(\frac{M^2}{2}\|x\|^2 + B\|x\| + A_0)} dx} \\
&= \frac{\beta^{-1} \int_{\mathbb{R}^d} 4\|ax\|^2 e^{-\beta(\frac{m}{3}\|ax\|^2 - \frac{b}{2}\log 3)} dx}{\int_{\mathbb{R}^d} (g_1(ax) - c_1)^2 e^{-\beta(\frac{M^2}{2}\|ax\|^2 + B\|ax\| + A_0)} dx} \\
&\leq \frac{\beta^{-1} \int_{\mathbb{R}^d} 4\|ax\|^2 e^{-\beta(\frac{m}{3}\|ax\|^2 - \frac{b}{2}\log 3)} dx}{\min_{a^2\underline{c}_1 \leq \tilde{c} \leq a^2 \bar{c}_1} \int_{\mathbb{R}^d} (a^2\|x\|^2 - \tilde{c})^2 e^{-\beta(\frac{M^2}{2}\|ax\|^2 + B\|ax\| + A_0)} dx} \\
&= a^{-2} \frac{\beta^{-1} \int_{\mathbb{R}^d} 4\|x\|^2 e^{-\beta(\frac{m_1}{3}\|x\|^2 - \frac{b}{2}\log 3)} dx}{\min_{\underline{c}_1 \leq c \leq \bar{c}_1} \int_{\mathbb{R}^d} (\|x\|^2 - c)^2 e^{-\beta(\frac{M_1^2}{2}\|x\|^2 + B_1\|x\| + A_0)} dx},
\end{aligned}$$

where we used $m = m_1 a^{-2}$, $M = M_1 a^{-2}$, $B = B_1 a^{-1}$ and $g_1(ax) = a^2 g_1(x) = a^2\|x\|^2$. Hence, we conclude that $\lambda_* = \mathcal{O}(a^{-2})$.

Next, let us prove that $\mu_* = \Theta(a^{-1})$. We recall that μ_* the convergence rate for underdamped Langevin dynamics is given by:

$$\mu_* = \frac{\gamma}{768} \min\left\{\lambda M\gamma^{-2}, \Lambda^{1/2} e^{-\Lambda} M\gamma^{-2}, \Lambda^{1/2} e^{-\Lambda}\right\},$$

where

$$\Lambda = \frac{12}{5}(1 + 2\alpha_1 + 2\alpha_1^2)(d + A)M\gamma^{-2}\lambda^{-1}(1 - 2\lambda)^{-1}, \qquad \alpha_1 = (1 + \Lambda^{-1})M\gamma^{-2},$$

where λ, A come from the drift condition (2.2), and from [GGZ20], we can take

$$\lambda = \frac{1}{2}\min\left(\frac{1}{4}, \frac{m}{M + \gamma^2/2}\right), \qquad A = \frac{\beta}{2}\frac{m}{M + \frac{1}{2}\gamma^2}\left(\frac{B^2}{2M + \gamma^2} + \frac{b}{m}\left(M + \frac{1}{2}\gamma^2\right) + A_0\right). \tag{F.3}$$

Note that μ_* depends on the objective function $F_{\mathbf{z}}$ only via the parameters from its properties, which is independent of $\mathbf{z}$. Recall that $m = m_1 a^{-2}$, $M = M_1 a^{-2}$, $B = B_1 a^{-1}$. We define $\gamma =: \gamma_1 a^{-1}$ so that γ_1 is independent of a and

$$\mu_* = a^{-1}\frac{\gamma_1}{768} \min\left\{\lambda M_1 \gamma_1^{-2}, \Lambda^{1/2} e^{-\Lambda} M_1 \gamma_1^{-2}, \Lambda^{1/2} e^{-\Lambda}\right\}, \tag{F.4}$$

where we can check that λ, Λ are independent of a. Then, we can see from (F.4) that μ_* is linear in a^{-1} so that we have $\mu_* = \Theta(a^{-1})$. The proof is complete.

G Explicit dependence of constants on key parameters

In this section we provide explicit dependence of constants on parameters $\beta, d, \mu_*, \lambda_*$ and n, which is used in Section 5. To simplify the presentation, we use the notation $\tilde{\mathcal{O}}$, $\tilde{\Theta}$ to hide factors that depend on other parameters.

We recall the constants from Table 1. It is easy to see that

$$A = \tilde{\Theta}(\beta), \qquad \alpha_1 = \tilde{\Theta}(1), \qquad \Lambda = \tilde{\Theta}(d + \beta),$$

where we take A as in (F.3) and

$$\mu_* = \tilde{\Theta}\left(\sqrt{d + \beta} e^{-\Lambda}\right) = \tilde{\Theta}\left(\sqrt{d + \beta} e^{-\tilde{\Theta}(d+\beta)}\right). \tag{G.1}$$

Since $\varepsilon_1 = \tilde{\mathcal{O}}(\mu_*/(d + \beta))$, and μ_* is exponentially small in $\beta + d$, we get that

$$\overline{\mathcal{H}}_\rho(\mu_0) = \tilde{\mathcal{O}}(R_1) = (1 + d/\beta)^{1/2}.$$

In addition, in view of (G.1), it follows that

$$C = \tilde{\mathcal{O}}\left(e^{\Lambda/2}(d+\beta)^{1/2}\beta^{-1/2}\mu_*^{-1/2}\right) = \tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/4}\beta^{-1/2}}{\mu_*}\right).$$

The structure of the initial distribution $\mu_0(dx, dv)$ would affect the overall dependence on β, d. Since we assumed in Section 5 that $\mu_0(dx, dv)$ is supported on a Euclidean ball with radius being a universal constant, then the Lyapunov function $\mathcal{V}$ in (2.1) is linear in β. We can then obtain

$$\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv) = \tilde{\mathcal{O}}(\beta), \qquad \int_{\mathbb{R}^{2d}} e^{\alpha\mathcal{V}(x,v)}\mu_0(dx,dv) = e^{\tilde{\mathcal{O}}(\beta)},$$

It follows that

$$C_x^d = \tilde{\mathcal{O}}\left((\beta+d)/\beta\right), \qquad C_v^d = \tilde{\mathcal{O}}\left((\beta+d)/\beta\right), \qquad \sigma = \tilde{\mathcal{O}}\left(\sqrt{(\beta+d)/\beta}\right).$$

Next, we have $\alpha_0 = \tilde{\mathcal{O}}(\beta)$ and $\alpha = \tilde{\mathcal{O}}(1)$, and

$$\hat{\gamma} = \tilde{\mathcal{O}}(\sqrt{(\beta+d)/\beta}),$$
$$C_0 = \tilde{\mathcal{O}}\left((d+\beta)/\sqrt{\beta}\right), \qquad C_1 = \tilde{\mathcal{O}}\left((d+\beta)/\sqrt{\beta}\right), \qquad C_2 = \tilde{\mathcal{O}}\left(\sqrt{(d+\beta)/\beta}\right).$$

Moreover, by the definition of $\hat{C}_1$ in (B.8), we get

$$\hat{C}_1 = \tilde{\mathcal{O}}\left((d+\beta)/\sqrt{\beta}\right).$$

Hence, from (3.6), we obtain

$$\overline{\mathcal{J}}_0(\varepsilon) = \tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon\right),$$

and from (3.2), we get

$$\mathcal{J}_1(\varepsilon) = \tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta(\mu_*)^{3/2}}\left((\log(1/\varepsilon))^{3/2}\delta^{1/4}+\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))} + \frac{d+\beta}{\beta}\frac{\varepsilon^2}{\mu_*(\log(1/\varepsilon))^2}\right).$$

Moreover, from (4.1), we get

$$\hat{\mathcal{J}}_1(\varepsilon) = \tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta\sqrt{\mu_*}}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}\right).$$

Finally, from (3.5) and (3.7), we have

$$\mathcal{J}_2 = \tilde{\mathcal{O}}\left(\frac{d}{\beta}\log(\beta+1)\right), \qquad \text{and} \qquad \mathcal{J}_3(n) = \tilde{\mathcal{O}}\left(\frac{1}{n}\frac{(\beta+d)^2}{\lambda_*}\right).$$

H Proof of Lemma 13 and Lemma 15

H.1 Proof of Lemma 13

Since the distribution of A_{in} has compact support, we have $\|a_i\| \le D$ for some $D > 0$. Let $s_i := \langle a_i, x\rangle$. By taking the gradient of $f(x, z_i)$ with respect to x, we obtain

$$\nabla f(x, z_i) = -2\left(y_i - \sigma(s_i)\right)\sigma'(s_i)a_i + \lambda_r x. \tag{H.1}$$

This implies

$$\begin{aligned}
\langle \nabla f(x, z_i), x\rangle &= -2\left(y_i - \sigma(s_i)\right)\sigma'(s_i)s_i + \lambda_r\|x\|^2 && \text{(H.2)}\\
&\ge \lambda_r\|x\|^2 - 2(1 + \|\sigma\|_\infty)\|\sigma'\|_\infty|s_i| && \text{(H.3)}\\
&\ge \lambda_r\|x\|^2 - 2(1 + \|\sigma\|_\infty)\|\sigma'\|_\infty D\|x\|\,, && \text{(H.4)}
\end{aligned}$$

where we used the triangle inequality and the Cauchy-Schwartz inequality. Then, it is straightforward to check that we obtain $\langle \nabla f(x, z_i), x\rangle \ge m\|x\|^2 - b$ for

$$m = \lambda_r/2, \quad b = 8(1 + \|\sigma\|_\infty)^2\|\sigma'\|_\infty^2 D^2/\lambda_r\,,$$

and therefore part (iii) of Assumption 1 holds. Also for any $z = (a, y)$, $|f(0, z)| = |(y - \sigma(0))^2| \le A_0$ for $A_0 = (1 + \|\sigma(0)\|)^2$. Similarly, $\|\nabla f(0, z)\| = \| - 2(y - \sigma(0))\sigma'(0)a\| \le B_1$ for

$$B_1 := 2(1 + |\sigma(0)|)|\sigma'(0)|D.$$

Therefore, part (i) of Assumption 1 holds for any $B \ge B_1$. We also have the Hessian matrix

$$\nabla^2 f(x, z_i) = 2[\sigma'(s_i)]^2 a_i a_i^T - 2(y_i - \sigma(s_i))\sigma''(s_i)a_i a_i^T + \lambda_r I_d\,,$$

where I_d is the $d \times d$ identity matrix. Hence, $\|\nabla^2 f(x, z_i)\| \le M_1$ where

$$M_1 = 2\|\sigma'\|_\infty^2 D^2 + 2\left(1 + \|\sigma\|_\infty\right)\|\sigma''\|_\infty D^2 + \lambda_r.$$

Therefore, part (ii) of Assumption 1 also holds for $M = M_1$. In particular, $\nabla f(x, z_j)$ is also i.i.d. and we have $\mathbb{E}[\nabla f(x, z_j)] = \nabla F_{\mathbf{z}}(x)$ for any $x \in \mathbb{R}^d$. Furthermore, it follows from (H.1) that

$$\|\nabla f(x, z_j)\| \le B_2 + \lambda_r\|x\|, \quad \text{where} \quad B_2 := 2\left(1 + \|\sigma\|_\infty\right)\|\sigma'\|_\infty D,$$

for any z_j. Therefore, if we let $u_j := \nabla f(x, z_j) - \nabla F_{\mathbf{z}}(x)$, then u_j are i.i.d. with $\mathbb{E}[u_j] = 0$ and

$$\begin{aligned}
\mathbb{E}\|u_j\|^2 &\le 2\mathbb{E}\left[\|\nabla f(x, z_j)\|^2\right] + 2\mathbb{E}\left[\|\nabla F_{\mathbf{z}}(x)\|^2\right]\\
&\le 4(B_2 + \lambda_r\|x\|)^2\\
&\le 8B_2^2 + 8\lambda_r^2\|x\|^2\,, && \text{(H.5)}
\end{aligned}$$

where we used Cauchy-Schwarz inequality. This implies

$$\mathbb{E}\left[\|g(x,U_{\mathbf{z}})-\nabla F_{\mathbf{z}}(x)\|^2\right]=\mathbb{E}\left\|\frac{1}{n_b}\sum_{j=1}^{n_b}u_j\right\|^2=\frac{1}{n_b^2}\sum_{j=1}^{n_b}\mathbb{E}\|u_j\|^2\leq 2\delta(M^2\|x\|^2+B^2) \quad \text{(H.6)}$$

for any $\delta\in[\frac{1}{4n_b},1)$, $M\geq M_2:=4\lambda_r$, $B\geq B_2:=4M_2$ where we used (H.5) and the fact that u_j are i.i.d. with mean zero. If we choose, for instance, $M=M_1+M_2$, $B=\max(B_1,B_2)=B_2$; we observe that part (i) and (iv) of Assumptions 1 hold.

H.2 Proof of Lemma 15

We set $r_i=y_i-\langle a_i,x\rangle$ and follow a similar approach to the proof of Lemma 13.

We can compute that

$$\nabla f(x,z_i)=-\rho'(r_i)a_i+\lambda_r x.$$

This leads to

$$\langle\nabla f(x,z_i),x\rangle=-\rho'(r_i)\langle a_i,x\rangle+\lambda_r\|x\|^2\geq\lambda_r\|x\|^2-\|\rho'\|D\|x\|\geq m\|x\|^2-b\,, \quad \text{(H.7)}$$

with

$$m=\lambda_r/2,\quad b=\frac{2\|\rho'\|_\infty^2D^2}{\lambda_r},$$

Therefore, part (iii) of Assumption 1 holds. We have also

$$|f(0,z_i)|\leq|\rho(y_i)|,\quad \nabla f(0,z_i)=-\|\rho'(y_i)a_i\|\leq\|\rho'\|_\infty D$$

for any z_i. Therefore, part (i) of Assumption 1 holds with $A_0=\|\rho\|_\infty$ and $B=\|\rho'\|_\infty D$. Since

$$\nabla^2 f(0,z_i)=\rho''(r_i)a_ia_i^T+\lambda_r I_d\,,$$

where I_d is the $d\times d$ identity matrix, we also have

$$\left\|\nabla^2 f(0,z_i)\right\|\leq\|\rho''\|_\infty D^2+\lambda_r.$$

Therefore, part (ii) of Assumption 1 holds for any $M\geq\|\rho''\|_\infty D^2+\lambda_r$. We have also

$$\|\nabla f(x,z_i)\|\leq\|\rho'\|_\infty D+\lambda_r\|x\|.$$

Therefore, if we let $v_j=\nabla f(x,z_j)-\nabla F_{\mathbf{z}}(x)$, then v_j are i.i.d. with $\mathbb{E}[v_j]=0$ and

$$\begin{aligned}\mathbb{E}\|v_j\|^2&\leq 2\mathbb{E}\left[\|\nabla f(x,z_j)\|^2\right]+2\mathbb{E}\left[\|\nabla F_{\mathbf{z}}(x)\|^2\right]\\&\leq 4\left(\|\rho'\|_\infty D+\lambda_r\|x\|\right)^2\\&\leq 8\left(\|\rho'\|_\infty^2D^2+8\lambda_r^2\|x\|^2\right),\end{aligned} \quad \text{(H.8)}$$

where we used Cauchy-Schwarz inequality. This implies

$$\mathbb{E}\left[\|g(x,U_{\mathbf{z}})-\nabla F_{\mathbf{z}}(x)\|^2\right]=\mathbb{E}\left\|\frac{1}{n_b}\sum_{j=1}^{n_b}v_j\right\|^2=\frac{1}{n_b^2}\sum_{j=1}^{n_b}\mathbb{E}\|v_j\|^2\leq 2\delta(M^2\|x\|^2+B^2) \quad \text{(H.9)}$$

for any $\delta\in[\frac{1}{4n_b},1)$ and $M_2\geq 4\lambda_r$ and $B\geq 4\|\rho'\|_\infty D$ where we used (H.8) and the fact that v_j are i.i.d. with mean zero. We conclude that Assumption 1 work for $M=\|\rho''\|_\infty D^2+5\lambda_r$ and $B=4\|\rho'\|_\infty D$.