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‡

September 14, 2018

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 provide finite-time performance bounds for the global convergence of SGHMC 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 $\epsilon$, on a class of non-convex problems, we obtain iteration complexity bounds for SGHMC that can be tighter than those for SGLD up to a square root factor. These results show that acceleration with momentum is possible in the context of non-convex optimization algorithms.

*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

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

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 samples \(Z = (Z_1, Z_2, \ldots, Z_n)\) 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. approximately minimize

\[
\mathbb{E}F(\hat{x}) - F^*,
\]

where the expectation is taken with respect to both \(Z\) and the randomness encountered (if any) during the iterations of the algorithm to compute \(\hat{x}\) and \(F^* = \min_x F(x)\) is the minimum value. 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, we consider the empirical risk minimization problem

\[
\min_{x \in \mathbb{R}^d} F^\sharp(x) := \frac{1}{n} \sum_{i=1}^{n} f(x, z_i),
\]

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

\[
\mathbb{E}F^\sharp(z) - \min_{x \in \mathbb{R}^d} F^\sharp(x)
\]

approximately, 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 its 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 \([\text{GLM17}, \text{DLT}^{+17}]\), however in general, methods with global convergence guarantees are also desirable and preferable in many settings \([\text{HLSS16}, \text{SYN}^{+18}]\).

\(^1\)We note that in our notation \(Z\) is a random vector, whereas \(z\) is deterministic vector associated to a dataset that corresponds to a realization of the random vector \(Z\).
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 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. [CCS17, Dal17, RRT17, CCG16, SBCR16, SYN+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, $\beta$ is the inverse temperature, $g_k$ is a conditionally unbiased estimate of the gradient of $F$ and $\xi_k \in \mathbb{R}^d$ is a 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)

$$dX(t) = -\nabla F(x(t)) dt + \sqrt{2\beta^{-1}} dB(t), \quad t \geq 0,$$  \hspace{1cm} (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 $\pi_F(dx) \propto \exp(-\beta F(x))$, also known as the Gibbs measure, under some assumptions on $F$ (see e.g. [CHS87, HKS89]). For $\beta$ chosen properly (large enough), it is easy to see that this distribution will concentrate around approximate global minimizers of $F$. Recently, Dalalyan 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 [Dal17]. In a seminal work, Raginsky et al. [RRT17] showed that SGLD iterates track the underdamped Langevin SDE closely and obtained finite-time performance bounds for SGLD. Their results show that SGLD converges to $\varepsilon$-approximate global minimizers after $O(\text{poly}(\frac{1}{\lambda_*}, \beta, d, \frac{1}{\varepsilon}))$ iterations where $\lambda_*$ is the spectral gap of the overdamped Langevin dynamics which is in general exponential in both $\beta$ and the dimension $d$ [RRT17, TLR18]. A related result of Zhang et al. [ZLC17] shows that a modified
version of the SGLD algorithm will find an $\varepsilon$-approximate local minimum after polynomial time (with respect to all parameters). Recently, Xu et al. [XCZG17] improved the $\varepsilon$ 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 (second-order) Langevin diffusion

$$dV(t) = -\gamma V(t)dt - \nabla F_z(X(t))dt + \sqrt{2\gamma\beta^{-1}} dB(t),$$  \hspace{1cm} (1.5)
$$dX(t) = V(t)dt,$$  \hspace{1cm} (1.6)

where $X(t), V(t) \in \mathbb{R}^d$ models the position and the momentum of a particle moving in a field of force (described by the gradient of $F_z$) plus a random (thermal) force described by Gaussian noise, first derived by Kramers [Kra40]. It is known that under some assumptions on $F_z$, the Markov process $(X, V)$ is ergodic and have a unique stationary distribution

$$\pi_z(dx, dv) = \frac{1}{\Gamma_z} \exp(-\beta \left( \frac{1}{2} \|v\|^2 + F_z(x) \right)) \, dxdv,$$  \hspace{1cm} (1.7)

(see e.g. [HN04a, Pav14]) where $\Gamma_z$ is the normalizing constant:

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

Hence, the $x$-marginal distribution of stationary distribution $\pi_z(dx, dv)$ is exactly the invariant distribution of the overdamped Langevin dynamics.\footnote{With slight abuse of notation, we use $\pi_z(dx)$ to denote the $x$-marginal of the equilibrium distribution $\pi_z(dx, dv)$.} 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 most common scheme among the practitioners [TTV16, CCG+16, CDC15]:

$$V_{k+1} = V_k - \eta[\gamma V_k + g(X_k, U_{z,k})] + \sqrt{2\gamma\beta^{-1}} \eta \xi_k,$$ \hspace{1cm} (1.8)
$$X_{k+1} = X_k + \eta V_k,$$ \hspace{1cm} (1.9)

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

$$\mathbb{E}g(x, U_{z,k}) = \nabla F_z(x) \quad \text{for any } x \in \mathbb{R}^d.$$
We also focus on the unadjusted dynamics (without Metropolis type of correction) that works well in many applications [CFG14, CDC15], as our primary goal is not to sample from a particular target distribution but rather to compute approximate global minimizers to the problem (1.1).

We note that if the term with $dB(t)$ involving the Gaussian noise is removed in the underdamped SDE (1.5)-(1.6), this results in a second-order ODE in $X(t)$. It is interesting to note that Polyak’s heavy ball method that accelerates gradient descent is based on the discretization of this ODE [Pol87]. Similarly, if the Gaussian noise is scaled with the step size $\eta$ (instead of $\sqrt{\eta}$), then one recovers the stochastic gradient method with momentum [CFG14]. There has been a number of works for understanding momentum-based acceleration in first-order convex optimization methods as discretizations of second-order differential equations [SRBd17, SBC14, AP16, FRV18, ZMSJ18, KBB15, WWJ16]. Recent results of [EGZ17] shows that underdamped SDE converges to its stationary distribution faster than the overdamped SDE in the 2-Wasserstein metric under some assumptions where $F_z$ can be non-convex. Similar acceleration behavior between underdamped and overdamped dynamics was also proven for a version of Hamiltonian Monte Carlo algorithm for sampling strongly log-concave densities [CCBJ17, MS17] as well as densities whose negative logarithm is strongly convex outside a ball of finite radius [CCA+18]. 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 [EGZ17, CDC15, CFG14]. Although asymptotic convergence guarantees for SGHMC exist [CFG14] [MSH02, Section 3], [LMS15]; there is a lack of finite-time explicit performance bounds for solving stochastic non-convex optimization problems with SGHMC in the literature including risk minimization problems. Our main contribution is to give finite-time guarantees to find approximate minimizers of both empirical and population risks with explicit constants, bridging a gap between the theory and the practice for the use of SGHMC algorithms in stochastic non-convex optimization as elaborated further in the next section.

1.1 Contributions

Our contributions can be summarized as follows:

- Under some regularity assumptions for the non-convex function $f$, we can show that SGHMC converges to an $\varepsilon-$approximate global minimizer of the empirical risk minimization problem (1.2) after $\text{poly}\left(\frac{1}{\mu^*}, \beta, d, \frac{1}{\varepsilon}\right)$ iterations in expectation where $\mu^*$ is a parameter of the underdamped SDE governing the speed of convergence of it to its stationary distribution with respect to the 2-Wasserstein distance. We make the constants and polynomial dependency of the parameters to our final iteration complexity estimate explicit in our analysis. We emphasize that we do not assume that $f$ is convex or strongly convex in any region. To our knowledge, this is the
first non-asymptotic provable guarantees for the SGHMC algorithm in the context of non-convex stochastic optimization with explicit constants. This is in contrast to \( \text{poly}(\frac{1}{\lambda_*}, \beta, d, \frac{1}{\varepsilon}) \) iterations needed for the SGLD algorithm [RRT17] where \( \lambda_* \) is a spectral gap parameter of the overdamped diffusion. For a class of non-convex objectives, there are tight characterizations available for the parameters \( \mu_* \) and \( \lambda_* \). We discuss some examples of such non-convex problem classes in Examples 6 and 7 when the convergence rate \( \mu_* \) of underdamped dynamics is faster than that of SGLD by a square root factor, i.e. \( \mu_* = \mathcal{O}(\sqrt{\lambda_*}) \) where \( \lambda_* \) is typically an exponentially small constant in \( \beta \) (see e.g. [BGK05]). For these problems, under some assumptions, we show that SGHMC can achieve an optimization error that is better than that of SGLD by a square root factor and this is achieved in less number of iterations compared to SGLD. Therefore, our results highlight that momentum-based acceleration is achievable at least for some classes of non-convex problems we precise. In particular, our analysis gives some theoretical justification into the success of momentum-based methods for solving non-convex machine learning problems, empirically observed in practice [SMDH13]. Our further analysis on some particular examples suggest that when the initialization is far away from a global minimizer at a distance \( R \), SGHMC needs \( \mathcal{O}(R) \) iterations to reach out to a neighborhood of the global minimizer whereas SGLD requires \( \Omega(R^2) \) iterations for the same task.

- On the technical side, in order to establish these results, we adapt the proof techniques of Raginsky et al. [RRT17] developed for the overdamped dynamics to the underdamped dynamics and combine it with the analysis of Eberle et al. [EGZ17] 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 SGHMC algorithm requires studying a different Lyapunov function \( V \) (that also depends on the objective \( f \)) as opposed to the classic Lyapunov function \( H(x) = ||x||^2 \) arising in the study of the SGLD algorithm [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 SGHMC 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 (see Lemma 8) exploiting the structure of the Lyapunov function \( V \) (which will be defined later in (2.1)). Second, we obtain a bound for the Kullback-Leibler divergence between the discrete and continuous underdamped dynamics making use of the Girsanov’s theorem, which is then converted to bounds in the 2-Wasserstein metric by an application of an optimal
transportation inequality of Bolley and Villani [BV05]. This step requires proving a certain exponential integrability property of the underdamped Langevin diffusion. We show in Lemma 9 that the exponential moments grow at most linearly in time, which is a strict improvement from the exponential growth in time in [RRT17]. The method that is used in the proof of Lemma 9 can indeed be adapted to improve the exponential integrability and hence the overall estimates in [RRT17] for overdamped dynamics as well.

– Second, we study the continuous-time underdamped Langevin SDE. We build on the seminal work of Eberle et al. [EGZ17] which showed that the underdamped SDE is geometrically ergodic with an explicit rate $\mu_*$ 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 explicit (see Lemma 12).

– As the $x$-marginal of the equilibrium distribution $\pi_z(dx, dv)$ of the underdamped Langevin SDE concentrates around the global minimizers of $F_z$ for $\beta$ appropriately chosen, and we can control the error between the discrete-time SGHMC dynamics and the underdamped SDE by choosing the step size accordingly; this leads to performance bounds for the empirical risk minimization provided in Corollary 3.

• If every sample is used once (in other words if we sample directly from the population distribution $D$), then the bounds we obtain for the empirical risk minimization will also lead to similar bounds for the population risk.\textsuperscript{3} In this case, for a fixed target accuracy $\varepsilon$, we can show that SGHMC generalizes better than SGLD in the following sense: For a fixed target accuracy $\varepsilon$, SGHMC iterates can achieve a generalization error that is (smaller) better than known guarantees for SGLD [RRT17, XCZG17] by a square root factor and this improvement requires relatively in less number of iterations compared to SGLD. Our performance bounds for SGHMC can also improve that of SGLD for computing the population risk up to a square root factor for fixed problem parameters $\beta, d$ and $\varepsilon$ on some non-convex problems. In the multi-pass regime, where the samples are not used only once, exploiting the fact that the $x$-marginal of the stationary distribution for the underdamped diffusion is the same as the overdamped diffusion, we show that the generalization error admits the same bounds known in the literature for the overdamped case, first derived in [RRT17]. In other words, SGHMC algorithm generalizes with a rate no worse than that of the known guarantees for the SGLD algorithm. The results are summarized in Corollary 4.

\textsuperscript{3}Because, in this case, we will not have to account for the suboptimality incurring due to optimizing the global decision variable with respect to a finite sample size.
1.2 Related Work

In a recent work, Simsekli et al. obtained a finite-time performance bound [SYN+18, Theorem 1] for the ergodic average of the SGHMC iterates in the presence of delays in gradient computations. Their analysis highlight the dependency of the optimization error on the delay in the gradient computations and the stepsize explicitly, however it hides some implicit constants $\beta$ and $d$ which can be exponential both in $\beta$ and $d$ in the worst case [SYN+18]. A comparison with the SGLD algorithm is also not given. On the contrary, in our paper, we make all the constants explicit, therefore the effect of acceleration compared to overdamped MCMC approaches such as SGLD is visible.

Cheng et al. [CCA+18] considered the problem of sampling from a target distribution $p(x) \propto \exp(-F(x))$ where $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is $L$-smooth everywhere and $m$-strongly convex outside a ball of finite radius $R$. They proved upper bounds for the time required to sample from a distribution that is within $\varepsilon$ of the target distribution with respect to the 1-Wasserstein distance for both underdamped and overdamped methods that scales polynomially in $\varepsilon$ and $d$. They also show that underdamped MCMC has a better dependency with respect to $\varepsilon$ and $d$ by a square root factor. In our analysis, we consider a larger class of non-convex functions since we do not assume strong convexity in any region, and therefore the distance to the invariant distribution scales exponentially with dimension $d$ in the worst-case. When $f$ is globally strongly convex (or equivalently when the target distribution is strongly log-concave), there is also a growing interesting literature that establish performance bounds for both overdamped MCMC [Dal17] and underdamped MCMC methods [CCBJ17, MS17]. Underdamped Langevin MCMC (also known as Hamiltonian MCMC) and its practical applications have also been analyzed further in a number of recent works [LV18, BBL+17, Bet17, BBG14].

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 $\varepsilon$ in norm) is also studied in the literature (see e.g. [CDHS18, Nes83, GL16, AZH16]). It has also been shown that under some assumptions momentum-based accelerated methods can escape saddle points faster [OW17, LCZZ18]. In contrast, in this work, our focus is obtaining performance guarantees for convergence to global minimizers instead. There is also an alternative approach for non-convex optimization based on graduated optimization techniques [HLSS16] that creates a sequence of smoothed approximations to an objective.

Su et al. [SBC14] shows that Nesterov’s accelerated gradient method [Nes83] tracks a second-order ODE (also referred to as the Nesterov’s ODE in the literature), closely whereas the first-order non-accelerated methods such as the classical gradient descent are known to track the first-order gradient flow dynamics. The authors show that for convex objectives, Nesterov’s ODE converges to its equilibrium faster (in time) than the first-order gradient flow ODE by a square root factor and show that the speed-up is also inherited by the discretized dynamics. 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 faster convergence guarantees compared to first-order overdamped dynamics for non-convex problems.

2 Main Result

In our analysis, we will use the following 2-Wasserstein distance: For any two probability measures $\nu_1, \nu_2$ on $\mathbb{R}^{2d}$, we define

$$W_2(\nu_1, \nu_2) = \left( \inf_{Y_1 \sim \mu, Y_2 \sim \nu} \mathbb{E} \left[ \|Y_1 - Y_2\|^2 \right] \right)^{1/2},$$

where $\| \cdot \|$ is the usual Euclidean norm, $\nu_1, \nu_2$ 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 first state the assumptions used in this paper below. Note that we do not assume $f$ to be convex or strongly convex in any region. 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 [MSH02], and the third assumption is known as the dissipativity condition [Hal88] and is standard in the literature to ensure the convergence of Langevin diffusions to the stationary distribution [RRT17, EGZ17, 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). Finally, the fifth assumption is a mild assumption saying that the initial distribution $\mu_0$ 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 $\mu_0$ is also necessary to achieve convergence to a stationary measure in continuous-time underdamped dynamics as well (see e.g. [HN04b]).

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

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

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

9
(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 $z$:
\[
\mathbb{E} [ \| g(x, U_z) - \nabla F_z(x) \|^2 ] \leq 2 \delta (M^2 \| x \|^2 + B^2).
\]

(v) The probability law $\mu_0$ of the initial state $(X_0, V_0)$ satisfies:
\[
\int_{\mathbb{R}^{2d}} e^{\alpha V(x, v)} \mu_0(dx, dv) < \infty,
\]
where $V$ is a Lyapunov function to be used repeatedly for the rest of the paper:
\[
V(x, v) := \beta F_z(x) + \frac{\beta}{4} \gamma^2 (\| x + \gamma^{-1} v \|^2 + \| \gamma^{-1} v \|^2 - \lambda \| x \|^2), \tag{2.1}
\]
and $\alpha$ is a positive constant that can be found in Table 1 and $\lambda$ is a positive constant less than $\min(1/4, m/(M + \gamma^2 / 2))$.

We note that the Lyapunov function $V$ is used in [EGZ17] to study the rate of convergence to equilibrium for underdamped Langevin diffusion and is motivated by [MSH02]. It follows from the above assumptions (applying Lemma 13) that there exist constants $\lambda \in (0, \min(1/4, m/(M + \gamma^2 / 2)))$ and $A \in (0, \infty)$ so that
\[
x \cdot \nabla F_z(x) \geq m \| x \|^2 - b \geq 2 \lambda (F_z(x) + \gamma^2 \| x \|^2 / 4) - 2A / \beta. \tag{2.2}
\]
This drift condition, which will be used later, guarantees the stability and the existence of Lyapunov function $V$ for the underdamped Langevin diffusion in (1.5)–(1.6), see [EGZ17].

Our first result shows SGHMC iterates $(X_k, V_k)$ track the underdamped Langevin SDE in the sense that the expectation of the empirical risk $F_z$ with respect to the probability law of $(X_k, V_k)$ conditional on the sample $z$, denoted by $\mu_{k,z}$, and the stationary distribution $\pi_z$ of the underdamped SDE is small when $k$ is large enough. The difference in expectations decomposes as a sum of two terms $\mathcal{J}_0(z, \varepsilon)$ and $\mathcal{J}_1(\varepsilon)$ while the former term quantifies the dependency on the initialization and the dataset $z$ whereas the latter term is controlled by the discretization error and the amount of noise parameter $\delta$ in the gradients. We also note that the parameter $\mu_*$ (formally defined later in Section 3.1) in our bounds governs the speed of convergence to the equilibrium of the underdamped Langevin diffusion.
Theorem 2. Consider the SGHMC 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 \(\mu_0\). If Assumption 1 is satisfied, then for \(\beta, \varepsilon > 0\), we have

\[
\|\mathbb{E} F_z(X_k) - \mathbb{E}_{(X,V) \sim \pi_z}(F_z(X))\| = \left| \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \mu_k \, dx \, dv - \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \pi_z \, dx \, dv \right| \\
\leq J_0(z, \varepsilon) + J_1(\varepsilon),
\]

where

\[
J_0(z, \varepsilon) := (M + B) \cdot C \sqrt{\mathcal{H}_\rho(\mu_0, \pi_z)} \cdot \varepsilon,
\]

(2.3)

\[
J_1(\varepsilon) := (M + B) \cdot \left( \frac{\hat{C}_0}{\sqrt{\mu_*}} \sqrt{\log(1/\varepsilon) \delta^{1/4}} + \frac{\hat{C}_1}{\sqrt{\mu_*}} \right) \sqrt{\log(\mu_*^{-1} \log(\varepsilon^{-1}))},
\]

(2.4)

provided that

\[
\eta \leq \min \left\{ \left( \frac{\varepsilon}{\sqrt{\log(1/\varepsilon)}} \right)^4, \frac{\gamma}{K_2} (d/\beta + A/\beta), \frac{\gamma \lambda}{2K_1} \right\},
\]

(2.5)

and

\[
k\eta = \frac{1}{\mu_*} \log \left( \frac{1}{\varepsilon} \right) \geq e.
\]

(2.6)

Here \(\mathcal{H}_\rho\) is a semi-metric for probability distributions defined by (3.11). All the constants are made explicit and are summarized in Table 1.

Proof. The proof of Theorem 2 will be presented in details in Section 3.

In the next subsections, we discuss that this theorem combined with some basic properties of the equilibrium distribution \(\pi_z\) leads to a number of results which provide performance guarantees for both the empirical risk and population risk minimization.

2.1 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

\[
\int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \pi_z \, dx \, dv - \min_{x \in \mathbb{R}^d} F_z(x),
\]

which is a measure of how much the \(x\)-marginal of the equilibrium distribution \(\pi_z\) concentrates around a global minimizer of the empirical risk. As \(\beta\) goes to infinity, it can be
verified that this quantity goes to zero. For finite $\beta$, it is also possible to derive explicit bounds (see Lemma 16) of the form
\[
\int_{\mathbb{R}^d \times \mathbb{R}^d} F_\pi(x) \pi(x,v) - \min_{x \in \mathbb{R}^d} F_\pi(x) \leq \mathcal{J}_2 := \frac{d}{2\beta} \log \left( \frac{eM}{m} \left( \frac{b\beta}{d} + 1 \right) \right). \tag{2.7}
\]
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). Under the setting of Theorem 2, the empirical risk minimization problem admits the performance bounds:
\[
\mathbb{E} F_\pi(X_k) - \min_{x \in \mathbb{R}^d} F_\pi(x) \leq \mathcal{J}_0(\varepsilon, z) + \mathcal{J}_1(\varepsilon) + \mathcal{J}_2,
\]
provided that conditions (2.5) and (2.6) hold where the terms $\mathcal{J}_0(\varepsilon, z)$, $\mathcal{J}_1(\varepsilon)$ and $\mathcal{J}_2$ are defined by (2.3), (2.4) and (2.7) respectively.

### 2.2 Performance bound for the population risk minimization

For controlling the population risk, in addition to the empirical risk, one has to account 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, it can be shown that the generalization error 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 3.

**Corollary 4.** Under the setting of Theorem 2, the expected population risk of $X_k$ (the iterates in (1.9)) is bounded by
\[
\mathbb{E} F(X_k) - F^* \leq \mathcal{J}_0(\varepsilon) + \mathcal{J}_1(\varepsilon) + \mathcal{J}_2 + \mathcal{J}_3(n),
\]
with
\[
\mathcal{J}_0(\varepsilon) := (M\sigma + B) \cdot C \cdot \sqrt{H}_\rho(\mu_0) \cdot \varepsilon, \tag{2.8}
\]
\[
\mathcal{J}_3(n) := \frac{4\beta c_{LS}}{n} \left( \frac{M^2}{m} (b + d/\beta) + B^2 \right), \tag{2.9}
\]
where $\mathcal{J}_1(\varepsilon)$ and $\mathcal{J}_2$ are defined by (2.4) and (2.7) respectively and $c_{LS}$ is a constant satisfying
\[
c_{LS} \leq \frac{2m^2 + 8M^2}{m^2 M \beta} + \frac{1}{\lambda_\ast} \left( \frac{6M(d + \beta)}{m} + 2 \right),
\]
and $\lambda_\ast$ is the uniform spectral gap for overdamped Langevin dynamics:
\[
\lambda_\ast := \inf_{z \in Z^n} \inf \left\{ \frac{\int_{\mathbb{R}^d} \| \nabla g \|^2 d\pi_z}{\int_{\mathbb{R}^d} g^2 d\pi_z} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_z), g \neq 0, \int_{\mathbb{R}^d} gd\pi_z = 0 \right\}.
\]
\[ C_x = \int_{\mathbb{R}^d} \mathcal{V}(x, v) \mu_0(dx, dv) + \frac{(d+A)}{\lambda}, \quad C_\nu = \int_{\mathbb{R}^d} \mathcal{V}(x, v) \mu_0(dx, dv) + \frac{(d+A)}{\lambda} \quad (3.1), (3.2) \]

\[ K_1 = \max \left\{ 32M^2 \left( \frac{1}{2} + \gamma + \delta \right), \frac{8}{3} \frac{1}{2} M + \frac{1}{4} \gamma^2 - \frac{1}{4} \gamma^2 \lambda + \gamma \right\} \beta(1-2\lambda) \quad (3.3) \]

\[ K_2 = B^2 (1+2\gamma+2\delta) \quad (3.4) \]

\[ C_x^d = \int_{\mathbb{R}^d} \mathcal{V}(x, v) \mu_0(dx, dv) + \frac{4(d+A)}{\lambda}, \quad C_\nu^d = \int_{\mathbb{R}^d} \mathcal{V}(x, v) \mu_0(dx, dv) + \frac{4(d+A)}{\lambda} \quad (3.5), (3.6) \]

\[ \sigma = \sqrt{C_x^d \vee C_\nu^d} = \sqrt{C_\nu^d} \quad (3.18) \]

\[ \hat{C}_0 \cdot \hat{C}_1 = \gamma \cdot \left( (M^2 C_x^d + B^2) \beta / \gamma + \sqrt{(M^2 C_x^d + B^2) \beta / \gamma} \right)^{1/2} \quad (3.8) \]

\[ \hat{C}_0 \cdot \hat{C}_1 = \gamma \cdot \left( \beta M^2 C_\nu^d / (2\gamma) + \sqrt{\beta M^2 C_\nu^d / (2\gamma)} \right)^{1/2} \quad (3.9) \]

\[ \hat{\gamma} = \frac{2\sqrt{2}}{\sqrt{\alpha_0}} \left( \frac{5}{2} + \log \left( \int_{\mathbb{R}^d} e^{\frac{1}{4} \alpha \mathcal{V}(x, v) \mu_0(dx, dv) + \frac{1}{4} e^{\frac{\alpha(d+A)}{\lambda} \alpha \gamma(d+A)}} \right) \right)^{1/2} \quad (3.10) \]

\[ \alpha_0 = \frac{\alpha(1-2\lambda)\beta \gamma^2}{64 + 32\gamma^2}, \quad \alpha = \frac{\lambda(1-2\lambda)}{12} \quad (3.7) \]

\[ \mu_* = \frac{\gamma}{768} \min \{ \lambda M \gamma^{-2}, \Lambda^{1/4} e^{-\Lambda} M \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda} \} \quad (3.12) \]

\[ C = \frac{(1+\gamma)\sqrt{2} e^{1+\frac{2}{\lambda}}}{\min \{ 1, \alpha_1 \}} \sqrt{\max \left\{ 1, 4(1+2\alpha_1+2\alpha_1^2)(d+A)\beta^{-1} \gamma^{-1} \mu_*^{-1} \right\} / \min \{ 1, R_1 \} } \quad (3.13) \]

\[ \Lambda = \frac{12}{5} (1+2\alpha_1+2\alpha_1^2)(d+A) M \gamma^{-2} \lambda^{-1}(1-2\lambda)^{-1}, \quad \alpha_1 = (1+\Lambda^{-1}) M \gamma^{-2} \quad (3.14) \]

\[ \varepsilon_1 = 4\gamma^{-1} \mu_*/(d+A) \quad (3.15) \]

\[ R_1 = 4 \cdot (6/5)^{1/2} (1+2\alpha_1+2\alpha_1^2) (d+A)^{1/2} \beta^{-1/2} \gamma^{-1} (\lambda - 2\lambda)^{-1/2} \quad (3.16) \]

\[ \mathcal{F}_\rho(\mu_0) = R_1 + R_1 \varepsilon_1 \max \left\{ M + \frac{1}{2} \beta \gamma^2, \frac{3}{4} \beta \right\} \| (x, v) \|_{L^2(\mu_0)}^2 + R_1 \varepsilon_1 \left( M + \frac{1}{2} \beta \gamma^2 \right) \frac{b + d/\beta}{m} + R_1 \varepsilon_1 \frac{3}{4} d + 2R_1 \varepsilon_1 \left( \beta A_0 + \frac{\beta B^2}{2M} \right) \quad (3.17) \]

Table 1: Summary of the constants and where they are defined in the text.
2.3 Generalization error of SGHMC in the one pass regime

The generalization error of the SGHMC algorithm after $k$ step is defined as $F(X_k) - F_Z(X_k)$ which is a measure of the differences between the objectives of the finite sample size problem (1.2) and the original problem (1.1). Since the predictor $X_k$ is random, it is natural to consider the expected generalization error $E[F(X_k) - F_Z(X_k)]$ (see e.g. [HRS16]) which admits the decomposition

$$E[F_Z(X_k) - F(X_k)] = (E[F_Z(X_k)] - E[F_Z(X^\pi)]) + (E[F_Z(X^\pi)] - E[F(X^\pi)])$$

where $X^\pi$ is the output of the underdamped Langevin dynamics, i.e. its conditional distribution $Z = z$ is given by $\pi_z$. If every sample is used once, i.e. if only one pass is made over the dataset, then the second term in (2.10) disappears. As a consequence, the generalization dynamics is controlled by the bound

$$|E[F_Z(X_k) - F(X_k)]| \leq |E[F_Z(X_k)] - E[F_Z(X^\pi)]| + |E[F(X^\pi)] - E[F(X_k)]|.$$  

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.** Under the setting of Theorem 2, we have

$$|E[F(X_k)] - E[F(X^\pi)]| \leq J_0(\varepsilon) + J_1(\varepsilon),$$

$$|E[F_Z(X_k)] - E[F_Z(X^\pi)]| \leq J_0(\varepsilon) + J_1(\varepsilon),$$

provided that (2.5) and (2.6) hold where $X^\pi$ is the output of the underdamped Langevin dynamics, i.e. its conditional distribution $Z = z$ is given by $\pi_z$ and $J_0(\varepsilon)$ is defined by (2.8). Then, it follows from (2.11) that if each data point is used once, the expected generalization error satisfies

$$|E[F_Z(X_k) - F(X_k)]| \leq 2J_0(\varepsilon) + 2J_1(\varepsilon).$$

2.4 Performance comparison with respect to SGLD algorithm

We use the notations $\tilde{O}$ and $\tilde{\Omega}$ to give explicit dependence on the parameters $d, \beta, \mu^*$ but it hides factors that depend (at worst polynomially) on other parameters $m, M, B, \lambda, \gamma, b$ and $A$.

**Generalization error in the one-pass setting.** A consequence of Theorem 5 is that the generalization error of the SGHMC iterates $|E[F_Z(X_k) - F(X_k)]|$ in the one-pass setting satisfy

$$\tilde{O}\left(\frac{d + \beta}{\mu^* \beta^{3/4} \varepsilon} + \frac{d + \beta}{\sqrt{\beta \mu^*}} \left(\sqrt{\log(1/\varepsilon)} \delta^{1/4} + \varepsilon\right) \sqrt{\log(\mu^* - 1 \log(1/\varepsilon))}\right),$$

(2.12)
for $k = K_{SGHMC} := \tilde{\Omega}\left(\frac{1}{\mu_*} \log^3(1/\varepsilon)\right)$ iterations (see also the discussion in Appendix G).

On the other hand, the results in [RRT17, Theorem 1] imply that the generalization error for the SGLD algorithm after $K_{SGLD}$ iterations in the one-pass setting scales as

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

(2.13)

The constants $\lambda_*$ and $\mu_*$ are exponential in both $\beta$ and $d$ in the worst case, but under some extra assumptions the dependency on $d$ can be polynomial (see e.g. [CCBJ17]) although the exponential dependence to $\beta$ is unavoidable in the presence of multiple minima in general (see [BGK05]).

In both expressions (2.12) and (2.13), we see a term scaling with $\delta^{1/4}$ due to the gradient noise level $\delta$. If we select $\delta$ so that the error term introduced by noisy gradients scaling as $\delta^{1/4}$ is on the same order as the other terms in (2.13) as suggested by the authors in [RRT17], and do the same for SGHMC, then we see that the generalization error for SGHMC (2.12) becomes

$$\tilde{O}\left(\frac{d + \beta}{\mu_* \beta^{3/4} \varepsilon} + \frac{d + \beta}{\sqrt{\beta} \mu_*} \varepsilon \sqrt{\log(\mu_*^{-1} \log(1/\varepsilon))}\right) \leq \tilde{O}\left(\max\left(\frac{d + \beta}{\beta^{3/4} \mu_*}, \frac{(d + \beta)}{\sqrt{\beta} \mu_*}\right) \varepsilon \left(\sqrt{\log(\mu_*^{-1}) + \sqrt{\log \log(1/\varepsilon)}}\right)\right)
= \tilde{O}\left(\frac{d + \beta}{\sqrt{\beta} \mu_*} \varepsilon \sqrt{\log \log(1/\varepsilon)}\right),$$

(2.14)

and if we ignore the $\sqrt{\log \log(1/\varepsilon)}$ factor \(^4\), then, we get

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

whereas the corresponding bound for SGLD from [RRT17, Theorem 1] is

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

(2.15)

Note that $K_{SGHMC}$ and $K_{SGLD}$ do not have the same dependency to $\varepsilon$ up to log factors (the former scales with $\varepsilon$ as $\log^3(1/\varepsilon)\varepsilon^{-4}$ and the latter as $\log^5(1/\varepsilon)\varepsilon^{-4}$). To make the comparison to SGLD simpler, if we run SGHMC for number of iterations $K_{SGHMC}$ proportional to $\log^5(1/\varepsilon)\varepsilon^{-4}$ instead, then the generalization error behaves as,

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

\(^4\)We emphasize that the effect of the last term $\sqrt{\log \log(1/\varepsilon)}$ appearing in (2.14) is typically negligible compared to other parameters. For instance even if $\varepsilon = 2^{-2^{16}}$ is double-exponentially small, we have $\sqrt{\log \log(1/\varepsilon)} \leq 4$. 

15
that is (ignoring the $\sqrt{\log \log (1/\varepsilon)}$ factor),
\[
\hat{O}\left(\frac{d + \beta}{\sqrt{\beta \mu_\ast}}\right) \quad \text{for} \quad K_{\text{SGHMC}} = \hat{\Omega}\left(\frac{1}{\mu_\ast \varepsilon^4} \log^5 (1/\varepsilon)\right).
\] (2.16)

Comparing $\lambda_\ast$ and $\mu_\ast$ on arbitrary non-convex functions seems not trivial, however we give a class of non-convex functions (see Examples 6 and 7) where $\frac{1}{\mu_\ast} = \hat{O}\left(\sqrt{\frac{1}{\lambda_\ast}}\right)$. For this class, comparing (2.16) and (2.15), we see that for $\varepsilon$ given and fixed, SGHMC generalization error upper bound is at least on the order of the square root of that of the SGLD (Note that the improvement with respect to $d$, $\beta$ and $\varepsilon$ dependency is more than a square root factor, but the improvement from $\lambda_\ast$ to $\mu_\ast$ can be a square root factor). Furthermore this can be achieved in significantly smaller number of iterations (up to a square root factor) if $\varepsilon$ is fixed, i.e. $K_{\text{SGHMC}}$ can be on the order of the square root of $K_{\text{SGLD}}$ if $\varepsilon$ is given and fixed in both (2.16) and (2.16).

**Empirical risk minimization.** The empirical risk minimization bound given in Corollary 3 has an additional term $J_2$ compared to the $J_0(\varepsilon)$ and $J_1(\varepsilon)$ terms appearing in the one-pass generalization bounds. Note also that $J_0(\varepsilon) \leq J_1(\varepsilon)$. As a consequence, SGHMC algorithm has expected empirical risk
\[
\hat{O}\left(\frac{d + \beta}{\mu_\ast^{\beta/4}} + \frac{d + \beta}{\sqrt{\beta \mu_\ast}} \left(\sqrt{\log (1/\varepsilon)} (\delta^{1/4} + \varepsilon) \sqrt{\log (\mu_\ast^{-1} \log (1/\varepsilon))) + d \cdot \frac{\log (1 + \beta)}{\beta}\right)\right),
\] (2.17)
after $K_{\text{SGHMC}} = \hat{\Omega}\left(\frac{1}{\mu_\ast \varepsilon^4} \log^3 (1/\varepsilon)\right)$ iterations as opposed to
\[
\hat{O}\left(\frac{(\beta(\beta + d)}{\lambda_\ast} \left(\delta^{1/4} \log (1/\varepsilon) + \varepsilon\right) + d \cdot \frac{\log (1 + \beta)}{\beta}\right),
\] (2.18)
after $K_{\text{SGLD}} = \hat{\Omega}\left(\frac{\beta(d + \beta)}{\lambda_\ast} \log^5 (1/\varepsilon)\right)$ iterations required in [RRT17]. Comparing (2.17) and (2.18), 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 SGHMC will be similar except that $K_{\text{SGHMC}}$ can be smaller than $K_{\text{SGLD}}$ for instance when $\frac{1}{\mu_\ast} = O\left(\sqrt{\frac{1}{\lambda_\ast}}\right)$. 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 (2.14) and (2.15) discussed above where a square root factor improvement for SGHMC compared to SGLD is possible.

Following [RRT17] and [XCGZ17], if we define an almost empirical risk minimizer (ERM) as a point which is within the ball of the global minimizer with radius $O(d \log (1 + \beta)/\beta)$. Xu et al. [XCGZ17] recently showed that it suffices to take
\[
\hat{K}_{\text{SGLD}} = \hat{O}\left(\frac{d^7}{\lambda^{5/2}}\right)
\] (2.19)
stochastic gradient computations to converge to an $\hat{\epsilon}$ neighborhood of an almost ERM where $\tilde{O}(\cdot)$ hides some factors in $\beta$ and $\hat{\lambda}$ is the spectral gap of the discrete underdamped Markov Chain. Our results show that (see e.g. (2.17)), it suffices to have

$$\hat{K}_{SGHMC} = \tilde{\Omega}\left(\frac{d^4}{\mu^3_*\hat{\epsilon}^8}\right)$$  \hspace{1cm} (2.20)

stochastic gradient computations, ignoring the log factors in the parameters $\hat{\epsilon}, \mu_*, d$ and hiding factors in $\beta$ that can be made explicit. It is hard to compare $\hat{\lambda}$ and $\lambda_*$ in general, the former being the spectral gap of the discretized overdamped Langevin dynamics and the latter being the spectral gap of the continuous-time overdamped dynamics. However, when the stepsize is small enough, $\hat{\lambda}$ will be similar to $\lambda_*$. As a consequence, when the stepsize $\eta$ is small enough (which is the case for instance, when target accuracy $\hat{\epsilon}$ is small enough or $\beta$ is large enough), we will have $\hat{\lambda} \approx \lambda_*$ and $\mu_* = O(\sqrt{\lambda_5}) \approx O(\sqrt{\hat{\lambda}})$ for the class of non-convex functions we discuss in Examples 6 and 7. For this class of problems, comparing (2.19) and (2.20), we see that we obtain an improvement in the dimension dependent factors in front of the spectral gap parameter and $\epsilon$ term (scaling $d^4$ vs. $d^7$) as well as an improvement in the spectral gap parameter ($\mu_*^3$ vs. $\hat{\lambda}^5$), however $\epsilon$ dependency of the bound (2.19) is better than (2.20) (scaling $1/\epsilon^5$ vs. $1/\epsilon^8$).

**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 $J_3$ that needs to be added to the bounds given in (2.17) and (2.18). This term satisfies

$$J_3 = \tilde{O}\left(\frac{(\beta + d)^2}{\lambda_* n}\right).$$

If this term is dominant compared to other terms $J_0, J_1$ and $J_2$, 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 SGHMC will be similar. Otherwise, if $n$ is large and $\beta$ is chosen in a way to keep the $J_2$ term on the order $J_0$, then the square root improvement can be achieved.

**Comparison of $\lambda_*$ and $\mu_*$.** The parameters $\lambda_*$ and $\mu_*$ govern the convergence rate to the equilibrium of the overdamped and underdamped Langevin SDE, they can be both exponentially large in dimension and in $\beta$. They appear naturally in the complexity estimates of SGHMC and SGLD method as they can be viewed as discretizations of these methods (when the discretization step is small and the gradient noise $\delta = 0$, the discrete dynamics will behave the same as the continuous dynamics) and arise naturally in the iteration complexity of both methods. Next, we discuss classes of non-convex functions in Examples 6 and 7 where $\mu_*$ is on the order of the square root of $\lambda_*$. As a consequence, for these examples if the other parameters ($\beta, d, \delta$) are fixed, then SGHMC can lead to an improvement upon the SGLD performance.
Example 6. Consider the following symmetric double-well potential in $\mathbb{R}^d$ studied previously in the context of Langevin diffusion:

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

where $a > 0$ is a scaling parameter which is illustrated in Figure 1. For this example, there are two minima that are apart at a distance $R = \mathcal{O}(a)$. Having access to a dataset $z$ and unbiased noisy gradients of this objective $\nabla f(x,z_i)$ satisfying $\mathbb{E}[\nabla f(x,z_i)] = \nabla F(x)$, we consider the stochastic optimization problem (1.1) with both the SGHMC algorithm and the SGLD algorithm. Eberle et al. [EGZ17] 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 $\Theta$ explicit. Our analysis shows that this contraction rate holds uniformly in $z$ under Assumption 1 for the underdamped diffusion (1.5)–(1.6). This shows that the contraction rate of the underdamped dynamics $\mu_*$ is (faster) larger than that of overdamped dynamics $\lambda_*$ by a square root factor when $a$ is large where all the constants can be made explicit. As a consequence, it follows from our previous discussion (see (2.14) and (2.15)) that the generalization error of SGHMC improves that of SGLD by a square root factor in the regime when $a$ is large. Roughly speaking, this result tells us that when there are minima that are far away from each other at a distance $R = \mathcal{O}(a)$, then the iteration complexity of SGHMC will grow linearly with $R$, whereas that of SGLD will grow quadratically proportional to $R^2$. In a sense, the iteration complexity of SGLD is more sensitive to the distance between the local minima. Such results extends 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 [EGZ17, Example 1.1, Example 1.13] for a further discussion).

For computing an $\epsilon$-approximate global minimizer of $f_a$ (or more generally for a non-convex function satisfying Assumption 1), $\beta$ is chosen large enough so that the stationary measure concentrates around the global minimizer, i.e. for instance $\beta$ should be large enough so that $\frac{1}{\beta}$ is on the order of $\epsilon$ in general. Using the tight characterization of $\lambda_*$ from [BGK05, Theorem 1.2] for $\beta$ 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 as we discuss in the next example.

Example 7. Consider the following asymmetric double-well potential in dimension one:

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

This is a continuously differentiable function with two local minimum at $\tilde{x}_1 = -1/4$ and $\tilde{x}_2 = 1$, where the latter is the (unique) global minimum. For $a > 0$, consider the scaled
functions $\tilde{f}_a(x) := \tilde{U}(x/a)$ with the same height but local minima are apart from each other at $O(a)$ distance. The right panel of Figure 1 illustrates the graph of $\tilde{f}_a(x)$ for $a = 4$.

Let $\beta$ be given and fixed. For this example, parts (ii) and (iii) of the Assumption 1 hold with parameters $M = 1/a^2$, $m = 15/R^2$ with $R = 4a$, $b = 4$. If $\gamma$ is chosen such that $M\gamma^{-2} \leq 1/30$, i.e. $\gamma \geq \sqrt{30}/a$, then $\Lambda = \Theta(\beta)$ and it follows from Theorem 11 that the contraction rate satisfies

$$\mu_* = \Theta(\beta) \left( \frac{1}{a} \right),$$

where $\Theta(\cdot)$ hides some constants that depend only on $\beta$ and that can be made explicit. On the other hand, it follows from Bovier et al. that [BGK05, Theorem 1.2] for the spectral gap of the overdamped dynamics, we have $\lambda_* = \Theta(1/a^2)$. This shows that when the separation between minima, or alternatively the scaling factor $a$ is large enough, $\mu_*$ is larger than $\lambda_*$ by a square factor up to constants.

3 Proof of Theorem 2 and Corollary 4

We first derive several technical lemmas that will be used in our analysis and review existing results for the underdamped SDE.

Our analysis for analyzing the convergence speed of the SHGMC 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 $V$ which is designed originally for the continuous-time process and show that the discrete dynamics can also be controlled by it.
Lemma 8 (Uniform $L^2$ bounds).

(i) It holds that
\[
\sup_{i \geq 0} \mathbb{E}_z \|X(t)\|^2 \leq C_c^x := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) d\mu_0(x,v) + \frac{d+A}{\lambda}}{1 \beta (1 - 2\lambda) \beta \gamma^2} < \infty, \tag{3.1}
\]
\[
\sup_{i \geq 0} \mathbb{E}_z \|V(t)\|^2 \leq C_c^v := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) d\mu_0(x,v) + \frac{d+A}{\lambda}}{\beta (1 - 2\lambda)} < \infty, \tag{3.2}
\]

(ii) For $0 < \eta \leq \min \left\{ \frac{\gamma}{K_2} (d/\beta + A/\beta), \frac{\gamma \lambda}{2K_1} \right\}$, where
\[
K_1 := \max \left\{ \frac{32M^2 (\frac{1}{2} + \gamma + \delta)}{(1 - 2\lambda) \beta \gamma^2}, \frac{8 \left( \frac{1}{2}M + \frac{1}{4} \gamma^2 - \frac{1}{4} \gamma^2 \lambda + \gamma \right)}{\beta (1 - 2\lambda)} \right\}, \tag{3.3}
\]
\[K_2 := 2B^2 \left( \frac{1}{2} + \gamma + \delta \right), \tag{3.4}\]
we have
\[
\sup_{j \geq 0} \mathbb{E}_z \|X_j\|^2 \leq C_c^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) d\mu_0(x,v) + \frac{4(d+A)}{\lambda}}{\beta (1 - 2\lambda) \beta \gamma^2} < \infty, \tag{3.5}
\]
\[
\sup_{j \geq 0} \mathbb{E}_z \|V_j\|^2 \leq C_c^d := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v) d\mu_0(x,v) + \frac{4(d+A)}{\lambda}}{\beta (1 - 2\lambda)} < \infty. \tag{3.6}
\]

Since SGHMC is a discretization of the underdamped SDE (except that noise is also added to the gradients), we expect SGHMC to follow the underdamped SDE dynamics. It is natural to seek for bounds between the probability law $\mu_{z,k}$ of the SGHMC algorithm at step $k$ with time step $\eta$ and that of the underdamped SDE at time $t = k \eta$ which we denote by $\nu_{z,k\eta}$. In our analysis, we first control the 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 Bolley and Villani [BV05].\footnote{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 10 about the diffusion approximation of the SGHMC iterates.
Lemma 9 (Exponential integrability). For every $t$,

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

where

$$
\alpha_0 := \frac{\alpha}{(1 - 2\lambda) \beta \gamma^2 + \frac{32}{\beta (1 - 2\lambda)}}, \quad \alpha := \frac{\lambda (1 - 2\lambda)}{12}.
$$

(3.7)

We showed in the above Lemma 9 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 10 for the diffusion approximation, our upper bound is of the order $\sqrt{k \eta} \cdot \sqrt{\log(k \eta)} (\frac{\delta}{4} + \eta \frac{1}{4})$ which strictly improves from $k \eta (\frac{\delta}{4} + \eta \frac{1}{4})$ in [RRT17].

Lemma 10 (Diffusion approximation). For any $k \in \mathbb{N}$ and any $\eta$, so that $k \eta \geq e$ and $\eta$ satisfies the condition in Part (ii) of Lemma 8, we have

$$
\mathcal{W}_2(\mu_{z,k}, \nu_{z,k}) \leq (\hat{C}_0 \delta^{1/4} + \hat{C}_1 \eta^{1/4}) \cdot \sqrt{k \eta} \cdot \sqrt{\log(k \eta)},
$$

where $\hat{C}_0$ and $\hat{C}_1$ are given by:

$$
\hat{C}_0 = \hat{\gamma} \cdot \left( (M^2 C_x^d + B^2) \frac{\beta}{\gamma} + \sqrt{(M^2 C_x^d + B^2) \frac{\beta}{\gamma}} \right)^{1/2},
$$

(3.8)

$$
\hat{C}_1 = \hat{\gamma} \cdot \left( \frac{\beta M^2 C_v^d}{2 \gamma} + \sqrt{\frac{\beta M^2 C_v^d}{2 \gamma}} \right)^{1/2},
$$

(3.9)

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

(3.10)

3.1 Convergence rate to the equilibrium of the underdamped SDE

We consider the underdamped SDE and bound the 2-Wasserstein distance $\mathcal{W}_2(\nu_{z,t}, \pi_z)$ to the equilibrium for a fix arbitrary time $t \geq 0$. Crucial to the analysis is Eberle et al. [EGZ17], which quantifies the convergence to equilibrium for underdamped Langevin diffusions. We first review the results from [EGZ17]. Let us recall from (2.1) the definition of the Lyapunov function $V(x, v)$:

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

The method that is used in the proof of Lemma 9 for the underdamped dynamics can indeed be adapted to the case of the overdamped dynamics to improve the results in [RRT17].
For any \((x, v), (x', v') \in \mathbb{R}^d\), we set:

\[
\begin{align*}
   r((x, v), (x', v')) &= \alpha_1 \|x - x'\| + \|x - x' + \gamma^{-1}(v - v')\|, \\
   \rho((x, v), (x', v')) &= h(r((x, v), (x', v'))),
\end{align*}
\]

where \(\alpha_1, \varepsilon > 0\) 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 \(R_1 \geq 0\) 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 \(\mu, \nu\) on \(\mathbb{R}^d\), we define

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

Note that \(\mathcal{H}_\rho\) is a semi-metric, but not necessarily a metric. A simplified version of the main result from [EGZ17] which will be used in our setting is given below.

**Theorem 11** (**Theorem 1.4.** and **Corollary 1.7.** [EGZ17]). There exist constants \(\alpha_1, \varepsilon_1 \in (0, \infty)\) and a continuous non-decreasing function \(h: [0, \infty) \to [0, \infty)\) with \(h(0) = 0\) such that we have

\[
W_2(\nu_{x, k\eta}, \pi_x) \leq C \sqrt{\mathcal{H}_\rho(\mu_0, \pi_x)} e^{-\mu_* k\eta}
\]

where

\[
\begin{align*}
   \mu_* &= \frac{\gamma}{738} \min\{\lambda M \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda} M \gamma^{-2}, \Lambda^{1/2} e^{-\Lambda}\}, \\
   C &= \beta^{-1} \gamma^{-1} \mu_* / \min\{1, R_1\}, \\
   \Lambda &= \frac{12}{5} (1 + 2\alpha_1 + 2\alpha_2^2) (d + A) \gamma^{-2} (1 - 2\lambda)^{-1}, \\
   \alpha_1 &= (1 + \omega^{-1}) M \gamma^{-2}, \\
   \varepsilon_1 &= 4\gamma^{-1} \mu_*/(d + A), \\
   R_1 &= 4 \cdot (6/5)^{1/2} (1 + 2\alpha_1 + 2\alpha_2^2) (d + A)^{1/2} \gamma^{-2} (1 - 2\lambda)^{-1/2}.
\end{align*}
\]

We remark that there are unique values \(\Lambda, \alpha_1 \in (0, \infty)\) such that (3.14) is satisfied, see [EGZ17]. In order to get explicit performance bounds, we also derive an upper bound for \(\mathcal{H}_\rho(\mu_0, \pi_x)\) in the next lemma. It is based on the (integrability properties) structure of the stationary distribution \(\pi_x\) and the Lyapunov function \(V\) that controls the \(L^2\) norm of the initial distribution \(\mu_0\).

**Lemma 12** (**Bounding initialization error**). If parts (i), (ii), (iii) and (iv) of **Assumption 1** hold, then we have

\[
\begin{align*}
   \mathcal{H}_\rho(\mu_0, \pi_x) &\leq \mathcal{H}_\rho(\mu_0) := R_1 + R_1 \varepsilon_1 \max\left\{ M + \frac{1}{2} \beta \gamma^2, \frac{3}{4} \beta \right\} \|(x, v)\|^2_{L^2(\mu_0)} \\
   &\quad + R_1 \varepsilon_1 \left( M + \frac{1}{2} \beta \gamma^2 \right) \frac{b + d/\beta}{m} + R_1 \varepsilon_1 \frac{3}{4} d + 2R_1 \varepsilon_1 \left( \beta A_0 + \frac{\beta B^2}{2M} \right),
\end{align*}
\]

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

As the function $F_z$ satisfies the conditions in Lemma 14 with $c_1 = M$ and $c_2 = B$ (Lemma 13), and the probability measures $\mu_{k,z}, \pi_z$ have finite second moments (Lemma 8), we can apply Lemma 14 and deduce that

$$\left| \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \mu_{k,z}(dx, dv) - \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \pi_z(dx, dv) \right| \leq (M + B) \cdot W_2(\mu_{k,z}, \pi_z).$$

Here, one can obtain from Lemma 8 and Theorem 11 (convergence in 2-Wasserstein distance implies convergence of second moments) that

$$\sigma^2 = C_x^0 \vee C_x^d = C_x^d. \quad (3.18)$$

Now, by Lemma 10 and Theorem 11, we have

$$W_2(\mu_{z,k}, \pi_z) \leq W_2(\mu_{z,k}, \nu_{z,k\eta}) + W_2(\nu_{z,k\eta}, \pi_z) \leq (\check{C}_0 \delta^{1/4} + \check{C}_1 \eta^{1/4}) \cdot \sqrt{k\eta} \cdot \sqrt{\log(k\eta)} + C \sqrt{\mathcal{H}_\rho(\mu_0, \pi_z)} e^{-\mu_* k\eta}.$$

It then follows that

$$\left| \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \mu_{k,z}(dx, dv) - \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \pi_z(dx, dv) \right| \leq (M + B) \cdot \left( C \sqrt{\mathcal{H}_\rho(\mu_0, \pi_z)} e^{-\mu_* k\eta} + \check{C}_0 \delta^{1/4} + \check{C}_1 \eta^{1/4} \right) \cdot \sqrt{k\eta} \cdot \sqrt{\log(k\eta)}.$$

Let $k\eta \geq e$, and

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

Then for any $\eta$ satisfying the condition in Lemma 8 and $\eta \leq \left( \frac{\varepsilon}{\sqrt{\log(1/\varepsilon)}} \right)^4$, we have

$$\left| \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \mu_{k,z}(dx, dv) - \int_{\mathbb{R}^d \times \mathbb{R}^d} F_z(x) \pi_z(dx, dv) \right| \leq (M + B) \cdot \left( C \sqrt{\mathcal{H}_\rho(\mu_0, \pi_z)} \varepsilon + \left( \frac{\check{C}_0}{\sqrt{\mu_*}} \sqrt{\log(1/\varepsilon)} \delta^{1/4} + \frac{\check{C}_1}{\sqrt{\mu_*}} \varepsilon \right) \sqrt{\log(\mu_*^{-1} \log(1/\varepsilon))} \right).$$

The proof is therefore complete.
3.3 Completing the proof of Corollary 4

Consider the random elements $(\hat{X}, \hat{V})$ and $(\hat{X}^*, \hat{V}^*)$ with $\mathcal{L}((\hat{X}, \hat{V})|Z = z) = \mu_{z,k}$ and $\mathcal{L}((\hat{X}^*, \hat{V}^*)|Z = z) = \pi_z$. Then we can decompose the expected population risk of $\hat{X}$ (which has the same distribution as $X_k$) as follows:

$$EF(\hat{X}) - F^* = \left( EF(\hat{X}) - EF(\hat{X}^*) \right) + \left( EF(\hat{X}^*) - EF_Z(\hat{X}^*) \right) + \left( EF_Z(\hat{X}^*) - F^* \right). \quad (3.19)$$

The first term in (3.19) can be written as:

$$EF(\hat{X}) - EF(\hat{X}^*) = \int_Z P^n(dz) \left( \int_{\mathbb{R}^{2d}} F_z(x) \mu_{k,z}(dx, dv) - \int_{\mathbb{R}^{2d}} F_z(x) \pi_z(dx, dv) \right),$$

where $P^n$ is the product measure of independent random variables $Z_1, \ldots, Z_n$. Then it follows from Theorem 2 and Lemma 12 that

$$EF(\hat{X}) - EF(\hat{X}^*) \leq \mathcal{J}_0(\varepsilon) + \mathcal{J}_1(\varepsilon).$$

Next, we bound the second and third terms in (3.19). Note that

$$\int_{\mathbb{R}^{2d}} F_z(x) \pi_z(dx, dv) = \int_{\mathbb{R}^{2d}} F_z(x) \pi_z(dx),$$

where $\pi_z(dx) = \Lambda_z e^{-\beta F_z(x)} dx$ and $\Lambda_z = \int_{\mathbb{R}^d} e^{-\beta F_z(x)} dx$. The distribution $\pi_z(dx)$, i.e., the $x$–marginal of $\pi_z(dx, dv)$, is the same as the stationary distribution of the overdamped Langevin SDE in (1.4). Therefore the second term and the third term in (3.19) can be bounded the same as in [RRT17] for the overdamped dynamics.

Specifically, the second term in (3.19) can be bounded as

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

by applying Lemma 15, and the last term in (3.19) can be bounded as

$$EF_Z(\hat{X}^*) - F^* = E \left[ F_Z(\hat{X}^*) - \min_{x \in \mathbb{R}^d} F_Z(x) \right] + E \left[ \min_{x \in \mathbb{R}^d} F_Z(x) - F_Z(x^*) \right] \leq \mathcal{J}_2,$$

where $x^*$ is any minimizer of $F(x)$, i.e., $F(x^*) = F^*$, and the last step is due to Lemma 16. The proof is complete.
4 Conclusion

SGHMC is a momentum-based popular variant of stochastic gradient where a controlled amount of anistropic 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 SGHMC to the $\varepsilon$-global minimizers under some regularity assumption on the non-convex objective $f$. Our analysis does not assume convexity or strong convexity of $f$ in any region of the domain.

We also show that on a class of non-convex problems, SGHMC can be faster than overdamped Langevin MCMC approaches such as SGLD by a square root factor on some problems. 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.

Acknowledgements

We thank 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 Proceedings of the 29th International Conference on International Conference on Machine Learning, pages 1771–1778. Omnipress, 2012.

[AP16] Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov’s accelerated forward-backward method is actually faster than $1/k^2$. SIAM Journal on Optimization, 26(3):1824–1834, 2016.

[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.

[BBL+17] Michael Betancourt, Simon Byrne, Sam Livingstone, Mark Girolami, et al. 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.

[BGMZ18] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre.-André Zitt. Functional inequalities for Gaussian convolutions of compactly supported measures: Explicit bounds and dimension dependence. *Bernoulli*, 24(1):333–353, 2018.

[BLNR15] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In Peter Grunwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 240–265, Paris, France, 03–06 Jul 2015. PMLR.

[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. In *Annales Faculté des sciences Toulouse Mathematiques*, volume 14, page 331. Université Paul Sabatier, 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 e-prints*, May 2018.

[CCBJ17] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[CCG+16] Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Artificial Intelligence and Statistics*, 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.

[CDH15] 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 non-convex optimization. SIAM Journal on Optimization, 28(2):1751–1772, 2018.

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

[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.

[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.

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

[DLT+17] Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer CNN: Don’t be afraid of spurious local minima. arXiv preprint arXiv:1712.00779, 2017.

[DT12] Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. Journal of Computer and System Sciences, 78(5):1423–1443, 2012.
[EGZ17] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

[FRV18] Guilherme Francha, Daniel P Robinson, and René Vidal. Relax, and accelerate: A continuous perspective on admm. *arXiv preprint arXiv:1808.04048*, 2018.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[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 non-convex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, March 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.

[Gyö86] István Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Itô differential. *Probability theory and related fields*, 7(4):501–516, 1986.

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

[Hal88] JK Hale. Asymptotic behavior of dissipative systems (vol 25). American Mathematical Society. 1988.

[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.

[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.
[HN04a] 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, Feb 2004.

[HN04b] 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 *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1225–1234. JMLR. org, 2016.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015.

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

[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. *arXiv preprint arXiv:1806.02476*, 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.

[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.
[MS17] O. Mangoubi and A. Smith. Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions. *ArXiv e-prints*, August 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.

[Oks03] B. K. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 2003.

[OW17] Michael O’Neill and Stephen J Wright. Behavior of accelerated gradient methods near critical points of nonconvex problems. *arXiv preprint arXiv:1706.07993*, 2017.

[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. translations series in mathematics and engineering. *Optimization Software*, 1987.

[PT13] Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 26*, pages 3102–3110. Curran Associates, Inc., 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.

[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 Nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.
[SBCR16] Umut Simsekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. Stochastic quasi-Newton Langevin Monte Carlo. In *International Conference on Machine Learning*, pages 642–651, 2016.

[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.

[SRBd17] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d’Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 1109–1118, 2017.

[SYN+18] U. Şimşekli, Ç. Yıldız, T. H. Nguyen, G. Richard, and A. Taylan Cemgil. Asynchronous Stochastic Quasi-Newton MCMC for Non-Convex Optimization. *arXiv preprint arXiv:1806.02617*, June 2018.

[TLR18] Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 857–875. PMLR, 06–09 Jul 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.

[Wib18] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. *arXiv preprint arXiv:1802.08089*, 2018.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[XCZG17] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *arXiv preprint arXiv:1707.06618*, 2017.
[ZLC17] Yuchen Zhang, Percy. Liang, and Moses. Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. *arXiv preprint arXiv:1702.05575*, February 2017.

[ZMSJ18] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. *arXiv preprint arXiv:1805.00521*, 2018.
Proof of Lemma 8

Proof. (i) We first prove the continuous–time case. The main idea is to use the following Lyapunov function (see (2.1)) introduced in [EGZ17] for the underdamped Langevin diffusion:

\[ V(x, v) = \beta F_z(x) + \frac{\beta}{4} \gamma^2 (\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2). \quad \text{(A.1)} \]

Lemma 1.3 in [EGZ17] showed that if the drift condition in (2.2) holds, then

\[ L V \leq \gamma (d + A - \lambda V), \quad \text{(A.2)} \]

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

\[ L V = -(\gamma v + \nabla F_z(x)) \nabla_v V + \gamma \beta^{-1} \Delta_v V + v \nabla_x V. \quad \text{(A.3)} \]

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

\[ V(x, v) \geq \beta F_z(x) + \frac{\beta}{4} (1 - 2\lambda) \gamma^2 (\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2) \]

\[ \geq \max \left\{ \frac{1}{8} (1 - 2\lambda) \beta \gamma^2 \|x\|^2, \frac{\beta}{4} (1 - 2\lambda) \|v\|^2 \right\}. \quad \text{(A.4)} \]

Now let us set for each \( t \geq 0 \),

\[ L(t) := \mathbb{E}_z[V(X(t), V(t))], \quad \text{(A.5)} \]

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

First, we can compute that

\[ \nabla_v V = \beta v + \frac{\beta \gamma}{2} x, \quad \text{(A.6)} \]

By Itô’s formula and (A.6),

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

which together with (A.2) implies that

\[
\begin{aligned}
    e^{\gamma t}V(X(t), V(t)) &\leq V(X(0), V(0)) + \gamma (d + A) \int_0^t e^{\gamma s} ds \\
    &- \int_0^t e^{\gamma s} \left( \beta V(s) + \frac{\beta \gamma}{2} X(s) \right) \cdot \sqrt{2\gamma \beta} dB(s). \quad \text{(A.7)}
\end{aligned}
\]
Note that $\nabla F_z(x)$ 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, 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 (A.7) and (A.5) that for any $t \geq 0$,

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

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

$$\frac{1}{8}(1 - 2\lambda)\beta\gamma^2 \mathbb{E}_z\|X(t)\|^2 \leq \mathbb{E}_z[\mathcal{V}(X_0, V_0)] + \frac{d + A}{\lambda},$$

$$\frac{\beta}{4}(1 - 2\lambda)\mathbb{E}_z\|V(t)\|^2 \leq \mathbb{E}_z[\mathcal{V}(X_0, V_0)] + \frac{d + A}{\lambda}.$$

The proof of part (i) is complete by noting that $\mathbb{E}_z[\mathcal{V}(X_0, V_0)]$ 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:

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

$$X_{k+1} = X_k + \eta V_k, \quad (A.9)$$

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

$$L_2(k) = \mathbb{E}_z\mathcal{V}(X_k, V_k)/\beta = \mathbb{E}_z \left[ F_z(X_k) + \frac{1}{4}\gamma^2(\|X_k + \gamma^{-1}V_k\|^2 + \|\gamma^{-1}V_k\|^2 - \lambda\|X_k\|^2) \right]. \quad (A.10)$$

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 $L_2(k + 1)$.
Firstly, by using the independence of $V_k - \eta[\gamma V_k + g_k(X_k, U_{z,k})]$ and $\xi_k$, we can obtain from (A.8) that
\[
\mathbb{E}_z\|V_{k+1}\|^2 = \mathbb{E}_z\|V_k - \eta[\gamma V_k + g_k(X_k, U_{z,k})]\|^2 + 2\gamma \beta^{-1} \eta \mathbb{E}_z\|\xi_k\|^2 = \mathbb{E}_z\|V_k - \eta[\gamma V_k + \nabla F_{z}(X_k)]\|^2 + 2\gamma \beta^{-1} \eta d = (1 - \eta\gamma)^2 \mathbb{E}_z\|V_k\|^2 - 2\eta(1 - \eta\gamma)\mathbb{E}_z[\langle V_k, \nabla F_{z}(X_k) \rangle] + \eta^2 \mathbb{E}_z\|\nabla F_{z}(X_k)\|^2 + 2\gamma \beta^{-1} \eta d + 2\delta \eta^2 M^2 \mathbb{E}_z\|X_k\|^2 + 2\delta \eta^2 B^2.
\]
where we have used part (iv) of Assumption 1 and Lemma 13 in Appendix E. By using $|x| \leq \frac{|x|^2 + 1}{2}$, we immediately get
\[
\mathbb{E}_z\|V_{k+1}\|^2 \leq (1 - \eta\gamma)^2 \mathbb{E}_z\|V_k\|^2 - 2\eta \mathbb{E}_z[\langle V_k, \nabla F_{z}(X_k) \rangle] + \eta^2 \mathbb{E}_z\|\nabla F_{z}(X_k)\|^2 + (\eta^2 MB + 2\gamma \beta^{-1} \eta d + 2\delta \eta^2 B^2).
\]

Secondly, we can compute from (A.9) that
\[
\mathbb{E}_z\|X_{k+1}\|^2 = \mathbb{E}_z\|X_k\|^2 + 2\eta \mathbb{E}_z[\langle X_k, V_k \rangle] + \eta^2 \mathbb{E}_z\|V_k\|^2.
\]

Thirdly, note that
\[
F_{z}(X_{k+1}) = F_{z}(X_{k} + \eta V_k) = F_{z}(X_{k}) + \int_0^1 \langle \nabla F_{z}(X_{k} + \tau \eta V_k), \eta V_k \rangle d\tau,
\]
which immediately suggests that
\[
|F_{z}(X_{k+1}) - F_{z}(X_{k}) - \langle \nabla F_{z}(X_{k}), \eta V_k \rangle| = \left| \int_0^1 \langle \nabla F_{z}(X_{k} + \tau \eta V_k) - \nabla F_{z}(X_{k}), \eta V_k \rangle d\tau \right| \\
\leq \int_0^1 \|\nabla F_{z}(X_{k} + \tau \eta V_k) - \nabla F_{z}(X_{k})\| \cdot \|\eta V_k\| d\tau \\
\leq \frac{1}{2} M \eta^2 \|V_k\|^2,
\]
where the last inequality is due to the $M$-smoothness of $F_{z}$. This implies
\[
\mathbb{E}_z F_{z}(X_{k+1}) - \mathbb{E}_z F_{z}(X_{k}) \leq \eta \mathbb{E}_z \langle \nabla F_{z}(X_{k}), V_k \rangle + \frac{1}{2} M \eta^2 \mathbb{E}_z \|V_k\|^2.
\]
Finally, we can compute that
\[
\mathbb{E}_z\|X_{k+1} + \gamma^{-1}V_{k+1}\|^2
\]
\[
= \mathbb{E}_z\|X_k + \gamma^{-1}V_k - \eta \gamma^{-1}g(X_k, U_{z,k}) + \sqrt{2\gamma^{-1}\beta^{-1}\eta k}\|^2
\]
\[
= \mathbb{E}_z\|X_k + \gamma^{-1}V_k - \eta \gamma^{-1}g(X_k, U_{z,k})\|^2 + 2\gamma^{-1}\beta^{-1}\eta d
\]
\[
= \mathbb{E}_z\|X_k + \gamma^{-1}V_k - \eta \gamma^{-1}\nabla F_z(X_k)\|^2 + 2\gamma^{-1}\beta^{-1}\eta d
\]
\[
+ \mathbb{E}_z[\eta \gamma^{-1}g(X_k, U_{z,k}) - \eta \gamma^{-1}\nabla F_z(X_k)]^2
\]
\[
\leq \mathbb{E}_z\|X_k + \gamma^{-1}V_k - \eta \gamma^{-1}\nabla F_z(X_k)\|^2 + 2\gamma^{-1}\beta^{-1}\eta d + 2\eta^2 \gamma^{-2}\delta(M^2\mathbb{E}_z\|X_k\|^2 + B^2)
\]
\[
= \mathbb{E}_z\|X_k + \gamma^{-1}V_k\|^2 - 2\eta \gamma^{-1}\mathbb{E}_z\langle X_k + \gamma^{-1}V_k, \nabla F_z(X_k) \rangle
\]
\[
+ \eta^2 \gamma^{-2}\mathbb{E}_z\|\nabla F_z(X_k)\|^2 + 2\gamma^{-1}\beta^{-1}\eta d + 2\eta^2 \gamma^{-2}\delta(M^2\mathbb{E}_z\|X_k\|^2 + B^2), \tag{A.14}
\]
where we have used part (iv) of Assumption 1 in the inequality above.

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

\[(L_2(k + 1) - L_2(k))/\eta\]
\[
= \left(\mathbb{E}_z[F_z(X_{k+1})] - \mathbb{E}_z[F_z(X_k)] + \frac{1}{4} \gamma^2 (\mathbb{E}_z\|X_{k+1} + \gamma^{-1}V_{k+1}\|^2 - \mathbb{E}_z\|X_k + \gamma^{-1}V_k\|^2)
\]
\[
+ \frac{1}{4} (\mathbb{E}_z\|V_{k+1}\|^2 - \mathbb{E}_z\|V_k\|^2) - \frac{1}{4} \gamma^2 \lambda (\mathbb{E}_z\|X_{k+1}\|^2 - \mathbb{E}_z\|X_k\|^2) \right)/\eta
\]
\[
\leq \mathbb{E}_z\langle \nabla F_z(X_k), V_k \rangle + \frac{1}{2} M\eta \mathbb{E}_z\|V_k\|^2 - \frac{1}{2} \gamma \mathbb{E}\langle X_k + \gamma^{-1}V_k, \nabla F_z(X_k) \rangle
\]
\[
+ \frac{1}{4} \eta \mathbb{E}\|\nabla F_z(X_k)\|^2 + \frac{1}{2} \gamma \beta^{-1} d + \frac{1}{2} \eta \delta(M^2\mathbb{E}\|X_k\|^2 + B^2)
\]
\[
+ \frac{1}{4} (-2\gamma + \eta \gamma^2) \mathbb{E}_z\|V_k\|^2 - \frac{1}{2} (1 - \eta \gamma) \mathbb{E}_z]\langle V_k, \nabla F_z(X_k) \rangle]
\]
\[
+ \frac{1}{4} \eta (M^2\mathbb{E}_z\|X_k\|^2 + B^2 + 2MB\mathbb{E}_z\|X_k\|) + \frac{1}{2} \gamma \beta^{-1} d
\]
\[
+ \frac{1}{2} \delta \eta M^2\mathbb{E}_z\|X_k\|^2 + \frac{1}{2} \delta \eta B^2 - \frac{1}{2} \gamma^2 \lambda \mathbb{E}_z(X_k, V_k) - \frac{1}{4} \gamma^2 \lambda \eta \mathbb{E}_z\|V_k\|^2
\]
\[
= -\frac{\gamma}{2} \mathbb{E}_z\langle \nabla F_z(X_k), X_k \rangle - \frac{\gamma}{2} \mathbb{E}_z\|V_k\|^2 - \frac{\gamma^2 \lambda}{2} \mathbb{E}_z(X_k, V_k) + \gamma \beta^{-1} d + \mathcal{E}_k \eta
\]
\[
\leq -\gamma \lambda \mathbb{E}_z[F_z(X_k)] - \frac{1}{4} \lambda \gamma^3 \mathbb{E}_z\|X_k\|^2 + \gamma A/\beta - \frac{\gamma^2 \lambda}{2} \mathbb{E}_z\|V_k\|^2 - \frac{\gamma^2 \lambda}{2} \mathbb{E}_z(X_k, V_k) + \gamma \beta^{-1} d + \mathcal{E}_k \eta, \tag{A.15}
\]
where we used the drift condition (2.2) in the last inequality, and
\[
\mathcal{E}_k := \left(\frac{1}{2} M + \frac{1}{4} \gamma^2 - \frac{1}{4} \lambda \gamma^2 \right) \mathbb{E}_z\|V_k\|^2 + \frac{1}{4} \mathbb{E}_z\|\nabla F_z(X_k)\|^2 + \delta(M^2\mathbb{E}_z\|X_k\|^2 + B^2)
\]
\[
+ \frac{1}{2} \gamma \mathbb{E}_z(\langle V_k, \nabla F_z(X_k) \rangle) + \frac{1}{4} (M^2\mathbb{E}_z\|X_k\|^2 + B^2 + 2MB\mathbb{E}_z\|X_k\|). \]
We can upper bound $\mathcal{E}_k$ as follows:

$$\begin{align*}
\mathcal{E}_k &\leq \left( \frac{1}{2} M + \frac{1}{4} \gamma^2 - \frac{1}{4} \gamma^2 \lambda \right) \mathbb{E}_x \|V_k\|^2 + \mathbb{E}_x \|\nabla F_x(X_k)\|^2 + \delta (M^2 \mathbb{E}_x \|X_k\|^2 + B^2) \\
&\quad + \gamma \mathbb{E}_x \|V_k\|^2 + \gamma \mathbb{E}_x \|\nabla F_x(X_k)\|^2 + \mathbb{E}_x (M \|X_k\| + B)^2 \\
&\leq \left( \frac{1}{2} M + \frac{1}{4} \gamma^2 - \frac{1}{4} \gamma^2 \lambda + \gamma \right) \mathbb{E}_x \|V_k\|^2 + \delta (M^2 \mathbb{E}_x \|X_k\|^2 + B^2) \\
&\quad + \left( \frac{1}{4} + \gamma \right) \mathbb{E}_x (M \|X_k\| + B)^2 + \mathbb{E}_x (M \|X_k\| + B)^2 \\
&\leq \left( \frac{1}{2} M + \frac{1}{4} \gamma^2 - \frac{1}{4} \gamma^2 \lambda + \gamma \right) \mathbb{E}_x \|V_k\|^2 \\
&\quad + 2M^2 \left( \frac{1}{2} + \gamma + \delta \right) \mathbb{E}_x ||X_k||^2 + 2B^2 \left( \frac{1}{2} + \gamma + \delta \right). 
\end{align*}$$

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

$$L_2(k) \geq \max \left\{ \frac{1}{8} (1 - 2\lambda) \gamma^2 \mathbb{E}_x ||X_k||^2, \frac{1}{4} (1 - 2\lambda) \mathbb{E}_x \|V_k\|^2 \right\}$$

(A.16)

$$\geq \frac{1}{16} (1 - 2\lambda) \gamma^2 \mathbb{E}_x ||X_k||^2 + \frac{1}{8} (1 - 2\lambda) \mathbb{E}_x \|V_k\|^2.$$ 

Therefore,

$$\mathcal{E}_k \leq K_1 L_2(k) + K_2,$$  \hspace{1cm} \text{(A.17)}

where we recall from (3.3) and (3.4) 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 $\lambda \leq \frac{1}{4}$, we infer from the definition of $L_2(k)$ in (A.10) that

$$L_2(k) = \mathbb{E}_x [F_x(X_k)] + \frac{1}{4} \gamma^2 (1 - \lambda) \mathbb{E}_x \|X_k\|^2 + \frac{1}{2} \gamma \mathbb{E}_x [\langle X_k, V_k \rangle] + \frac{1}{2} \mathbb{E}_x \|V_k\|^2 \\
\leq \mathbb{E}_x [F_x(X_k)] + \frac{1}{4} \gamma^2 \mathbb{E}_x \|X_k\|^2 + \frac{1}{2} \gamma \mathbb{E}_x [\langle X_k, V_k \rangle] + \frac{1}{2} \mathbb{E}_x \|V_k\|^2.$$ 

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

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

37
For $0 < \eta \leq \min \left\{ \frac{\sqrt{2}}{K_2} (d/\beta + A/\beta), \frac{\gamma \lambda}{2K_1} \right\}$, we get

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

which implies

\[
L_2(k + 1) \leq \rho L_2(k) + K,
\]

where

\[
\rho := 1 - \eta \gamma \lambda / 2, \quad K := 2\eta \gamma (d/\beta + A/\beta).
\]

It follows that

\[
L_2(k) \leq L_2(0) + \frac{K}{1 - \rho} = \mathbb{E}_x [\mathcal{V}(X_0, V_0)/\beta] + \frac{4(d/\beta + A/\beta)}{\lambda}.
\]

The result then follows from the inequality above and (A.16). \hfill \Box

### B Proof of Lemma 9

**Proof.** From (A.1)–(A.3), we can directly obtain that

\[
\mathcal{L} e^{\alpha \mathcal{V}} = \left[ - (\gamma v + \nabla F_x(x)) \alpha \nabla_v \mathcal{V} + \gamma \beta^{-1} \alpha \Delta_v \mathcal{V} + \gamma \beta^{-1} \alpha^2 \| \nabla_v \mathcal{V} \|^2 + v \alpha \nabla_x \mathcal{V} \right] e^{\alpha \mathcal{V}}
\]

\[
= \left[ \alpha \mathcal{L} \mathcal{V} + \gamma \beta^{-1} \alpha^2 \| \nabla_v \mathcal{V} \|^2 \right] e^{\alpha \mathcal{V}}
\]

\[
\leq \alpha \gamma d + \alpha \gamma A - \alpha \gamma \lambda \mathcal{V} + \alpha^2 \gamma \beta^{-1} \| \nabla_v \mathcal{V} \|^2 \right] e^{\alpha \mathcal{V}}. 
\]  

(B.1)

Moreover, we recall from (A.6) that

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

and thus

\[
\| \nabla_v \mathcal{V} \|^2 \leq 2\beta^2 \| v \|^2 + \frac{\beta^2 \gamma^2}{2} \| x \|^2.
\]

We recall from (A.4) that

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

Therefore, we have

\[
\| \nabla_v \mathcal{V} \|^2 \leq \left[ \frac{8\beta^2}{\beta(1 - 2\lambda)} + \frac{4\beta^2 \gamma^2}{(1 - 2\lambda) \beta \gamma^2} \right] \mathcal{V} = \frac{12\beta}{1 - 2\lambda} \mathcal{V}.
\]  

(B.2)
By choosing:
\[
\alpha = \frac{\lambda \beta}{1 - 2\lambda} = \frac{\lambda(1 - 2\lambda)}{12},
\] (B.3)
we get
\[
\mathcal{L}e^{\alpha V} \leq \alpha \gamma(d + A)e^{\alpha V}.
\] (B.4)

Since \(\mathcal{L}e^{\alpha V} = [\mathcal{L}e^{\alpha V} + \gamma \beta^{-1} \|\nabla_v \alpha V\|^2]e^{\alpha V}\), we have showed that
\[
\mathcal{L}e^{\alpha V} + \gamma \beta^{-1} \|\nabla_v \alpha V\|^2 \leq \alpha \gamma(d + A).
\]

Applying an exponential integrability result, e.g. Corollary 2.4 [CHJ13], we get
\[
\mathbb{E}[e^{\alpha V(X(t),V(t))}] \leq \mathbb{E}[e^{\alpha V(X(0),V(0))}]e^{\alpha \gamma(d + A)t}.
\] (B.5)

That is,
\[
\mathbb{E}_z[e^{\alpha V(X(t),V(t))}] \leq \int_{\mathbb{R}^{2d}} e^{\alpha V(x,v) + \alpha \gamma(d + A)t} \mu_0(dx, dv) < \infty.
\] (B.5)

Next, applying Itô’s formula to \(e^{\frac{1}{2} \alpha V(X(t),V(t))}\), we obtain
\[
e^{\frac{1}{2} \alpha V(X(t),V(t))} = e^{\frac{1}{2} \alpha V(X(0),V(0))} + \int_0^t \mathcal{L}e^{\frac{1}{2} \alpha V(X(s),V(s))} ds
\[
+ \int_0^t \frac{1}{2} \left( \beta V(s) + \frac{\beta \gamma}{2} X(s) \right) e^{\frac{1}{2} \alpha V(X(s),V(s))} dB(s). \] (B.6)

For every \(T > 0\),
\[
\int_0^T \mathbb{E} \left\| \frac{1}{2} \left( \beta V(s) + \frac{\beta \gamma}{2} X(s) \right) e^{\frac{1}{2} \alpha V(X(s),V(s))} \right\|^2 ds
\[
\leq \frac{\beta^2}{2} \int_0^T \mathbb{E} \left[ \left( \|V(s)\|^2 + \frac{\beta^2}{4} \|X(s)\|^2 \right) e^{\frac{1}{2} \alpha V(X(s),V(s))} \right] ds
\[
\leq \frac{6\beta}{1 - 2\lambda} \int_0^T \mathbb{E} \left[ V(X(s), V(s))e^{\frac{1}{2} \alpha V(X(s),V(s))} \right] ds
\[
\leq \frac{12\beta}{1 - 2\lambda} \int_0^T \mathbb{E} \left[ e^{\alpha V(X(s),V(s))} \right] ds < \infty,
\]
where we used (A.4) and (B.5). Thus, \(\int_0^T \frac{1}{2} \left( \beta V(s) + \frac{\beta \gamma}{2} X(s) \right) e^{\frac{1}{2} \alpha V(X(s),V(s))} dB(s)\) is a martingale. By taking expectations on both hand sides of (B.6), we get
\[
\mathbb{E}[e^{\frac{1}{2} \alpha V(X(s),V(s))}] = \mathbb{E}[e^{\frac{1}{2} \alpha V(X(0),V(0))}] + \int_0^t \mathbb{E}[\mathcal{L}e^{\frac{1}{2} \alpha V(X(s),V(s))}] ds. \] (B.7)
From (B.1), (B.2) and (B.3), we can infer that

\[
\mathcal{L} e^{\frac{1}{4} \alpha V} \leq \left( \frac{1}{4} \alpha \gamma (d + A) - \frac{1}{4} \alpha \gamma \lambda V + \gamma \beta^{-1} \frac{\alpha^2}{16} \|\nabla_v V\|^2 \right) e^{\frac{1}{4} \alpha V} \\
\leq \left( \frac{1}{4} \alpha \gamma (d + A) - \frac{3}{16} \alpha \gamma \lambda V \right) e^{\frac{1}{4} \alpha V} \\
\leq \frac{1}{4} \alpha \gamma (d + A) e^{\frac{\alpha(d + A)}{4\lambda}}.
\]

where in the last inequality we used the facts that \( V \geq 0 \) and \( \frac{1}{4} \alpha \gamma (d + A) - \frac{3}{16} \alpha \gamma \lambda V \geq 0 \) if and only if \( V \leq \frac{4(d + A)}{3\lambda} \). Therefore, it follows from (B.7) that

\[
E[e^{\frac{1}{4} \alpha V(X_t, V_t)}] \leq E[e^{\frac{1}{4} \alpha V(X_0, V_0)}] + \frac{1}{4} e^{\frac{\alpha(d + A)}{4\lambda}} \alpha \gamma (d + A) t.
\]

Finally, by (A.4) again,

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

Hence, the conclusion follows. \( \square \)

C Proof of Lemma 10

Proof of Lemma 10. We follow similar steps as in the proof of Lemma 7 in Raginsky et al. [RRT17]. Consider the following continuous-time interpolation of \((X_k, V_k)\):

\[
\begin{align*}
\overline{V}(t) &= V_0 - \int_0^t \gamma \overline{V}([s/\eta] \eta) ds - \int_0^t g(\overline{X}([s/\eta] \eta), \overline{U}_z(s)) ds + \sqrt{2\gamma \beta^{-1}} \int_0^t dB(s), \\
\overline{X}(t) &= X_0 + \int_0^t \overline{V}([s/\eta] \eta) ds,
\end{align*}
\]

where \( \overline{U}_z(t) := U_{z,k} \) 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 \( \mu_{z,k} \) for each \( k \geq 0 \). In addition, note that the (non-Markov) process \((\overline{V}, \overline{X})\) has the same marginals as the Markov process \((\tilde{V}, \tilde{X})\) (see e.g. [Gyö86]):

\[
\begin{align*}
\tilde{V}(t) &= V_0 - \int_0^t \gamma \tilde{V}(s) ds - \int_0^t g_{z,t}(\tilde{V}(s), \tilde{X}(s)) ds + \sqrt{2\gamma \beta^{-1}} \int_0^t dB(s), \quad (C.1) \\
\tilde{X}(t) &= X_0 + \int_0^t \tilde{V}(s) ds, \quad (C.2)
\end{align*}
\]

where

\[
g_{z,t}(v, x) := \mathbb{E}_z \left[ g(\overline{X}([t/\eta] \eta), \overline{U}_z(t)) | \overline{V}(t) = v, \overline{X}(t) = x \right]. \quad (C.3)
\]
Let $\mathbb{P}$ be the probability measure associated with the undamped 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.1)–(C.2). 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 formula (see e.g. Section 7.6.4. in [LS13]):

$$
\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}}|_{\mathcal{F}_t} = e^{\frac{\beta}{2\gamma} \int_0^t (\nabla F_x(\tilde{X}(s)) - g_{x,s}(\tilde{V}(s), \tilde{X}(s))) dB(s) - \frac{\beta}{2\gamma} \int_0^t \|\nabla F_x(\tilde{X}(s)) - g_{x,s}(\tilde{V}(s), \tilde{X}(s))\|^2 ds}.
$$

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

$$
D(\tilde{\mathbb{P}}_t\|\mathbb{P}_t) := - \int d\tilde{\mathbb{P}}_t \log \frac{d\mathbb{P}_t}{d\tilde{\mathbb{P}}_t}
\quad = \frac{\beta}{4\gamma} \int_0^t \mathbb{E}_x \|\nabla F_x(\tilde{X}(s)) - g_{x,s}(\tilde{V}(s), \tilde{X}(s))\|^2 ds
\quad = \frac{\beta}{4\gamma} \int_0^t \mathbb{E}_x \|\nabla F_x(\tilde{X}(s)) - g_{x,s}(\tilde{V}(s), \tilde{X}(s))\|^2 ds.
$$

Then, we get

$$
D(\tilde{\mathbb{P}}_t\|\mathbb{P}_t) = \frac{\beta}{4\gamma} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}_x \|\nabla F_x(\tilde{X}(s)) - g_{x,s}(\tilde{V}(s), \tilde{X}(s))\|^2 ds,
$$

where in the first inequality above, we used the definition (C.3) and Jensen’s inequality for conditional expectations. We first bound the first term in (C.4):

$$
\frac{\beta}{2\gamma} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}_x \|\nabla F_x(\tilde{X}(s)) - \nabla F_x(\tilde{X}([s/\eta]\eta))\|^2 ds
\quad \leq \frac{\beta M^2}{2\gamma} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}_x \|\tilde{X}(s) - \tilde{X}([s/\eta]\eta)\|^2 ds.
$$

41
Consider some \( j\eta \leq s < (j + 1)\eta \), then we have \( \overline{X}(s) - \overline{X}(j\eta) = (s - j\eta)\overline{V}(j\eta) \), which has the same distribution as \( (s - j\eta)\overline{V}_j \). Hence, we obtain

\[
\frac{\beta}{2\gamma} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}_{\mathbf{z}} \left\| \nabla F_{\mathbf{z}}(\overline{X}(s)) - \nabla F_{\mathbf{z}}(\overline{X}([s/\eta]\eta)) \right\|^2 ds \leq \frac{\beta M^2}{2\gamma} \sup_{j \geq 0} \mathbb{E}_{\mathbf{z}}[\|V_j\|^2] \eta^2 k. \tag{C.5}
\]

We can also bound the second term in \((C.4)\):

\[
\frac{\beta}{2\gamma} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}_{\mathbf{z}} \left\| \nabla F_{\mathbf{z}}(\overline{X}([s/\eta]\eta)) - g(\overline{X}([s/\eta]\eta), U_{\mathbf{z}}(s)) \right\|^2 ds
\]

\[
= \frac{\beta}{2\gamma} \eta \sum_{j=0}^{k-1} \mathbb{E}_{\mathbf{z}} \left\| \nabla F_{\mathbf{z}}(X_j) - g(X_j, U_{\mathbf{z},j}) \right\|^2
\]

\[
\leq \frac{\beta}{2\gamma} \eta \delta \sum_{j=0}^{k-1} 2(M^2\mathbb{E}_{\mathbf{z}}\|X_j\|^2 + B^2)
\]

\[
\leq \left( M^2 \sup_{j \geq 0} \mathbb{E}_{\mathbf{z}}\|X_j\|^2 + B^2 \right) \frac{\beta}{\gamma} \eta \delta,
\]

where the first inequality follows from part (iv) of Assumption 1. Hence, we conclude that

\[
D(\tilde{\mathbb{P}}_{k\eta} \| \mathbb{P}_{k\eta}) \leq \frac{\beta M^2}{2\gamma} \sup_{j \geq 0} \mathbb{E}_{\mathbf{z}}[\|V_j\|^2] \eta^2 k + \left( M^2 \sup_{j \geq 0} \mathbb{E}_{\mathbf{z}}\|X_j\|^2 + B^2 \right) \frac{\beta}{\gamma} k\eta \delta.
\]

Hence, it follows from Lemma 8 that

\[
D(\mu_{z,k} \| \nu_{z,k\eta}) \leq \frac{\beta M^2}{2\gamma} C^d k\eta^2 + \left( M^2 C^d_x + B^2 \right) \frac{\beta}{\gamma} k\eta \delta.
\]

We can then apply the following result of Bolley and Villani [BV05], that is, for any two Borel probability measures \( \mu, \nu \) on \( \mathbb{R}^d \) with finite second moments,

\[
W_2(\mu, \nu) \leq C_{\nu} \left[ \sqrt{D(\mu \| \nu)} + \left( \frac{D(\mu \| \nu)}{2} \right)^{1/4} \right],
\]

where

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

From the exponential integrability of the measure \( \nu_{z,k\eta} \) in Lemma 9, we have

\[
C_{\nu_{z,k\eta}} \leq 2 \left( \frac{1}{\alpha_0} \left( \frac{3}{2} + \log \int_{\mathbb{R}^d} e^{\alpha_0 \|x,v\|^2} \nu_{z,k\eta}(dx,dv) \right) \right)^{1/2}
\]

\[
\leq 2 \left( \frac{1}{\alpha_0} \left( \frac{3}{2} + \log \left( \int_{\mathbb{R}^d} e^{\frac{1}{4}\alpha_0 \|x,v\|^2} \mu_0(dx,dv) + \frac{1}{4} e^{\frac{\alpha_0 (d+A)}{4} } \alpha_0 (d + A) k\eta \right) \right) \right)^{1/2}.
\]
Hence
\[ W_2^2(\mu_{z,k}, \nu_{z,k}) \leq \frac{4}{\alpha_0} \left( \frac{3}{2} + \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{4} \alpha V(x,v) \mu_0(dx, dv) + \frac{1}{4} e^{-\frac{\alpha(d+A)}{\alpha} \alpha \gamma(d + A)k \eta}} \right) \right) \]
\[ \cdot \left[ \sqrt{D(\mu_{z,k} \| \nu_{z,k})} + \left( \frac{D(\mu_{z,k} \| \nu_{z,k})}{2} \right)^{1/4} \right]^2, \tag{C.6} \]
where
\[ D(\mu_{z,k} \| \nu_{z,k}) \leq \frac{\beta M^2}{2\gamma} C_\nu^d k \eta^2 + \left( M^2 C_x^d + B^2 \right) \frac{\beta}{\gamma} k \eta \delta. \tag{C.7} \]
By using \((x + y)^2 \leq 2(x^2 + y^2)\), we get
\[ W_2^2(\mu_{z,k}, \nu_{z,k}) \leq \frac{8}{\alpha_0} \left( \frac{3}{2} + \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{4} \alpha V(x,v) \mu_0(dx, dv) + \frac{1}{4} e^{-\frac{\alpha(d+A)}{\alpha} \alpha \gamma(d + A)k \eta}} \right) \right) \]
\[ \cdot \left[ D(\mu_{z,k} \| \nu_{z,k}) + \sqrt{D(\mu_{z,k} \| \nu_{z,k})} \right], \tag{C.8} \]
where
\[ D(\mu_{z,k} \| \nu_{z,k}) \leq \left( M^2 C_x^d + B^2 \right) \frac{\beta}{\gamma} \delta + \frac{\beta M^2}{2\gamma} C_\nu^d \right) k \eta. \]
Since \( k \eta \geq e > 1 \), we get
\[ \frac{8}{\alpha_0} \left( \frac{3}{2} + \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{4} \alpha V(x,v) \mu_0(dx, dv) + \frac{1}{4} e^{-\frac{\alpha(d+A)}{\alpha} \alpha \gamma(d + A)k \eta}} \right) \right) \]
\[ \leq \frac{8}{\alpha_0} \left( \frac{3}{2} + \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{4} \alpha V(x,v) \mu_0(dx, dv) + \frac{1}{4} e^{-\frac{\alpha(d+A)}{\alpha} \alpha \gamma(d + A)k \eta}} \right) \right) + \log(k \eta) \]
\[ \leq \frac{8}{\alpha_0} \left( \frac{3}{2} + \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{4} \alpha V(x,v) \mu_0(dx, dv) + \frac{1}{4} e^{-\frac{\alpha(d+A)}{\alpha} \alpha \gamma(d + A)k \eta}} \right) \right) + 1 \right) \log(k \eta) \tag{C.9} \]
and
\[ D(\mu_{z,k} \| \nu_{z,k}) + \sqrt{D(\mu_{z,k} \| \nu_{z,k})} \]
\[ \leq \left( \frac{\beta M^2}{2\gamma} C_\nu^d + \sqrt{\frac{\beta M^2}{2\gamma} C_\nu^d} \right) k \eta^{3/2} + \left( M^2 C_x^d + B^2 \right) \frac{\beta}{\gamma} \delta + \sqrt{M^2 C_x^d + B^2} \frac{\beta}{\gamma} \right) k \eta \delta, \]
which implies that
\[ W_2^2(\mu_{z,k}, \nu_{z,k}) \leq (\hat{C}_0^2 \sqrt{\delta} + \hat{C}_1^2 \sqrt{\eta})(k \eta) \log(k \eta), \]
where \( \hat{C}_0 \) and \( \hat{C}_1 \) are defined in (3.8) and (3.9). The result then follows from the fact that 
\( \sqrt{x + y} \leq \sqrt{x} + \sqrt{y} \) for non-negative real numbers \( x \) and \( y \). \( \Box \)
D Proof of Lemma 12

Proof. We recall first from (A.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 [EGZ17])
\[ h(r) \leq \min\{r, h(R_1)\} \leq \min\{r, R_1\}, \quad \text{for any } r \geq 0. \]

It follows that
\[ \rho((x,v),(x',v')) \leq \min\{r((x,v),(x',v'))(1+\varepsilon_1)\mathcal{V}(x,v) + \varepsilon_1 \mathcal{V}(x',v')\} \leq R_1(1+\varepsilon_1)\mathcal{V}(x,v) + \varepsilon_1 \mathcal{V}(x',v'). \]

Moreover, by the definition of \( \mathcal{V} \) in (2.1) and Lemma 13, we deduce that
\[ \mathcal{V}(x,v) \leq \beta \left( \frac{M}{2} \|x\|^2 + B \|x\| + A_0 \right) + \frac{1}{4}\beta \gamma^2 (\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda \|x\|^2) \]
\[ \leq \beta \left( \frac{M}{2} \|x\|^2 + B \|x\| + A_0 \right) + \frac{1}{4}\beta \gamma^2 (2\|x\|^2 + 2\gamma^{-2}\|v\|^2 + \|\gamma^{-1}v\|^2 - \lambda \|x\|^2) \]
\[ \leq \beta \left( \frac{M}{2} \|x\|^2 + A_0 + \frac{B^2}{2M} \right) + \frac{1}{4}\beta \gamma^2 (2\|x\|^2 + 2\gamma^{-2}\|v\|^2 + \|\gamma^{-1}v\|^2 - \lambda \|x\|^2) \]
\[ \leq \left( \beta \frac{M}{2} + \frac{1}{2}\beta \gamma^2 \right) \|x\|^2 + \frac{3}{4}\beta \|v\|^2 + \beta A_0 + \frac{\beta B^2}{2M}. \]

Therefore, we obtain
\[ \mathcal{H}_\rho(\mu_0, \pi_x) \]
\[ \leq R_1 + R_1 \varepsilon_1 \left( \left( \frac{M}{2} + \frac{1}{2}\beta \gamma^2 \right) \int_{\mathbb{R}^{2d}} \|x\|^2 \mu_0(dx, dv) + \frac{3}{4}\beta \int_{\mathbb{R}^{2d}} \|v\|^2 \mu_0(dx, dv) + \beta A_0 + \frac{\beta B^2}{2M} \right) \]
\[ + R_1 \varepsilon_1 \left( \left( \frac{M}{2} + \frac{1}{2}\beta \gamma^2 \right) \int_{\mathbb{R}^{2d}} \|x\|^2 \pi_x(dx, dv) + \frac{3}{4}\beta \int_{\mathbb{R}^{2d}} \|v\|^2 \pi_x(dx, dv) + \beta A_0 + \frac{\beta B^2}{2M} \right). \]

It has been shown in [RRT17, Section 3.5] that
\[ \int_{\mathbb{R}^{2d}} \|x\|^2 \pi_x(dx, dv) \leq \frac{b + d/\beta}{m}. \]

In addition, from the explicit expression of \( \pi_x(dx, dv) \) in (1.7), we have
\[ \int_{\mathbb{R}^{2d}} \|v\|^2 \pi_x(dx, dv) = (2\pi \beta^{-1})^{-d/2} \int_{\mathbb{R}^d} \|v\|^2 e^{-\|v\|^2/(2\beta^{-1})} dv = d/\beta. \]

Hence, the conclusion follows from (D.1).  \( \square \)
E Supporting Lemmas

The first lemma shows that $f$ admits lower and upper bounds that are quadratic functions.

**Lemma 13** (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 Raginsky et al. [RRT17] to study the SGLD dynamics.

**Lemma 14** (See [PW16]). Let $\mu, \nu$ 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,
$$\left| \int_{\mathbb{R}^{2d}} Gd\mu - \int_{\mathbb{R}^{2d}} Gd\nu \right| \leq (c_1\sigma + c_2)W_2(\mu, \nu),$$
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 $\pi_z$. Note that the $x-$ marginal of $\pi_z(dx, dv)$ for the underdamped diffusion is the same as the stationary distribution for the overdamped diffusion studied in [RRT17]. For two $n-$tuples $z = (z_1, \ldots, z_n), \bar{z} = (\bar{z}_1, \ldots, \bar{z}_n) \in \mathbb{Z}^n$, we say $z$ and $\bar{z}$ differ only in a single coordinate if $\operatorname{card}\{|i : z_i \neq \bar{z}_i| = 1\}$.

**Lemma 15** (Proposition 12, [RRT17]). For any two $z, \bar{z} \in \mathbb{Z}^n$ that differ only in a single coordinate,
$$\sup_{z \in \mathbb{Z}} \left| \int_{\mathbb{R}^{2d}} f(x, z)\pi_z(dx, dv) - \int_{\mathbb{R}^{2d}} f(x, \bar{z})\pi_{\bar{z}}(dx, dv) \right| \leq \frac{4\beta c_{LS}}{n} \left( \frac{M^2}{m}(b + d/\beta) + B^2 \right),$$
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 $\lambda_*$ is the uniform spectral gap for overdamped Langevinian dynamics:
$$\lambda_* = \inf_{z \in \mathbb{Z}^n} \inf \left\{ \frac{\int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi_z}{\int_{\mathbb{R}^d} g^2 d\pi_z} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_z), g \neq 0, \int_{\mathbb{R}^d} g d\pi_z = 0 \right\}.$$
The next lemma shows that for large values of \( \beta \), the \( x \)-marginal of the stationary distribution \( \pi_z(dx, dv) \) is concentrated at the minimizer of \( F_z \).

**Lemma 16** (Proposition 11, [RRT17]). *It holds that*

\[
\int_{\mathbb{R}^{2d}} F_z(x) \pi_z(dx, dv) - \min_{x \in \mathbb{R}^d} F_z(x) \leq \frac{d}{2\beta} \log \left( \frac{eM}{m} \left( \frac{b\beta}{d} + 1 \right) \right).
\]

**F Further Results on Diffusion Approximation**

In this section, we discuss the improvement on the diffusion approximation result in Lemma 10 under some stronger assumptions, and hence we will get better upper bounds for the main results. Indeed, we can improve the diffusion approximation result in Lemma 10 to get

\[
W_2(\mu_{z,k}, \nu_{z,k}) \leq C_0 \cdot (k\eta)^{1/4} \sqrt{\log(k\eta)} \cdot \delta^{1/4} + C_1 \cdot (k\eta)^{1/4} \sqrt{\log(k\eta)} \cdot \eta^{1/4}, \tag{F.1}
\]

where

\[
C_0 = \sqrt{2\hat{\gamma}} \left( M^2 C_x^d + B^2 \right)^{1/4} \left( \frac{\beta}{\gamma} \right)^{1/4}, \quad C_1 = \sqrt{2\hat{\gamma}} \left( \frac{\beta M^2 C_v^d}{2\gamma} \right)^{1/4},
\]

where \( \hat{\gamma} \) is defined in (3.10) provided that \( \delta, \eta \) are small enough so that the relative entropy \( D(\mu_{z,k} \parallel \nu_{z,k}) \leq 1 \). Next, let us prove (F.1). We recall the bounds on \( W_2(\mu_{z,k}, \nu_{z,k}) \) in (C.6) and \( D(\mu_{z,k} \parallel \nu_{z,k}) \) in (C.7). We choose \( \eta \) and \( \delta \) sufficiently small so that

\[
D(\mu_{z,k} \parallel \nu_{z,k}) \leq \left[ \left( M^2 C_x^d + B^2 \right) \frac{\beta}{\gamma} \delta + \frac{\beta M^2 C_v^d}{2\gamma} \eta \right] k\eta \leq 1.
\]

Then it follows from (C.8) and (C.9) that

\[
W_2^2(\mu_{z,k}, \nu_{z,k}) \\
\leq \frac{8}{\alpha_0} \left( \frac{3}{2} + \frac{1}{2} \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{2} \alpha V(x,v)} \mu_0(dx, dv) + \frac{1}{4} e^{\frac{\alpha(d+A)}{3\lambda}} \alpha \gamma (d + A) \right) + 1 \right) \log(k\eta)
\]

\[
\cdot \frac{1}{2} \left( M^2 C_x^d + B^2 \right)^{1/2} \frac{\beta}{\gamma} \delta + \frac{\beta M^2 C_v^d}{2\gamma} \eta \sqrt{k\eta},
\]

\[
\leq \frac{8}{\alpha_0} \left( \frac{3}{2} + \frac{1}{2} \log \left( \int_{\mathbb{R}^{2d}} e^{\frac{1}{2} \alpha V(x,v)} \mu_0(dx, dv) + \frac{1}{4} e^{\frac{\alpha(d+A)}{3\lambda}} \alpha \gamma (d + A) \right) + 1 \right) \log(k\eta)
\]

\[
\cdot \frac{1}{2} \left( M^2 C_x^d + B^2 \right)^{1/2} \sqrt{\frac{\beta}{\gamma} \delta + \sqrt{\frac{\beta M^2 C_v^d}{2\gamma}}} \sqrt{k\eta}.
\]

\footnote{In Proposition 11, [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.}
Hence the bound (F.1) follows.

As a result, given $k\eta = \mu_*^{-1}\log(\varepsilon^{-1}) \geq \varepsilon$ and $\eta \leq \min \left\{ \frac{\varepsilon^4}{\log(1/\varepsilon)}, \frac{\gamma}{K_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2K_1} \right\}$.

In addition to Assumption 1, if we assume that

$$\frac{\beta M^2}{2\gamma} C_d^d \frac{1}{\mu_*} \log \left( \frac{1}{\varepsilon} \right) \frac{\varepsilon^4}{\log(1/\varepsilon)} + \left( M^2 C_d^d + B^2 \right) \frac{\beta}{\gamma} \frac{1}{\mu_*} \log \left( \frac{1}{\varepsilon} \right) \delta \leq 1, \quad (F.2)$$

(so that $D(\mu_{z,k}||\nu_{z,k\eta}) \leq 1$, that is, the relative entropy (i.e. KL-divergence) between the law $\mu_{z,k}$ of $(X_k, V_k)$ and the law $\nu_{z,k\eta}$ of $(X(k\eta), V(k\eta))$ is no greater than 1) then, the bound on the both empirical risk and population risk can be improved by replacing $J_1(\varepsilon)$ by $J_1(\varepsilon)$ as below:

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

where

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

and $C_0, C_1$ are defined as:

$$C_0 = \sqrt{2\hat{\gamma}} \left( M^2 C_d^d + B^2 \right)^{1/4} \left( \frac{\beta}{\gamma} \right)^{1/4}, \quad C_1 = \sqrt{2\hat{\gamma}} \left( \frac{\beta M^2}{2\gamma} C_d^d \right)^{1/4},$$

where $\hat{\gamma}$ is defined in (3.10).

### 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 2.4. To simplify the presentation, we use the notation $\hat{O}$ to hide factors that depend on other parameters.

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

$$A = \hat{O}(\beta), \quad \alpha_1 = \hat{O}(1), \quad \Lambda = \hat{O}(d + \beta),$$

and

$$\mu_* = \hat{O}(\sqrt{d + \beta} e^{-\Lambda}) = \hat{O}(\sqrt{d + \beta} e^{-\hat{O}(d + \beta)}). \quad (G.1)$$

---

Note that in general, the condition (F.2) is very restrictive since the 1/$\mu_*$ factor on the LHS of (F 2) is exponential in dimension $d$. Nevertheless, for special classes of functions $F_z$, e.g. $F_z$ is a convolution of a Gaussian measure and a compactly supported probability measure (see e.g. [BGMZ18]), then 1/$\mu_*$ can be bounded dimension-free and (F.2) becomes less restrictive.
Since \( \varepsilon_1 = \tilde{O}(\mu_*/(d + \beta)) \), and \( \mu_* \) is exponentially small in \( \beta + d \), we get that
\[
\overline{R}_\rho(\mu_0) = \tilde{O}(R_1) = (1 + d/\beta)^{1/2}.
\]
In addition, in view of (G.1), it follows that
\[
C = \tilde{O}\left( e^{\lambda/2}(d + \beta)^{1/2} \beta^{-1/2} \mu_*^{-1/2} \right) = \tilde{O}\left( \frac{(d + \beta)^{3/4} \beta^{-1/2}}{\mu_*} \right).
\]
The structure of the initial distribution \( \mu_0(dx, dv) \) would affect the overall dependence on \( \beta, d \). If we assume that \( \mu_0(dx, dv) \) is supported on a Euclidean ball with radius being a universal constant, then the Lyapunov function \( V \) in (2.1) is linear in \( \beta \), we can then obtain
\[
\int_{\mathbb{R}^2} V(x, v) \mu_0(dx, dv) = \tilde{O}(\beta), \quad \int_{\mathbb{R}^2} e^{\alpha V(x, v)} \mu_0(dx, dv) = e^{\tilde{O}(\beta)},
\]
It follows that
\[
C^d_x = \tilde{O}\left( (\beta + d)/\beta \right), \quad C^d_v = \tilde{O}\left( (\beta + d)/\beta \right), \quad \sigma = \tilde{O}\left( \sqrt{(\beta + d)/\beta} \right).
\]
As in [RRT17], we will use \( d/\beta = \tilde{O}(1) \) here, so that \( \sigma = \tilde{O}(1) \) in Theorem 2 and in the performance comparison with respect to SGLD algorithm in Section 2.4.

Next, we have \( \alpha_0 = \tilde{O}(\beta) \) and \( \alpha = \tilde{O}(1) \), and
\[
\hat{\gamma} = \tilde{O}\left( \sqrt{(\beta + d)/\beta} \right), \quad \hat{C}_0 = \tilde{O}\left( (d + \beta)/\sqrt{\beta} \right), \quad \hat{C}_1 = \tilde{O}\left( (d + \beta)/\sqrt{\beta} \right).
\]
Hence, from (2.8), we obtain
\[
\overline{J}_0(\varepsilon) = \tilde{O}\left( \frac{d + \beta}{\mu_*^{3/4} \varepsilon} \right),
\]
and from (2.4) and (F.3), we get
\[
\overline{J}_1(\varepsilon) = \tilde{O}\left( \frac{d + \beta}{\sqrt{\beta \mu_*}} \left( \sqrt{\log(1/\varepsilon) \delta^{1/4} + \varepsilon} \right) \sqrt{\log(\mu_*^{-1} \log(\varepsilon^{-1}))} \right),
\]
and
\[
\underline{J}_1(\varepsilon) = \tilde{O}\left( \frac{\beta^{1/4}}{\mu_*^{1/4}} \left( \log^{1/4}(1/\varepsilon) \delta^{1/4} + \varepsilon \right) \sqrt{\log(\mu_*^{-1} \log(\varepsilon^{-1}))} \right),
\]
Finally, from (2.7) and (2.9), we have
\[
\overline{J}_2 = \tilde{O}\left( \frac{d}{\beta} \log(\beta + 1) \right), \quad \text{and} \quad \overline{J}_3(n) = \tilde{O}\left( \frac{1}{n} \frac{(\beta + d)^2}{\lambda_*} \right).
\]

48
Choice of $\beta$. We recall from (2.17) that SGHMC algorithm has expected empirical risk

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

after $K_{SGHMC} = \tilde{\Omega}\left( \frac{1}{\mu_* \varepsilon} \log^3(1/\varepsilon) \right)$ iterations. Note that in the worst-case, $\frac{1}{\mu_*} = e^{\tilde{O}(\beta + d)}$. This suggests the choice of $\beta \approx c_1 \log(1/\varepsilon)$, for some $c_1 > 0$ and if we have $\delta \approx \varepsilon^{c_2}$ for some $c_2 > 0$, then the first two terms in (G.2) go to zero faster than the third term and the empirical risk decays like $\tilde{O}(1/\log(1/\varepsilon))$ and the number of iterations scales like $\exp(\tilde{O}(d + \log(1/\varepsilon)))$. 