Breaking Reversibility Accelerates Langevin Dynamics for Global Non-Convex Optimization

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

Abstract

Langevin dynamics (LD) has been proven to be a powerful technique for optimizing a non-convex objective as an efficient algorithm to find local minima while eventually visiting a global minimum on longer time-scales. LD is based on the first-order Langevin diffusion which is reversible in time. We study two variants that are based on non-reversible Langevin diffusions: the underdamped Langevin dynamics (ULD) and the Langevin dynamics with a non-symmetric drift (NLD).

Adopting the techniques of Tzen, Liang and Raginsky [TLR18] for LD to non-reversible diffusions, we show that for a given local minimum that is within an arbitrary distance from the initialization, with high probability, either the ULD trajectory ends up somewhere outside a small neighborhood of this local minimum within a recurrence time which depends on the smallest eigenvalue of the Hessian at the local minimum or they enter this neighborhood by the recurrence time and stay there for a potentially exponentially long escape time. The ULD algorithms improve upon the recurrence time obtained for LD in [TLR18] with respect to the dependency on the smallest eigenvalue of the Hessian at the local minimum. Similar result and improvement are obtained for the NLD algorithm.

We also show that non-reversible variants can exit the basin of attraction of a local minimum faster in discrete time when the objective has two local minima separated by a saddle point and quantify the amount of improvement. Our analysis suggests that non-reversible Langevin algorithms are more efficient to locate a local minimum as well as exploring the state space. Our analysis is based on the quadratic approximation of the objective around a local minimum. As a by-product of our analysis, we obtain optimal mixing rates for quadratic objectives in the 2-Wasserstein distance for two non-reversible Langevin algorithms we consider.

[*]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 optimization problem:

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

where $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is a real-valued, smooth, possibly non-convex objective function with two inputs, the decision vector $x \in \mathbb{R}^d$ and a random vector Z with probability distribution P defined on a set $\mathcal{Z}$. A standard approach for solving stochastic optimization problems is to approximate the expectation as an average over independent observations $z = (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$ and to solve the problem:

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

Such problems with finite-sum structure arise frequently in many applications including data analysis and machine learning. For example, in the context of stochastic learning problems, if $z = (a, b)$ is an input-output pair of data sampled from an unknown underlying joint distribution, the function f corresponds to the loss function $f(x, z)$ of using the decision variable x and input a to predict b. This formulation encompasses a number of regression and classification problems where f can be both convex and non-convex with respect to its first argument. For example, for linear and logistic regressions, f is convex whereas f is typically non-convex for non-linear regression problems and for optimization problems arising in training of neural networks where the decision variable x corresponds to the choice of the parameters of the neural network we want to fit to the dataset z (see e.g. [GBCB16]). In this work, our primary focus will be non-convex objectives.

First-order methods such as gradient descent and stochastic gradient descent and their variants with momentum have been popular for solving large-scale optimization problems due to their empirical performance, their scalability and cheaper iteration and storage cost compared to second-order methods in high dimensions [Ber15, Bub14]. These first-order methods admit some theoretical guarantees to locate a local minimizer, however their convergence depends strongly on the initialization (a different initial point can result in convergence to a different stationary point) and they do not have guarantees to visit a global minimum. The (gradient descent) Langevin dynamics (LD) is a variant of gradient descent where a properly scaled Gaussian noise is added to the gradients:

$$X_{k+1} = X_k - \eta\nabla F(X_k) + \sqrt{2\eta\beta^{-1}}\xi_k,$$

where $\eta > 0$ is the stepsize, ξ_k is a d-dimensional anistropic Gaussian noise with distribution $\mathcal{N}(0, I)$ where for every k, the noise ξ_k is independent of the (filtration) past up to time k and β is called the *inverse temperature* parameter. With proper choice of the parameters and under mild assumptions, LD algorithm converges to a stationary distribution that concentrates around a global minimum (see e.g. [RRT17, BM99, GM91]) from an arbitrary initial point. Therefore, LD algorithm has a milder dependency on the initialization, visiting a global minimum eventually.[1]

The analysis of the convergence behavior of LD is often based on viewing LD as a discretization of the associated stochastic differential equation (SDE), known as the *overdamped Langevin* diffusion or the *first-order Langevin* diffusion,

$$dX(t) = -\nabla F(X(t))dt + \sqrt{2\beta^{-1}}dB_t, \tag{1.3}$$

where B_t is a d-dimensional standard Brownian motion (see e.g. [GM91, BM99]). Under some mild assumptions on $F : \mathbb{R}^d \to \mathbb{R}$, this SDE admits a unique stationary distribution:

$$\pi(dx) = \frac{1}{\Gamma} e^{-\beta F(x)} dx, \tag{1.4}$$

where $\Gamma > 0$ is a normalizing constant. Without the noise term, overdamped Langevin SDE reduces to the gradient descent dynamics

$$x'(t) = -\nabla F(x(t)), \tag{1.5}$$

which is an ordinary differential equation (ODE) that arises naturally in the study of LD (see e.g. [GM91], [FGQ97, Sec. 4]). This ODE is the continuum limit of the gradient descent algorithm as the stepsize goes to zero (see e.g. [SRBd17]).

Langevin dynamics has a long history, and has also been studied under simulated annealing algorithms in the optimization, physics and statistics literature and its asymptotic convergence guarantees are well known (see e.g. [Gid85, Haj85, GM91, KGV83, BT93, BLNR15, BM99]). However, finite-time performance guarantees for LD have not been studied until more recently. In particular, Raginsky *et al.* [RRT17] showed that Langevin dynamics with stochastic gradients converges to an ε-neighborhood of a global minimizer of (1.2) in poly$(\beta, d, \frac{1}{\lambda_*}, \frac{1}{\varepsilon})$ iterations where λ_* is a spectral gap parameter related to the overdamped Langevin SDE. Zhang *et al.* [ZLC17] showed that LD with stochastic gradients is able to escape shallow local minima, hitting an ε-neighborhood of a local minimizer in time polynomial in the variables β, d and ε. More recently, Tzen *et al.* [TLR18] showed that for a given local optimum x_*, with high probability and arbitrary initialization, either LD iterates arrive at a point outside an ε-neighborhood of this local minimum within a recurrence time $\mathcal{T}_{\text{rec}} = \mathcal{O}\left(\frac{1}{m}\log(\frac{1}{\varepsilon})\right)$ where m is smallest eigenvalue of the Hessian $\nabla^2 F(x_*)$ at the local minimum or they enter this ε-neighborhood by the recurrence time and stay there until a potentially exponentially long escape time $\mathcal{T}_{\text{esc}}$. The escape time $\mathcal{T}_{\text{esc}}$ measures how quickly the LD algorithm can get away from a given neighborhood around a

[1] In the worst case, the number of iterations required to converge to a global minimum can be exponential in the dimension for an objective with multiple minima, in particular finding a global minimum of a non-convex objective is hard in general. However, when the objective has further structure such as a growth condition, the dependency can also be polynomial in dimension.

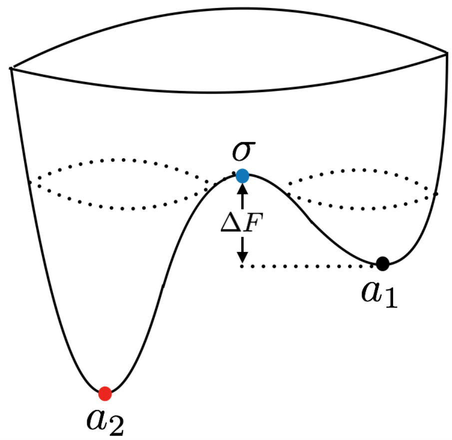

Figure 1: Double-well example. Here, $\Delta F = F(\sigma) - F(a_1)$. There are exactly two local minima a_1 and a_2 which are separated with a saddle point σ.

local minimum, therefore it can be viewed as a measure of the effectiveness of LD for the search of a global minimizer, whereas the recurrence time $\mathcal{T}_{\rm rec}$ can be viewed as the order of the time-scale for which search for local minima in the basin of attraction of that minimum happens [TLR18]. Typically, we have $\mathcal{T}_{\rm rec} \ll \mathcal{T}_{\rm esc}$, both time-scales are associated to the discrete-time LD dynamics, however they are often analyzed and estimated by considering the recurrence time and the exit time of the continuous-time overdamped Langevin diffusion [RRT17, GM91]. The simplest non-convex function studied frequently in the literature that sheds light into the exit time behavior of overdamped Langevin dynamics is the *double-well example* demonstrated in Figure 1, where the objective F has two local minima a_1, a_2 separated by a saddle point σ in between. For the overdamped Langevin diffusion, it is known that the expected time of the process starting from a_1 and hitting a small neighborhood of a_2 is given by

$$\mathbb{E}\left[\theta^{\beta}_{a_1\to a_2}\right] = [1 + o_\beta(1)] \cdot \frac{2\pi}{\mu^*(\sigma)} \cdot e^{\beta[F(\sigma)-F(a_1)]} \cdot \sqrt{\frac{|\det \mathrm{Hess}\ F(\sigma)|}{\det \mathrm{Hess}\ F(a_1)}}. \tag{1.6}$$

Here, $o_\beta(1) \to 0$ as $\beta \to \infty$, $\det \mathrm{Hess}\ F(x)$ stands for the determinant of the Hessian of F at x, and $-\mu^*(\sigma)$ is the unique negative eigenvalue of the Hessian of F at the saddle point σ. This formula is known as the Eyring-Kramers formula for reversible diffusions. Its rigorous proof was first obtained by [BGK04] by a potential analysis approach, and then by [HKN04] through Witten Laplacian analysis. We refer to [Ber13] for a survey on mathematical approaches to the Eyring-Kramers formula.

We note that in many practical applications, for instance in the training of neural networks, the eigenvalues of the Hessian at local extrema concentrate around zero and the magnitude of the eigenvalues m and $\mu^*(\sigma)$ can often be very small [SBL16, CCS+17]. In this case, the recurrence time $\mathcal{T}_{\rm rec}$ and the expected hitting time $\mathbb{E}\left[\theta^{\beta}_{a_1\to a_2}\right]$, being quantities that scale inversely with m and $\mu^*(\sigma)$ respectively, can be quite large. This results in slow convergence of the overdamped process to its equilibrium [BGK05]. It is also known that the overdamped Langevin diffusion is a reversible[2] Markov process, and reversible processes can be much slower than their non-reversible variants that admit the same equilibrium distribution in terms of their convergence rate to the equilibrium (see e.g. [S+09, LNP13, DHN00, HHMS93, HHMS05, EGZ17, GO07]).

There are two popular non-reversible variants of overdamped Langevin that can improve its performance in practice for a variety of applications [LNP13, Sec. 4]. The first variant is based on the *underdamped Langevin* diffusion, also known as the *second-order Langevin* diffusion,

$$dV(t) = -\gamma V(t)dt - \nabla F(X(t))dt + \sqrt{2\gamma\beta^{-1}}dB_t, \tag{1.7}$$

$$dX(t) = V(t)dt, \tag{1.8}$$

where $X(t), V(t) \in \mathbb{R}^d$ for each t, and γ is known as the *friction coefficient*, and B_t is a standard d-dimensional Brownian motion. This diffusion goes back to Kramers [Kra40] and was derived in the physics literature to model particles moving in a potential subject to random noise. It is known that under mild assumption on F, the Markov process (X, V) is ergodic and have a unique stationary distribution

$$\pi(dx, dv) = \frac{1}{\Gamma^U}\exp\left(-\beta\left(\frac{1}{2}\|v\|^2 + F(x)\right)\right)dxdv, \tag{1.9}$$

where $\Gamma^U > 0$ is a normalizing constant. Hence, the marginal distribution in X of the Gibbs distribution $\pi(dx, dv)$ is exactly the same as the invariant distribution (1.4) of the overdamped Langevin dynamics (1.3). We refer the reader to [Pav14] for more on underdamped Langevin diffusions. The discretized dynamics are called the *underdamped Langevin dynamics* (ULD):[3]

$$V_{k+1} = V_k - \eta[\gamma V_k + \nabla F(X_k)] + \sqrt{2\gamma\beta^{-1}\eta}\xi_k, \tag{1.10}$$

$$X_{k+1} = X_k + \eta V_k, \tag{1.11}$$

where $(\xi_k)_{k\geq 0}$ is a sequence of i.i.d standard Gaussian random vectors in $\mathbb{R}^d$. Recent work [GGZ18] showed that ULD admits better non-asymptotic performance guarantees compared to known guarantees for LD in the context of non-convex optimization when the

[2] We note that overdamped Langevin SDE is *reversible* in the sense that if X_0 is distributed according to the stationary measure π, then $(X_t)_{0\leq t\leq T}$ and $(X_{T-t})_{0\leq t\leq T}$ have the same law. Mathematically, this is equivalent to the infinitesimal generator of X_t process being (self-adjoint) symmetric in $L^2(\pi)$ [LNP13, RW00]. Roughly speaking, reversibility of a Markov process says that statistical properties of the process are preserved if the process is run backwards in time.

[3] This algorithm is also known as the inertial Langevin dynamics or the Hamiltonian Langevin Monte Carlo algorithm.

objective satisfies a dissipativity condition. Recent work also showed that ULD or alternative discretizations of the underdamped diffusion can sample from the Gibbs distribution more efficiently than LD when F is globally strongly convex [CCBJ17, DRD18, MS17] or strongly convex outside a compact set [CCA+18].

The second variant of overdamped Langevin involves adding a drift term to make the dynamics non-reversible:

$$dX(t) = -A_J(\nabla F(X(t)))dt + \sqrt{2\beta^{-1}}dB_t, \quad A_J = I + J, \tag{1.12}$$

where $J \neq 0$ is a $d \times d$ anti-symmetric matrix, i.e. $J^T = -J$ and I is the $d \times d$ identity matrix, and B_t is a standard d-dimensional Brownian motion. It can be shown that such a drift preserves the stationary distribution (1.4) (Gibbs distribution) of the overdamped Langevin dynamics [HHMS05, LNP13, Pav14, GM16]. Note that the non-reversible Langevin diffusion reduces to the overdamped Langevin diffusion (1.3) when $J = 0$. When F is quadratic, the $X(t)$ process in (1.12) is Gaussian. Using the rate of convergence of the covariance of $X(t)$ as the criterion, [HHMS93] showed that $J = 0$ is the worst choice, and improvement is possible if and only if the eigenvalues of the matrix associated with the linear drift term are not identical. [LNP13] proved the existence of the optimal J such that the rate of convergence to equilibrium is maximized, and provided an easily implementable algorithm for constructing them. [WHC14] proposed two approaches to obtain the optimal rate of Gaussian diffusion and they also made the comparison with [LNP13]. [GM16] studied optimal linear drift for the speed of convergence in the hypoelliptic case. For more general non-quadratic F, [HHMS05] showed by comparing the spectral gaps that by adding $J \neq 0$, the convergence to the Gibbs distribution is at least as fast as the overdamped Langevin diffusion ($J = 0$), and is strictly faster except for some rare situations. [DLP16] showed that the asymptotic variance can also be reduced by using the non-reversible Langevin diffusion.

Finally, the discretization of the non-reversible Langevin diffusion (1.12) leads to

$$X_{k+1} = X_k - \eta A_J(\nabla F(X_k)) + \sqrt{2\eta\beta^{-1}}\xi_k, \tag{1.13}$$

which we refer to as the *non-reversible Langevin dynamics* (NLD).

2 Contributions

In this paper, we study two non-reversible variants of the gradient Langevin dynamics, the underdamped Langevin dynamics (ULD) and the non-reversible Langevin dynamics (NLD).

First, we consider the special case of quadratic objectives and derive an explicit characterization of the rate of convergence to the stationary distribution in the 2-Wasserstein metric. Our exponential rate is optimal and unimprovable in the sense that it is achieved for some quadratic functions and initialization. Our results show that both non-reversible diffusions mix faster than the reversible Langevin diffusion. Our mixing rate results improve upon the existing results by Cheng *et al.* [CCBJ17] and Dalalyan and Riou-Durand [DRD18] developed earlier for the more general case when F is a strongly convex function with Lipschitz gradients.

Second, we investigate the metastability behavior of non-reversible Langevin algorithms for non-convex objectives. We extend the results of [TLR18] obtained for Langevin dynamics to non-reversible Langevin dynamics and show that for a given local minimum that is within an arbitrary distance r from the initialization, with high probability, either ULD trajectory ends up somewhere outside an ε-neighborhood of this local minimum within a recurrence time $\mathcal{T}_{\text{rec}}^U = \mathcal{O}\left(\frac{|\log(m)|}{\sqrt{m}} \log(r/\varepsilon)\right)$ or they enter this neighborhood by the recurrence time and stay there for a potentially exponentially long escape time. The analogous result shown in [TLR18] for reversible LD requires a recurrence time of $\mathcal{T}_{\text{rec}} = \mathcal{O}\left(\frac{1}{m}\log(r/\varepsilon)\right)$. This shows that underdamped dynamics requires a smaller recurrence time by a square root factor in m (ignoring a $\log(m)$ factor). Since the recurrence time can be viewed as a measure of the efficiency of the search of a local minimum [TLR18], our results suggest that underdamped Langevin dynamics operate on a faster time-scale to locate a local minimum. This is also consistent with our results for the quadratic objectives. We also show analogous results for NLD proving that its recurrence time improves upon that of LD under some assumptions.

Third, we consider the mean exit time from the basin of attraction of a local minimum for non-reversible algorithms. We focus on the double-well example (illustrated in Figure 1) which has been the recent focus of the literature [BR16, LMS17] as it is the simplest non-convex function that gives intuition about the more general case and for which mean exit time has been studied in continuous time.[4] We show that these results can be translated into discrete time if the stepsize parameter is small enough and the inverse temperature is large enough. Our analysis shows that non-reversible dynamics can exit the basin of attraction of a local minimum faster under some conditions and characterizes the improvement for both ULD and NLD compared to LD when the parameters of these algorithms are chosen appropriately. These results support the numerical evidence that non-reversible algorithms can explore the state space more efficiently (see e.g. [CDC15, CFG14, GM16]) and bridges a gap between the theory and practice of Langevin algorithms.

Notations

Throughout the paper, for any two real numbers x, y, we use the notation $x \wedge y$ to denote $\min\{x, y\}$ and $x \vee y$ to denote $\max\{x, y\}$. For any $n \times n$ matrix A, we use $\lambda_i(A)$, $1 \leq i \leq n$, to denote the n eigenvalues of A. We also assume that H is the Hessian matrix $\nabla^2 F$

[4] To our knowledge, a rigorous mathematical theory that characterizes the mean escape time for non-reversible Langevin algorithms beyond the double well example for general non-convex functions is non-existent at the moment.

evaluated at the local minimum x_*, and is positive definite. The norm $\|\cdot\|$ denotes the 2-norm of a vector and the (spectral norm) 2-norm of a matrix.

In our analysis, we will use the following 2-Wasserstein distance. For any two Borel probability measures ν_1 and ν_2 with finite second moments, the 2-Wasserstein distance is defined as

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

where the infimum is taken over all the random couples (Y_1, Y_2) taking values in $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ν_1 and ν_2. We refer the reader to [Vil09] for more on Wasserstein distances.

3 Special case: Quadratic objective

We first consider the case when F is a strongly convex quadratic function

$$F(x) = \frac{1}{2}x^T Hx - b^T x + c, \tag{3.1}$$

where H is a $d \times d$ symmetric positive definite matrix with eigenvalues $\{\lambda_i\}_{i=1}^d$ in increasing order, i.e. $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$, and b is a d-dimensional vector and c is a scalar. Let

$$m := \lambda_1 \leq M := \lambda_d$$

be the lower and upper bounds on the eigenvalues, where m is the strong convexity constant and M is the Lipschitz constant of the gradient

$$\nabla F(x) = Hx - b\,.$$

In this case, the underdamped SDE corresponds to a particular Gaussian process known as the Ornstein-Uhlenbeck process. If the objective gradient is linearized around a point, such dynamics would naturally arise. Understanding the behavior of Langevin dynamics for this simpler special case have proven to be useful for analyzing the more general case when F can be non-convex [RRT17], [Pav14, Sec. 6.3], [HN04] as well as shedding light into the convergence behavior of Langevin algorithms, see e.g. [DNP17, BG02]. Our analysis for characterizing the recurrence time of ULD and NLD will also build on our results in this section developed for quadratic objectives.

It is known that the overdamped Langevin diffusion mixes at the exponential rate e^{-mt} with respect to the 2-Wasserstein metric [BGG12]. We next show that underdamped Langevin and non-reversible Langevin diffusions mix with a faster exponential rate.

3.1 Underdamped Langevin diffusion

By strong convexity, the quadratic function $F(x)$ given in (3.1) has a unique minimum at $x_H^* := H^{-1}b$. We write

$$X(t) = Y(t) + x_H^*.$$

Thus, we have the decomposition

$$\nabla F(X(t)) = HY(t).$$

Then, we can write the underdamped diffusion in matrix form as

$$d \begin{bmatrix} V(t) \\ Y(t) \end{bmatrix} = -H_\gamma \begin{bmatrix} V(t) \\ Y(t) \end{bmatrix} dt + \sqrt{2\gamma\beta^{-1}} I^{(2)} dB_t^{(2)},$$

where $B_t^{(2)}$ is a $2d$-dimensional standard Brownian motion and

$$H_\gamma := \begin{bmatrix} \gamma I & H \\ -I & 0 \end{bmatrix}, \qquad I^{(2)} := \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.2}$$

This process is an Ornstein-Uhlenbeck process with a solution

$$\begin{bmatrix} V(t) \\ Y(t) \end{bmatrix} = e^{-tH_\gamma} \begin{bmatrix} V(0) \\ Y(0) \end{bmatrix} + \sqrt{2\gamma\beta^{-1}} \int_0^t e^{(s-t)H_\gamma} I^{(2)} dB_s^{(2)}, \tag{3.3}$$

where $[V(0), Y(0)]$ is the initial point (see e.g. [Øks98]). Then it follows that the vector $[V(t), Y(t)]^T$ follows a $2d$-dimensional Gaussian distribution for all t with mean μ_t and covariance Σ_t given by

$$\mu_t := e^{-tH_\gamma}[V(0), Y(0)]^T, \tag{3.4}$$

$$\Sigma_t := 2\gamma\beta^{-1} \int_0^t e^{-sH_\gamma} I^{(2)} e^{-sH_\gamma^T} ds. \tag{3.5}$$

It follows from (1.9) that the stationary distribution of $[Y(t), V(t)]$ is

$$\pi(dy, dv) \propto e^{-\frac{1}{2}\beta y^T H y - \frac{\beta}{2}\|v\|^2} dy dv,$$

which is a Gaussian distribution with mean and covariance given by

$$\mu_\infty = [0_d,\ 0_d]^T, \qquad \Sigma_\infty = \beta^{-1} \begin{bmatrix} H^{-1} & 0_d \\ 0_d & I_d \end{bmatrix},$$

where 0_d and I_d are $d \times d$ zero and identity matrices, and the marginal stationary distribution of the $Y(t)$ process is given by

$$\pi(dy) \propto e^{-\frac{1}{2}\beta y^T H y} dy.$$

Natural questions to ask are the following:

Q1. Let $\nu_t(dy)$ be the probability law of the underdamped process $Y(t)$ at time t. How fast does the underdamped process converge to its equilibrium $\pi(dy)$ in the $\mathcal{W}_2$ metric?

Q2. How does the convergence depend on the constants m and M?

These questions are naturally related to how fast μ_t and Σ_t converges to its equilibrium values μ_∞ and Σ_∞. Dalalyan and Riou-Durand [DRD18] showed recently that

$$\mathcal{W}_2\big(\nu_t(dy), \pi(dy)\big) \le C_0(\beta) e^{\frac{-m}{\sqrt{M+m}} t} \tag{3.6}$$

for some explicit constant C_0 that depends on β. In [DRD18], the bound in (3.6) holds for $\beta = 1$, and the result in (3.6) can be extended to general $\beta > 0$ by applying Lemma 1 in [DRD18]. For general $\beta > 0$, we should replace t by $\sqrt{\beta^{-1}}t$ and m and M by βm and βM to apply (3.6), and the exponent in (3.6) remains invariant. This improves upon the mixing rate $e^{-(m/2M)t}$ obtained in [CCBJ17] for strongly convex F for $\gamma = 2$ and $\beta = M$ in the underdamped SDE (see Theorem 5 [5] in [CCBJ17]). It is known that overdamped Langevin diffusion mixes at rate e^{-mt} in the 2-Wasserstein metric [DRD18], the rate in (3.6) improves upon this rate when $M + m \ge 1$. Next, in Theorem 3, we show a stronger result that shows that the underdamped diffusion mixes at rate $e^{-\sqrt{m}t}$ when F is a strongly convex quadratic.

Let $\pi_t(dv, dy)$ denote the probability law of the underdamped process $(V(t), Y(t))$ at time t. The 2-Wasserstein distance between the two Gaussian distributions $\pi_t(dv, dy) \sim \mathcal{N}(\mu_t, \Sigma_t)$ and $\pi(dy, dv) \sim \mathcal{N}(\mu_\infty, \Sigma_\infty)$ admits an explicit formula (see e.g. [Gel90])

$$\mathcal{W}_2^2(\pi_t(dv, dy), \pi(dv, dy)) = \|\mu_t - \mu_\infty\|_2^2 + \mathrm{Tr}\left(\Sigma_t + \Sigma_\infty - 2\left(\Sigma_t^{1/2}\Sigma_\infty\Sigma_t^{1/2}\right)^{1/2}\right). \tag{3.7}$$

Inserting formulas (3.4)–(3.5) into (3.7) leads to an explicit expression for the Wasserstein distance. In particular we see from (3.4)–(3.5) that the first term $\|\mu_t - \mu_\infty\|$ is controlled by $\|e^{-tH_\gamma}\|$ where $\|\cdot\|$ denotes the spectral norm (2-norm) of a matrix. In the next lemma, we provide an estimate on $\|e^{-tH_\gamma}\|$, in particular achieving the fastest mixing rate $e^{-\sqrt{m}t}$ requires tuning the parameter γ.

Lemma 1. *(i) If $\gamma \in (0, 2\sqrt{m})$, then*

$$\left\|e^{-tH_\gamma}\right\| \le C_{\hat\varepsilon} e^{-\sqrt{m}(1-\hat\varepsilon)t},$$

where

$$C_{\hat\varepsilon} := \frac{1+M}{\sqrt{m(1-(1-\hat\varepsilon)^2)}}, \qquad \hat\varepsilon := 1 - \frac{\gamma}{2\sqrt{m}} \in (0,1). \tag{3.8}$$

(ii) If $\gamma = 2\sqrt{m}$, then,

$$\left\|e^{-tH_\gamma}\right\| \le \sqrt{C_H + 2 + (m+1)^2 t^2} \cdot e^{-\sqrt{m}t}$$

where C_H is a constant that depends only on the eigenvalues of H and is defined as

$$C_H := \max_{i:\lambda_i > m} \frac{(1+\lambda_i)^2}{\lambda_i - m}. \tag{3.9}$$

[5] In our notation, our F is the $\beta^{-1}F$ in [CCBJ17], but the rate $e^{-(m/2M)t}$ is preserved.

Remark 2. *In Lemma 1, when $\lambda = 2\sqrt{m}$, the matrix H_γ has double eigenvalues that are not simple (the eigenvalues have a Jordan block of size 2). As a consequence, in this case, there is a factor $\sqrt{C_H + 2 + (m+1)^2 t^2}$ in front of the exponential $e^{-\sqrt{m}t}$ that grows with t and it is not possible to replace this factor with a universal constant that does not depend on t. This behavior is also the reason why the constant $C_{\hat{\varepsilon}} \uparrow \infty$ in part (i) of Lemma 1 as $\gamma \uparrow 2\sqrt{m}$.*

We recall from (1.4) that π denotes the stationary distribution of the X process, that is, $\pi(dx) \propto e^{-\beta F(x)} dx$. Based on the previous lemma and a standard coupling approach, we obtain the following convergence result in the 2-Wasserstein distance.

Theorem 3. *Consider the underdamped Langevin diffusion with parameter γ for a quadratic objective given by (3.1). Let $\pi_{t,\gamma}$ denote the law of $X(t)$ at time t for the non-reversible Langevin diffusion described by (1.12). We have the following:*

(i) *If $\gamma < 2\sqrt{m}$, then for every $t \geq 0$,*

$$\mathcal{W}_2(\pi_{t,\gamma}, \pi) \leq C_{\hat{\varepsilon}} e^{-\sqrt{m}(1-\hat{\varepsilon})t} \cdot \mathcal{W}_2(\pi_{0,\gamma}, \pi),$$

where $C_{\hat{\varepsilon}}$ and $\hat{\varepsilon}$ are defined in (3.8).

(ii) *If $\gamma = 2\sqrt{m}$, then for every $t \geq 0$,*

$$\mathcal{W}_2(\pi_{t,\gamma}, \pi) \leq \sqrt{C_H + 2 + (m+1)^2 t^2} \cdot e^{-\sqrt{m}t} \cdot \mathcal{W}_2(\pi_{0,\gamma}, \pi).$$

where C_H is a constant defined by (3.9).

(iii) *For any $\gamma > 0$, we have*

$$\mathcal{W}_2(\pi_{0,\gamma}, \pi) \leq \left(2R^2 + \frac{2}{m}(b + d/\beta)\right)^{1/2}.$$

Remark 4. *Following the proof technique of Theorem 3, it can be shown that if $\gamma > 2\sqrt{m}$, then $\mathcal{W}_2(\pi_{t,\gamma}, \pi)$ decays with an exponential rate $e^{-t\frac{\gamma-\sqrt{\gamma^2-4m}}{2}}$ which is a slower decay than the $e^{-\sqrt{m}t}$ rate achieved for $\gamma = 2\sqrt{m}$. In this sense, the choice of $\gamma = 2\sqrt{m}$ optimizes the convergence rate to the stationary distribution. This is illustrated in Figure 3.1. A similar result was known in dimension one (see [Ris89, Sec. 10.2], [Pav14, Section 6.3], [EGZ17, Sec. 1.6]), our result generalizes this result to arbitrary dimension d and gives explicit constants.*

Theorem 3 shows that a convergence rate of $e^{-\sqrt{m}t}$ is achievable by tuning the parameter γ and specifies all the constants and dependence to initialization explicitly. We note that the exponent in t in our analysis is optimal and is achieved for some quadratic functions.

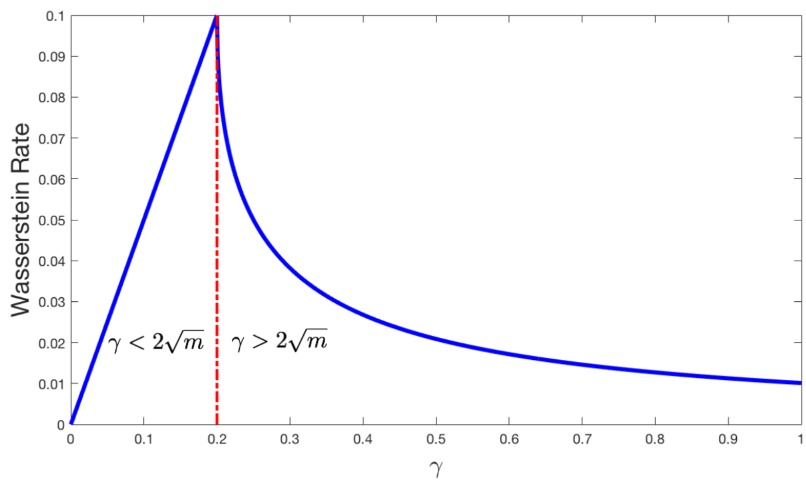

Figure 2: The exponent of the Wasserstein rate is optimized for the choice of $\gamma = 2\sqrt{m}$. This is illustrated in the figure for $m = 0.01$.

3.2 Non-reversible Langevin diffusion

We will show that the non-reversible Langevin diffusion given by (1.12) can converge to its equilibrium with a faster exponential rate compared to the e^{-mt} rate of overdamped Langevin diffusion. With a change of variable $X(t) = Y(t) + x_H^*$, the non-reversible Langevin dynamics given in (1.12) become

$$dY(t) = -A_J HY(t)dt + \sqrt{2\beta^{-1}}dB_t, \quad A_J = I + J,$$

which admits the solution

$$Y(t) = e^{-tA_J H}Y(0) + \sqrt{2\beta^{-1}}\int_0^t e^{(s-t)A_J H}dB_s.$$

We observe that, the convergence behavior of the $Y(t)$ process to its equilibrium is controlled by the decay of $\left\|e^{-tA_J H}\right\|$ in time t. It is expected that the decay of $\|e^{-tA_J H}\|$ in time t is related to the real part of the eigenvalues

$$\lambda_i^J := \mathrm{Re}\left(\lambda_i(A_J H)\right)$$

indexed with increasing order and their multiplicity. We note that when $J = 0$, we have $A_J = I$ and $\lambda_i^J = \lambda_i$ which are the eigenvalues of H. Hwang *et al.* [HHMS93] showed that for any anti-symmetric J, we have the following result.

Lemma 5 ([HHMS93, Theorem 3.3])**.** *For any anti-symmetric matrix J, we have*

$$m = \lambda_1 \leq \lambda_1^J \leq \lambda_d^J \leq \lambda_d = M,$$

and $m = \lambda_1 = \lambda_1^J$ if and only if the following condition holds:

(C1) *There exists non-zero vectors u and v in $\mathbb{R}^d$ and $\rho \in \mathbb{R}$ such that $u + iv$ is an eigenvector of $A_J H$ with eigenvalue $m + i\rho$ and $u + iv$ is an eigenvector of J with eigenvalue $i\rho$ and $Ju = -\rho v$, $Jv = \rho u$.*

A consequence of Lemma 5 is that $\lambda_1^J \geq \lambda_1 = m$ with the equality being satisfied only if a special (non-generic) condition **(C1)** given in Lemma 5 holds. Considering the Jordan normal form of $A_J H$, there exists an invertible matrix S such that

$$S^{-1}(A_J H)S = \operatorname{diag}\left(\hat{J}_1, \hat{J}_2, \ldots, \hat{J}_m\right),$$

where the latter is a block diagonal matrix with diagonals $\hat{J}_i$, with $\hat{J}_i$ being the Jordan block corresponding to the i-th eigenvalue. This leads to the immediate bound

$$\left\|e^{-tA_J H}\right\| = \left\|Se^{-t\hat{J}}S^{-1}\right\| \leq \|S\| \left\|S^{-1}\right\| \left\|e^{-t\hat{J}}\right\| \leq C_J(1 + t^{n_1 - 1})e^{-t\lambda_1^J},$$

where C_J is a universal constant that depends on J and n_1 is the size of the Jordan block $\hat{J}_1$ corresponding to the eigenvalue λ_1^J (with the convention that a simple eigenvalue is corresponding to a Jordan block of size one). This is summarized in the following lemma:

Lemma 6. *There exists a positive constant C_J that depends on J such that*

$$\left\|e^{-tA_J H}\right\| \leq C_J(1 + t^{n_1 - 1})e^{-t\lambda_1^J},$$

where n_1 is the maximal size of a Jordan block of $A_J H$ corresponding to the eigenvalue λ_1^J. It follows that for any $\tilde{\varepsilon} > 0$, there exist some constant $C_J(\tilde{\varepsilon})$ that depends on $\tilde{\varepsilon}$ and J such that for every $t \geq 0$,

$$\|e^{-tA_J H}\| \leq C_J(\tilde{\varepsilon})e^{-tm_J(\tilde{\varepsilon})}, \qquad m_J(\tilde{\varepsilon}) := \lambda_1^J - \tilde{\varepsilon}. \tag{3.10}$$

Remark 7. *When $J = 0$, we have $A_J H = H$ which has the simple eigenvalue $m = \lambda_1^J = \lambda_1$. Since H is positive definite, and $\|e^{-tH}\| = e^{-t\lambda_1^J} = e^{-tm}$ and we can take $n_1 = 1$ and $C_J = 1/2$ in Lemma 6. A consequence of Lemmas 5 and 6 is that unless the condition (C1) holds, non-reversible Langevin leads to a faster exponential decay of compared to reversible Langevin, i.e $\lambda_1^J > \lambda_1$.*

One could also ask what is the choice of the matrix J that can maximize the exponent λ_1^J that appears in Lemma 6, i.e., let

$$J_{opt} := \arg \max_{J=-J^T} \lambda_1^J .$$

A formula for J_{opt} and an algorithm to compute it is known (see [LNP13, Fig. 1]), however this is not practical to compute for optimization purposes as it requires the knowledge of the eigenvectors and eigenvalues of the matrix H which is unknown. Nevertheless, J_{opt} gives information about the extent of acceleration that can be obtained. It is known that

$$\lambda_1^{J_{opt}} = \frac{\mathrm{Tr}(H)}{d},$$

as well as a characterization of the constants $C_{J_{opt}}$ and n_1 arising in Lemma 6 when $J = J_{opt}$ [LNP13, equation (46)]. We see that $md \leq \mathrm{Tr}(H) \leq M(d-1) + m$ as the smallest and the largest eigenvalue of H is m and M. Therefore, we have

$$1 \leq \frac{\lambda_1^{J_{opt}}}{\lambda_1} \leq \frac{M(d-1)+m}{md}.$$

The acceleration is not possible (the ratio above is 1) if and only if all the eigenvalues of H are the same and are equal to m; i.e. when $M = m$ and $\mathrm{Tr}(H) = md$. Otherwise, J_{opt} can accelerate by a factor of $\frac{M(d-1)+m}{md}$ which is on the order of the condition number $\kappa := M/m$ up to a constant $\frac{d-1}{d}$ which is close to one for d large.

We recall that π denotes the stationary distribution (1.4) of the Langevin diffusion, that is, $\pi(dx) \propto \exp(-\beta F(x))dx$. Let $\pi_{t,J}$ denote the law of $X(t)$ at time t. Using Lemma 6, we can show the following convergence rate in Wasserstein distances using a proof technique similar to the proof of Theorem 3.

Theorem 8. *For every $t \geq 0$,*

$$\mathcal{W}_2(\pi_{t,J}, \pi) \leq C_J(1 + t^{n_1-1})e^{-t\lambda_1^J} \cdot \left(2R^2 + \frac{2}{m}(b + d/\beta)\right)^{1/2},$$

where C_J, n_1, λ_1^J are defined in Lemma 6.

In the next section, we will discuss non-convex objectives. The results from this section will play a crucial role to understand the behavior of Langevin algorithms near a local minimum which can be approximated by an Ornstein-Uhlenbeck process.

4 Recurrence and escape times for non-reversible Langevin dynamics

It is known that the reversible Langevin algorithm converges to a local minimum in time polynomial with respect to parameters β and d, the intuition being that the expectation of the iterates follows the gradient descent dynamics which converges to a local minimum [ZLC17, FGQ97]. It is also known that once Langevin algorithms arrive to a neighborhood of a local optimum, they can spend exponentially many iterations in dimension to escape from the basin of attraction of this local minimum. This behavior is known as "metastability" and has been studied well [BGK05, BGK04, Ber13, TLR18].

Recently, Tzen *et al.* [TLR18] provided a finer characterization of this metastability phenomenon and showed that for a particular local minimum, if the stepsize η is small enough and the inverse temperature β is large enough, with an arbitrary initialization, at least one of the following two events will occur with high probability: (1) The Langevin trajectory will end up being outside of the ε-neighborhood of this particular optimum within a short *recurrence time* $\mathcal{T}_{\text{rec}} = \mathcal{O}(\frac{1}{m}\log(\frac{1}{\varepsilon}))$ where $m > 0$ is the smallest eigenvalue of the Hessian at the local minimum. (2) The Langevin dynamics' trajectory enters this ε-neighborhood by the recurrence time and stays there for a potentially exponentially long *escape time* $\mathcal{T}_{\text{esc}}$. In Section 3, we observed that both underdamped Langevin and reversible Langevin concentrates around the local minimum faster in continuous-time. When the stepsize is small, we also expect similar results to hold in discrete time. In this section, we will study the recurrence time and escape time $\mathcal{T}^U_{\text{rec}}$ and $\mathcal{T}^U_{\text{esc}}$ of underdamped Langevin dynamics (ULD) and the corresponding time-scales $\mathcal{T}^J_{\text{rec}}$ and $\mathcal{T}^J_{\text{esc}}$ for non-reversible Langevin dynamics (NLD). We will show that recurrence time of underdamped and non-reversible Langevin algorithms will improve upon that of reversible Langevin algorithms in terms of its dependency to the smallest eigenvalue of the Hessian at a local minimum. Our analysis is based on linearizing the gradient of the objective around a local minimum and is based on the results derived in the previous section on quadratic objectives. We will also show in Section 5 that for the double-well potential, the mean exit times from *the basin of attraction of a local minimum*[6] for non-reversible Langevin dynamics will improve upon that of reversible Langevin dynamics in terms of dependency to the curvature at the saddle point.

Throughout the rest of the paper, we impose the following assumptions.

Assumption 9. *We impose the following assumptions.*

(i) The functions $f(\cdot, z)$ are twice continuously differentiable, non-negative valued, and

$$|f(0,z)| \le A, \quad \|\nabla f(0,z)\| \le B, \quad \left\|\nabla^2 f(0,z)\right\| \le C,$$

uniformly in $z \in \mathcal{Z}$ for some $A, B, C > 0$.

(ii) $f(\cdot, z)$ have Lipschitz-continuous gradients and Hessians, uniformly in $z \in \mathcal{Z}$, there exist constants $L, M > 0$ so that for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned} &\|\nabla f(x,z) - \nabla f(y,z)\| \le M\|x-y\|, \\ &\left\|\nabla^2 f(x,z) - \nabla^2 f(y,z)\right\| \le L\|x-y\|. \end{aligned} \tag{4.1}$$

[6]Formally, we define the basin of attraction of a local minimum x_* as the set of all initial points $x_0 \in \mathbb{R}^d$ such that the *gradient flow ODE* given in (1.5) with initial point x_0 converges to x_* as time t goes to infinity.

(iii) The empirical risk $F(\cdot)$ is (m, b)-dissipative [7]*:*

$$\langle x, \nabla F(x) \rangle \geq m\|x\|^2 - b. \tag{4.2}$$

(iv) The initialization satisfies $\|X_0\| \leq R := \sqrt{b/m}$.

The first assumption and the second assumption on gradient Lipschitzness is standard (see e.g. [MSH02, TLR18, EGZ17, CCBJ17]). The second assumption on the Lipschitzness of the Hessian is also frequently made in the literature [TLR18, CFM+18, DRD18], this allows a more accurate local approximation of the Langevin dynamics as an Ornstein-Uhlenbeck process. The third assumption on dissipativity is also standard in the literature to ensure convergence of Langevin diffusions to the stationary measure. It is not hard to see from dissipativity condition (4.2) in Assumption 9 (iii) above that for any local minimum x_* of the Hessian matrix $H = \nabla^2 F$, which is positive definite and the minimum eigenvalue of H is m, then we have $\|x_*\| \leq R = \sqrt{b/m}$. This indicates that we can choose an initialization X_0 to satisfy $\|X_0\| \leq R = \sqrt{b/m}$ as in part (iv) of Assumption 9.

4.1 Underdamped Langevin dynamics

In this section, we investigate the behavior around local minima for the underdamped Langevin dynamics (1.10)-(1.11) by studying recurrence and escape times with the choice of the friction coefficient $\gamma = 2\sqrt{m}$ which is optimal for the convergence rate as discussed in Section 3. We also give an analogous theorem for the case $\gamma < 2\sqrt{m}$ in Section E of the appendix; the remaining case $\gamma > 2\sqrt{m}$ can also be treated similarly.

Before we state the main result, we summarize the technical constants that will be used in Table 1 [8]. In the following result and the rest of the paper unless we specify otherwise, the notation $\mathcal{O}(\cdot)$ hides dependency to the constants C_H, M, L. We give explicit expressions for all the constants in the proofs.

Theorem 10. *Fix $\gamma = 2\sqrt{m}$, $\delta \in (0, 1)$ and $r > 0$. For a given ε satisfying*

$$0 < \varepsilon < \overline{\varepsilon}^U = \min\left\{\mathcal{O}(r), \mathcal{O}(m)\right\},$$

where $\overline{\varepsilon}^U$ is more formally defined in Table 1, and we define the recurrence time

$$\mathcal{T}_{rec}^U = -\frac{1}{\sqrt{m}} W_{-1}\left(\frac{-\varepsilon^2\sqrt{m}}{8r^2\sqrt{C_H + 2 + (m+1)^2}}\right) = \mathcal{O}\left(\frac{|\log(m)|}{\sqrt{m}} \log\left(\frac{r}{\varepsilon}\right)\right), \tag{4.3}$$

[7]In terms of notations, the dissipativity constant m here is taken to be the same as the minimum eigenvalue of the Hessian matrix H in this section, as well as for the quadratic case in Section 3. With abuse of notations, we use m both for non-convex and convex cases.

[8]In Table 1, C_x^c, C_v^c give the uniform L^2 bounds for the continuous-time processes $X(t)$ and $V(t)$ defined in (1.8) and (1.7) respectively, and C_x^d, C_v^d give the uniform L^2 bounds for the discrete-time processes X_k and V_k defined in (1.11) and (1.10) respectively, and K_1 and K_2 are used in the upper bound on the stepsize η under which the uniform L^2 bounds for the discrete-time processes are valid, see Lemma 24 for details.

Constants	*Source*
$\bar{\varepsilon}_1^U = \sqrt{\frac{C_H + 2 + (m+1)^2}{(C_H+2)m + (m+1)^2}}\, r, \quad \bar{\varepsilon}_2^U = 2\sqrt{2}(C_H + 2 + (m+1)^2)^{1/4}\frac{e^{-1/2}r}{m^{1/4}}$	(B.1), (B.2)
$\bar{\varepsilon}_3^U = \frac{\sqrt{m}}{4L\left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}} + \frac{\sqrt{(C_H+2)m+(m+1)}}{8\sqrt{C_H+2+(m+1)^2}}\right)}, \qquad \bar{\varepsilon}^U := \min\left\{\bar{\varepsilon}_1^U, \bar{\varepsilon}_2^U, \bar{\varepsilon}_3^U\right\}$	(B.3)
$\bar{\eta}_1^U = \frac{\varepsilon e^{-(1+\gamma+M)}}{8B}, \quad \bar{\eta}_2^U = \frac{\delta\varepsilon^2 e^{-2(1+\gamma+M)}}{384(M^2 C_x^c + (1+\gamma)^2 C_v^c)\mathcal{T}_{\mathrm{rec}}^U}$	(B.22), (B.20)
$\bar{\eta}_3^U = \frac{4\gamma\delta^2}{9\beta M^2 C_v^d \mathcal{T}_{\mathrm{esc}}^U}$	(B.9)
$\bar{\eta}_4^U = \min\left\{\frac{\gamma}{K_2}(d/\beta + \overline{A}/\beta), \frac{\gamma\lambda}{2K_1}\right\}, \qquad \bar{\eta}^U := \min\left\{1, \bar{\eta}_1^U, \bar{\eta}_2^U, \bar{\eta}_3^U, \bar{\eta}_4^U\right\}$	(B.36)
$\underline{\beta}_1^U = \frac{256(2C_H m + 4m + (m+1)^2)}{m\varepsilon^2}\left(d\log(2) + \log\left(\frac{6\sqrt{4m + M^2 + 1}\,\mathcal{T} + 3}{\delta}\right)\right)$	(B.11)
$\underline{\beta}_2^U = \frac{512 d\eta\gamma\log(2^{1/4}e^{1/4}6\delta^{-1}\mathcal{T}_{\mathrm{rec}}^U/\eta)}{\varepsilon^2 e^{-2(1+\gamma+M)\eta}}, \qquad \underline{\beta}^U := \max\{\underline{\beta}_1^U, \underline{\beta}_2^U\}$	(B.21)
$C_x^c := \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right)R^2 + \beta B R + \beta A + \frac{3}{4}\beta\lVert V(0)\rVert^2 + \frac{d+\overline{A}}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}$	(B.34)
$C_v^c := \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right)R^2 + \beta B R + \beta A + \frac{3}{4}\beta\lVert V(0)\rVert^2 + \frac{d+\overline{A}}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}$	(B.35)
$K_1 = \max\left\{\frac{32M^2\left(\frac{1}{2}+\gamma\right)}{(1-2\lambda)\beta\gamma^2}, \frac{8\left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)}{\beta(1-2\lambda)}\right\}$	(B.37)
$K_2 = 2B^2\left(\frac{1}{2}+\gamma\right)$	(B.38)
$C_x^d = \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right)R^2 + \beta B R + \beta A + \frac{3}{4}\beta\lVert V(0)\rVert^2 + \frac{4(d+\overline{A})}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}$	(B.39)
$C_v^d = \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right)R^2 + \beta B R + \beta A + \frac{3}{4}\beta\lVert V(0)\rVert^2 + \frac{4(d+\overline{A})}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}$	(B.40)
$\lambda = \frac{1}{8} \wedge \frac{m}{2M+\gamma^2}, \quad \overline{A} := \frac{\beta m}{2M+\gamma^2}\left(\frac{B^2}{2M+\gamma^2} + \frac{b}{m}\left(M + \frac{1}{2}\gamma^2\right) + A\right)$	(B.31), (B.32)

Table 1: Summary of the constants used in Section 4.1 and where they are defined.

and the escape time

$$\mathcal{T}_{esc}^U := \mathcal{T}_{rec}^U + \mathcal{T},$$

for any arbitrary $\mathcal{T} > 0$, where W_{-1} is the lower branch of the Lambert W function [9].

Consider an arbitrary initial point x for the underdamped Langevin dynamics and a local minimum x_* at a distance at most r. Assume that the stepsize η satisfies

$$\eta \leq \overline{\eta}^U = \min\left\{\mathcal{O}(\varepsilon), \mathcal{O}\left(\frac{m^2\beta\delta\varepsilon^2}{(md+\beta)\mathcal{T}_{rec}^U}\right), \mathcal{O}\left(\frac{m^{3/2}\delta^2}{(md+\beta)\mathcal{T}_{esc}^U}\right), \mathcal{O}\left(\frac{m^{1/2}\beta}{d+\beta}\right)\right\},$$

where more formally $\overline{\eta}^U$ is defined in Table 1 and β satisfies

$$\beta \geq \underline{\beta}^U = \max\left\{\Omega\left(\frac{d+\log((\mathcal{T}+1)/\delta)}{m\varepsilon^2}\right), \Omega\left(\frac{d\eta m^{1/2}\log(\delta^{-1}\mathcal{T}_{rec}^U/\eta)}{\varepsilon^2}\right)\right\},$$

where more formally $\underline{\beta}^U$ is defined in Table 1, for any realization of training data z, with probability at least $1-\delta$ w.r.t. the Gaussian noise, at least one of the following events will occur:

1. $\|X_k - x_*\| \geq \frac{1}{2}\left(\varepsilon + re^{-\sqrt{m}k\eta}\right)$ *for some* $k \leq \eta^{-1}\mathcal{T}_{rec}^U$.
2. $\|X_k - x_*\| \leq \varepsilon + re^{-\sqrt{m}k\eta}$ *for every* $\eta^{-1}\mathcal{T}_{rec}^U \leq k \leq \eta^{-1}\mathcal{T}_{esc}^U$.

Remark 11. *Notice that in Theorem 10, the definition of η and β are coupled since $\overline{\eta}^U$ depends on β and $\underline{\beta}^U$ depends on η. A closer look reveals that when η is sufficiently small, the first term in the definition of $\underline{\beta}^U$ dominates the second term and $\underline{\beta}^U$ is independent of η. So to satisfy the constraints in Theorem 10, it suffices to first choose β to be larger than the first term in $\underline{\beta}^U$ and then choose η to be sufficiently small.*

Remark 12. *Theorem 10 is about the empirical risk but the ideas can be generalized to population risk problems. More specifically, we can apply Theorem 10 to obtain Theorem 39 for the population risk in the Appendix, similar to Theorem 3 in [TLR18].*

Remark 13. *In [TLR18], the overdamped Langevin algorithm is used and the recurrence time $\mathcal{T}_{rec} = \mathcal{O}\left(\frac{1}{m}\log(\frac{r}{\varepsilon})\right)$, and thus our recurrence time $\mathcal{T}_{rec}^U = \mathcal{O}\left(\frac{|\log(m)|}{\sqrt{m}}\log(\frac{r}{\varepsilon})\right)$ for the underdamped Langevin algorithm, which has a square root factor improvement.*

[9] The Lambert W function $W(x)$ is defined via the solution of the algebraic equation $W(x)e^{W(x)} = x$. When $x \geq 0$, $W(x)$ is uniquely defined. When $-e^{-1} \leq x < 0$, $W(x)$ has two branches, the upper branch $W_0(x)$ and the lower branch $W_{-1}(x)$, see e.g. [CGH+96].

Remark 14. *The recurrence time* $\mathcal{T}_{rec}^U$ *defined in* (4.3) *using the Lambert W function is indeed the unique solution greater than* $\frac{1}{\sqrt{m}}$ *of the algebraic equation* [10]*:*

$$\mathcal{T}_{rec}^U e^{-\sqrt{m}\mathcal{T}_{rec}^U} = \frac{\varepsilon^2}{8r^2\sqrt{C_H + 2 + (m+1)^2}}.$$

Since $\varepsilon < \overline{\varepsilon}^U$ *and by the definition of* $\overline{\varepsilon}^U$ *in Table 1, we have* $\varepsilon < \overline{\varepsilon}^U \leq 2\sqrt{2}(C_H + 2 + (m+1)^2)^{1/4}\frac{e^{-1/2}r}{m^{1/4}}$ *so that* $\frac{\varepsilon^2}{8r^2\sqrt{C_H+2+(m+1)^2}} < \frac{1}{\sqrt{m}}e^{-1}$ *and thus* $\mathcal{T}_{rec}^U \geq \frac{1}{\sqrt{m}}$ *is well-defined and moreover for every* $x \geq \mathcal{T}_{rec}^U$*, the function* $x \mapsto xe^{-\sqrt{m}x}$ *is decreasing in* x*.*

4.2 Non-reversible Langevin dynamics

In this section, we investigate the behavior around local minima for the non-reversible Langevin dynamics (1.13) by studying recurrence and escape times. We recall from (3.10) in Lemma 6 that for any $\tilde{\varepsilon} > 0$, there exists some constant $C_J(\tilde{\varepsilon})$ that may depend on J and $\tilde{\varepsilon}$ such that for every $t \geq 0$:

$$\|e^{-tA_J H}\| \leq C_J(\tilde{\varepsilon})e^{-tm_J(\tilde{\varepsilon})}, \qquad m_J(\tilde{\varepsilon}) = (\lambda_1^J - \tilde{\varepsilon}).$$

Note that by considering $t = 0$, it is clear $C_J(\tilde{\varepsilon}) \geq 1$.

Before we state the main result, we summarize the technical constants that will be used in Table 2. [11]

Theorem 15. *Fix* $\delta \in (0,1)$ *and* $r > 0$*. Given by* $\varepsilon > 0$ *satisfying*

$$\varepsilon < \overline{\varepsilon}^J = \min\left\{\mathcal{O}\left(\frac{m_J(\tilde{\varepsilon})}{C_J(\tilde{\varepsilon})}\right), \mathcal{O}(rC_J(\tilde{\varepsilon}))\right\},$$

where $\overline{\varepsilon}^J$ *is more formally defined in Table 2, we define the recurrence time*

$$\mathcal{T}_{rec}^J := \frac{2}{m_J(\tilde{\varepsilon})}\log\left(\frac{8r}{C_J(\tilde{\varepsilon})\varepsilon}\right) = \mathcal{O}\left(\frac{1}{m_J(\tilde{\varepsilon})}\log\left(\frac{r}{C_J(\tilde{\varepsilon})\varepsilon}\right)\right), \tag{4.4}$$

and the escape time $\mathcal{T}_{esc}^J := \mathcal{T}_{rec}^J + \mathcal{T}$ *for any arbitrary* $\mathcal{T} > 0$*.*

For any initial point x *and a local minimum* x_* *at a distance at most* r*. Assume the stepsize* η *satisfies*

$$\eta \leq \overline{\eta}^J = \min\left\{\mathcal{O}(\varepsilon), \mathcal{O}\left(\frac{\delta\varepsilon^2 m^3}{(m+\beta^{-1}d)\mathcal{T}_{rec}^J}\right), \mathcal{O}\left(\frac{\delta^2 m^3}{(d+m\beta+dm^3)\mathcal{T}_{esc}^J}\right)\right\},$$

[10] We used the fact that the solution to the algebraic equation $ax = p^x$, $p > 0$, $p \neq 1$, $a \neq 0$, can be expressed as $x = -\frac{1}{\log p}W(-\frac{1}{a}\log p)$, where W is the Lambert W function.

[11] In Table 2, C_c gives the uniform L^2 bound for $X(t)$ in the continuous time dynamics (1.12) (see Corollary 29) and C_d gives the uniform L^2 bound for X_k in the discrete time dynamics (1.13) (see Corollary 31), and C_1 is used in the upper bound on the relative entropy between the probability measures of the discrete and continuous dynamics (see (C.4)).

Constants	*Source*
$\underline{\varepsilon}_1^J = \frac{m_J(\tilde{\varepsilon})}{4C_J(\tilde{\varepsilon})(1+\|J\|)L(1+\frac{1}{64C_J(\tilde{\varepsilon})^2})}, \qquad \underline{\varepsilon}_2^J = 8rC_J(\tilde{\varepsilon})$	(C.1)
$\underline{\varepsilon}^J = \min\{\underline{\varepsilon}_1^J, \underline{\varepsilon}_2^J\}$	
$\overline{\eta}_1^J = \frac{\varepsilon e^{-(1+\|J\|)M}}{8(1+\|J\|)B}$	(C.10)
$\overline{\eta}_2^J = \frac{\delta\varepsilon^2 e^{-2(1+\|J\|)M}}{384(1+\|J\|)^2 M^2 C_c \mathcal{T}_{\mathrm{rec}}^J}$	(C.11)
$\overline{\eta}_3^J = \frac{2\delta^2}{9C_1\mathcal{T}_{\mathrm{esc}}^J}$	(C.7)
$\overline{\eta}_4^J = \frac{1}{M(1+\|J\|)^2}$	(C.3)
$\overline{\eta}^J := \min\{1, \overline{\eta}_1^J, \overline{\eta}_2^J, \overline{\eta}_3^J, \overline{\eta}_4^J\}$	
$\underline{\beta}_1^J = \frac{128C_J(\tilde{\varepsilon})^2}{m_J(\tilde{\varepsilon})\varepsilon^2}\left(\frac{d}{2}\log(2) + \log\left(\frac{6(1+\|J\|)M\mathcal{T}+3}{\delta}\right)\right)$	(C.9)
$\underline{\beta}_2^J = \frac{512d\eta\log(2^{1/4}e^{1/4}6\delta^{-1}\mathcal{T}_{\mathrm{rec}}^J/\eta)}{\varepsilon^2 e^{-2(1+\|J\|)M\eta}}$	(C.12)
$\underline{\beta}^J := \max\left\{\underline{\beta}_1^J, \underline{\beta}_2^J\right\}$	
$C_c := \frac{MR^2+2BR+B+4A}{m} + \frac{2b(M+B)}{m^2} + \frac{4M\beta^{-1}d(M+B)}{m^3} + \frac{b}{m}\log 3$	(C.18)
$C_d := \frac{MR^2+2BR+B+4A}{m} + \frac{8(M+B)M\beta^{-1}d}{m^3} + \frac{2(M+B)b}{m^2} + \frac{b}{m}\log 3$	(C.19)
$C_1 := 6(\beta((1+\|J\|)^2M^2C_d+B^2)+d)(1+\|J\|)^2M^2$	(C.5)

Table 2: Summary of the constants used in Section 4.2 and where they are defined.

where more formally $\overline{\eta}^J$ is defined in Table 2 and β satisfies

$$\beta \geq \underline{\beta}^J = \max\left\{\Omega\left(\frac{C_J(\tilde{\varepsilon})^2}{m_J(\tilde{\varepsilon})\varepsilon^2}\left(d+\log\left(\frac{\mathcal{T}+1}{\delta}\right)\right)\right),\Omega\left(\frac{d\eta\log(\delta^{-1}\mathcal{T}^J_{rec}/\eta)}{\varepsilon^2}\right)\right\},$$

where more formally $\underline{\beta}^J$ is defined in Table 2, for any realization of training data z, with probability at least $1-\delta$ w.r.t. the Gaussian noise, at least one of the following events will occur:

1. $\|X_k - x_*\| \geq \frac{1}{2}\left(\varepsilon + re^{-m_J(\tilde{\varepsilon})k\eta}\right)$ *for some* $k \leq \eta^{-1}\mathcal{T}^J_{rec}$.
2. $\|X_k - x_*\| \leq \varepsilon + re^{-m_J(\tilde{\varepsilon})k\eta}$ *for every* $\eta^{-1}\mathcal{T}^J_{rec} \leq k \leq \eta^{-1}\mathcal{T}^J_{esc}$.

where $C_J(\tilde{\varepsilon}$ and $m_J(\tilde{\varepsilon}$ are given by (3.10).

Remark 16. *Notice that in Theorem 15, the definition of η and β are coupled since $\overline{\eta}^J$ depends on β and $\underline{\beta}^J$ depends on η. A closer look reveals that when η is sufficiently small, the first term in the definition of $\underline{\beta}^J$ dominates the second term and $\underline{\beta}^J$ is independent of η. So to satisfy the constraints in Theorem 15, it suffices to first choose β to be larger than the first term in $\underline{\beta}^J$ and then choose η to be sufficiently small.*

Remark 17. *Theorem 15 is about the empirical risk but the ideas can be generalized to population risk problems. In particular, we can apply Theorem 15 to obtain Theorem 40 for the population risk in the Appendix, similar to Theorem 3 in [TLR18].*

Remark 18. *In [TLR18], the overdamped Langevin algorithm is used and the recurrence time $\mathcal{T}_{rec} = \mathcal{O}\left(\frac{1}{m}\log(\frac{r}{\varepsilon})\right)$, and thus our recurrence time $\mathcal{T}^J_{rec} = \mathcal{O}\left(\frac{1}{m_J(\tilde{\varepsilon})}\log(\frac{r}{C_J(\tilde{\varepsilon})\varepsilon})\right)$ for the non-reversible Langevin algorithm, and if $C_J(\tilde{\varepsilon}) = \mathcal{O}(1)$, then $\mathcal{T}^J_{rec} = \mathcal{O}\left(\frac{1}{m_J(\tilde{\varepsilon})}\log(\frac{r}{\varepsilon})\right)$ which has the improvement over the overdamped Langevin algorithm since $m_J(\tilde{\varepsilon}) > m$ in general.*

5 Exit time for non-reversible Langevin dynamics

The escape time studied in the previous section quantifies the amount of time spent in a neighborhood of a local optimum. However, for convergence to the stationary distribution or to a small neighborhood of the global minimum, Langevin trajectory needs to not only escape from the neighborhood of a local optimum but also exit the basin of attraction of the current minimum and transition to the basin of attraction of other local minima including the global minima. In particular, the convergence rate to a global minimum is controlled by the mean *exit time* of a Langevin diffusion from the basin of attraction of a local minima in a potential landscape (i.e. $F(\cdot)$ in (1.3)) [BGK05]. The mean exit time is described by the celebrated Eyring-Kramers formula in statistical physics. The formula is named after Eyring and Kramers' respective papers [Eyr35, Kra40]. For surveys and recent developments, see, e.g. [Ber13, BR16, LMS17]. Although Kramer's formula is relatively well understood for reversible Langevin algorithms, it is not available for non-reversible Langevin diffusions for a general non-convex objective except for the double-well example [BR16, LMS17] described in Figure 1. Nevertheless, the double-well example is the simplest non-convex function that is considered as an important benchmark which provides intuition about the more general but much harder case of arbitrary smooth and non-convex objectives. Furthermore, the existing results are available in continuous time but not in discrete time.

Throughout this section, we consider a double-well potential $F : \mathbb{R}^d \to \mathbb{R}$, which has two local minima $a_1 < a_2$. The two local minima are separated by a saddle point σ. We will show that non-reversible Langevin dynamics in discrete-time can lead to faster (smaller) exit times. In addition to Assumption 9 (i)-(iii), we make generic assumptions that $F \in C^3$, the Hessian of F at each of the local minima is positive definite, and that the Hessian of F at the saddle point σ has exactly one strictly negative eigenvalue (denoted as $-\mu^*(\sigma) < 0$) and other eigenvalues are all positive.

5.1 Underdamped Langevin dynamics

Recall the underdamped Langevin diffusion defined in (1.7)–(1.8) and the underdamped Langevin dynamics (ULD) defined in (1.10)–(1.11). As ULD tracks the underdamped diffusion closely when the stepsize is small, we first discuss the exit times for the underdamped diffusion to get some intuition for the accerelation compared with the overdamped Langevin diffusion.

5.1.1 Acceleration of exit times in continuous time dynamics

Denote $\Theta^{\beta}_{a_1 \to a_2}$ as the first time of that the underdamped diffusion (1.7)–(1.8) starting from a_1 and hitting a small neighborhood of a_2. When dimension $d = 1$, it was derived originally in [Kra40] that

$$\mathbb{E}\left[\Theta^{\beta}_{a_1 \to a_2}\right] = [1 + o_{\beta}(1)] \cdot \frac{2\pi}{\mu_*} \cdot e^{\beta[F(\sigma) - F(a_1)]} \cdot \sqrt{\frac{|F''(\sigma)|}{F''(a_1)}}, \tag{5.1}$$

where

$$\mu_* = \frac{1}{2} \cdot \left(\sqrt{\gamma^2 - 4F''(\sigma)} - \gamma\right).$$

In view of (1.6), we can deduce that if

$$\gamma - F''(\sigma) < 1,$$

we have when $d = 1$,

$$\lim_{\beta\to\infty} \frac{\mathbb{E}\left[\Theta^{\beta}_{a_1\to a_2}\right]}{\mathbb{E}\left[\theta^{\beta}_{a_1\to a_2}\right]} = \frac{-F''(\sigma)}{\mu_*} < 1.$$

That is, for large β, the mean exit time of one dimensional underdamped diffusion is strictly smaller than that of the overdamped diffusion when $\mu_* > -F''(\sigma)$.

For general dimension d, Remark 5.2 in [BR16] suggests that[12] we have

$$\mathbb{E}\left[\Theta^{\beta}_{a_1\to a_2}\right] = [1 + o_\beta(1)] \cdot \frac{2\pi}{\mu_*} \cdot e^{\beta[F(\sigma)-F(a_1)]} \cdot \sqrt{\frac{|\det \mathrm{Hess}\ F(\sigma)|}{\det \mathrm{Hess}\ F(a_1)}},$$

where, with slight abuse of notations, μ_* is the unique positive eigenvalue of the matrix

$$\hat{H}_\gamma(\sigma) = \begin{bmatrix} -\gamma I & -\mathbb{L}^\sigma \\ I & 0 \end{bmatrix}, \tag{5.2}$$

where $\mathbb{L}^\sigma$ is the Hessian matrix of F at the saddle point σ. So to compare with the expected exit time for the overdamped diffusion in (1.6), we need to compare μ_* and $\mu^*(\sigma)$. One can readily show that[13] the unique positive eigenvalue μ_* of the matrix $\hat{H}_\gamma(\sigma)$ is given by the positive eigenvalue of the 2×2 matrix

$$\begin{bmatrix} -\gamma & \mu^*(\sigma) \\ 1 & 0 \end{bmatrix},$$

which suggests that

$$\mu_* = \frac{1}{2} \cdot \left(\sqrt{\gamma^2 + 4\mu^*(\sigma)} - \gamma\right). \tag{5.3}$$

So if

$$\gamma + \mu^*(\sigma) < 1, \tag{5.4}$$

where we recall that $-\mu^*(\sigma)$ is the unique negative eigenvalue of the Hessian of F at the saddle point σ, then we have $\mu_* \geq \mu^*(\sigma)$ and therefore for $d \geq 2$,

$$\lim_{\beta\to\infty} \frac{\mathbb{E}\left[\Theta^{\beta}_{a_1\to a_2}\right]}{\mathbb{E}\left[\theta^{\beta}_{a_1\to a_2}\right]} = \frac{\mu^*(\sigma)}{\mu_*} < 1. \tag{5.5}$$

That is, for sufficiently large β, we achieve acceleration; i.e. the mean exit time for the underdamped diffusion in arbitrary dimension is smaller compared with that of the overdamped diffusion. Roughly speaking, the condition (5.4) says that if the curvature of the

[12] For the underdamped diffusion, the diffusion matrix is not invertible. It is argued in Remark 5.2 in [BR16] that the analogue of (1.6), i.e., the Eyring-Kramers formula, can still be expected to hold in this underdamped case. A rigorous mathematical proof for general dimension is not known in the literature to the best of our knowledge.

[13] The argument of the proof is similar as that in Section 3.

saddle point in the negative descent direction is not too steep (i.e. if $\mu^*(\sigma) < 1$), we can choose γ small enough to accelerate the exit time of the reversible Langevin dynamics. Intuitively speaking, it can be argued that the underdamped process can climb hills faster and can explore the state space faster as it is less likely to go back to the recent states visited due to the momentum term [BR16]. However, an explanation of why and when underdamped process can accelerate overdamped dynamics is far from being understood. However, it is expected that γ should not be too large to achieve acceleration, because it is known that when γ gets larger, underdamped diffusion behaves like the overdamped diffusion [BR16, Section 5.5], [LMS15, Section 2.6], [LRS10]. When f is a quadratic, our results from Section 3 also show that γ needs to be properly chosen to get the fastest convergence rate and as γ gets larger than the threshold $2\sqrt{m}$, the convergence rate to the equilibrium deteriorates (see Figure 3.1).

In particular, if we choose $\gamma = \mathcal{O}(\sqrt{\mu^*(\sigma)})$, we see from (5.3) that $\mu_* = \mathcal{O}(\sqrt{\mu^*(\sigma)})$ and therefore, it follows from (5.5) that

$$\lim_{\beta\to\infty} \frac{\mathbb{E}\left[\Theta^{\beta}_{a_1\to a_2}\right]}{\mathbb{E}\left[\theta^{\beta}_{a_1\to a_2}\right]} = \frac{\mu^*(\sigma)}{\mu_*} = \mathcal{O}\left(\sqrt{\mu^*(\sigma)}\right) < 1. \tag{5.6}$$

This shows that amount of acceleration with underdamped dynamics obtained can be arbitrarily large and treating other parameters as a constant, underdamped dynamics' exit time improves upon that of overdamped dynamics by a square root factor in m (ignoring a $|\log(m)|$ term).

5.1.2 Acceleration of exit times in discrete time dynamics

Now we consider the discrete time dynamics (1.10)–(1.11). Consider appropriate choices of a bounded domain $\tilde{D}_R$ so that the underdamped SDE is non-degenerate along the normal direction to the boundary of $\tilde{D}_R$. The region $\tilde{D}_R$ contains a_1, a_2, σ, and as the parameter R grows to infinity, $\tilde{D}_R$ increases to the set $O^c(a_2)$, which is the complement of the set $O(a_2)$, i.e. $O^c(a_2) = \mathbb{R}^d \backslash O(a_2)$, and $O(a_2)$ denotes a small neighborhood of a_2. Denote $\hat{\Theta}^{\beta,R}_{a_1\to a_2}$ be the exit time of X_k (from the ULD dynamics) starting from a_1 and exiting domain $\tilde{D}_R$. Then one can expect that with a sufficiently large R and small stepsize η, the mean exit time for the discrete underdamped dynamics is close to that for the continuous dynamics. Specifically, fix the parameters β and γ in the underdamped Langevin dynamics, and fix $\epsilon > 0$. One can choose a sufficiently large R and choose a constant $\tilde{\eta}(\epsilon, R, \gamma, \beta)$ so that for stepsize $\eta \leq \tilde{\eta}(\epsilon, R, \gamma, \beta)$, we have [14]

$$\left| \mathbb{E}\left[\hat{\Theta}^{\beta,R}_{a_1\to a_2}\right] - \mathbb{E}\left[\Theta^{\beta}_{a_1\to a_2}\right] \right| < 2\epsilon. \tag{5.7}$$

Write $\hat{\theta}^{\beta,R}_{a_1\to a_2}$ as the exit time of X_k (from the overdamped discrete dynamics) starting from a_1 and exiting domain $\tilde{D}_R$. It then follows from (5.6), (5.7) and Proposition 20 that for large enough β and sufficiently small stepsize η, we obtain the acceleration in discrete time:

$$\frac{\mathbb{E}\left[\hat{\Theta}^{\beta,R}_{a_1\to a_2}\right]}{\mathbb{E}\left[\hat{\theta}^{\beta,R}_{a_1\to a_2}\right]} = \mathcal{O}\left(\sqrt{\mu^*(\sigma)}\right) < 1.$$

5.2 Non-reversible Langevin dynamics

Recall the non-reversible Langevin dynamics defined in (1.13) and the corresponding continuous time dynamics defined in (1.12).

5.2.1 Acceleration of exit times for continuous time dynamics

We first discuss the continuous-time dynamics (1.12). Theorem 5.2 in [LMS17] (see also [BR16]) showed[15] that the expected time of the diffusion $X(t)$ in (1.12) starting from a_1 and hitting a small neighborhood of a_2 is given by

$$\mathbb{E}\left[\tau^{\beta}_{a_1\to a_2}\right] = [1 + o_\beta(1)] \cdot \frac{2\pi}{\mu^*_J} \cdot e^{\beta[F(\sigma)-F(a_1)]} \cdot \sqrt{\frac{|\det \text{Hess } F(\sigma)|}{\det \text{Hess } F(a_1)}}. \tag{5.8}$$

Here, $o_\beta(1) \to 0$ as $\beta \to \infty$, $\det \text{Hess } F(x)$ stands for the determinant of the Hessian of F at x, and $-\mu^*_J$ is the unique negative eigenvalue of the matrix $A_J \cdot \mathbb{L}^\sigma$, where $\mathbb{L}^\sigma := \text{Hess } F(\sigma)$, the Hessian of F at the saddle point σ. The existence and uniqueness of such a negative eigenvalue $-\mu^*_J$ was proved in Lemma 11.1 in [LS18]. To facilitate the presentation, we denote u for the corresponding eigenvector of $A_J\mathbb{L}^\sigma$ for the eigenvalue $-\mu^*_J < 0$, i.e., we have

$$A_J\mathbb{L}^\sigma u = -\mu^*_J u. \tag{5.9}$$

In addition, since $\mathbb{L}^\sigma$ is a real symmetric matrix, we have

$$\mathbb{L}^\sigma = S^T D S, \tag{5.10}$$

[14] We apply Theorems 3.9 and 3.11 in [BGG17]. It suffices to verify their Assumptions 3.5 and 3.8. Assumption 3.5 in [BGG17] is automatically satisfied as the region $\tilde{D}_R$ is bounded, and the drift and the diffusion coefficients of underdamped dynamics are continuous. Assumption 3.8 in [BGG17] requires a mild non-characteristic boundary condition for the underdamped dynamics exiting $\tilde{D}_R$, but with our choice of $\tilde{D}_R$, the underdamped SDE is non-degenerate along the normal direction to the boundary of $\tilde{D}_R$, so this assumption is satisfied.

[15] Assumptions (P1)-(P4) in [LMS17] can be readily verified in our setting.

for a real orthogonal matrix S, where $D = \mathrm{diag}(\mu_1, \mu_2, \dots, \mu_n)$ with $\mu_1 < 0 < \mu_2 < \dots < \mu_n$ being the eigenvalues of $\mathbb{L}^\sigma$. As $A_J = I + J$, we have $-\mu_1 = \mu^*(\sigma)$.

In the case $d = 2$, it was checked in Section 5.1.4 of [BR16] that $\mu_J^* \geq \mu^*(\sigma)$. From (5.8), this suggests that $\mathbb{E}\left[\tau^\beta_{a_1 \to a_2}\right]$ is smaller for non-reversible Langevin diffusions compared with reversible Langevin diffusions with $J = 0$. That is, the non-reversible diffusion exits a local minimum faster. We now present a result showing that this holds for general dimension $d \geq 2$ as well.

Proposition 19. *We have $\mu_J^* \geq \mu^*(\sigma)$. As a consequence,*

$$\lim_{\beta \to \infty} \frac{\mathbb{E}\left[\tau^\beta_{a_1 \to a_2}\right]}{\mathbb{E}\left[\tau^\beta_{a_1 \to a_2}\right]_{J=0}} = \frac{\mu^*(\sigma)}{\mu_J^*} \leq 1. \tag{5.11}$$

The equality is attained if and only if $(Su)_i = 0$ for $i = 1, 2, \dots, n$ where u and S are defined by (5.9)–(5.10), which occurs if and only if u is a singular vector of J satisfying $Ju = 0$.

Proposition 19 shows that if J is not singular, the non-reversible dynamics is always faster. A natural question to ask is how much acceleration can be achieved for the exit time with the non-reversible dynamics. The approach we studied in Section 3 for quadratic functions does not apply here, because strongly convex quadratics cannot have negative eigenvalues in their Hessian, whereas at the saddle point the acceleration is achieved by making the negative eigenvalue larger in magnitude as shown in Proposition 19. We give a characterization of the speed-up in (D.3) in terms of the eigenvalues μ_i of the matrix $\mathbb{L}^\sigma$ as well as the eigenvector u of $A_J\mathbb{L}^\sigma$ in the Appendix. However, in principle the speed-up can be arbitrarily large. For instance, in the toy example in dimension two when

$$\mathbb{L}^\sigma = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix} \quad \text{with } a \in \mathbb{R},$$

after a simple computation we see that $\mu_J^* = \sqrt{1 + a^2}$ and $\mu^*(\sigma) = 1$ so that the ratio $\mu^*(\sigma)/\mu_J^* = 1/\sqrt{1 + a^2}$ in (5.11) can be made arbitrarily small.

5.2.2 Acceleration of exit times for discrete time dynamics

Next let us discuss the discrete dynamics (1.13). Let B_R be the ball centered at zero with radius R in $\mathbb{R}^d$. For R sufficiently large, we always have $a_1, a_2 \in B_R$. Let $D_R = B_R \backslash O(a_2)$, where $O(a_2)$ denotes a small neighborhood of a_2. Then it follows that the set D_R is bounded for each R, and it increases to the set $O^c(a_2)$ as R is sent to infinity.

Write $\hat{\tau}^{\beta, R}_{a_1 \to a_2}$ for the first time that the discrete-time dynamics starting from a_1 to exit the region D_R. Then we have the following result.

Proposition 20. *Fix the antisymmetric matrix J, the temperature parameter β, and $\epsilon > 0$. One can choose a sufficiently large R and a constant $\bar{\eta}(\epsilon, R, \beta)$ so that for stepsize $\eta \leq \bar{\eta}(\epsilon, R, \beta)$, we have*

$$\left| \mathbb{E}\left[\hat{\tau}^{\beta,R}_{a_1 \to a_2}\right] - \mathbb{E}\left[\tau^{\beta}_{a_1 \to a_2}\right] \right| < 2\epsilon.$$

It then follows from Proposition 19 that for large β we have

$$\frac{\mathbb{E}\left[\hat{\tau}^{\beta,R}_{a_1 \to a_2}\right]}{\mathbb{E}\left[\hat{\tau}^{\beta,R}_{a_1 \to a_2}\right]_{J=0}} < 1, \tag{5.12}$$

provided that $(Su)_i \neq 0$ for some $i \in \{2, \dots, d\}$ which occurs if and only if u is a singular vector of J satisfying $Ju = 0$.

6 Conclusion

Langevin Monte Carlo has been proven to be a powerful technique for sampling from a target distribution as well as for optimizing a non-convex objective. The classic Langevin algorithm is based on the first-order Langevin diffusion which is reversible in time. We study two variants that are based on non-reversible Langevin diffusions: the underdamped Langevin diffusion and the Langevin diffusion with a non-symmetric drift.

First, we consider the special case of quadratic objectives and derive an explicit characterization of the rate of convergence of the non-reversible Langevin diffusions to the stationary distribution in the 2-Wasserstein metric. Our exponential rate is optimal and unimprovable in the sense that it is achieved for some quadratic functions and initialization. Our results show that both non-reversible diffusions mix faster than the reversible Langevin, and we characterize the amount of improvement. In particular, the underdamped diffusion mixes with a rate that is faster than that of the reversible Langevin by a square root factor in m (ignoring a $|\log(m)|$ term) by tuning the friction coefficient γ properly.

Second, we give a refined analysis of non-reversible Langevin dynamics around a local minimum, by linearizing the gradient of the objective and building on our results for quadratic functions. Our results show that iterates for the non-reversible dynamics can both escape a local minima and exit from the basin of the attraction of a local minimum faster. Our results quantify the improvement that can be obtained in performance.

Third, we show that non-reversible Langevin dynamics in discrete time can exit the basin of attraction of a local minimum to find the global minimum faster when the objective has two local minima separated by a saddle point. We also discuss the amount of improvement.

By breaking the reversibility in the Langevin dynamics, our results quantify the improvement in performance and fill a gap between the theory and practice of non-reversible Langevin algorithms.

Acknowledgements

The authors are very grateful to Emmanuel Gobet, Claudio Landim, Insuk Seo and Gabriel Stoltz for helpful comments and 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

[Ber13] Nils Berglund. Kramers' law: Validity, derivations and generalisations. *Markov Processes and Related Fields*, 19(3):459–490, 2013.

[Ber15] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, Nashua, NH, 2015.

[BG02] Nils Berglund and Barbara Gentz. Beyond the Fokker-Planck equation: Pathwise control of noisy bistable systems. *Journal of Physics A: Mathematical and General*, 35(9):2057, 2002.

[BG03] Nils Berglund and Barbara Gentz. Geometric singular perturbation theory for stochastic differential equations. *Journal of Differential Equations*, 191:1–54, 2003.

[BGG12] François Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in wasserstein distance for fokker–planck equations. *Journal of Functional Analysis*, 263(8):2430–2457, 2012.

[BGG17] B. Bouchard, G. Geiss, and E. Gobet. First time to exit of a continuous Itô process: General moment estimates and L_1-convergence rate for discrete time approximatnions. *Bernoulli*, 23(3):1631–1662, 2017.

[BGK04] Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

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

[BLNR15] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *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.

[BR16] Freddy Bouchet and Julien Reygner. Generalisation of the Eyring-Kramers transition rate formula to irreversible diffusion processes. *Annales Henri Poincaré*, 17(12):3499–3532, 2016.

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

[Bub14] S. Bubeck. Theory of Convex Optimization for Machine Learning. *arXiv preprint arXiv:1405.4980*, May 2014.

[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.

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

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

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

[CFM+18] Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*, 2018.

[CGH+96] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996.

[DHN00] Persi Diaconis, Susan Holmes, and Radford M. Neal. Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.*, 10(3):726–752, 08 2000.

[DLP16] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of Statistical Physics*, 163(3):457–491, 2016.

[DNP17] A. B. Duncan, N. Nüsken, and G. A. Pavliotis. Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, 169(6):1098–1131, Dec 2017.

[DRD18] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

[EGZ17] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

[Eyr35] H. Eyring. The activated complex in chemical reactions. *J. Chem. Phys.*, 3(2):107–115, 1935.

[FGQ97] Haitao Fang, Guanglu Gong, and Minping Qian. Annealing of iterative stochastic schemes. *SIAM Journal on Control and Optimization*, 35(6):1886–1907, 1997.

[GBCB16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[Gel90] M. Gelbrich. On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[GGZ18] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of Stochastic Gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv:1809.04618*, 2018.

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

[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.

[GM05] E. Gobet and S. Maire. Sequential control variates for functionals of Markov processes. *SIAM Journal on Numerical Analysis*, 43(3):1256–1275, 2005.

[GM16] Arnaud Guillin and Pierre Monmarch. Optimal linear drift for the speed of convergence of an hypoelliptic diffusion. *Electron. Commun. Probab.*, 21:14 pp., 2016.

[GO07] Kranthi K Gade and Michael L Overton. Optimizing the asymptotic convergence rate of the Diaconis–Holmes–Neal sampler. *Advances in Applied Mathematics*, 38(3):382–403, 2007.

[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.

[HHMS93] C.-R. Hwang, S.-Y. Hwang-Ma, and S.-J. Sheu. Accelerating Gaussian diffusions. *Annals of Applied Probability*, 3:897–913, 1993.

[HHMS05] C.-R. Hwang, S.-Y. Hwang-Ma, and S.-J. Sheu. Accelerating diffusions. *Annals of Applied Probability*, 15:1433–1444, 2005.

[HKN04] B. Helffer, M. Klein, and F. Nier. Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Mat. Contemp.*, 26:41–86, 2004.

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

[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.

[Lin92] T. Lindvall. *Lectures on the Coupling Method.* Dover Publications, Mineola, NY, 1992.

[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.

[LMS17] C. Landim, M. Mariani, and I. Seo. Dirichlet's and Thomson's principles for non-selfadjoint elliptic operators with application to non-reversible metastable diffusion processes. *arXiv:1701.00985*, 2017.

[LNP13] T. Lelièvre, F. Nier, and G. A. Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, Jul 2013.

[LRS10] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. *Free energy computations: A mathematical perspective.* Imperial College Press, London, 2010.

[LS18] C. Landim and I. Seo. Metastability of nonreversible random walks in a potential field and the Eyring–Kramers transition rate formula. *Communications on Pure and Applied Mathematics*, 71(2):203–266, 2018.

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

[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.

[MT92] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes I: Criteria for discrete-time chains. *Advances in Applied Probability*, 24(3):542–574, 1992.

[Nes13] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013.

[Øks98] B. K. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag, New York, 5th edition, 1998.

[Pav14] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

[Ris89] H Risken. The fokker-planck equation. methods of solution and applications. *Springer Series in Synergetics, Berlin, New York: Springer,— c1989, 2nd ed.*, 1989.

[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.

[RW00] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Itô calculus*, volume 2. Cambridge University Press, 2000.

[S+09] Devavrat Shah et al. Gossip algorithms. *Foundations and Trends® in Networking*, 3(1):1–125, 2009.

[SBL16] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[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.

[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.

[Vil09] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

[WHC14] Sheng-Jhih Wu Wu, Chii-Ruey Hwang, and Moody T. Chu. Attaining the optimal Gaussian diffusion acceleration. *Journal of Statistical Physics*, 155(3):571–590, 2014.

[ZLC17] Yuchen Zhang, Percy Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. *arXiv preprint arXiv:1702.05575*, February 2017.

A Proof of results in Section 3

Proof of Lemma 1. Let H be a symmetric positive definite matrix with eigenvalue decomposition

$$H = QDQ^T \,,$$

where D is diagonal consisting of the eigenvalues in increasing order

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$$

of the matrix H. Let

$$m := \lambda_1 \leq M := \lambda_d \,,$$

be the lower and upper bounds on the eigenvalues, where we used the fact that L is the largest eigenvalue of the Hessian at the local minimum. Recall that

$$H_\gamma = \begin{bmatrix} \gamma I & H \\ -I & 0 \end{bmatrix}.$$

Note that

$$H_\gamma = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} G_\gamma \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix}, \quad G_\gamma := \begin{bmatrix} \gamma I & D \\ -I & 0 \end{bmatrix}.$$

Therefore H_γ and G_γ have the same eigenvalues. Due to the structure of G_γ, it can be seen that there exists a permutation matrix P such that

$$T_\gamma := PG_\gamma P^T = \begin{bmatrix} T_1(\gamma) & 0 & 0 & 0 \\ 0 & T_2(\gamma) & 0 & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & T_d(\gamma) \end{bmatrix}, \tag{A.1}$$

where

$$T_i(\gamma) = \begin{bmatrix} \gamma & \lambda_i \\ -1 & 0 \end{bmatrix}, \qquad i = 1, 2, \ldots, d \,,$$

are 2×2 block matrices with the eigenvalues:

$$\mu_{i,\pm} := \frac{\gamma \pm \sqrt{\gamma^2 - 4\lambda_i}}{2} \,. \qquad i = 1, 2, \ldots, d \,.$$

We observe that T_γ and G_γ (and therefore H_γ) have the same eigenvalues and the eigenvalues of T_γ are determined by the eigenvalues of the 2×2 block matrices $T_i(\gamma)$.

Since H_γ is unitarily equivalent to the matrix T_γ, i.e. there exists a unitary matrix U such that $H_\gamma = UT_\gamma U^*$, we have

$$R(t) := \left\| e^{-tH_\gamma} \right\| = \left\| U e^{-tT_\gamma} U^* \right\| = \left\| e^{-tT_\gamma} \right\| .$$

Since T_γ is a block diagonal matrix with 2×2 blocks $T_i(\gamma)$ we have

$$\left\|e^{-tT_\gamma}\right\| = \max_{1\le i\le d} \left\|e^{-tT_i(\gamma)}\right\|.$$

Assume that $\gamma^2 - 4\lambda_1 = \gamma^2 - 4m \le 0$ so that the eigenvalues of $T_i(\gamma)$

$$\mu_{i,\pm} := \frac{\gamma \pm \sqrt{\gamma^2 - 4\lambda_i}}{2}, \qquad 1 \le i \le d,$$

are real when $\gamma = 2\sqrt{m}$ and complex when $\lambda < 2\sqrt{m}$. Indeed, let us define

$$\tilde{T}_i(\gamma) := T_i(\gamma) - \frac{\gamma}{2} I, \qquad 1 \le i \le d.$$

Note that

$$\left\|e^{-tT_i(\gamma)}\right\| = e^{-t\gamma/2} \left\|e^{-t\tilde{T}_i(\gamma)}\right\| , \tag{A.2}$$

We consider $\gamma \in (0, 2\sqrt{m}]$. Depending on the value of λ_i and γ, there are two cases:

(i) If $\gamma < 2\sqrt{m}$ or ($\lambda_i > m$ and $\gamma = 2\sqrt{m}$), then $\tilde{T}_i(\gamma)$ has purely imaginary eigenvalues that are complex conjugates which we denote by

$$\tilde{\mu}_{i,\pm} = \pm i \frac{\sqrt{4\lambda_i - \gamma^2}}{2}, \qquad 1 \le i \le d.$$

We will show that the last term in (A.2) stays bounded due to the imaginariness of the eigenvalues of $\tilde{T}_i(\gamma)$. It is easy to check that 2×2 matrix $\tilde{T}_i(\gamma)$ have the eigenvectors $v_{i,\pm} = [\mu_{i,\pm}, -1]^T$. If we set

$$G_i = \begin{bmatrix} v_{i,+} & v_{i,-} \end{bmatrix} \in \mathbb{C}^{2\times 2},$$

the eigenvalue decomposition of $\tilde{T}_i(\gamma)$ is given by

$$\tilde{T}_i(\gamma) = G_i \begin{bmatrix} \tilde{\mu}_{i,+} & 0 \\ 0 & \tilde{\mu}_{i,-} \end{bmatrix} G_i^{-1}, \quad \text{where} \quad G_i^{-1} = \frac{1}{\det G_i} \begin{bmatrix} -1 & -\mu_{i,-} \\ 1 & \mu_{i,+} \end{bmatrix},$$

and

$$\det G_i = i\sqrt{4\lambda_i - \gamma^2}.$$

We can compute that

$$\begin{aligned} e^{-t\tilde{T}_i(\gamma)} &= G_i \begin{bmatrix} e^{-it\sqrt{4\lambda_i - \gamma^2}/2} & 0 \\ 0 & e^{it\sqrt{4\lambda_i-\gamma^2}/2} \end{bmatrix} G_i^{-1} \\ &= \frac{1}{\det G_i} \begin{bmatrix} \mu_{i,+} & \mu_{i,-} \\ -1 & -1 \end{bmatrix} \begin{bmatrix} -e^{-it\sqrt{4\lambda_i-\gamma^2}/2} & -\mu_{i,-} e^{-it\sqrt{4\lambda_i-\gamma^2}/2} \\ e^{it\sqrt{4\lambda_i-\gamma^2}/2} & \mu_{i,+} e^{it\sqrt{4\lambda_i-\gamma^2}/2} \end{bmatrix} \\ &= \frac{1}{i\sqrt{4\lambda_i - \gamma^2}} \begin{bmatrix} 2\text{Imag}\left(\mu_{i,-} e^{it\sqrt{4\lambda_i-\gamma^2}/2}\right) & 2i|\mu_{i,+}|^2 \sin\left(t\sqrt{4\lambda_i - \gamma^2}/2\right) \\ -2i \sin\left(t\sqrt{4\lambda_i-\gamma^2}/2\right) & 2\text{Imag}\left(\mu_{i,+} e^{it\sqrt{4\lambda_i-\gamma^2}/2}\right) \end{bmatrix}, \end{aligned}$$

where $\mathrm{Imag}(a + ib) := ib$ denotes the imaginary part of a complex number. As a consequence, by taking componentwise absolute values

$$
\begin{aligned}
\left\| e^{-t\tilde{T}_i(\gamma)} \right\| &\leq \frac{1}{\sqrt{4\lambda_i - \gamma^2}} \left\| \begin{bmatrix} 2|\mu_{i,-}| & 2|\mu_{i,+}|^2 \\ 2 & 2|\mu_{i,+}| \end{bmatrix} \right\| \\
&= \frac{1}{\sqrt{4\lambda_i - \gamma^2}} \left\| \begin{bmatrix} 2\sqrt{\lambda_i} & 2\lambda_i \\ 2 & 2\sqrt{\lambda_i} \end{bmatrix} \right\| \\
&= \frac{1}{\sqrt{4\lambda_i - \gamma^2}} \left\| \begin{bmatrix} 2\sqrt{\lambda_i} \\ 2 \end{bmatrix} \begin{bmatrix} 1 & \sqrt{\lambda_i} \end{bmatrix} \right\| \\
&= \frac{1}{\sqrt{4\lambda_i - \gamma^2}} \left\| \begin{bmatrix} 2\sqrt{\lambda_i} \\ 2 \end{bmatrix} \right\| \left\| \begin{bmatrix} 1 & \sqrt{\lambda_i} \end{bmatrix} \right\| \\
&= \frac{2(1 + \lambda_i)}{\sqrt{4\lambda_i - \gamma^2}} ,
\end{aligned}
\tag{A.3}
$$

where in the second from last inequality we used the fact that the 2-norm of a rank-one matrix is equal to its Frobenius norm. [16] Then, it follows from (A.2) that

$$
\left\| e^{-tT_i(\gamma)} \right\| = e^{-t\gamma/2} \left\| e^{-t\tilde{T}_i(\gamma)} \right\| \leq \frac{2(1 + \lambda_i)}{\sqrt{4\lambda_i - \gamma^2}} e^{-t\gamma/2} ,
$$

which implies

$$
\left\| e^{-tH_\gamma} \right\| = \left\| e^{-tT_\gamma} \right\| \leq \max_{1 \leq i \leq d} \left\| e^{-tT_i(\gamma)} \right\| \leq \frac{2(1 + M)}{\sqrt{4m - \gamma^2}} e^{-t\gamma/2} ,
$$

provided that $\gamma^2 - 4m < 0$. In particular, if we choose $\hat{\varepsilon} = 1 - \frac{\gamma}{2\sqrt{m}}$ for any $\hat{\varepsilon} > 0$, we obtain

$$
\left\| e^{-tH_\gamma} \right\| \leq \frac{1 + M}{\sqrt{m(1 - (1 - \hat{\varepsilon})^2)}} e^{-\sqrt{m}(1-\hat{\varepsilon})t} .
$$

The proof for (i) is complete.

(ii) If $\gamma = 2\sqrt{m}$ and $\lambda_i = m$, then $\tilde{T}_i(\gamma)$ has double eigenvalues at zero and is not diagonalizable. It admits the Jordan decomposition

$$
\tilde{T}_i(\gamma) = G_i \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} G_i^{-1} \quad \text{with} \quad G_i = \begin{bmatrix} \sqrt{m} & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad G_i^{-1} = \begin{bmatrix} 0 & -1 \\ 1 & \sqrt{m} \end{bmatrix} .
$$

[16] The 2-norm of a rank-one matrix $R = uv^*$ should be exactly equal to $\sigma = \|u\|\|v\|$. This follows from the fact that we can write $R = \sigma \tilde{u}\tilde{v}^T$ where $\tilde{u}$ and $\tilde{v}$ have unit norm. This would be a singular value decomposition of R, showing that all the singular values are zero except a singular value at σ. Because the 2-norm is equal to the largest singular value, the 2-norm of R is equal to σ.

By a direct computation, we obtain

$$e^{-t\tilde{T}_i(\gamma)} = G_i \begin{bmatrix} 1 & -t \\ 0 & 1 \end{bmatrix} G_i^{-1} = \begin{bmatrix} 1 - t\sqrt{m} & -tm \\ t & 1 + t\sqrt{m} \end{bmatrix}.$$

A simple computation reveals

$$\left\| e^{-t\tilde{T}_i(\gamma)} \right\| \leq \sqrt{\operatorname{Tr}\left(e^{-t\tilde{T}_i(\gamma)} e^{-t\tilde{T}_i(\gamma)^T} \right)} = \sqrt{2 + (m+1)^2 t^2}. \tag{A.4}$$

To finish the proof of part (ii), let $\gamma = 2\sqrt{m}$. We compute

$$\begin{aligned} \max_{1 \leq i \leq d} \left\| e^{-t\tilde{T}_i(\gamma)} \right\| &= \max \left\{ \max_{i:\lambda_i = m} \left\| e^{-t\tilde{T}_i(\gamma)} \right\|, \max_{i:\lambda_i > m} \left\| e^{-t\tilde{T}_i(\gamma)} \right\| \right\} \\ &\leq \max \left\{ \sqrt{2 + (m+1)^2 t^2}, \max_{i:\lambda_i > m} \frac{(1 + \lambda_i)}{\sqrt{\lambda_i - m}} \right\}, \end{aligned}$$

where we used (A.3) and (A.4) in the last inequality. We conclude from (A.2) for part (ii).
□

Proof of Theorem 3. Consider a coupling $(V(t), X(t))$ and $(\tilde{V}(t), \tilde{X}(t))$ starting at $(V(0), X(0))$ and $(V(0), \tilde{X}(0))$ respectively in the same probability space with the same Brownian motion B_t. This implies that

$$d \begin{bmatrix} X(t) - \tilde{X}(t) \\ V(t) - \tilde{V}(t) \end{bmatrix} = -H_\gamma \begin{bmatrix} X(t) - \tilde{X}(t) \\ V(t) - \tilde{V}(t) \end{bmatrix} dt.$$

It follows that

$$\begin{bmatrix} X(t) - \tilde{X}(t) \\ V(t) - \tilde{V}(t) \end{bmatrix} = e^{-H_\gamma t} \begin{bmatrix} X(0) - \tilde{X}(0) \\ 0 \end{bmatrix},$$

and thus

$$\|X(t) - \tilde{X}(t)\| \leq \left\| \begin{bmatrix} X(t) - \tilde{X}(t) \\ V(t) - \tilde{V}(t) \end{bmatrix} \right\| \leq \|X(0) - \tilde{X}(0)\| \cdot \left\| e^{-H_\gamma t} \right\|.$$

Assume that $\tilde{X}$ is stationary and follows π, and recall that $X(0)$ is deterministic and satisfies $\|X(0)\| \leq R$. Then, we get

$$\mathcal{W}_2(\pi_{t,\gamma}, \pi) \leq \left\| e^{-H_\gamma t} \right\| \left(\int_{\mathbb{R}^d} \|X(0) - x\|^2 \pi(dx) \right)^{1/2}.$$

Taking an infimum of the left-hand side over all couplings leads to

$$\mathcal{W}_2(\pi_{t,\gamma}, \pi) \leq \left\| e^{-H_\gamma t} \right\| \mathcal{W}_2(\pi_{0,\gamma}, \pi),$$

which using Lemma 1, proves parts (i) and (ii). To prove part (iii), we estimate

$$\mathcal{W}_2(\pi_{0,\gamma},\pi) \leq \left(2R^2 + 2\int_{\mathbb{R}^d} \|x\|^2 \pi(dx)\right)^{1/2} \leq \left(2R^2 + \frac{2}{m}(b + d/\beta)\right)^{1/2},$$

where we used (3.19) in [RRT17] for the last step. By Lemma 1, the conclusion follows. □

Proof of Theorem 8. Consider a coupling $X(t)$ and $\tilde{X}(t)$ starting at $X(0)$ and $\tilde{X}(0)$ respectively in the same probability space with the same Brownian motion B_t. This implies that

$$d(X(t) - \tilde{X}(t)) = -A_J H(X(t) - \tilde{X}(t))dt.$$

It follows that

$$X(t) - \tilde{X}(t) = (X(0) - \tilde{X}(0))e^{-A_J H t},$$

and thus

$$\|X(t) - \tilde{X}(t)\| \leq \|X(0) - \tilde{X}(0)\| \cdot \left\|e^{-A_J H t}\right\|.$$

Assume that $\tilde{X}$ is stationary and follows π, and $X(0)$ is deterministic with $\|X(0)\| R$. Following the same argument as in the proof of Theorem 3, we get

$$\mathcal{W}_2(\pi_{t,J},\pi) \leq \left\|e^{-A_J H t}\right\| \cdot \left(2R^2 + \frac{2}{m}(b + d/\beta)\right)^{1/2}.$$

By Lemma 6, the conclusion follows. □

B Proof of results in Section 4.1

B.1 Proof of Theorem 10

The main result we use to prove Theorem 10 is the following proposition. The proof of the following result will be presented later in Section B.1.2.

Proposition 21. *Assume $\gamma = 2\sqrt{m}$. Fix any $r > 0$ and*

$$0 < \varepsilon < \min\left\{\overline{\varepsilon}_1^U, \overline{\varepsilon}_2^U, \overline{\varepsilon}_3^U\right\},$$

where

$$\overline{\varepsilon}_1^U := \sqrt{\frac{C_H + 2 + (m+1)^2}{(C_H + 2)m + (m+1)^2}}r, \tag{B.1}$$

$$\overline{\varepsilon}_2^U := 2\sqrt{2}\left(C_H + 2 + (m+1)^2\right)^{1/4} \frac{e^{-1/2}r}{m^{1/4}}, \tag{B.2}$$

$$\overline{\varepsilon}_3^U := \frac{\sqrt{m}}{4L\left(\sqrt{C_H + 2} + \frac{m+1}{\sqrt{m}} + \frac{\sqrt{(C_H+2)m + (m+1)}}{8\sqrt{C_H + 2 + (m+1)^2}}\right)}. \tag{B.3}$$

Consider the stopping time:

$$\tau := \inf\left\{t \geq 0 : \|X(t) - x_*\| \geq \varepsilon + re^{-\sqrt{m}t}\right\}.$$

For any initial point $X(0) = x$ with $\|x - x_*\| \leq r$, and

$$\beta \geq \frac{256(2C_H m + 4m + (m+1)^2)}{m\varepsilon^2}\left(d\log(2) + \log\left(\frac{2\|H_{2\sqrt{m}}\|\mathcal{T} + 1}{\delta}\right)\right),$$

we have

$$\mathbb{P}_x\left(\tau \in [\mathcal{T}^U_{rec}, \mathcal{T}^U_{esc}]\right) \leq \delta.$$

We are now ready to complete the proof of Theorem 10.

B.1.1 Completing the proof of Theorem 10

Assume that $\gamma = 2\sqrt{m}$. Let us compare the discrete dynamics (1.10)-(1.11) and the continuous dynamics (1.7)-(1.8). Define:

$$\tilde{V}(t) = V_0 - \int_0^t \gamma\tilde{V}\left(\lfloor s/\eta\rfloor\eta\right)ds - \int_0^t \nabla F\left(\tilde{X}\left(\lfloor s/\eta\rfloor\eta\right)\right)ds + \sqrt{2\gamma\beta^{-1}}\int_0^t dB_s, \tag{B.4}$$

$$\tilde{X}(t) = X_0 + \int_0^t \tilde{V}\left(\lfloor s/\eta\rfloor\eta\right)ds. \tag{B.5}$$

The process $(\tilde{V}, \tilde{X})$ defined in (B.4) and (B.5) is the continuous-time interpolation of the iterates $\{(V_k, X_k)\}$. In particular, the joint distribution of $\{(V_k, X_k) : k = 1, 2, \ldots, K\}$ is the same as $\{(\tilde{V}(t), \tilde{X}(t)) : t = \eta, 2\eta, \ldots, K\eta\}$ for any positive integer K.

It is derived in [GGZ18] that the relative entropy $D(\cdot\|\cdot)$ between the law $\tilde{\mathbb{P}}^{K\eta}$ of $((\tilde{V}(t), \tilde{X}(t)) : t \leq K\eta)$ and the law $\mathbb{P}^{K\eta}$ of $((V(t), X(t)) : t \leq K\eta)$ is upper bounded as follows:

$$D\left(\tilde{\mathbb{P}}^{K\eta}\middle\|\mathbb{P}^{K\eta}\right) \leq \frac{\beta M^2}{2\gamma}C_v^d K\eta^2,$$

provided that $\eta \leq \min\left\{\frac{\gamma}{K_2}(d/\beta + \overline{A}/\beta), \frac{\gamma\lambda}{2K_1}\right\}$, where C_v^d is defined in Lemma 24. Using Pinsker's inequality, we obtain an upper bound on the total variation $\|\cdot\|_{TV}$:

$$\left\|\tilde{\mathbb{P}}^{K\eta} - \mathbb{P}^{K\eta}\right\|_{TV}^2 \leq \frac{\beta M^2}{4\gamma}C_v^d K\eta^2.$$

Using a result about an optimal coupling (Theorem 5.2., [Lin92]), that is, given any two random elements $\mathcal{X}, \mathcal{Y}$ of a common standard Borel space, there exists a coupling $\mathcal{P}$ of $\mathcal{X}$ and $\mathcal{Y}$ such that

$$\mathcal{P}(\mathcal{X} \neq \mathcal{Y}) \leq \|\mathcal{L}(\mathcal{X}) - \mathcal{L}(\mathcal{Y})\|_{TV}.$$

Hence, given any $\beta > 0$ and $K\eta \le \mathcal{T}_{\mathrm{esc}}^U$, we can choose

$$\eta \le \frac{4\gamma\delta^2}{\beta M^2 C_v^d \mathcal{T}_{\mathrm{esc}}^U}, \tag{B.6}$$

so that there is a coupling of $\{(V(k\eta), X(k\eta)) : k = 1, 2, \dots, K\}$ and $\{(V_k, X_k) : k = 1, 2, \dots, K\}$ such that

$$\mathcal{P}(((V(\eta), X(\eta)), \dots, (V(K\eta), X(K\eta))) \ne ((V_1, X_1), \dots, (V_K, X_K)) \le \delta. \tag{B.7}$$

It follows that

$$\mathbb{P}(((V_1, X_1), \dots, (V_K, X_K)) \in \cdot) \le \mathbb{P}(((V(\eta), X(\eta)), \dots, (V(K\eta), X(K\eta))) \in \cdot) + \delta.$$

Let us now complete the proof of Theorem 10. We need to show that

$$\mathbb{P}\left((X_1, \dots, X_K) \in \mathcal{A}\right) \le \delta,$$

where $K = \lfloor \eta^{-1}\mathcal{T}_{\mathrm{esc}}^U \rfloor$ and $\mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2$, where

$$\mathcal{A}_1 := \left\{ (x_1, \dots, x_K) \in (\mathbb{R}^d)^K : \max_{k \le \eta^{-1}\mathcal{T}_{\mathrm{rec}}^U} \frac{\|x_k - x_*\|}{\varepsilon + re^{-\sqrt{m}k\eta}} \le \frac{1}{2} \right\},$$
$$\mathcal{A}_2 := \left\{ (x_1, \dots, x_K) \in (\mathbb{R}^d)^K : \max_{\eta^{-1}\mathcal{T}_{\mathrm{rec}}^U \le k \le K} \frac{\|x_k - x_*\|}{\varepsilon + re^{-\sqrt{m}k\eta}} \ge 1 \right\}.$$

We can choose β sufficiently large so that with probability at least $1 - \delta/3$, we have either $\|X(t) - x_*\| \ge \varepsilon + re^{-\sqrt{m}t}$ for some $t \le \mathcal{T}_{\mathrm{rec}}^U$ or $\|X(t) - x_*\| \le \varepsilon + re^{-\sqrt{m}t}$ for all $t \le \mathcal{T}_{\mathrm{esc}}^U$. Moreover, for any K, η and β satisfying the conditions of the theorem, there exists a coupling of $(X(\eta), \dots, X(K\eta))$ and $(X_1, \dots, X_K)$ so that with probability $1 - \delta/3$, $X_k = X(k\eta)$ for all $k \in [K]$. Then, by (B.6) and (B.7), we get

$$\mathbb{P}((X_1, \dots, X_K) \in \mathcal{A}) \le \mathbb{P}((X(\eta), \dots, X(K\eta)) \in \mathcal{A}) + \frac{\delta}{3}, \tag{B.8}$$

provided that

$$\eta \le \overline{\eta}_3^U := \frac{4\gamma\delta^2}{9\beta M^2 C_v^d \mathcal{T}_{\mathrm{esc}}^U}. \tag{B.9}$$

It remains to estimate the probability of $\mathbb{P}((X(\eta), \dots, X(K\eta)) \in \mathcal{A}_1 \cap \mathcal{A}_2)$ for the underdamped Langevin diffusion. Partition the interval $[0, \mathcal{T}_{\mathrm{rec}}^U]$ using the points $0 = t_1 < t_1 < \dots < t_{\lceil \eta^{-1}\mathcal{T}_{\mathrm{rec}}^U \rceil} = \mathcal{T}_{\mathrm{rec}}^U$ with $t_k = k\eta$ for $k = 0, 1, \dots, \lceil \eta^{-1}\mathcal{T}_{\mathrm{rec}}^U \rceil - 1$, and consider the event:

$$\mathcal{B} := \left\{ \max_{0 \le k \le \lceil \eta^{-1}\mathcal{T}_{\mathrm{rec}}^U \rceil - 1} \max_{t \in [t_k, t_{k+1}]} \|X(t) - X(t_{k+1})\| \le \frac{\varepsilon}{2} \right\}.$$

On the event $\{(X(\eta),\dots,X(K\eta))\in\mathcal{A}_1\}\cap\mathcal{B}$,

$$\sup_{t\in[0,\mathcal{T}^U_{\mathrm{rec}}]}\frac{\|X(t)-x_*\|}{\varepsilon+re^{-\sqrt{m}t}}=\max_{0\le k\le\lceil\eta^{-1}\mathcal{T}^U_{\mathrm{rec}}\rceil-1}\sup_{t\in[t_k,t_{k+1}]}\frac{\|X(t)-x_*\|}{\varepsilon+re^{-\sqrt{m}t}}$$
$$\le\frac{1}{2}+\max_{0\le k\le\lceil\eta^{-1}\mathcal{T}^U_{\mathrm{rec}}\rceil-1}\max_{t\in[t_k,t_{k+1}]}\frac{1}{\varepsilon}\|X(t)-X(t_{k+1})\|<1,$$

and thus

$$\begin{aligned}\mathbb{P}((X(\eta),\cdots,X(K\eta))\in\mathcal{A})&\le\mathbb{P}(\{(X(\eta),\cdots,X(K\eta))\in\mathcal{A}\}\cap\mathcal{B})+\mathbb{P}(B^c)\\&\le\mathbb{P}(\tau\in[\mathcal{T}^U_{\mathrm{rec}},\mathcal{T}^U_{\mathrm{esc}}])+\mathbb{P}(\mathcal{B}^c)\\&\le\frac{\delta}{3}+\mathbb{P}(\mathcal{B}^c)\,,\end{aligned}\tag{B.10}$$

provided that (by applying Proposition 21 and Lemma 41) (with $\gamma=2\sqrt{m}$):

$$\beta\ge\underline{\beta}^U_1:=\frac{256(2C_Hm+4m+(m+1)^2)}{m\varepsilon^2}\left(d\log(2)+\log\left(\frac{6\sqrt{4m+M^2+1}\mathcal{T}+3}{\delta}\right)\right).\tag{B.11}$$

To complete the proof, we need to show that $\mathbb{P}(\mathcal{B}^c)\le\frac{\delta}{3}$ in view of (B.8) and (B.10). For any $t\in[t_k,t_{k+1}]$, where $t_{k+1}-t_k=\eta$, we have

$$\|X(t)-X(t_{k+1})\|\le\int_t^{t_{k+1}}\|V(s)\|ds\le\eta\|V(t_{k+1})\|+\int_t^{t_{k+1}}\|V(s)-V(t_{k+1})\|ds,\tag{B.12}$$

and

$$\begin{aligned}&\|V(t)-V(t_{k+1})\|\\&\le\gamma\int_t^{t_{k+1}}\|V(s)\|ds+\int_t^{t_{k+1}}\|\nabla F(X(s))\|ds+\sqrt{2\gamma\beta^{-1}}\|B_t-B_{t_{k+1}}\|\\&\le\gamma\eta\|V(t_{k+1})\|+\gamma\int_t^{t_{k+1}}\|V(s)-V(t_{k+1})\|ds\\&\qquad+M\int_t^{t_{k+1}}\|X(s)-X(t_{k+1})\|ds+\eta\|\nabla F(X(t_{k+1}))\|+\sqrt{2\gamma\beta^{-1}}\|B_t-B_{t_{k+1}}\|\\&\le\gamma\eta\|V(t_{k+1})\|+\gamma\int_t^{t_{k+1}}\|V(s)-V(t_{k+1})\|ds\\&\qquad+M\int_t^{t_{k+1}}\|X(s)-X(t_{k+1})\|ds+M\eta\|X(t_{k+1})\|+B\eta+\sqrt{2\gamma\beta^{-1}}\|B_t-B_{t_{k+1}}\|\,,\end{aligned}\tag{B.13}$$

where the second inequality above used M-Lipschitz property of ∇F and the last inequality above used Lemma 43. By adding the above two inequalities (B.12) and (B.13) together, we get

$$
\begin{aligned}
&\|X(t) - X(t_{k+1})\| + \|V(t) - V(t_{k+1})\| \\
&\leq (1+\gamma)\eta\|V(t_{k+1})\| + (1+\gamma)\int_t^{t_{k+1}} \|V(s) - V(t_{k+1})\| ds \\
&\qquad + M\int_t^{t_{k+1}} \|X(s) - X(t_{k+1})\| ds + M\eta\|X(t_{k+1})\| + B\eta + \sqrt{2\gamma\beta^{-1}}\|B_t - B_{t_{k+1}}\| \\
&\leq (1+\gamma+M)\int_t^{t_{k+1}} \left(\|V(s) - V(t_{k+1})\| + \|X(s) - X(t_{k+1})\|\right) ds \\
&\qquad + (1+\gamma)\eta\|V(t_{k+1})\| + M\eta\|X(t_{k+1})\| + B\eta + \sqrt{2\gamma\beta^{-1}} \sup_{t\in[t_k,t_{k+1}]} \|B_t - B_{t_{k+1}}\|.
\end{aligned}
$$

By applying Gronwall's inequality, we get

$$
\begin{aligned}
&\sup_{t\in[t_k,t_{k+1}]} \left[\|X(t) - X(t_{k+1})\| + \|V(t) - V(t_{k+1})\|\right] \\
&\leq e^{(1+\gamma+M)\eta}\left[(1+\gamma)\eta\|V(t_{k+1})\| + M\eta\|X(t_{k+1})\| + B\eta + \sqrt{2\gamma\beta^{-1}} \sup_{t\in[t_k,t_{k+1}]} \|B_t - B_{t_{k+1}}\|\right].
\end{aligned}
\tag{B.14}
$$

We have from Lemma 24 that for any $u > 0$,

$$
\mathbb{P}(\|V(t_{k+1})\| \geq u) \leq \frac{\sup_{t>0}\mathbb{E}\|V(t)\|^2}{u^2} \leq \frac{C_v^c}{u^2}, \tag{B.15}
$$

and

$$
\mathbb{P}(\|X(t_{k+1})\| \geq u) \leq \frac{\sup_{t>0}\mathbb{E}\|X(t)\|^2}{u^2} \leq \frac{C_x^c}{u^2}, \tag{B.16}
$$

where C_v^c, C_x^c are defined in Lemma 24. By Lemma 42, we have

$$
\mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}]} \|B_t - B_{t_{k+1}}\| \geq u\right) \leq 2^{1/4} e^{1/4} e^{-\frac{u^2}{4d\eta}}.
$$

Therefore, we can infer from (B.14) that with $K_0 := \lceil \eta^{-1} \mathcal{T}_{\rm rec}^U \rceil$,

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{B}^c\right) \\
&\leq \sum_{k=0}^{K_0-1} \mathbb{P}\left(\|X(t_{k+1})\| \geq \frac{\varepsilon e^{-(1+\gamma+M)\eta}}{8M\eta}\right) + \sum_{k=0}^{K_0-1} \mathbb{P}\left(\|V(t_{k+1})\| \geq \frac{\varepsilon e^{-(1+\gamma+M)\eta}}{8(1+\gamma)\eta}\right) \\
&\qquad + \sum_{k=0}^{K_0-1} \mathbb{P}\left(B \geq \frac{\varepsilon e^{-(1+\gamma+M)\eta}}{8\eta}\right) + \sum_{k=0}^{K_0-1} \mathbb{P}\left(\sup_{t\in[t_k,t_{k+1}]} \|B_t - B_{t_{k+1}}\| \geq \frac{\varepsilon e^{-(1+\gamma+M)\eta}\sqrt{\beta}}{8\sqrt{2\gamma}}\right) \\
&\leq \frac{64K_0}{\varepsilon^2}\left(M^2 C_x^c + (1+\gamma)^2 C_v^c\right)\cdot \eta^2 e^{2(1+\gamma+M)\eta} \qquad \text{(B.17)} \\
&\qquad\qquad + 2^{1/4}e^{1/4}K_0 \cdot \exp\left(-\frac{1}{4d\eta}\frac{\varepsilon^2 e^{-2(1+\gamma+M)\eta}\beta}{128\gamma}\right) \qquad \text{(B.18)} \\
&\qquad\qquad + K_0 \mathbb{P}\left(B \geq \frac{\varepsilon e^{-(1+\gamma+M)\eta}}{8\eta}\right), \qquad \text{(B.19)}
\end{aligned}
$$

where the last inequality follows from (B.15), (B.16) and Lemma 42. We can choose $\eta \leq 1$ so that

$$
\eta \leq \overline{\eta}_2^U := \frac{\delta \varepsilon^2 e^{-2(1+\gamma+M)}}{384(M^2 C_x^c + (1+\gamma)^2 C_v^c)\mathcal{T}_{\rm rec}^U}, \tag{B.20}
$$

so that the term in (B.17) is less than $\delta/6$, where C_v^c, C_x^c are defined in Lemma 24, and then we choose β so that

$$
\beta \geq \underline{\beta}_2^U := \frac{512 d\eta\gamma \log(2^{1/4}e^{1/4}6\delta^{-1}\mathcal{T}_{\rm rec}^U/\eta)}{\varepsilon^2 e^{-2(1+\gamma+M)\eta}}, \tag{B.21}
$$

so that the term in (B.18) is also less than $\delta/6$, and we can choose η so that $\eta \leq 1$ and

$$
\eta \leq \overline{\eta}_1^U := \frac{\varepsilon e^{-(1+\gamma+M)}}{8B}, \tag{B.22}
$$

so that the term in (B.19) is zero.

To complete the proof, let us work on the leading orders of the constants. For the sake of convenience, we hide the dependence on M and L and assume that $M, L = \mathcal{O}(1)$. We also assume that $C_H = \mathcal{O}(1)$. Recall that $0 < \varepsilon \leq \min\{\overline{\varepsilon}_1^U, \overline{\varepsilon}_2^U, \overline{\varepsilon}_3^U\}$, where it is easy to check that It is easy to check that

$$
\overline{\varepsilon}_1^U = \sqrt{\frac{C_H + 2 + (m+1)^2}{(C_H+2)m + (m+1)^2}} r \geq \Omega\left(\frac{C_H^{1/2} r}{C_H^{1/2} m^{1/2} + m + 1}\right) \geq \Omega(r),
$$

where we used $m \leq M = \mathcal{O}(1)$ and

$$
\overline{\varepsilon}_2^U = 2\sqrt{2}(C_H + 2 + (m+1)^2)^{1/4}\frac{e^{-1/2}r}{m^{1/4}} \geq \Omega\left(\frac{(1+C_H^{1/4})r}{m^{1/4}}\right) \geq \Omega\left(\frac{r}{m^{1/4}}\right),
$$

and

$$\overline{\varepsilon}_3^U = \frac{\sqrt{m}}{4L\left(\sqrt{C_H+2}+\frac{m+1}{\sqrt{m}}+\frac{\sqrt{(C_H+2)m}+(m+1)}{8\sqrt{C_H+2+(m+1)^2}}\right)} \geq \Omega\left(\frac{\sqrt{m}}{L\left(1+\frac{m+1}{\sqrt{m}}+\frac{\sqrt{m}}{m+1}\right)}\right) \geq \Omega(m),$$

where we used the fact that $m+1 \geq 2\sqrt{m}$. Hence, we can take

$$\varepsilon \leq \min\left\{\mathcal{O}\left(r\right), \mathcal{O}\left(\frac{r}{m^{1/4}}\right), \mathcal{O}(m)\right\}.$$

Moreover, $m \leq M = \mathcal{O}(1)$. Hence, we can take

$$\varepsilon \leq \min\left\{\mathcal{O}(r), \mathcal{O}(m)\right\}.$$

Next, we recall the recurrence time:

$$\mathcal{T}_{\text{rec}}^U = -\frac{1}{\sqrt{m}}W_{-1}\left(\frac{-\varepsilon^2\sqrt{m}}{8r^2\sqrt{C_H+2+(m+1)^2}}\right),$$

and since $W_{-1}(-x) \sim \log(1/x)$ for $x \to 0^+$, and we assume $C_H = \mathcal{O}(1)$, we get

$$\mathcal{T}_{\text{rec}}^U = \mathcal{O}\left(\frac{1}{\sqrt{m}}\log\left(\frac{r}{\varepsilon m}\right)\right) \leq \mathcal{O}\left(\frac{|\log(m)|}{\sqrt{m}}\log\left(\frac{r}{\varepsilon}\right)\right).$$

Next, we recall that stepsize η satisfies $\eta \leq \min\{1, \overline{\eta}_1^U, \overline{\eta}_2^U, \overline{\eta}_3^U, \overline{\eta}_4^U\}$ and it is easy to check that

$$\overline{\eta}_1^U = \frac{\varepsilon e^{-(1+2\sqrt{m}+M)}}{8B} \geq \Omega\left(\varepsilon e^{-(2m^{1/2}+M)}\right) \geq \Omega(\varepsilon),$$

and

$$\overline{\eta}_2^U = \frac{\delta\varepsilon^2 e^{-2(1+2\sqrt{m}+M)}}{384(M^2C_x^c+(1+2\sqrt{m})^2C_v^c)\mathcal{T}_{\text{rec}}^U} \geq \Omega\left(\frac{\delta\varepsilon^2 e^{-(4m^{1/2}+2M)}}{(M^2C_x^c+(1+m)C_v^c)\mathcal{T}_{\text{rec}}^U}\right).$$

Moreover, we have (note that $R = \sqrt{b/m}$ in the definition of C_x^c, C_v^c)

$$C_x^c \leq \mathcal{O}\left(\frac{1+\frac{1}{m}+\frac{d}{\beta}}{m}\right), \qquad C_v^c \leq \mathcal{O}\left(1+\frac{1}{m}+\frac{d}{\beta}\right),$$

together with $m \leq M = \mathcal{O}(1)$ implies that

$$\overline{\eta}_2^U = \frac{\delta\varepsilon^2 e^{-2(1+2\sqrt{m}+M)}}{384(M^2C_x^c+(1+2\sqrt{m})^2C_v^c)\mathcal{T}_{\text{rec}}^U} \geq \Omega\left(\frac{m^2\beta\delta\varepsilon^2}{(md+\beta)\mathcal{T}_{\text{rec}}^U}\right).$$

Moreover,

$$\overline{\eta}_3^U = \frac{8\sqrt{m}\delta^2}{9\beta M^2 C_v^d \mathcal{T}_{\rm esc}^U} \geq \Omega\left(\frac{m^{3/2}\delta^2}{(md+\beta)\mathcal{T}_{\rm esc}^U}\right),$$

and

$$\overline{\eta}_4^U = \min\left\{\frac{2\sqrt{m}}{K_2}\frac{d+\overline{A}}{\beta}, \frac{\sqrt{m}\lambda}{K_1}\right\} \geq \min\left\{\Omega\left(\frac{m^{1/2}\beta}{d+\beta}\right), \Omega(m^{1/2}\beta)\right\}$$

and the minimum between $\frac{m^{1/2}\beta}{d+\beta}$ and $m^{1/2}\beta$ is $\frac{m^{1/2}\beta}{d+\beta}$. Hence, we can take

$$\eta \leq \min\left\{\mathcal{O}(\varepsilon), \mathcal{O}\left(\frac{m^2\beta\delta\varepsilon^2}{(md+\beta)\mathcal{T}_{\rm rec}^U}\right), \mathcal{O}\left(\frac{m^{3/2}\delta^2}{(md+\beta)\mathcal{T}_{\rm esc}^U}\right), \mathcal{O}\left(\frac{m^{1/2}\beta}{d+\beta}\right)\right\}$$

Finally, β satisfies $\beta \geq \max\{\underline{\beta}_1^U, \underline{\beta}_2^U\}$, and We have

$$\begin{aligned}\underline{\beta}_1^U &= \frac{256(2C_H m + 4m + (m+1)^2)}{m\varepsilon^2}\left(d\log(2) + \log\left(\frac{6(4m+M^2+1)^{1/2}\mathcal{T}+3}{\delta}\right)\right) \\ &\leq \mathcal{O}\left(\frac{d+\log((\mathcal{T}+1)/\delta)}{m\varepsilon^2}\right),\end{aligned}$$

and

$$\underline{\beta}_2^U = \frac{1024 d\eta\sqrt{m}\log(2^{1/4}e^{1/4}6\delta^{-1}\mathcal{T}_{\rm rec}^U/\eta)}{\varepsilon^2 e^{-2(1+2\sqrt{m}+M)\eta}} \leq \mathcal{O}\left(\frac{d\eta m^{1/2}\log(\delta^{-1}\mathcal{T}_{\rm rec}^U/\eta)}{\varepsilon^2}\right),$$

where we used $e^{2(1+2\sqrt{m}+M)\eta} = e^{\mathcal{O}(\varepsilon)} = \mathcal{O}(1)$.

Hence, we can take

$$\beta \geq \max\left\{\Omega\left(\frac{d+\log((\mathcal{T}+1)/\delta)}{m\varepsilon^2}\right), \Omega\left(\frac{d\eta m^{1/2}\log(\delta^{-1}\mathcal{T}_{\rm rec}^U/\eta)}{\varepsilon^2}\right)\right\}.$$

The proof is now complete.

B.1.2 Proof of Proposition 21

In this section, we focus on the proof of Proposition 21. We adopt some ideas from [BG03, TLR18]. We recall x_* is a local minimum of F and H is the Hessian matrix: $H = \nabla^2 F(x_*)$, and we write

$$X(t) = Y(t) + x_*.$$

Thus, we have the decomposition

$$\nabla F(X(t)) = HY(t) - \rho(Y(t)),$$

where $\|\rho(Y(t))\| \leq \frac{1}{2}L\|Y(t)\|^2$ since the Hessian of F is L-Lipschitz (Lemma 1.2.4. [Nes13]). Then, we have

$$\begin{aligned}
&dV(t) = -\gamma V(t)dt - (H(Y(t)) - \rho(Y(t)))dt + \sqrt{2\gamma\beta^{-1}}dB_t, \\
&dY(t) = V(t)dt.
\end{aligned}$$

We can write it in terms of matrix form as:

$$d\begin{bmatrix} V(t) \\ Y(t) \end{bmatrix} = \begin{bmatrix} -\gamma I & -H \\ I & 0 \end{bmatrix}\begin{bmatrix} V(t) \\ Y(t) \end{bmatrix}dt + \sqrt{2\gamma\beta^{-1}}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}dB_t^{(2)} + \begin{bmatrix} \rho(V(t)) \\ 0 \end{bmatrix}dt,$$

where $B_t^{(2)}$ is a $2d$-dimensional standard Brownian motion. Therefore, we have

$$\begin{bmatrix} V(t) \\ Y(t) \end{bmatrix} = e^{-tH_\gamma}\begin{bmatrix} V(0) \\ Y(0) \end{bmatrix} + \sqrt{2\gamma\beta^{-1}}\int_0^t e^{(s-t)H_\gamma}I^{(2)}dB_s^{(2)} + \int_0^t e^{(s-t)H_\gamma}\begin{bmatrix} \rho(V(s)) \\ 0 \end{bmatrix}ds,$$

where

$$H_\gamma = \begin{bmatrix} \gamma I & H \\ -I & 0 \end{bmatrix}, \qquad I^{(2)} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \tag{B.23}$$

Given $0 \leq t_0 \leq t_1$, we define the matrix flow

$$Q_{t_0}(t) := e^{(t_0-t)H_\gamma} \tag{B.24}$$

and we also define

$$Z(t) := e^{(t-t_0)H_\gamma}\begin{bmatrix} V(t) \\ Y(t) \end{bmatrix} = Z_t^0 + Z_t^1,$$

where

$$Z_t^0 = e^{-t_0H_\gamma}\begin{bmatrix} V(0) \\ Y(0) \end{bmatrix} + \sqrt{2\gamma\beta^{-1}}\int_0^t e^{(s-t_0)H_\gamma}I^{(2)}dB_s^{(2)}, \tag{B.25}$$

$$Z_t^1 = \int_0^t e^{(s-t_0)H_\gamma}\begin{bmatrix} \rho(V(s)) \\ 0 \end{bmatrix}ds. \tag{B.26}$$

Note that

$$Q_{t_0}(t_1)Z_t^0 = e^{-t_1H_\gamma}\begin{bmatrix} V(0) \\ Y(0) \end{bmatrix} + \sqrt{2\gamma\beta^{-1}}\int_0^t e^{(s-t_1)H_\gamma}I^{(2)}dB_s^{(2)}$$

is a martingale. Before we proceed to the proof of Proposition 21, we state the following lemma, which will be used in the proof of Proposition 21.

Lemma 22. *Assume $\gamma = 2\sqrt{m}$. Define:*

$$\mu_t := e^{-tH_\gamma}(V(0), Y(0))^T, \tag{B.27}$$

$$\Sigma_t := 2\gamma\beta^{-1} \int_0^t e^{(s-t)H_\gamma} I^{(2)} e^{(s-t)H_\gamma^T} ds. \tag{B.28}$$

For any $\theta \in \left(0, \frac{2m\sqrt{m}}{\gamma(2C_H m + 4m + (m+1)^2)}\right)$, and $h > 0$ and any $(V(0), Y(0))$,

$$\mathbb{P}\left(\sup_{t_0 \le t \le t_1} \|Q_{t_0}(t_1) Z_t^0\| \ge h\right)$$
$$\le \left(1 - \theta \frac{\gamma(2C_H m + 4m + (m+1)^2)}{2m\sqrt{m}}\right)^{-d} e^{-\frac{\beta\theta}{2}[h^2 - \langle \mu_{t_1}, (I - \beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle]}.$$

Notice that Z_t^0 process is exactly the Ornstein-Uhlenbeck process (3.3) that we have studied in Section 3. There, the emphasis is the convergence speed of this Ornstein-Uhlenbeck process as time t goes to infinity, and the above Lemma 22 is about the tail estimate on a finite time interval.

Finally, let us complete the proof of Proposition 21.

Proof of Proposition 21. Since $\|Y(0)\| = \|X(0) - x_*\| \le r$, we know that $\tau > 0$. Fix some $\mathcal{T}_{\text{rec}}^U \le t_0 \le t_1$, such that $t_1 - t_0 \le \frac{1}{2\|H_\gamma\|}$. Then, for every $t \in [t_0, t_1]$,

$$\|Y(t)\| \le \left\| e^{(t_1 - t)H_\gamma} Q_{t_0}(t_1) Z_t \right\| \le e^{\frac{1}{2}} \|Q_{t_0}(t_1) Z_t\|.$$

It follows that (with $e^{-1/2} \ge 1/2$)

$$\begin{aligned}
&\mathbb{P}(\tau \in [t_0, t_1]) \\
&= \mathbb{P}\left(\sup_{t_0 \le t \le t_1 \wedge \tau} \frac{\|Y(t)\|}{\varepsilon + re^{-\sqrt{m}t}} \ge 1, \tau \ge t_0\right) \\
&\le \mathbb{P}\left(\sup_{t_0 \le t \le t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t\|}{\varepsilon + re^{-\sqrt{m}t}} \ge \frac{1}{2}, \tau \ge t_0\right) \\
&\le \mathbb{P}\left(\sup_{t_0 \le t \le t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t^0\|}{\varepsilon + re^{-\sqrt{m}t}} \ge c_0, \tau \ge t_0\right) + \mathbb{P}\left(\sup_{t_0 \le t \le t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t^1\|}{\varepsilon + re^{-\sqrt{m}t}} \ge c_1, \tau \ge t_0\right),
\end{aligned} \tag{B.29}$$

where $c_0 + c_1 = \frac{1}{2}$ and $c_0, c_1 > 0$. We will first bound the second term in (B.29) which will turn out to be zero, and then use Lemma 22 to bound the first term in (B.29).

First, notice that $Z_t^1 \equiv 0$ in the quadratic case (Sec. 3) and the second term in (B.29) is automatically zero. In the more general case, we will show that the second term in (B.29) is also zero. On the event $\tau \in [t_0, t_1]$, for any $0 \le s \le t_1 \wedge \tau$, we have

$$\|\rho(Y(s))\| \le \frac{L}{2}\|Y(s)\|^2 \le \frac{L}{2}\left(\varepsilon + re^{-\sqrt{m}s}\right)^2.$$

Therefore, for any $t \in [t_0, t_1 \wedge \tau]$, by Lemma 1, we get

$$
\begin{aligned}
&\left\|Q_{t_0}(t_1) Z_t^1\right\| \\
&\leq \int_0^t \left\|e^{(s-t_1)H_\gamma}\right\| \cdot \|\rho(Y(s))\| ds \\
&\leq \frac{L}{2} \int_0^t \sqrt{C_H + 2 + (m+1)^2 (t_1 - s)^2} e^{(s-t_1)\sqrt{m}} \left(\varepsilon + r e^{-\sqrt{m}s}\right)^2 ds \\
&\leq L \int_0^t \left(\sqrt{C_H + 2} + (m+1)(t_1 - s)\right) e^{(s-t_1)\sqrt{m}} \left(\varepsilon^2 + r^2 e^{-2\sqrt{m}s}\right) ds \\
&\leq L \int_0^{t_1} \left(\sqrt{C_H + 2} + (m+1)(t_1 - s)\right) e^{(s-t_1)\sqrt{m}} \left(\varepsilon^2 + r^2 e^{-2\sqrt{m}s}\right) ds \\
&\leq \frac{L}{\sqrt{m}} \left(\left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}}\right)\varepsilon^2 + \sqrt{C_H+2}\, r^2 e^{-\sqrt{m}t_1}\right) \\
&\qquad\qquad + L(m+1) r^2 \int_0^{t_1} (t_1 - s) e^{(s-t_1)\sqrt{m}} e^{-2\sqrt{m}s} ds \\
&\leq \frac{L}{\sqrt{m}} \left(\left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}}\right)\varepsilon^2 + \sqrt{C_H+2}\, r^2 e^{-\sqrt{m}t_1} + (m+1) r^2 t_1 e^{-t_1\sqrt{m}}\right) \\
&\leq \frac{L}{\sqrt{m}} \left(\left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}}\right)\varepsilon^2 + \left(\sqrt{(C_H+2)m} + (m+1)\right) r^2 t_1 e^{-t_1\sqrt{m}}\right) \\
&\leq \frac{L}{\sqrt{m}} \left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}} + \frac{\sqrt{(C_H+2)m} + (m+1)}{8\sqrt{C_H + 2 + (m+1)^2}}\right)\varepsilon^2
\end{aligned}
$$

where we used $t_1 \geq t \geq t_0 \geq \mathcal{T}_{\mathrm{rec}}^U \geq \frac{1}{\sqrt{m}}$, and $t_1 e^{-t_1\sqrt{m}} \leq \mathcal{T}_{\mathrm{rec}}^U e^{-\mathcal{T}_{\mathrm{rec}}^U \sqrt{m}}$ and the definition of $\mathcal{T}_{\mathrm{rec}}^U$:

$$
\sqrt{C_H + 2 + (m+1)^2}\, \mathcal{T}_{\mathrm{rec}}^U e^{-\sqrt{m}\mathcal{T}_{\mathrm{rec}}^U} = \frac{\varepsilon^2}{8r^2}.
$$

Consequently, if we take $c_1 = \frac{L}{\sqrt{m}} \left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}} + \frac{\sqrt{(C_H+2)m}+(m+1)}{8\sqrt{C_H+2+(m+1)^2}}\right)\varepsilon$, then,

$$
\sup_{t_0 \leq t \leq t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t\|}{\varepsilon + r e^{-\sqrt{m}t}} \leq \frac{1}{\varepsilon} \sup_{t_0 \leq t \leq t_1 \wedge \tau} \|Q_{t_0}(t_1) Z_t\| \leq c_1,
$$

which implies that

$$
\mathbb{P}\left(\sup_{t_0 \leq t \leq t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t^1\|}{\varepsilon + r e^{-\sqrt{m}t}} \geq c_1, \tau \geq t_0\right) = 0.
$$

Moreover, $c_0 = \frac{1}{2} - c_1 = \frac{1}{2} - \frac{L}{\sqrt{m}} \left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}} + \frac{\sqrt{(C_H+2)m}+(m+1)}{8\sqrt{C_H+2+(m+1)^2}}\right)\varepsilon > \frac{1}{4}$ since it is assumed that $\varepsilon < \frac{\sqrt{m}}{4L\left(\sqrt{C_H+2} + \frac{m+1}{\sqrt{m}} + \frac{\sqrt{(C_H+2)m}+(m+1)}{8\sqrt{C_H+2+(m+1)^2}}\right)}$.

Second, we will apply Lemma 22 to bound the first term in (B.29). By using $V(0)=0$ and $\|Y(0)\|\leq r$ and the definition of μ_{t_1} and Σ_{t_1} in (B.27) and (B.28), we get

$$\begin{aligned}
&\left\langle \mu_{t_1},(I-\beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\right\rangle \\
&=\left\langle e^{-t_1H_\gamma}(V(0),Y(0))^T,(I-\beta\theta\Sigma_{t_1})^{-1}e^{-t_1H_\gamma}(V(0),Y(0))^T\right\rangle \\
&\leq\left(1-\theta\frac{\gamma(2C_Hm+4m+(m+1)^2)}{2m\sqrt{m}}\right)^{-1}\left(C_H+2+(m+1)^2t_1^2\right)e^{-2\sqrt{m}t_1}r^2 \\
&\leq 2\left((C_H+2)m+(m+1)^2\right)t_1^2e^{-2\sqrt{m}t_1}r^2 \\
&\leq\frac{1}{32}\frac{(C_H+2)m+(m+1)^2}{C_H+2+(m+1)^2}\frac{\varepsilon^4}{r^2}\leq\frac{1}{32}\varepsilon^2,
\end{aligned}$$

by choosing $\theta=\frac{m\sqrt{m}}{\gamma(2C_Hm+4m+(m+1)^2)}$ and $t_1\geq\mathcal{T}_{\mathrm{rec}}^U\geq\frac{1}{\sqrt{m}}$, and $t_1e^{-t_1\sqrt{m}}\leq\mathcal{T}_{\mathrm{rec}}^Ue^{-\mathcal{T}_{\mathrm{rec}}^U}$, and using the definition $\sqrt{C_H+2+(m+1)^2}\mathcal{T}_{\mathrm{rec}}^Ue^{-\sqrt{m}\mathcal{T}_{\mathrm{rec}}^U}=\frac{\varepsilon^2}{8r^2}$, and we also used $\varepsilon\leq\sqrt{\frac{C_H+2+(m+1)^2}{(C_H+2)m+(m+1)^2}}r$.

Then with the choice of $h=(\varepsilon+re^{-\sqrt{m}t_1})c_0$ and $\theta=\frac{m\sqrt{m}}{\gamma(2C_Hm+4m+(m+1)^2)}$ in Lemma 22, and using the fact that $h=(\varepsilon+re^{-\sqrt{m}t_1})c_0\geq\varepsilon c_0$, we get

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t_0\leq t\leq t_1\wedge\tau}\frac{\|Q_{t_0}(t_1)Z_t^0\|}{\varepsilon+re^{-\sqrt{m}t}}\geq c_0,\tau\geq t_0\right) \\
&\leq\mathbb{P}\left(\sup_{t_0\leq t\leq t_1}\|Q_{t_0}(t_1)Z_t^0\|\geq\left(\varepsilon+re^{-\sqrt{m}t_1}\right)c_0\right) \\
&\leq\left(1-\theta\frac{\gamma(2C_Hm+4m+(m+1)^2)}{2m\sqrt{m}}\right)^{-\frac{2d}{2}}\cdot\exp\left(-\frac{\beta\theta}{2}\left[h^2-\langle\mu_{t_1},(I-\beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle\right]\right) \\
&\leq 2^d\cdot\exp\left(-\frac{\beta\gamma^{-1}m\sqrt{m}\varepsilon^2}{2(2C_H+4m+(m+1)^2)}\left(c_0^2-\frac{1}{32}\right)\right) \\
&\leq 2^d\cdot\exp\left(-\frac{\beta\gamma^{-1}m\sqrt{m}\varepsilon^2}{128(2C_H+4m+(m+1)^2)}\right).
\end{aligned}$$

Thus for any $t_0\geq\mathcal{T}_{\mathrm{rec}}^U$ and $t_0\leq t_1\leq t_0+\frac{1}{2\|H_\gamma\|}$,

$$\mathbb{P}(\tau\in[t_0,t_1])\leq 2^d\cdot\exp\left(-\frac{\beta\gamma^{-1}m\sqrt{m}\varepsilon^2}{128(2C_Hm+4m+(m+1)^2)}\right).$$

Fix any $\mathcal{T}>0$ and recall the definition of the escape time $\mathcal{T}_{\mathrm{esc}}^U=\mathcal{T}+\mathcal{T}_{\mathrm{rec}}^U$. Partition the interval $[\mathcal{T}_{\mathrm{rec}}^U,\mathcal{T}_{\mathrm{esc}}^U]$ using the points $\mathcal{T}_{\mathrm{rec}}^U=t_0<t_1<\cdots<t_{\lceil 2\|H_\gamma\|\mathcal{T}\rceil}=\mathcal{T}_{\mathrm{esc}}^U$ with

$t_j = j/(2\|H_\gamma\|)$, then we have

$$\begin{aligned}\mathbb{P}\left(\tau \in \left[\mathcal{T}_{\mathrm{rec}}^U, \mathcal{T}_{\mathrm{esc}}^U\right]\right) &= \sum_{j=0}^{\lceil 2\|H_\gamma\|\mathcal{T}\rceil} \mathbb{P}(\tau \in [t_j, t_{j+1}]) \\ &\leq (2\|H_\gamma\|\mathcal{T} + 1) \cdot 2^d \cdot \exp\left(-\frac{\beta\gamma^{-1} m\sqrt{m}\varepsilon^2}{128(2C_H m + 4m + (m+1)^2)}\right) \leq \delta,\end{aligned}$$

provided that

$$\beta \geq \frac{128(2C_H m + 4m + (m+1)^2)\gamma}{m\sqrt{m}\varepsilon^2}\left(d\log(2) + \log\left(\frac{2\|H_\gamma\|\mathcal{T} + 1}{\delta}\right)\right).$$

Finally, plugging $\gamma = 2\sqrt{m}$ into the above formulas and applying the bound on $\|H_\gamma$ from Lemma 41, the conclusion follows. □

B.1.3 Uniform L^2 bounds

In this section, we state the uniform L^2 bounds for the continuous time underdamped Langevin dynamics ((1.7) and (1.8)) and the discrete time iterates ((1.10) and (1.11)) in Lemma 24, which is a modification of Lemma 8 in [GGZ18]. The uniform L^2 bound for the discrete dynamics (1.10)-(1.11) is used to derive the relative entropy to compare the laws of the continuous time dynamics and the discrete time dynamics, and the uniform L^2 bound for the continuous dynamics (1.7)-(1.8) is used to control the tail of the continuous dynamics in Section B.1.1.

Before we proceed, let us first introduce the following Lyapunov function (from the paper [EGZ17]) which will be used in the proof the uniform L^2 boundedness results for both the continuous and discrete underdamped Langevin dynamics. We define the Lyapunov function $\mathcal{V}$ as:

$$\mathcal{V}(x,v) := \beta F(x) + \frac{\beta}{4}\gamma^2(\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2), \tag{B.30}$$

and λ is a positive constant less than $1/4$ according to [EGZ17].

We will first show in the following lemma that we can find explicit constants $\lambda \in (0, \min(1/4, m/(M + \gamma^2/2)))$ and $\overline{A} \in (0,\infty)$ so that the drift condition (B.33) is satisfied. The drift condition is needed in [EGZ17], which is applied to obtain the uniform L^2 bounds in [GGZ18] (Lemma 8) that implies the uniform L^2 bounds in our current setting (the following Lemma 24).

Lemma 23. *Let us define:*

$$\lambda = \frac{1}{2}\min(1/4, m/(M + \gamma^2/2)), \tag{B.31}$$

$$\overline{A} = \frac{\beta}{2}\frac{m}{M + \frac{1}{2}\gamma^2}\left(\frac{B^2}{2M + \gamma^2} + \frac{b}{m}\left(M + \frac{1}{2}\gamma^2\right) + A\right), \tag{B.32}$$

then the following drift condition holds:

$$x \cdot \nabla F(x) \geq 2\lambda(F(x) + \gamma^2\|x\|^2/4) - 2\overline{A}/\beta\,. \tag{B.33}$$

The following lemma provides uniform L^2 bounds for the continuous-time underdamped Langevin diffusion process $(X(t), V(t))$ defined in (1.7)-(1.8) and discrete-time underdamped Langevin dynamics (X_k, V_k) defined in (1.10)-(1.11).

Lemma 24 (Uniform L^2 bounds)**.** *Suppose parts (i), (ii), (iii), (iv) of Assumption 9 and the drift condition (B.33) hold. $\gamma > 0$ is arbitrary and λ, $\overline{A}$ are defined in (B.31) and (B.32).*

(i) It holds that

$$\sup_{t\geq 0} \mathbb{E}\|X(t)\|^2 \leq C_x^c := \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right) R^2 + \beta BR + \beta A + \frac{3}{4}\beta\|V(0)\|^2 + \frac{d+\overline{A}}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \tag{B.34}$$

$$\sup_{t\geq 0} \mathbb{E}\|V(t)\|^2 \leq C_v^c := \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right) R^2 + \beta BR + \beta A + \frac{3}{4}\beta\|V(0)\|^2 + \frac{d+\overline{A}}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}, \tag{B.35}$$

(ii) For any stepsize η satisfying:

$$0 < \eta \leq \overline{\eta}_4^U := \min\left\{\frac{\gamma}{K_2}(d/\beta + \overline{A}/\beta), \frac{\gamma\lambda}{2K_1}\right\}, \tag{B.36}$$

where

$$K_1 := \max\left\{\frac{32M^2\left(\frac{1}{2}+\gamma\right)}{(1-2\lambda)\beta\gamma^2}, \frac{8\left(\frac{1}{2}M + \frac{1}{4}\gamma^2 - \frac{1}{4}\gamma^2\lambda + \gamma\right)}{\beta(1-2\lambda)}\right\}, \tag{B.37}$$

$$K_2 := 2B^2\left(\frac{1}{2}+\gamma\right), \tag{B.38}$$

and we have

$$\sup_{j\geq 0} \mathbb{E}\|X_j\|^2 \leq C_x^d := \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right) R^2 + \beta BR + \beta A + \frac{3}{4}\beta\|V(0)\|^2 + \frac{4(d+\overline{A})}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \tag{B.39}$$

$$\sup_{j\geq 0} \mathbb{E}\|V_j\|^2 \leq C_v^d := \frac{\left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right) R^2 + \beta BR + \beta A + \frac{3}{4}\beta\|V(0)\|^2 + \frac{4(d+\overline{A})}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}. \tag{B.40}$$

B.1.4 Proofs of auxiliary results

Proof of Lemma 22. Note that $Q_{t_0}(t_1)Z_t^0$ is a $2d$-dimensional martingale and by Doob's martingale inequality, for any $h>0$,

$$\mathbb{P}\left(\sup_{t_0\leq t\leq t_1}\|Q_{t_0}(t_1)Z_t^0\|\geq h\right)\leq e^{-\beta\theta h^2/2}\mathbb{E}\left[e^{(\beta\theta/2)\|Q_{t_0}(t_1)Z_{t_1}^0\|^2}\right]$$
$$=e^{-\beta\theta h^2/2}\frac{1}{\sqrt{\det(I-\beta\theta\Sigma_{t_1})}}e^{\frac{\beta\theta}{2}\langle\mu_{t_1},(I-\beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle},\quad\text{(B.41)}$$

where the last line above uses the fact that $Q_{t_0}(t_1)Z_{t_1}$ is a Gaussian random vector with mean

$$\mu_{t_1}=e^{-t_1H_\gamma}(V(0),Y(0))^T,$$

and covariance matrix

$$\Sigma_{t_1}=2\gamma\beta^{-1}\int_0^{t_1}\left(e^{(s-t_1)H_\gamma}I^{(2)}\right)\left(e^{(s-t_1)H_\gamma}I^{(2)}\right)^T ds$$
$$=2\gamma\beta^{-1}\int_0^{t_1}e^{-sH_\gamma}I^{(2)}e^{-sH_\gamma^T}ds.$$

We next estimate $\det(I-\beta\theta\Sigma_{t_1})$ fron (B.41). Let us recall from Lemma 1 that if $\gamma=2\sqrt{m}$, then we recall from Lemma 1 that,

$$\left\|e^{-tH_\gamma}\right\|\leq\sqrt{C_H+2+(m+1)^2t^2}\cdot e^{-\sqrt{m}t},$$

and thus, we have

$$\|\Sigma_{t_1}\|\leq 2\gamma\beta^{-1}\int_0^{t_1}\left(C_H+2+(m+1)^2t^2\right)e^{-2\sqrt{m}t}dt\leq\gamma\beta^{-1}\frac{2C_Hm+4m+(m+1)^2}{2m\sqrt{m}}.$$

Therefore we infer that the eigenvalues of $I-\beta\theta\Sigma$ are bounded below by $1-\theta\frac{\gamma(2C_Hm+4m+(m+1)^2)}{2m\sqrt{m}}$. The conclusion then follows from (B.41). □

Proof of Lemma 23. By Assumption 9 (iii), $x\cdot\nabla F(x)\geq m\|x\|^2-b$. Thus in order to show the drift condition (B.33), it suffices to show that

$$m\|x\|^2-b-2\lambda(F(x)+\gamma^2\|x\|^2/4)\geq -2\overline{A}/\beta.\quad\text{(B.42)}$$

Given the definition of λ in (B.31), by Lemma 43, we get

$$\begin{aligned}
&m\|x\|^2 - b - 2\lambda(F(x) + \gamma^2\|x\|^2/4) \\
&\geq m\|x\|^2 - b - \frac{m}{M+\frac{1}{2}\gamma^2}(F(x) + \gamma^2\|x\|^2/4) \\
&\geq \frac{mM + \frac{1}{4}m\gamma^2}{M+\frac{1}{2}\gamma^2}\|x\|^2 - b - \frac{m}{M+\frac{1}{2}\gamma^2}\left(\frac{M}{2}\|x\|^2 + B\|x\| + A\right) \\
&= \frac{m}{M+\frac{1}{2}\gamma^2}\left(\frac{1}{2}M\|x\|^2 + \frac{1}{4}\gamma^2\|x\|^2 - B\|x\| - \frac{b}{m}\left(M + \frac{1}{2}\gamma^2\right) - A\right) \\
&\geq \frac{m}{M+\frac{1}{2}\gamma^2}\left(-\frac{B^2}{2M+\gamma^2} - \frac{b}{m}\left(M + \frac{1}{2}\gamma^2\right) - A\right) = -2\overline{A}/\beta,
\end{aligned}$$

by the definition of $\overline{A}$ in (B.32). Hence, (B.42) holds and the proof is complete. $\square$

Proof of Lemma 24. According to Lemma 8 in [GGZ18],

$$\begin{aligned}
\sup_{t\geq 0}\mathbb{E}\|X(t)\|^2 &\leq \frac{\int_{\mathbb{R}^{2d}}\mathcal{V}(x,v)d\mu_0(x,v) + \frac{d+\overline{A}}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \\
\sup_{t\geq 0}\mathbb{E}\|V(t)\|^2 &\leq \frac{\int_{\mathbb{R}^{2d}}\mathcal{V}(x,v)d\mu_0(x,v) + \frac{d+\overline{A}}{\lambda}}{\frac{\beta}{4}(1-2\lambda)},
\end{aligned}$$

where $\mathcal{V}$ is the Lyapunov function defined in (B.30) and μ_0 is the initial distribution of $(X(0),V(0))$ and in our case, $\mu_0 = \delta_{(X(0),V(0))}$ and $\|X(0)\| \leq R$ and $V(0) \in \mathbb{R}^d$, and for any $0 < \eta \leq \min\left\{\frac{\gamma}{K_2}(d/\beta + \overline{A}/\beta), \frac{\gamma\lambda}{2K_1}\right\}$with K_1 and K_2 given in (B.37) and (B.38), [17] and according to Lemma 8 in [GGZ18], we also have

$$\begin{aligned}
\sup_{j\geq 0}\mathbb{E}\|X_j\|^2 &\leq \frac{\int_{\mathbb{R}^{2d}}\mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+\overline{A})}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \\
\sup_{j\geq 0}\mathbb{E}\|V_j\|^2 &\leq \frac{\int_{\mathbb{R}^{2d}}\mathcal{V}(x,v)\mu_0(dx,dv) + \frac{4(d+\overline{A})}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}.
\end{aligned}$$

We recall from (B.30) that $\mathcal{V}(x,v) = \beta F(x) + \frac{\beta}{4}\gamma^2(\|x+\gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2)$, and $\|X(0)\| \leq R$ and $V(0) \in \mathbb{R}^d$. By Lemma 43, we get

$$\mathcal{V}(x,v) \leq \frac{\beta M}{2}\|x\|^2 + \beta B\|x\| + \beta A + \frac{\beta}{4}\gamma^2(\|x+\gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2)\,,$$

[17]Note that in the definition of K_1, K_2 in [GGZ18], there is a constant δ, which is simply zero, in the context of the current paper.

so that

$$
\begin{aligned}
&\mathcal{V}(X(0),V(0)) \\
&= \frac{\beta M}{2}\|X(0)\|^2 + \beta B\|X(0)\| + \beta A + \frac{\beta}{4}\gamma^2(2\|X(0)\|^2 + 3\gamma^{-2}\|V(0)\|^2 - \lambda\|X(0)\|^2) \\
&\leq \left(\frac{\beta M}{2} + \frac{\beta\gamma^2(2-\lambda)}{4}\right) R^2 + \beta BR + \beta A + \frac{3}{4}\beta\|V(0)\|^2.
\end{aligned}
$$

Hence, the conclusion follows. □

C Proofs of results in Section 4.2

C.1 Proof of Theorem 15

The proof of Theorem 15 is similar to the proof of Theorem 10. For brevity, we omit some of the details, and only outline the key steps and the propositions and lemmas used for the proof of Theorem 15.

Proposition 25. *Fix any $r > 0$ and $0 < \varepsilon < \min\{\bar{\varepsilon}_1^J, \bar{\varepsilon}_2^J\}$, where*

$$
\bar{\varepsilon}_1^J := \frac{m_J(\tilde{\varepsilon})}{4C_J(\tilde{\varepsilon})(1+\|J\|)L(1+\frac{1}{64C_J(\tilde{\varepsilon})^2})}, \qquad \bar{\varepsilon}_2^J := 8rC_J(\tilde{\varepsilon}). \tag{C.1}
$$

Consider the stopping time:

$$
\tau := \inf\left\{t \geq 0 : \|X(t) - x_*\| \geq \varepsilon + re^{-m_J(\tilde{\varepsilon})t}\right\}.
$$

For any initial point $X(0) = x$ with $\|x - x_*\| \leq r$, and

$$
\beta \geq \frac{128C_J(\tilde{\varepsilon})^2}{m_J(\tilde{\varepsilon})\varepsilon^2}\left(\frac{d}{2}\log(2) + \log\left(\frac{2(1+\|J\|)M\mathcal{T}+1}{\delta}\right)\right),
$$

we have

$$
\mathbb{P}_x\left(\tau \in [\mathcal{T}_{rec}^J, \mathcal{T}_{esc}^J]\right) \leq \delta.
$$

C.1.1 Completing the proof of Theorem 15

We first compare the discrete dynamics (1.13) and the continuous dynamics (1.12). Define:

$$
\tilde{X}(t) = X_0 - \int_0^t A_J\left(\nabla F(\tilde{X}(\lfloor s/\eta\rfloor\eta))\right)ds + \sqrt{2\gamma\beta^{-1}}\int_0^t dB_s. \tag{C.2}
$$

The process $\tilde{X}$ defined in (C.2) is the continuous-time interpolation of the iterates $\{X_k\}$. In particular, the joint distribution of $\{X_k : k = 1, 2, \ldots, K\}$ is the same as $\{\tilde{X}(t) : t = \eta, 2\eta, \ldots, K\eta\}$ for any positive integer K.

By following Lemma 7 in [RRT17] and apply the uniform L^2 bounds for X_k in Corollary 31 provided that the stepsize η is sufficiently small (we apply the bound $\|A_J\| \leq 1 + \|J\|$ to Corollary 31)

$$\eta \leq \overline{\eta}_4^J := \frac{1}{M(1 + \|J\|)^2}, \tag{C.3}$$

we will obtain an upper bound on the relative entropy $D(\cdot\|\cdot)$ between the law $\tilde{\mathbb{P}}^{K\eta}$ of $(\tilde{X}(t) : t \leq K\eta)$ and the law $\mathbb{P}^{K\eta}$ of $(X(t) : t \leq K\eta)$, and by Pinsker's inequality an upper bound on the total variation $\| \cdot \|_{TV}$ as well. More precisely, we have

$$\left\|\tilde{\mathbb{P}}^{K\eta} - \mathbb{P}^{K\eta}\right\|_{TV}^2 \leq \frac{1}{2} D\left(\tilde{\mathbb{P}}^{K\eta}\middle\|\mathbb{P}^{K\eta}\right) \leq \frac{1}{2} C_1 K\eta^2, \tag{C.4}$$

where (we use the bound $\|A_J\| \leq 1 + \|J\|$)

$$C_1 := 6(\beta((1 + \|J\|)^2 M^2 C_d + B^2) + d)(1 + \|J\|)^2 M^2, \tag{C.5}$$

where C_d is defined in (C.19).

Let us now complete the proof of Theorem 15. We need to show that

$$\mathbb{P}\left((X_1, \ldots, X_K) \in \mathcal{A}\right) \leq \delta,$$

where $K = \lfloor \eta^{-1} \mathcal{T}_{\text{esc}}^J \rfloor$ and $\mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2$:

$$\mathcal{A}_1 := \left\{(x_1, \ldots, x_K) \in (\mathbb{R}^d)^K : \max_{k \leq \eta^{-1}\mathcal{T}_{\text{rec}}^J} \frac{\|x_k - x_*\|}{\varepsilon + re^{-m_J(\tilde{\varepsilon})k\eta}} \leq \frac{1}{2}\right\},$$
$$\mathcal{A}_2 := \left\{(x_1, \ldots, x_K) \in (\mathbb{R}^d)^K : \max_{\eta^{-1}\mathcal{T}_{\text{rec}}^J \leq k \leq K} \frac{\|x_k - x_*\|}{\varepsilon + re^{-m_J(\tilde{\varepsilon})k\eta}} \geq 1\right\}.$$

Similar to the proof in Section B.1.1 and by (C.4), we get

$$\mathbb{P}((X_1, \ldots, X_K) \in \mathcal{A}) \leq \mathbb{P}((X(\eta), \ldots, X(K\eta)) \in \mathcal{A}) + \frac{\delta}{3}, \tag{C.6}$$

provided that

$$\eta \leq \overline{\eta}_3^J := \frac{2\delta^2}{9C_1 \mathcal{T}_{\text{esc}}^J}. \tag{C.7}$$

It remains to estimate the probability of $\mathbb{P}((X(\eta), \ldots, X(K\eta)) \in \mathcal{A}_1 \cap \mathcal{A}_2)$ for the non-reversible Langevin diffusion. Partition the interval $[0, \mathcal{T}_{\text{rec}}^J]$ using the points $0 = t_1 < t_1 < \cdots < t_{\lceil \eta^{-1}\mathcal{T}_{\text{rec}}^J \rceil} = \mathcal{T}_{\text{rec}}^J$ with $t_k = k\eta$ for $k = 0, 1, \ldots, \lceil \eta^{-1}\mathcal{T}_{\text{rec}}^J \rceil - 1$, and consider the event:

$$\mathcal{B} := \left\{\max_{0 \leq k \leq \lceil \eta^{-1}\mathcal{T}_{\text{rec}}^J \rceil - 1} \max_{t \in [t_k, t_{k+1}]} \|X(t) - X(t_{k+1})\| \leq \frac{\varepsilon}{2}\right\}.$$

Similar to the proof in Section B.1.1, we get

$$\mathbb{P}((X(\eta), \cdots, X(K\eta)) \in \mathcal{A}) \le \frac{\delta}{3} + \mathbb{P}(\mathcal{B}^c)\,, \tag{C.8}$$

provided that (by applying Proposition 25):

$$\beta \ge \underline{\beta}_1^J := \frac{128 C_J(\tilde{\varepsilon})^2}{m_J(\tilde{\varepsilon})\varepsilon^2}\left(\frac{d}{2}\log(2) + \log\left(\frac{6(1+\|J\|)M\mathcal{T}+3}{\delta}\right)\right). \tag{C.9}$$

To complete the proof, we need to show that $\mathbb{P}(\mathcal{B}^c) \le \frac{\delta}{3}$ in view of (C.6) and (C.8). For any $t \in [t_k, t_{k+1}]$, where $t_{k+1} - t_k = \eta$, we have

$$\begin{aligned}
&\|X(t) - X(t_{k+1})\| \\
&\le \int_t^{t_{k+1}} \|A_J \nabla F(X(s))\| ds + \sqrt{2\beta^{-1}}\|B_t - B_{t_{k+1}}\| \\
&\le \|A_J\| M \int_t^{t_{k+1}} \|X(s) - X(t_{k+1})\| ds + \eta \|A_J \nabla F(X(t_{k+1}))\| + \sqrt{2\beta^{-1}}\|B_t - B_{t_{k+1}}\| \\
&\le \|A_J\| M \int_t^{t_{k+1}} \|X(s) - X(t_{k+1})\| ds \\
&\qquad\qquad + \eta \|A_J\| \cdot (M\|X(t_{k+1})\| + B) + \sqrt{2\beta^{-1}}\|B_t - B_{t_{k+1}}\|\,.
\end{aligned}$$

By Gronwall's inequality, we get the key estimate:

$$\begin{aligned}
&\sup_{t \in [t_k, t_{k+1}]} \|X(t) - X(t_{k+1})\| \\
&\le e^{\eta\|A_J\|M}\left[\eta\|A_J\| \cdot (M\|X(t_{k+1})\| + B) + \sqrt{2\beta^{-1}} \sup_{t \in [t_k, t_{k+1}]} \|B_t - B_{t_{k+1}}\|\right].
\end{aligned}$$

Then, by following the same argument as in Section B.1.1 and also apply $\|A_J\| \le 1 + \|J\|$, we can show that $\mathbb{P}(\mathcal{B}^c) \le \frac{\delta}{3}$ provided that $\eta \le 1$ and

$$\eta \le \overline{\eta}_1^J := \frac{\varepsilon e^{-(1+\|J\|)M}}{8(1+\|J\|)B}, \tag{C.10}$$

and

$$\eta \le \overline{\eta}_2^J := \frac{\delta \varepsilon^2 e^{-2(1+\|J\|)M}}{384(1+\|J\|)^2 M^2 C_c \mathcal{T}_{\mathrm{rec}}^J}\,, \tag{C.11}$$

where C_c is defined in (C.18) and

$$\beta \ge \underline{\beta}_2^J := \frac{512 d\eta \log(2^{1/4} e^{1/4} 6\delta^{-1}\mathcal{T}_{\mathrm{rec}}^J/\eta)}{\varepsilon^2 e^{-2(1+\|J\|)M\eta}}\,. \tag{C.12}$$

To complete the proof, we need work on the leading orders of the constants. We treat $\|J\|$, M, L as constant. The argument is similar to the argument in the proof of Theorem 10 and is thus omitted here. The proof is now complete.

C.1.2 Proof of Proposition 25

Before we proceed to the proof of Proposition 25, let us first state the following two lemmas that will be used in the proof of Proposition 25.

Lemma 26. *For any $\theta \in (0, \frac{\lambda_1^J - \tilde{\varepsilon}}{(C_J(\tilde{\varepsilon}))^2})$, $h > 0$ and $y_0 \in \mathbb{R}^d$,*

$$\mathbb{P}\left(\sup_{t_0 \le t \le t_1} \left\|Q_{t_0}(t_1) Z_t^0\right\| \ge h\right) \le \left(1 - \theta \frac{(C_J(\tilde{\varepsilon}))^2}{\lambda_1^J - \tilde{\varepsilon}}\right)^{-d/2} e^{-\frac{\beta\theta}{2}[h^2 - \langle \mu_{t_1}, (I - \beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle]},$$

where $Q_{t_0}(t_1)$ is defined in (C.14), Z_t^0 is defined in (C.15), and

$$\mu_t := e^{-tA_J H} y_0, \qquad \Sigma_t := 2\beta^{-1} \int_0^t e^{-s(A_J H)} e^{-s(A_J H)^T} ds. \tag{C.13}$$

Lemma 27. *Given $t_0 \le t \le (t_1 \wedge \tau)$, where τ is the stopping time defined in Proposition 25, we have*

$$\left\|Q_{t_0}(t_1) Z_t^1\right\| \le \frac{C_J(\tilde{\varepsilon}) \|A_J\| L}{2} \int_0^t e^{(s - t_1) m_J(\tilde{\varepsilon})} \left(\varepsilon + r e^{-m_J(\tilde{\varepsilon}) s}\right)^2 ds,$$

where $Q_{t_0}(t_1)$ is defined in (C.14), and Z_t^1 is defined in (C.16).

Proof of Proposition 25. We recall x_* is a local minimum of F and H is the Hessian matrix: $H = \nabla^2 F(x_*)$, and we write

$$X(t) = Y(t) + x_*.$$

Thus, we have the decomposition

$$\nabla F(X(t)) = HY(t) - \rho(Y(t)),$$

where $\|\rho(Y(t))\| \le \frac{1}{2} L \|Y(t)\|^2$ since the Hessian of F is L-Lipschitz (Lemma 1.2.4. [Nes13]). This implies that

$$dY(t) = -A_J H Y(t) dt + A_J \rho(Y(t)) dt + \sqrt{2\beta^{-1}} dB_t.$$

Thus, we get

$$Y(t) = e^{-tA_J H} Y(0) + \sqrt{2\beta^{-1}} \int_0^t e^{(s-t)A_J H} dB_s + \int_0^t e^{(s-t)A_J H} A_J \rho(Y(s)) ds.$$

Given $0 \le t_0 \le t_1$, we define the matrix flow

$$Q_{t_0}(t) := e^{(t_0 - t)A_J H}, \tag{C.14}$$

and $Z_t := e^{(t-t_0)A_J H} Y_t$ so that

$$Z_t = e^{-t_0 A_J H} Y(0) + \sqrt{2\beta^{-1}} \int_0^t e^{(s-t_0)A_J H} dB_s + \int_0^t e^{(s-t_0)A_J H} A_J \rho(Y(s)) ds.$$

We define the decomposition $Z_t = Z_t^0 + Z_t^1$, where

$$Z_t^0 = e^{-t_0 A_J H} Y(0) + \sqrt{2\beta^{-1}} \int_0^t e^{(s-t_0)A_J H} dB_s, \tag{C.15}$$

$$Z_t^1 = \int_0^t e^{(s-t_0)A_J H} A_J \rho(Y(s)) ds. \tag{C.16}$$

It follows that for any $t_0 \leq t \leq t_1$,

$$Q_{t_0}(t_1) Z_t^1 = \int_0^t e^{(s-t_1)A_J H} A_J \rho(Y(s)) ds,$$

$$Q_{t_0}(t_1) Z_t^0 = e^{-t_1 A_J H} Y(0) + \sqrt{2\beta^{-1}} \int_0^t e^{(s-t_1)A_J H} dB_s.$$

The rest of the proof is similar to the proof of Proposition 21. We apply Lemma 27 to bound the term $Q_{t_0}(t_1) Z_t^1$ and apply Lemma 26 to bound the term $Q_{t_0}(t_1) Z_t^0$. By letting $\gamma = 1$ in Proposition 21 and replacing d by $d/2$ due to Lemma 26, and $\|H_\gamma\|$ by $\|A_J H\|$ and using the bounds $\|A_J\| \leq (1 + \|J\|)$ and $\|A_J H\| \leq (1 + \|J\|)M$, we obtain the desired result in Proposition 25. □

C.1.3 Uniform L^2 bounds

In this section we establish uniform L^2 bounds for both the continuous time dynamics (1.12) and discrete time dynamics (1.13). The main idea of the proof is to use Lyapunov functions. Our local analysis result relies on the approximation of the continuous time dynamics (1.12) by the discrete time dynamics (1.13). The uniform L^2 bound for the discrete dynamics (1.13) is used to derive the relative entropy to compare the laws of the continuous time dynamics and the discrete time dynamics, and the uniform L^2 bound for the continuous dynamics (1.12) is used to control the tail of the continuous dynamics in Section C.1.1. We first recall the continuous-time dynamics from (1.12):

$$dX(t) = -A_J(\nabla F(X(t)))dt + \sqrt{2\beta^{-1}} dB_t, \qquad A_J = I + J,$$

where J is a $d \times d$ anti-symmetric matrix, i.e. $J^T = -J$. The generator of this continuous time process is given by

$$\mathcal{L} = -A_J \nabla F \cdot \nabla + \beta^{-1} \Delta \tag{C.17}$$

Lemma 28. *Given $X(0) = x \in \mathbb{R}^d$,*

$$\mathbb{E}[F(X(t))] \leq F(x) + \frac{B}{2} + A + \frac{b(M+B)}{m} + \frac{2M\beta^{-1} d(M+B)}{m^2}.$$

Since F has at most the quadratic growth (due to Lemma 43), we immediately have the following corollary.

Corollary 29. *Given* $\|X(0)\| \leq R = \sqrt{b/m}$*,*

$$\mathbb{E}[\|X(t)\|^2] \leq C_c := \frac{MR^2 + 2BR + B + 4A}{m} + \frac{2b(M+B)}{m^2} + \frac{4M\beta^{-1}d(M+B)}{m^3} + \frac{b}{m}\log 3. \tag{C.18}$$

We next show uniform L^2 bounds for the discrete iterates X_k, where we recall from (1.13) that the non-reversible Langevin dynamics is given by:

$$X_{k+1} = X_k - \eta A_J(\nabla F(X_k)) + \sqrt{2\eta\beta^{-1}}\xi_k.$$

Lemma 30. *Given that* $\eta \leq \frac{1}{M\|A_J\|^2}$*, we have*

$$\mathbb{E}_x[F(X_k)] \leq F(x) + \frac{B}{2} + A + \frac{4(M+B)M\beta^{-1}d}{m^2} + \frac{(M+B)b}{m}.$$

Since F has at most the quadratic growth (due to Lemma 43), we immediately have the following corollary.

Corollary 31. *Given that* $\eta \leq \frac{1}{M\|A_J\|^2}$ *and* $\|X(0)\| \leq R = \sqrt{b/m}$*, we have*

$$\mathbb{E}[\|X_k\|^2] \leq C_d := \frac{MR^2 + 2BR + B + 4A}{m} + \frac{8(M+B)M\beta^{-1}d}{m^3} + \frac{2(M+B)b}{m^2} + \frac{b}{m}\log 3. \tag{C.19}$$

C.1.4 Proofs of auxiliary results

Proof of Lemma 26. By following the proof of Lemma 22. We get

$$\mathbb{P}\left(\sup_{t_0 \leq t \leq t_1} \left\|Q_{t_0}(t_1)Z_t^0\right\| \geq h\right) \leq \frac{1}{\sqrt{\det(I - \beta\theta\Sigma_{t_1})}} e^{-\frac{\beta\theta}{2}[h^2 - \langle \mu_{t_1}, (I-\beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle]},$$

Recall from Lemma 6 that for any $\tilde{\varepsilon} > 0$, there exists some $C_J(\tilde{\varepsilon})$ such that for every $t \geq 0$,

$$\left\|e^{-tA_J H}\right\| \leq C_J(\tilde{\varepsilon}) e^{-(\lambda_1^J - \tilde{\varepsilon})t},$$

Hence, by the definition of Σ_t from (C.13), we get

$$\|\Sigma_t\| \leq 2\beta^{-1} \int_0^\infty (C_J(\tilde{\varepsilon}))^2 e^{-2(\lambda_1^J - \tilde{\varepsilon})t} dt = \frac{\beta^{-1}(C_J(\tilde{\varepsilon}))^2}{\lambda_1^J - \tilde{\varepsilon}}.$$

The rest of the proof follows similarly as in the proof of Lemma 22. $\square$

Proof of Lemma 27. Note that

$$\left\|Q_{t_0}(t_1)Z_t^1\right\| \leq \int_0^t \left\|e^{(s-t_1)A_J H}\right\| \|A_J\| \, \|\rho(Y(s))\| \, ds,$$

and by applying $\|\rho(Y(t))\| \leq \frac{1}{2}L\|Y(t)\|^2$ and Lemma 6, and $t_0 \leq t \leq (t_1 \wedge \tau)$ and the definition of the stopping time τ in Proposition 25, we get the desired result. □

Proof of Lemma 28. Note that if we can show that $F(x)$ is a Lyapunov function for $X(t)$:

$$\mathcal{L}F(x) \leq -\epsilon_1 F(x) + b_1, \tag{C.20}$$

for some $\epsilon_1, b_1 > 0$, then

$$\mathbb{E}[F(X(t))] \leq F(x) + \frac{b_1}{\epsilon_1}.$$

Let us first prove this. Applying Ito formula to $e^{\epsilon_1 t}F(X(t))$, we obtain from Dynkin formula and the drift condition (C.20) that for $t_K := \min\{t, \tau_K\}$ with τ_K be the exit time of $X(t)$ from a ball centered at 0 with radius K with $X(0) = x$,

$$\mathbb{E}[e^{\epsilon_1 t_K}F(X(t_K))] \leq F(x) + \mathbb{E}\left[\int_0^{t_K} b_1 e^{\epsilon_1 s} ds\right] \leq F(x) + \int_0^t b_1 e^{\epsilon_1 s} ds \leq F(x) + \frac{b_1}{\epsilon_1} \cdot e^{\epsilon_1 t}.$$

Let $K \to \infty$, then we can infer from Fatou's lemma that for any t:

$$\mathbb{E}\left[e^{\epsilon_1 t}F(X(t))\right] \leq F(x) + \frac{b_1}{\epsilon_1} \cdot e^{\epsilon_1 t}.$$

Hence, we have

$$\mathbb{E}[F(X(t))] \leq F(x) + \frac{b_1}{\epsilon_1}.$$

Next, let us prove (C.20). By the definition of $\mathcal{L}$ in (C.17), we can compute that

$$\begin{aligned}\mathcal{L}F(x) &= -A_J \nabla F(x) \cdot \nabla F(x) + \beta^{-1}\Delta F(x) \\ &= -\|\nabla F(x)\|^2 + \beta^{-1}\Delta F(x),\end{aligned}$$

since J is anti-symmetric so that $\langle \nabla F(x), J\nabla F x\rangle = 0$. Moreover,

$$\|x\| \cdot \|\nabla F(x)\| \geq \langle x, \nabla F(x)\rangle \geq m\|x\|^2 - b, \tag{C.21}$$

implies that

$$\|\nabla F(x)\| \geq m\|x\| - \frac{b}{\|x\|} \geq \frac{1}{2}m\|x\|, \tag{C.22}$$

provided that $\|x\| \geq \sqrt{2b/m}$, and thus

$$\mathcal{L}F(x) \leq -\frac{m^2}{4}\|x\|^2 + \beta^{-1}\Delta F(x) \leq -\frac{m^2}{4}\|x\|^2 + \frac{mb}{2} + \beta^{-1}\Delta F(x), \tag{C.23}$$

for any $\|x\| \geq \sqrt{2b/m}$. On the other hand, for any $\|x\| \leq \sqrt{2b/m}$, we have

$$\mathcal{L}F(x) \leq \beta^{-1}\Delta F(x) \leq -\frac{m^2}{4}\|x\|^2 + \frac{mb}{2} + \beta^{-1}\Delta F(x). \tag{C.24}$$

Hence, for any $x \in \mathbb{R}^d$,

$$\mathcal{L}F(x) \leq -\frac{m^2}{4}\|x\|^2 + \frac{mb}{2} + \beta^{-1}\Delta F(x). \tag{C.25}$$

Next, recall that F is M-smooth, and thus

$$\Delta F(x) \leq Md.$$

Finally, by Lemma 43,

$$F(x) \leq \frac{M}{2}\|x\|^2 + B\|x\| + A \leq \frac{M+B}{2}\|x\|^2 + \frac{B}{2} + A.$$

Therefore, we have

$$\mathcal{L}F(x) \leq -\frac{m^2}{2(M+B)}F(x) + \frac{m^2(\frac{B}{2}+A)}{2(M+B)} + \frac{mb}{2} + M\beta^{-1}d.$$

Hence, the proof is complete. $\square$

Proof of Corollary 29. Recall from Lemma 43 that

$$F(x) \geq \frac{m}{2}\|x\|^2 - \frac{b}{2}\log 3,$$

which implies that

$$\|x\|^2 \leq \frac{2}{m}F(x) + \frac{b}{m}\log 3.$$

It then follows from Lemma 28 that

$$\mathbb{E}[\|X(t)\|^2] \leq \frac{2}{m}F(x) + \frac{B}{m} + \frac{2A}{m} + \frac{2b(M+B)}{m^2} + \frac{4M\beta^{-1}d(M+B)}{m^3} + \frac{b}{m}\log 3.$$

Recall that $\|X(0)\| = \|x\| \leq R$ and by Lemma 43 we get $F(x) \leq \frac{M}{2}\|x\|^2 + B\|x\| + A$, and thus

$$\mathbb{E}[\|X(t)\|^2] \leq C_c = \frac{MR^2 + 2BR + B + 4A}{m} + \frac{2b(M+B)}{m^2} + \frac{4M\beta^{-1}d(M+B)}{m^3} + \frac{b}{m}\log 3.$$

$\square$

Proof of Lemma 30. Suppose we have

$$\frac{\mathbb{E}_x[F(X_1)] - F(x)}{\eta} \leq -\epsilon_2 F(x) + b_2, \tag{C.26}$$

uniformly for small η, where ϵ_2, b_2 are postive constants that are independent of η, then we will first show below that

$$\mathbb{E}_x[F(X_k)] \leq F(x) + \frac{b_2}{\epsilon_2}.$$

We will use the discrete Dynkin's formula (see, e.g. [MT92, Section 4.2]). Let $\mathbb{F}_i$ denote the filtration generated by $X_0, \dots, X_i$. Note $\{X_k : k \geq 0\}$ is a time-homogeneous Markov process, so the drift condition (C.26) implies that

$$\mathbb{E}[F(X_i)|\mathbb{F}_{i-1}] \leq (1 - \eta\epsilon_2)F(X_{i-1}) + b_2.$$

Then by letting $r = 1/(1 - \eta\epsilon_2)$, we obtain

$$\mathbb{E}\left[rF(X_i)|\mathbb{F}_{i-1}\right] \leq F(X_{i-1}) + rb_2.$$

Then we can compute that

$$\mathbb{E}\left[r^i F(X_i)|\mathbb{F}_{i-1}\right] - r^{i-1}F(X_{i-1}) = r^{i-1} \cdot \left[\mathbb{E}[rF(X_i)|\mathbb{F}_{i-1}] - F(X_{i-1})\right] \leq r^i b_2. \tag{C.27}$$

Define the stopping time $\tau_{k,K} = \min\{k, \inf\{i : |X_i| \geq K\}\}$, where K is a positive integer, so that X_i is essentially bounded for $i \leq \tau_{k,K}$. Applying the discrete Dynkin's formula (see, e.g. [MT92, Section 4.2]), we have

$$\mathbb{E}_x\left[r^{\tau_{k,K}} F(X_{\tau_{k,K}})\right] = \mathbb{E}_x\left[F(X_0)\right] + \mathbb{E}\left[\sum_{i=1}^{\tau_{k,K}} \left(\mathbb{E}[r^i F(X_i)|\mathbb{F}_{i-1}] - r^{i-1}F(X_{i-1})\right)\right].$$

Then it follows from (C.27) that

$$\mathbb{E}_x\left[r^{\tau_{k,K}} F(X_{\tau_{k,K}})\right] \leq F(x) + b_2\eta \sum_{i=1}^{k} r^i.$$

As $\tau_{k,K} \to k$ almost surely as $K \to \infty$, we infer from Fatou's Lemma that

$$\mathbb{E}_x\left[r^k F(X_k)\right] \leq F(x) + b_2\eta \sum_{i=1}^{k} r^i,$$

which implies that for all k,

$$\mathbb{E}_x\left[F(X_k)\right] \leq F(x) + \frac{b_2\eta}{r-1} = F(x) + \frac{b_2(1 - \eta_2\epsilon_2)}{\epsilon_2} \leq F(x) + \frac{b_2}{\epsilon_2},$$

as $r = 1/(1-\eta_2\epsilon_2)$. Hence we have

$$\mathbb{E}_x\left[F(X_k)\right] \le F(x) + \frac{b_2}{\epsilon_2}.$$

It remains to prove (C.26). Note that as ∇F is Lipschitz continuous with constant M so that:

$$F(y) \le F(x) + \nabla F(x)(y-x) + \frac{M}{2}\|y-x\|^2.$$

Therefore,

$$\begin{aligned}
\frac{\mathbb{E}_x[F(X_1)] - F(x)}{\eta} &= \frac{1}{\eta}\left(\mathbb{E}_x\left[F(x - \eta A_J(\nabla F(x)) + \sqrt{2\eta\beta^{-1}}\xi_0)\right] - F(x)\right) \\
&\le -\nabla F(x) A_J \nabla F(x) + \frac{M}{2\eta}\mathbb{E}_x\left[\left\|-\eta A_J(\nabla F(x)) + \sqrt{2\eta\beta^{-1}}\xi_0\right\|^2\right] \\
&= -\|\nabla F(x)\|^2 + \frac{M}{2}\eta\|A_J\nabla F(x)\|^2 + M\beta^{-1}d \\
&\le -\frac{1}{2}\|\nabla F(x)\|^2 + M\beta^{-1}d\,,
\end{aligned}$$

provided that $\frac{M}{2}\|A_J\|^2\eta \le \frac{1}{2}$. Similar to the arguments in (C.21)-(C.25), we get

$$\frac{\mathbb{E}_x[F(X_1)] - F(x)}{\eta} \le -\frac{m^2}{8}\|x\|^2 + M\beta^{-1}d + \frac{mb}{4}.$$

Finally, by Lemma 43,

$$F(x) \le \frac{M}{2}\|x\|^2 + B\|x\| + A \le \frac{M+B}{2}\|x\|^2 + \frac{B}{2} + A.$$

Therefore, we have

$$\frac{\mathbb{E}_x[F(X_1)] - F(x)}{\eta} \le -\frac{m^2}{4(M+B)}F(x) + \frac{m^2(\frac{B}{2}+A)}{4(M+B)} + M\beta^{-1}d + \frac{mb}{4}.$$

Hence, the proof is complete. □

Proof of Corollary 31. The proof is similar to the proof of Corollary 29 and is thus omitted. □

D Proofs of results in Section 5

Proof of Proposition 19. Write u as the corresponding eigenvector of $A_J\mathbb{L}^\sigma$ for the eigenvalue $-\mu_J^* < 0$, so we have

$$A_J\mathbb{L}^\sigma u = -\mu_J^* u. \tag{D.1}$$

Then it follows that

$$(-\mu_J^*)u^*\mathbb{L}^\sigma u = u^*\mathbb{L}^\sigma(-\mu_J^* u) = u^*\mathbb{L}^\sigma A_J \mathbb{L}^\sigma u = u^*(\mathbb{L}^\sigma)^T A_J \mathbb{L}^\sigma u = |\mathbb{L}^\sigma u|^2 + u^*(\mathbb{L}^\sigma)^T J \mathbb{L}^\sigma u,$$

where u^* denotes the conjugate transpose of u, $(\mathbb{L}^\sigma)^T$ denotes the transpose of $\mathbb{L}^\sigma$, and $(\mathbb{L}^\sigma)^T = \mathbb{L}^\sigma$ as $\mathbb{L}^\sigma$ is a real symmetric matrix. It is easy to see that $u^*\mathbb{L}^\sigma u$ is a real number as $(u^*\mathbb{L}^\sigma u)^* = u^*\mathbb{L}^\sigma u$. In addition, $u^*(\mathbb{L}^\sigma)^T J\mathbb{L}^\sigma u$ is pure imaginary, since $(u^*(\mathbb{L}^\sigma)^T J \mathbb{L}^\sigma u)^* = u^*(\mathbb{L}^\sigma)^T J^T \mathbb{L}^\sigma u = -u^*(\mathbb{L}^\sigma)^T J \mathbb{L}^\sigma u$ by the fact that J is an anti-symmetric real matrix. Hence, we deduce that

$$u^*(\mathbb{L}^\sigma)^T J \mathbb{L}^\sigma u = 0,$$

and it implies that

$$(-\mu_J^*)u^*\mathbb{L}^\sigma u = |\mathbb{L}^\sigma u|^2. \tag{D.2}$$

Note $u^*\mathbb{L}^\sigma u \neq 0$ as otherwise 0 becomes an eigenvalue of $\mathbb{L}^\sigma$ from (D.2), which is a contradiction. In fact, we obtain from (D.2) that $-u^*\mathbb{L}^\sigma u > 0$ as $\mu_J^* > 0$ and $|\mathbb{L}^\sigma u|^2 > 0$.

Then together with the decomposition $\mathbb{L}^\sigma = S^T D S$ from (5.10) we obtain

$$\mu_J^* = \frac{|\mathbb{L}^\sigma u|^2}{-u^*\mathbb{L}^\sigma u} = \frac{u^* S^* D^2 S u}{-u^* S^* D S u} = \frac{\sum_{i=1}^n \mu_i^2 |(Su)_i|^2}{\sum_{i=1}^n -\mu_i |(Su)_i|^2}, \tag{D.3}$$

where $(Su)_i$ denotes the i-th component of the vector Su. Since $\mu_1 < 0 < \mu_2 < \ldots < \mu_n$, we then have $(Su)_1 \neq 0$ as otherwise $-u^*\mathbb{L}^\sigma u = \sum_{i=1}^n -\mu_i |(Su)_i|^2 \leq 0$, which is a contradiction. Therefore, we conclude from (D.3) that

$$\mu_J^* \geq |\mu_1| = \mu^*(\sigma). \tag{D.4}$$

The equality $\mu_J^* = |\mu_1| = \mu^*(\sigma)$ is attained if and only if $(Su)_i = 0$ for $i = 2, \ldots, n$. Or equivalently if and only if the vector $Su = ae_1$ where a is a non-zero constant and $e_1 = [1\ 0\ \ldots\ 0]^T$ is the first basis vector. Since $S^{-1} = S^T$, this is also equivalent to $u = av$ where $v = S^T e_1$ is an eigenvector of $\mathbb{L}^\sigma$ corresponding to the eigenvalue λ_1. Since u and v are related up to a constant, this is the same as saying v is an eigenvector of $A_J\mathbb{L}^\sigma$ satisfying (D.1). Since v is also an eigenvalue of $\mathbb{L}^\sigma$ and J being anti-symmetric, has only purely imaginary eigenvalues except a zero eigenvalue, this is if and only if $Jv = 0$. In other words, the equality $\mu_J^* = |\mu_1| = \mu^*(\sigma)$ is attained if and only if the eigenvector of $\mathbb{L}^\sigma$ corresponding to the negative eigenvalue μ_1 is an eigenvector of J for the eigenvalue 0.

We note finally that Equation (5.11) then readily follows from (5.8) and (D.4). □

Proof of Proposition 20. Write $\tau_{a_1 \to a_2}^{\beta,R}$ for the first time that the continuous-time dynamics $\{X(t)\}$ starting from a_1 to exit the region D_R. Then by monotone convergence theorem, we have

$$\lim_{R\to\infty} \mathbb{E}\left[\tau_{a_1\to a_2}^{\beta,R}\right] = \mathbb{E}\left[\tau_{a_1\to a_2}^{\beta}\right].$$

Hence, for fixed $\epsilon > 0$, one can choose a sufficiently large R such that

$$\left|\mathbb{E}\left[\tau_{a_1\to a_2}^{\beta,R}\right] - \mathbb{E}\left[\tau_{a_1\to a_2}^{\beta}\right]\right| < \epsilon. \tag{D.5}$$

We next control the expected difference between the exit times $\hat{\tau}_{a_1\to a_2}^{\beta,R}$ of the discrete dynamics, and $\tau_{a_1\to a_2}^{\beta,R}$ of the continuous dynamics, from the bounded domain D_R. For fixed ϵ and large R, we can infer from [GM05, Theorem 4.2] that[18], for sufficiently small stepsize $\eta \leq \bar{\eta}(\epsilon, R, \beta)$,

$$\left|\mathbb{E}\left[\hat{\tau}_{a_1\to a_2}^{\beta,R}\right] - \mathbb{E}\left[\tau_{a_1\to a_2}^{\beta,R}\right]\right| < \epsilon. \tag{D.6}$$

Together with (D.5), we obtain for η sufficiently small,

$$\left|\mathbb{E}\left[\hat{\tau}_{a_1\to a_2}^{\beta,R}\right] - \mathbb{E}\left[\tau_{a_1\to a_2}^{\beta}\right]\right| < 2\epsilon.$$

The proof is therefore complete. □

E Recurrence and escape times for underdamped Langevin dynamics with small friction

In this section, we investigate the local analysis results for the underdamped Langevin dynamics (1.10)-(1.11) when the friction coefficient γ is small, and in particular, we assume that $0 < \gamma < 2\sqrt{m}$.

E.1 Main results

Theorem 32. *Fix $\gamma < 2\sqrt{m}$, $\delta \in (0,1)$ and $r > 0$. Assume*

$$0 < \varepsilon < \overline{\varepsilon}^U := \min\left\{\frac{\sqrt{m}(1-\hat{\varepsilon})}{4C_{\hat{\varepsilon}}L(1+\frac{1}{64C_{\hat{\varepsilon}}^2})}, 8rC_{\hat{\varepsilon}}\right\} = \min\left\{\mathcal{O}\left(m\right), \mathcal{O}\left(\frac{r}{\sqrt{m}}\right)\right\},$$

where $\hat{\varepsilon}$ and $C_{\hat{\varepsilon}}$ are defined in (3.8), and we assumed that $\hat{\varepsilon} = \Omega(1)$ and thus $C_{\hat{\varepsilon}}$ is of order $m^{-1/2}$. Define the recurrence time

$$\mathcal{T}_{rec}^U := \frac{2}{\sqrt{m}(1-\hat{\varepsilon})}\log\left(\frac{8rC_{\hat{\varepsilon}}}{\varepsilon}\right) = \mathcal{O}\left(\frac{1}{\sqrt{m}}\log\left(\frac{r}{\varepsilon m}\right)\right), \tag{E.1}$$

and we also define the escape time

$$\mathcal{T}_{esc}^U := \mathcal{T}_{rec}^U + \mathcal{T},$$

[18] The Assumption (H2') in Theorem 4.2 of [GM05] can be readily verified in our setting: for both reversible and non-reversible SDE, the drift and diffusion coefficients are clearly Lipschitz; the diffusion matrix is uniformly elliptic; and the domain D_R is bounded and it satisfies the exterior cone condition.

for any arbitrary $\mathcal{T} > 0$.

Consider an arbitrary initial point x for the underdamped Langevin dynamics and a local minimum x_* at a distance at most r. Assume that the stepsize η satisfies

$$\eta \leq \overline{\eta}^U = \min\left\{\mathcal{O}(\varepsilon), \mathcal{O}\left(\frac{m^2\beta\delta\varepsilon^2}{(md+\beta)\mathcal{T}_{rec}^U}\right), \mathcal{O}\left(\frac{m^{3/2}\delta^2}{(md+\beta)\mathcal{T}_{esc}^U}\right), \mathcal{O}\left(\frac{m^{1/2}\beta}{d+\beta}\right)\right\},$$

where $\overline{\eta}^U$ is more formally defined in Table 1 and β satisfies

$$\beta \geq \underline{\beta}^U = \max\left\{\Omega\left(\frac{(d+\log((\mathcal{T}+1)/\delta))C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}\right), \Omega\left(\frac{d\eta m^{1/2}\log(\delta^{-1}\mathcal{T}_{rec}^U/\eta)}{\varepsilon^2}\right)\right\},$$

where more formally $\underline{\beta}^U = \max\left\{\underline{\beta}_3^U, \underline{\beta}_2^U\right\}$ and $\underline{\beta}_2^U$ is defined in Table 1 and

$$\underline{\beta}_3^U := \frac{128\gamma C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}\left(d\log(2) + \log\left(\frac{6(\gamma^2+M^2+1)^{1/2}\mathcal{T}+3}{\delta}\right)\right),$$

for any realization of Z, with probability at least $1-\delta$ w.r.t. the Gaussian noise, at least one of the following events will occur:

1. $\|X_k - x_*\| \geq \frac{1}{2}\left(\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})k\eta}\right)$ *for some $k \leq \eta^{-1}\mathcal{T}_{rec}^U$.*
2. $\|X_k - x_*\| \leq \varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})k\eta}$ *for every $\eta^{-1}\mathcal{T}_{rec}^U \leq k \leq \eta^{-1}\mathcal{T}_{esc}^U$.*

Remark 33. *Notice that in Theorem 32, the definition of η and β are coupled since $\overline{\eta}^U$ depends on β and $\underline{\beta}^U$ depends on η. A closer look reveals that when η is sufficiently small, the first term in the definition of $\underline{\beta}^U$ dominates the second term and $\underline{\beta}^U$ is independent of η. So to satisfy the constraints in Theorem 32, it suffices to first choose β to be larger than the first term in $\underline{\beta}^U$ and then choose η to be sufficiently small.*

Remark 34. *We can apply Theorem 10 to obtain Theorem 39 for the population risk in the Appendix, similar to Theorem 3 in [TLR18].*

Remark 35. *In [TLR18], the overdamped Langevin algorithm is used and the recurrence time $\mathcal{T}_{rec} = \mathcal{O}\left(\frac{1}{m}\log(\frac{r}{\varepsilon})\right)$, and thus our recurrence time $\mathcal{T}_{rec}^U = \mathcal{O}\left(\frac{1}{\sqrt{m}}\log\left(\frac{r}{\varepsilon m}\right)\right)$ for the underdamped Langevin algorithm with the choice of $\gamma < 2\sqrt{m}$, which has a square root factor improvement ignoring the logarithmic factor. This recurrence time is worse than $\mathcal{T}_{rec}^U = \mathcal{O}\left(\frac{1}{\sqrt{m}}\log\left(\frac{r}{\varepsilon}\right)\right)$ for the underdamped Langevin algorithm with the choice of $\gamma = 2\sqrt{m}$ by a logarithmic factor assuming $C_H = \mathcal{O}(1)$.*

Remark 36. *Let us compare the case $\gamma = 2\sqrt{m}$ with the case $\gamma < 2\sqrt{m}$ (to be discussed in Section E). When $\gamma = 2\sqrt{m}$, since $W_{-1}(-x) \sim \log(1/x)$ for $x \to 0^+$, assuming $r, \varepsilon, C_H = \mathcal{O}(1)$, we have*

$$\mathcal{T}_{rec}^U \sim \frac{1}{2\sqrt{m}} \log(1/m), \qquad \text{as } m \to 0^+.$$

When $\gamma < 2\sqrt{m}$, we have $\mathcal{T}_{rec}^U = \frac{2}{\sqrt{m}(1-\hat{\varepsilon})} \log(8rC_{\hat{\varepsilon}}/\varepsilon)$, where $C_{\hat{\varepsilon}} = \frac{1+M}{\sqrt{m(1-(1-\hat{\varepsilon})^2)}}$, and $\hat{\varepsilon} = 1 - \frac{\gamma}{2\sqrt{m}} \in (0,1)$. For example, if $\gamma = m^\chi$, where $\chi \in (0, 1/2)$ and $m \to 0^+$, then $\mathcal{T}_{rec}^U \sim \frac{2}{m^\chi} \log(m)$, as $m \to 0^+$, and if $\gamma = \gamma_0 \sqrt{m}$, where $\gamma_0 \in (0,2)$, then $\mathcal{T}_{rec}^U \sim \frac{2}{\gamma_0\sqrt{m}} \log(m)$, as $m \to 0^+$. To summarize, with all the parameters fixed, as $m \to 0^+$, the choice $\gamma = 2\sqrt{m}$ is more optimal than the choice $\gamma < 2\sqrt{m}$. On the other hand, when the second smallest eigenvalue is close to the smallest eigenvalue m, such that $C_H = \max_{i:\lambda_i > m} \frac{(1+\lambda_i)^2}{\lambda_i - m}$ is large, it is more desirable to use the underdamped Langevin algorith with $\gamma < 2\sqrt{m}$ instead.

E.2 Proof of Theorem 32

The proof of Theorem 32 is similar to the proof of Theorem 10 and the following proposition, and the similar arguments in Section B.1.1.

Proposition 37. *Assume $\gamma < 2\sqrt{m}$. Fix any $r > 0$ and*

$$\varepsilon < \min \left\{ \frac{\sqrt{m}(1-\hat{\varepsilon})}{4C_{\hat{\varepsilon}} L (1 + \frac{1}{64 C_{\hat{\varepsilon}}^2})}, 8rC_{\hat{\varepsilon}} \right\}.$$

Consider the stopping time:

$$\tau := \inf \left\{ t \geq 0 : \|X(t) - x_*\| \geq \varepsilon + r e^{-\sqrt{m}(1-\hat{\varepsilon})t} \right\}.$$

For any initial point $X(0) = x$ with $\|x - x_*\| \leq r$, and

$$\beta \geq \frac{128 \gamma C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2} \left(d \log(2) + \log\left(\frac{2\|H_\gamma\|\mathcal{T} + 1}{\delta} \right) \right),$$

we have

$$\mathbb{P}_x \left(\tau \in \left[\mathcal{T}_{rec}^U, \mathcal{T}_{esc}^U \right] \right) \leq \delta.$$

The term $\|H_\gamma\|$ in Proposiiton 37 can be bounded using Lemma 41. Based on Proposition 37, the proof of Theorem 32 is similar to the proof of Theorem 10. So in the rest of the section, we will only focus on the proof of Proposition 37.

E.2.1 Proof of Proposition 37

In this section, we focus on the proof of Proposition 37. We recall some definitions from Section B.1.2. We recall the matrices H_γ and $I^{(2)}$ from (B.23), the matrix flow $Q_{t_0}(t)$ from (B.24) and the processes Z_t^0 and Z_t^1 from (B.25)-(B.26), and also μ_t and Σ_t from (B.27)-(B.28).

Lemma 38. *Assume $\gamma < 2\sqrt{m}$. For any $\theta \in \left(0, \frac{\sqrt{m}(1-\hat{\varepsilon})}{\gamma C_{\hat{\varepsilon}}^2}\right)$, and $h > 0$ and any $(V(0), Y(0))$,*

$$\mathbb{P}\left(\sup_{t_0 \le t \le t_1} \|Q_{t_0}(t_1) Z_t^0\| \ge h\right) \le \left(1 - \frac{\theta \gamma C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})}\right)^{-d} e^{-\frac{\beta\theta}{2}[h^2 - \langle \mu_{t_1}, (I - \beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle]}.$$

Proof of Lemma 38. The proof is similar to the proof of Lemma 22. Let us recall from Lemma 1 that if $\gamma < 2\sqrt{m}$, then

$$\left\|e^{-tH_\gamma}\right\| \le C_{\hat{\varepsilon}} e^{-\sqrt{m}(1-\hat{\varepsilon})t},$$

where $C_{\hat{\varepsilon}}$ and $\hat{\varepsilon}$ are defined in (3.8). Therefore, we have

$$\|\Sigma_{t_1}\| \le 2\gamma\beta^{-1} \int_0^{t_1} C_{\hat{\varepsilon}}^2 e^{-2\sqrt{m}(1-\hat{\varepsilon})t} dt \le \frac{\gamma\beta^{-1} C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})}.$$

Therefore we infer that the eigenvalues of $I - \beta\theta\Sigma$ are bounded below by $1 - \theta \frac{\gamma C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})}$. The conclusion then follows from (B.41). $\square$

Proof of Proposition 37. Since $\|Y(0)\| \le r$, we know that $\tau > 0$. Fix some $\mathcal{T}_{\mathrm{rec}}^U \le t_0 \le t_1$, such that $t_1 - t_0 \le \frac{1}{2\|H_\gamma\|}$. Then, for every $t \in [t_0, t_1]$,

$$\|Y(t)\| \le \left\|e^{(t_1 - t)H_\gamma} Q_{t_0}(t_1) Z_t\right\| \le e^{\frac{1}{2}} \left\|Q_{t_0}(t_1) Z_t\right\|.$$

Recall that $\gamma < 2\sqrt{m}$. Similar to the derivations in (B.29), we get

$$\begin{aligned}\mathbb{P}(\tau \in [t_0, t_1]) \le\ & \mathbb{P}\left(\sup_{t_0 \le t \le t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t^0\|}{\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})t}} \ge c_0, \tau \ge t_0\right) \\ & + \mathbb{P}\left(\sup_{t_0 \le t \le t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t^1\|}{\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})t}} \ge c_1, \tau \ge t_0\right), \end{aligned} \tag{E.2}$$

where $c_0 + c_1 = \frac{1}{2}$ and $c_0, c_1 > 0$.

First, we show that the second term in (E.2) is zero. On the event $\tau \in [t_0, t_1]$, for any $0 \le s \le t_1 \wedge \tau$, we have

$$\|\rho(Y(s))\| \le \frac{L}{2} \|Y(s)\|^2 \le \frac{L}{2} \left(\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})s}\right)^2.$$

Therefore, for any $t \in [t_0, t_1 \wedge \tau]$, since $\gamma < 2\sqrt{m}$, by Lemma 1, we get

$$\begin{aligned}
\|Q_{t_0}(t_1)Z_t^1\| &\leq \int_0^t \left\| e^{(s-t_1)H_\gamma} \right\| \cdot \|\rho(Y(s))\| ds \\
&\leq \frac{C_{\hat{\varepsilon}} L}{2} \int_0^t e^{(s-t_1)\sqrt{m}(1-\hat{\varepsilon})} \left(\varepsilon + r e^{-\sqrt{m}(1-\hat{\varepsilon})s} \right)^2 ds \\
&\leq C_{\hat{\varepsilon}} L \int_0^t e^{(s-t)\sqrt{m}(1-\hat{\varepsilon})} \left(\varepsilon^2 + r^2 e^{-2\sqrt{m}(1-\hat{\varepsilon})s} \right) ds \\
&\leq \frac{C_{\hat{\varepsilon}} L}{\sqrt{m}(1-\hat{\varepsilon})} \left(\varepsilon^2 + r^2 e^{-\sqrt{m}(1-\hat{\varepsilon})t} \right) \\
&\leq \frac{C_{\hat{\varepsilon}} L}{\sqrt{m}(1-\hat{\varepsilon})} \varepsilon^2 \left(1 + \frac{1}{64 C_{\hat{\varepsilon}}^2} \right),
\end{aligned}$$

since $t \geq t_0 \geq \mathcal{T}_{\text{rec}}^U = \frac{2}{\sqrt{m}(1-\hat{\varepsilon})} \log\left(\frac{8rC_{\hat{\varepsilon}}}{\varepsilon}\right)$. Consequently, if we take $c_1 = \frac{C_{\hat{\varepsilon}} L}{\sqrt{m}(1-\hat{\varepsilon})} \varepsilon (1 + \frac{1}{64 C_{\hat{\varepsilon}}^2})$, then,

$$\sup_{t_0 \leq t \leq t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t\|}{\varepsilon + r e^{-\sqrt{m}(1-\hat{\varepsilon})t}} \leq \frac{1}{\varepsilon} \sup_{t_0 \leq t \leq t_1 \wedge \tau} \|Q_{t_0}(t_1) Z_t\| \leq c_1,$$

which implies that

$$\mathbb{P}\left(\sup_{t_0 \leq t \leq t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1) Z_t^1\|}{\varepsilon + r e^{-\sqrt{m}(1-\hat{\varepsilon})t}} \geq c_1, \tau \geq t_0 \right) = 0.$$

Moreover, $c_0 = \frac{1}{2} - c_1 = \frac{1}{2} - \frac{C_{\hat{\varepsilon}} L}{\sqrt{m}(1-\hat{\varepsilon})}(1 + \frac{1}{64 C_{\hat{\varepsilon}}^2})\varepsilon > \frac{1}{4}$ since it is assumed $\varepsilon < \frac{\sqrt{m}(1-\hat{\varepsilon})}{4 C_{\hat{\varepsilon}} L (1 + \frac{1}{64 C_{\hat{\varepsilon}}^2})}$.

Second, we apply Lemma 38 to bound the first term in (E.2). By using $V(0) = 0$ and $\|Y(0)\| \leq r$ and the definition of μ_{t_1} and Σ_{t_1} in (B.27) and (B.28), we get

$$\begin{aligned}
\left\langle \mu_{t_1}, (I - \beta\theta\Sigma_{t_1})^{-1} \mu_{t_1} \right\rangle &= \left\langle e^{-t_1 H_\gamma}(V(0), Y(0))^T, (I - \beta\theta\Sigma_{t_1})^{-1} e^{-t_1 H_\gamma}(V(0), Y(0))^T \right\rangle \\
&\leq \frac{1}{1 - \frac{\theta\gamma C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})}} C_{\hat{\varepsilon}}^2 e^{-2\sqrt{m}(1-\hat{\varepsilon})t_1} r^2 \leq 2 C_{\hat{\varepsilon}}^2 e^{-2\sqrt{m}(1-\hat{\varepsilon})t_1} r^2,
\end{aligned}$$

by choosing $\theta = \frac{1}{2}\gamma^{-1} C_{\hat{\varepsilon}}^{-2} \sqrt{m}(1-\hat{\varepsilon})$, Finally, since

$$t_1 \geq \mathcal{T}_{\text{rec}}^U = \frac{2}{\sqrt{m}(1-\hat{\varepsilon})} \log\left(8rC_{\hat{\varepsilon}}/\varepsilon\right) \geq \frac{1}{\sqrt{m}(1-\hat{\varepsilon})} \log\left(8rC_{\hat{\varepsilon}}/\varepsilon\right),$$

so that $2 C_{\hat{\varepsilon}}^2 e^{-2\sqrt{m}(1-\hat{\varepsilon})t_1} r^2 \leq \frac{1}{32}\varepsilon^2$, and thus

$$\left\langle \mu_{t_1}, (I - \beta\theta_{t_1}\Sigma_{t_1})^{-1} \mu_{t_1} \right\rangle \leq \frac{1}{32}\varepsilon^2.$$

Then with the choice of $h = (\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})t_1})c_0$ and $\theta = \frac{1}{2}\gamma^{-1}C_{\hat{\varepsilon}}^{-2}\sqrt{m}(1-\hat{\varepsilon})$ in Lemma 22, and notice that $h = (\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})t_1})c_0 \geq \varepsilon c_0$ we get

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t_0 \leq t \leq t_1 \wedge \tau} \frac{\|Q_{t_0}(t_1)Z_t^0\|}{\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})t}} \geq c_0, \tau \geq t_0\right) \\
&\leq \mathbb{P}\left(\sup_{t_0 \leq t \leq t_1} \|Q_{t_0}(t_1)Z_t^0\| \geq (\varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})t_1})c_0\right) \\
&\leq 2^d \cdot \exp\left(-\frac{\beta\gamma^{-1}C_{\hat{\varepsilon}}^{-2}\sqrt{m}(1-\hat{\varepsilon})}{4}[h^2 - \langle \mu_{t_1}, (I - \beta\theta\Sigma_{t_1})^{-1}\mu_{t_1}\rangle]\right) \\
&\leq 2^d \cdot \exp\left(-\frac{\beta\gamma^{-1}C_{\hat{\varepsilon}}^{-2}\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}{4}\left(c_0^2 - \frac{1}{32}\right)\right) \leq 2^d \cdot \exp\left(-\frac{\beta\gamma^{-1}C_{\hat{\varepsilon}}^{-2}\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}{128}\right).
\end{aligned}$$

Thus for any $t_0 \geq \mathcal{T}_{\text{rec}}^U$ and $t_0 \leq t_1 \leq t_0 + \frac{1}{2\|H_\gamma\|}$,

$$\mathbb{P}(\tau \in [t_0, t_1]) \leq 2^d \cdot \exp\left(-\frac{\beta\gamma^{-1}C_{\hat{\varepsilon}}^{-2}\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}{128}\right).$$

Fix any $\mathcal{T} > 0$ and recall the definition of the escape time $\mathcal{T}_{\text{esc}}^U = \mathcal{T} + \mathcal{T}_{\text{rec}}^U$. Partition the interval $[\mathcal{T}_{\text{rec}}^U, \mathcal{T}_{\text{esc}}^U]$ using the points $\mathcal{T}_{\text{rec}}^U = t_0 < t_1 < \cdots < t_{\lceil 2\|H_\gamma\|\mathcal{T}\rceil} = \mathcal{T}_{\text{esc}}^U$ with $t_j = j/(2\|H_\gamma\|)$, then we have

$$\begin{aligned}
&\mathbb{P}\left(\tau \in \left[\mathcal{T}_{\text{rec}}^U, \mathcal{T}_{\text{esc}}^U\right]\right) \\
&= \sum_{j=0}^{\lceil 2\|H_\gamma\|\mathcal{T}\rceil} \mathbb{P}(\tau \in [t_j, t_{j+1}]) \leq (2\|H_\gamma\|\mathcal{T} + 1) \cdot 2^d \cdot \exp\left(-\frac{\beta\gamma^{-1}C_{\hat{\varepsilon}}^{-2}\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}{128}\right) \leq \delta,
\end{aligned}$$

provided that

$$\beta \geq \frac{128\gamma C_{\hat{\varepsilon}}^2}{\sqrt{m}(1-\hat{\varepsilon})\varepsilon^2}\left(d\log(2) + \log\left(\frac{2\|H_\gamma\|\mathcal{T} + 1}{\delta}\right)\right).$$

The proof is complete. $\square$

F Generalization to population risk

In this section, we apply Theorem 10, Theorem 32 and Theorem 15 to study the population risk. We recall that the population risk is denoted by $\overline{F}$, and the empirical risk is denoted by F. First, we need the assumption that the population risk $\overline{F}$ is $(2\varepsilon_0, 2m)$-strongly Morse (see e.g. [MBM18]), that is, $\left\|\nabla\overline{F}(x)\right\| \leq 2\varepsilon_0$ implies $\min_{j\in[d]}\left|\lambda_j(\nabla^2\overline{F}(x))\right| \geq 2m$.

F.1 Underdamped Langevin dynamics

Theorem 39. *Suppose the assumptions in Theorem 10 holds for $\gamma = 2\sqrt{m}$ case and the assumptions in Theorem 32 holds for $\gamma < 2\sqrt{m}$ case. Assume that $\frac{n}{\log n} \geq \frac{c\sigma_0^2 d}{(\varepsilon_0 \wedge m)^2}$ and $\varepsilon \leq \frac{3m}{(1+\frac{\sqrt{m}}{8\sqrt{(C_H+2)m+(m+1)^2}})L}$ for the case $\gamma = 2\sqrt{m}$ and $\varepsilon \leq \frac{3m}{(1+\frac{1}{8C_{\widetilde{\varepsilon}}})L}$ for the case $\gamma < 2\sqrt{m}$. With probability at least $1-\delta$, w.r.t. both the training data z and the Gaussian noise, for any local minimum x_*^z of F* [19]*, either $\|X_k - x_*^z\| \geq \varepsilon/2$ for some $k \leq \lceil \eta^{-1}\mathcal{T}_{esc}^U \rceil$ or*

$$\overline{F}(x_*^z) \leq \min_{\eta^{-1}\mathcal{T}_{rec}^U \leq k \leq \eta^{-1}\mathcal{T}_{esc}^U} F(X_k) + \sigma_0\sqrt{\frac{cd\log n}{n}},$$

where

$$c := c_0(1 \vee \log((M \vee L \vee (B+MR))R\sigma_0/\delta)), \tag{F.1}$$

$$\sigma_0 := (A + (B+MR)R) \vee (B+MR) \vee (C+LR). \tag{F.2}$$

Proof. Let us first assume that $\gamma = 2\sqrt{m}$. Let x_*^z be a local minimum of the empirical risk F. By Lemma 45, all eigenvalues of of the Hessian $H = \nabla^2 F(x_*^z)$ are at least m and therefore the norm $\|\cdot\|_H = \|H^{1/2}\cdot\|$ is well defined and $\|\cdot\|_H \geq \sqrt{m}\|\cdot\|$.

We can decompose (letting $K_0 = \lceil \eta^{-1}\mathcal{T}_{\text{rec}}^U \rceil$ and $K = \lceil \eta^{-1}\mathcal{T}_{\text{esc}}^U \rceil$)

$$\overline{F}(x_*^z) - \min_{K_0 \leq k \leq K} F(X_k) = \left(\overline{F}(x_*^z) - F(x_*^z)\right) + \left(F(x_*^z) - \min_{K_0 \leq k \leq K} F(X_k)\right).$$

From Lemma 44, we know with probability at least $1-\delta$,

$$\overline{F}(x_*^z) - F(x_*^z) \leq \sigma_0\sqrt{(cd/n)\log n}.$$

In addition, we can infer from (4.1) that for any x,

$$\left| F(x) - F(x_*^z) - \frac{1}{2}\|x - x_*^z\|_H^2 \right| \leq \frac{L}{6}\|x - x_*^z\|^3,$$

Hence, we obtain

$$\begin{aligned} F(x_*^z) - \min_{K_0 \leq k \leq K} F(X_k) &= \max_{K_0 \leq k \leq K}(F(x_*^z) - F(X_k)) \\ &\leq \max_{K_0 \leq k \leq K}\left(\frac{L}{6}\|X_k - x_*^z\|^3 - \frac{1}{2}\|X_k - x_*^z\|_H^2\right) \\ &\leq \max_{K_0 \leq k \leq K}\left(\frac{L}{6}\|X_k - x_*^z\|^3 - \frac{m}{2}\|X_k - x_*^z\|^2\right). \end{aligned}$$

[19] With the notation x_*^z emphasizing the dependence on the training data z

By Theorem 10, with probability $1-\delta$, either $\|X_k - x_*^z\| \geq \varepsilon/2$ for some $k \leq \lceil \eta^{-1}\mathcal{T}_{\text{rec}}^U \rceil$ or

$$\begin{aligned} \|X_k - x_*^z\| \leq \varepsilon + re^{-\sqrt{m}\mathcal{T}_{\text{rec}}^U} &= \varepsilon + \frac{\varepsilon^2}{\mathcal{T}_{\text{rec}}^U 8r\sqrt{C_H + 2 + (m+1)^2}} \\ &\leq \varepsilon + \frac{\varepsilon\sqrt{m}}{8\sqrt{(C_H+2)m + (m+1)^2}}, \end{aligned} \tag{F.3}$$

for all $K_0 \leq k \leq K$, where we used the definition of $\mathcal{T}_{\text{rec}}^U$ for the case $\gamma = 2\sqrt{m}$ in (4.3), and the assumption $\varepsilon < \sqrt{\frac{C_H+2+(m+1)^2}{(C_H+2)m+(m+1)^2}}r$ and the property $\mathcal{T}_{\text{rec}}^U > \frac{1}{\sqrt{m}}$. If the latter occurs, then by (F.3), we have

$$\begin{aligned} &F(x_*^z) - \min_{K_0 \leq k \leq K} F(X_k) \\ &\leq \max_{K_0 \leq k \leq K}\left(\frac{L}{6}\|X_k - x_*^z\|^3 - \frac{m}{2}\|X_k - x_*^z\|^2\right) \\ &\leq \max_{K_0 \leq k \leq K}\|X_k - x_*^z\|^2\left(\frac{L}{6}\cdot\left(1 + \frac{\sqrt{m}}{8\sqrt{(C_H+2)m+(m+1)^2}}\right)\varepsilon - \frac{m}{2}\right) \leq 0\,, \end{aligned}$$

for $\varepsilon \leq \frac{3m}{(1+\frac{\sqrt{m}}{8\sqrt{(C_H+2)m+(m+1)^2}})L}$. The proof for the case $\gamma = 2\sqrt{m}$ is therefore complete.

The proof for the case $\gamma < 2\sqrt{m}$ is similar. The only difference is that we replace (F.3) by the following estimate

$$\|X_k - x_*^z\| \leq \varepsilon + re^{-\sqrt{m}(1-\hat{\varepsilon})\mathcal{T}_{\text{rec}}^U} \leq \varepsilon + re^{-\frac{1}{2}\sqrt{m}(1-\hat{\varepsilon})\mathcal{T}_{\text{rec}}^U} = \varepsilon + \frac{\varepsilon}{8C_{\hat{\varepsilon}}},$$

for all $K_0 \leq k \leq K$, where we used the definition of $\mathcal{T}_{\text{rec}}^U$ for the $\gamma < 2\sqrt{m}$ case in (E.1). $\square$

F.2 Non-reversible Langevin dynamics

Theorem 40. *Suppose the assumptions in Theorem 15 holds. Assume that* $\frac{n}{\log n} \geq \frac{c\sigma_0^2 d}{(\varepsilon_0 \wedge m)^2}$ *and* $\varepsilon \leq \frac{3m}{(1+\frac{1}{8C_J(\tilde{\varepsilon})})L}$. *With probability at least* $1-\delta$, *w.r.t. both the training data* Z *and the Gaussian noise, for any local minimum* x_*^z *of* F, *either* $\|X_k - x_*^z\| \geq \varepsilon/2$ *for some* $k \leq \eta^{-1}\mathcal{T}_{rec}^J$ *or*

$$\overline{F}(x_*^z) \leq \min_{\eta^{-1}\mathcal{T}_{rec}^J \leq k \leq \eta^{-1}\mathcal{T}_{esc}^J} F(X_k) + \sigma_0\sqrt{\frac{cd\log n}{n}},$$

where c *and* σ_0 *are defined in (F.1) and (F.2).*

Proof. The proof is similar to Theorem 39 and Theorem 3 in [TLR18]. The only difference is that we replace (F.3) by the following estimate

$$\|X_k - x_*^z\| \leq \varepsilon + re^{-m_J(\tilde{\varepsilon})\mathcal{T}_{\text{rec}}^J} \leq \varepsilon + re^{-\frac{1}{2}m_J(\tilde{\varepsilon})\mathcal{T}_{\text{rec}}^J} = \varepsilon + \frac{\varepsilon}{8C_J(\tilde{\varepsilon})},$$

for all $\lceil \eta^{-1}\mathcal{T}_{\text{rec}}^J \rceil \leq k \leq \lceil \eta^{-1}\mathcal{T}_{\text{esc}}^J \rceil$, where we used the definition of $\mathcal{T}_{\text{rec}}^J$ in (4.4). $\square$

G Supporting technical lemmas

Lemma 41. *Consider the square matrix H_γ defined by (3.2). We have*

$$\|H_\gamma\| \le \sqrt{\gamma^2 + M^2 + 1}.$$

Proof. It follows from (A.1) that

$$\|H_\gamma\| = \|T_\gamma\| = \max_i \|T_i(\gamma)\|. \tag{G.1}$$

We also compute

$$\|T_i(\gamma)\|^2 = \lambda_{\max}\left(T_i(\gamma)T_i(\gamma)^T\right) = \lambda_{\max}\left(\begin{bmatrix} \gamma^2 + \lambda_i^2 & -\gamma \\ -\gamma & 1 \end{bmatrix}\right),$$

where $\lambda_{\max}$ denotes the largest real part of the eigenvalues. This leads to

$$\|T_i(\gamma)\|^2 = \frac{\gamma^2 + \lambda_i^2 + 1 + \sqrt{(\gamma^2 + \lambda_i^2 + 1)^2 - 4\lambda_i^2}}{2} \le \gamma^2 + \lambda_i^2 + 1.$$

Since $m \le \lambda_i \le M$ for every i, we obtain

$$\max_i \|T_i(\gamma)\|^2 \le \max_i \left(\gamma^2 + \lambda_i^2 + 1\right) = \gamma^2 + M^2 + 1.$$

We conclude from (G.1). □

Lemma 42. *Let B_t be a standard d-dimensional Brownian motion. For any $u > 0$ and any $t_1 > t_0 \ge 0$ with $t_1 - t_0 = \eta > 0$, we have*

$$\mathbb{P}\left(\sup_{t\in[t_0,t_1]} \|B_t - B_{t_1}\| \ge u\right) \le 2^{1/4} e^{1/4} e^{-\frac{u^2}{4d\eta}}.$$

Proof. Also, by the time reversibility, stationarity of time increments of Brownian motion and Doob's martingale inequality, for any $\theta > 0$ so that $2\theta\eta < 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\in[t_0,t_1]} \|B_t - B_{t_1}\| \ge u\right) &= \mathbb{P}\left(\sup_{t\in[0,\eta]} \|B_t - B_0\| \ge u\right) \\
&\le e^{-\theta u^2} \mathbb{E}\left[e^{\theta \|B_\eta - B_0\|^2}\right] \\
&= e^{-\theta u^2} (1 - 2\theta\eta)^{-d/2}.
\end{aligned}$$

By choosing $\theta = 1/(4d\eta)$, we get

$$\mathbb{P}\left(\sup_{t\in[t_0,t_1]} \|B_t - B_{t_1}\| \ge u\right) \le \left(1 - \frac{1}{2d}\right)^{-\frac{d}{2}} e^{-\frac{u^2}{4d\eta}}.$$

Note that for any $x > 0$, $(1 + \frac{1}{x})^x < e$. Let us define $x > 0$ via

$$1 - \frac{1}{2d} = \frac{1}{1+x}.$$

Then, we get $d = \frac{1+x}{2x}$ and $x = \frac{1}{1-\frac{1}{2d}} - 1 \leq 1$, and

$$\left(1 - \frac{1}{2d}\right)^{-\frac{d}{2}} = \left(\frac{1}{1+x}\right)^{-\frac{1+x}{4x}} = (1+x)^{\frac{1}{4}}(1+x)^{\frac{1}{4x}} \leq 2^{1/4}e^{1/4}.$$

Hence,

$$\mathbb{P}\left(\sup_{t\in[t_0,t_1]} \|B_t - B_{t_1}\| \geq u\right) \leq 2^{1/4}e^{1/4}e^{-\frac{u^2}{4d\eta}}.$$

□

Lemma 43 (See [RRT17, Lemma 2]). *If parts (i) and (ii) of Assumption 9 hold, then for all $x \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

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

and

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

Lemma 44 (Lemma 6 in [TLR18], Uniform Deviation Guarantees). *Under (i) and (ii) of Assumption 9, there exists an absolute constant c_0 such that for:*

$$c := c_0(1 \vee \log((M \vee L \vee (B + MR))R\sigma_0/\delta)),$$
$$\sigma_0 := (A + (B + MR)R) \vee (B + MR) \vee (C + LR),$$

we have, if $n \geq cd\log d$, then with probability at least $1 - \delta$:

$$\sup_{x\in\mathbb{R}^d} \left|F(x) - \overline{F}(x)\right| \leq \sigma_0\sqrt{\frac{cd\log n}{n}},$$
$$\sup_{x\in\mathbb{R}^d} \left\|\nabla F(x) - \nabla\overline{F}(x)\right\| \leq \sigma_0\sqrt{\frac{cd\log n}{n}},$$
$$\sup_{x\in\mathbb{R}^d} \left\|\nabla^2 F(x) - \nabla^2\overline{F}(x)\right\| \leq \sigma_0\sqrt{\frac{cd\log n}{n}}.$$

Lemma 45 (Proposition 7 in [TLR18]). *If the population risk $\overline{F}(x)$ is $(2\varepsilon_0, 2m)$-strongly Morse, then provided that $n \geq cd\log d$ and $\frac{n}{d\log n} \geq \frac{c\sigma_0^2}{(\varepsilon_0 \wedge m)^2}$, the empirical risk $F(x)$ is (ε_0, m)-strongly Morse with probability at least $1 - \delta$.*