# Is There an Analog of Nesterov Acceleration for MCMC?

Yi-An Ma*[a], Niladri S. Chatterji†[b], Xiang Cheng‡[a], Nicolas Flammarion§[a], Peter L. Bartlett¶[a, c], and Michael I. Jordan‖[a, c]

[a]Department of Electrical Engineering and Computer Sciences
[b]Department of Physics
[c]Department of Statistics, University of California, Berkeley, CA 94720

October 23, 2019

**Abstract**

We formulate gradient-based Markov chain Monte Carlo (MCMC) sampling as optimization on the space of probability measures, with Kullback-Leibler (KL) divergence as the objective functional. We show that an underdamped form of the Langevin algorithm performs accelerated gradient descent in this metric. To characterize the convergence of the algorithm, we construct a Lyapunov functional and exploit hypocoercivity of the underdamped Langevin algorithm. As an application, we show that accelerated rates can be obtained for a class of nonconvex functions with the Langevin algorithm.

## 1 Introduction

While optimization methodology has provided much of the underlying algorithmic machinery that has driven the theory and practice of machine learning in recent years, sampling-based methodology, in particular Markov chain Monte Carlo (MCMC), remains of critical importance, given its role in linking algorithms to statistical inference and, in particular, its ability to provide notions of confidence that are lacking in optimization-based methodology. However, the classical theory of MCMC is largely asymptotic and the theory has not developed as rapidly in recent years as the theory of optimization.

Recently, however, a literature has emerged that derives nonasymptotic rates for MCMC algorithms [see, e.g., 9, 12, 10, 8, 6, 14, 27, 28, 2, 5]. This work has explicitly aimed at making use of ideas from optimization; in particular, whereas the classical literature on MCMC focused on reversible Markov chains, the recent literature has focused on non-reversible stochastic processes that are built on gradients [see, e.g., 24, 26, 3, 1]. In particular, the gradient-based Langevin algorithm [39, 38, 13] has been shown to be a form of gradient descent on the space of probabilities [see, e.g., 19, 44].

What has not yet emerged is an analog of acceleration. Recall that the notion of acceleration has played a key role in gradient-based optimization methods [32]. In particular, Nesterov's accelerated gradient descent (AGD) method, an instance of the general family of "momentum methods," provably achieves a faster convergence rate than gradient descent (GD) in a variety of settings [31]. Moreover, it achieves the optimal convergence rate under an oracle model of optimization complexity in the convex setting [30].

*yianma@berkeley.edu
†chatterji@berkeley.edu
‡x.cheng@berkeley.edu
§flammarion@berkeley.edu
¶peter@berkeley.edu
‖jordan@cs.berkeley.edu

This motivates us to ask: Is there an analog of Nesterov acceleration for gradient-based MCMC algorithms? And does it provably accelerate the convergence rate of these algorithms?

This paper answers these questions in the affirmative by showing that an underdamped form of the Langevin algorithm performs accelerated gradient descent. Critically, our work is based on the use of Kullback-Leibler (KL) divergence as the metric. We build on previous work that has studied the underdamped Langevin algorithm and has used coupling methods to establish convergence of the algorithm in the Wasserstein distance [see, e.g., 8, 7, 11]. Our work establishes a direct linkage between the underdamped Langevin algorithm and Nesterov acceleration by working directly in the objective functional, the KL divergence. Combining ideas from optimization theory and diffusion processes, we construct a Lyapunov functional that couples the convergence in the momentum and the original variables. We then prove the overall convergence rate by leveraging the hypocoercivity structure of the underdamped Langevin algorithm [42]. For target distributions satisfying a log-Sobolev inequality, we find that the underdamped Langevin algorithm accelerates the convergence rate of the classical Langevin algorithm from $d/\epsilon$ to $\sqrt{d/\epsilon}$ in terms of KL divergence (See Theorem 1 for formal statement).

# 2 Preliminaries

We start by laying out the problem setting, including our assumptions on the target distribution that we sample from, properties of the KL divergence with respect to other measure of differences between probability distributions, and the notion of gradient on the space of probabilities.

## 2.1 Problem setting

Assume that we wish to sample from a target (posterior) probability density, $\mathbf{p}^*(\theta)$, where $\theta \in \mathbb{R}^d$. Consider the KL divergence to this target:

$$\mathrm{KL}\left(\mathbf{p}\|\mathbf{p}^*\right) = \int \mathbf{p}(\theta) \ln\left(\frac{\mathbf{p}(\theta)}{\mathbf{p}^*(\theta)}\right) \mathrm{d}\theta.$$

We use this KL divergence as an objective functional in an optimization-theoretic formulation of convergence to $\mathbf{p}^*(\theta)$.

We assume that $\mathbf{p}^*$ satisfies the following conditions.

**A1** The target density $\mathbf{p}^*$ satisfies a log-Sobolev inequality with constant $\rho$ [18, 34]. That is, for any smooth function $g : \mathbb{R}^d \to \mathbb{R}$, we have

$$\int g(\theta) \ln g(\theta) \cdot \mathbf{p}^*(\theta)\mathrm{d}\theta - \int g(\theta)\ \mathbf{p}^*(\theta)\mathrm{d}\theta \cdot \ln\left(\int g(\theta)\ \mathbf{p}^*(\theta)\mathrm{d}\theta\right) \le \frac{1}{2\rho} \int \frac{\left\|\nabla g(\theta)\right\|^2}{g(\theta)} \mathbf{p}^*(\theta)\mathrm{d}\theta.$$

**A2** For $\mathbf{p}^* \propto e^{-U}$, the potential function $U$ is $L_G$-gradient Lipschitz and is $L_H$-Hessian Lipschitz; that is, for $U \in C^2(\mathbb{R}^d)$ and for all $\theta, \vartheta \in \mathbb{R}^d$:[1]

$$\left\|\nabla U(\theta) - \nabla U(\vartheta)\right\| \le L_G \left\|\theta - \vartheta\right\|;$$
$$\left\|\nabla^2 U(\theta) - \nabla^2 U(\vartheta)\right\|_F \le L_H \left\|\theta - \vartheta\right\|.$$

[1] It is worth noting that this definition of Hessian Lipschitzness with respect to the Frobenius norm is stronger than that with respect to the spectral norm. We postulate here that the requirement of a Hessian Lipschitz condition is an artifact of our particular choice of Lyapunov functional $\mathcal{L}$ and can possibly be removed in future work.

**A3** Without loss of generality, for $\mathbf{p}^*(\theta) \propto e^{-U(\theta)}$, let $\nabla U(0) = 0$ and $U(0) = 0$ (which can be achieved by shifting the potential function $U$). Further assume that the normalization constant for $e^{-U(\theta)}$ is bounded and scales at most exponentially with dimension $d$: $\ln\left(\int \exp(-U(\theta))\mathrm{d}\theta\right) \leq C_N \cdot d + C_M$.

As a concrete example, these assumptions are satisfied in the "locally nonconvex" case studied by [25], with nonconvex region of radius $R$ and strong convexity $m$; see also Assumption (a)–(c) in Appendix A. Note that [25] instantiates both the log-Sobolev constant $\rho$ and the normalization constants $C_N$ in terms of the smoothness and conditioning of $U$, showing that $\rho \geq \frac{m}{2} e^{-16 L_G R^2}$. Here we additionally establish (see Fact 1) that $C_N \leq \frac{1}{2} \ln \frac{4\boldsymbol{\pi}}{m}$, and $C_M \leq 32 \frac{L_G^2}{m^2} L_G R^2$.

## 2.2 KL divergence and relation to other metrics

Our convergence result is expressed in terms of the KL Divergence. In this section, we recall that $\mathrm{KL}\left(\mathbf{p} \| \mathbf{p}^*\right)$ upper bounds a number of other metrics of interest.

1. By Pinsker's inequality, we can upper bound the total variation distance by the KL divergence:

$$\mathrm{TV}\left(\mathbf{p}, \mathbf{p}^*\right) \leq \sqrt{2\mathrm{KL}\left(\mathbf{p} \| \mathbf{p}^*\right)}.$$

2. Since $\mathbf{p}^*$ satisfies the log-Sobolev inequality (**A1**) with constant $\rho$ and has a Lipschitz smoothness property, by the Talagrand inequality (Theorem 1 of [34]), we can upper bound the Wasserstein-2 distance (defined in Eq. (2)) by the KL divergence:

$$W_2(\mathbf{p}, \mathbf{p}^*) \leq \sqrt{\frac{2\mathrm{KL}\left(\mathbf{p} \| \mathbf{p}^*\right)}{\rho}}. \tag{1}$$

## 2.3 Gradients on the space of probabilities

Given an iterative algorithm that generates a random vector $\theta^{(k)}$ at each step $k$, we are interested in the convergence of the law of $(\theta^{(k)}, \boldsymbol{\pi}^{(k)})$ to the measure $\boldsymbol{\pi}^*$ associated with the target density $\mathbf{p}^*$. In this paper, we consider the space of probability measures that are absolutely continuous with respect to the Lebesgue measure (have density functions) and have finite second moments, $\mathcal{P}_2(\mathbb{R}^d)$. It will become clear later in the paper (in Theorem 1) that when the target density $\mathbf{p}^*$ satisfies Assumptions **A1**–**A3**, the measure $\boldsymbol{\pi}^{(k)}$ belongs to $\mathcal{P}_2$, for any $k > 0$. For this reason, we can always analyze behaviors of the distributions in terms of their density functions.

In order to define a notion of "gradient" for accelerated gradient descent on the space of probabilities, $\mathcal{P}_2(\mathbb{R}^d)$, we first need to equip $\mathcal{P}_2(\mathbb{R}^d)$ with a metric. To this end, we use the Wasserstein-2 distance, defined in terms of couplings as follows [43]. For a pair of distributions $\mathbf{p}$ and $\mathbf{q}$ on $\mathbb{R}^d$, a coupling $\boldsymbol{\gamma}$ is a joint measure over the product space $\mathbb{R}^d \times \mathbb{R}^d$ that has $\mathbf{p}$ and $\mathbf{q}$ as its two marginal densities. We let $\Gamma(\mathbf{p}, \mathbf{q})$ denote the space of all possible couplings of $\mathbf{p}$ and $\mathbf{q}$. With this notation, the Wasserstein-2 distance is given by

$$W_2^2(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \inf_{\boldsymbol{\gamma} \in \Gamma(\mathbf{p}, \mathbf{q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \vartheta\|_2^2 \, \mathrm{d}\boldsymbol{\gamma}(\theta, \vartheta), \tag{2}$$

where the set of $\boldsymbol{\gamma}$ that attains the infimum above is denoted $\Gamma_{\mathrm{opt}}$.

On the space of $\mathcal{P}_2(\mathbb{R}^d)$ with Wasserstein-2 metric, there is also an optimal transport picture of the coupling. Namely, for the measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ corresponding to the densities $\mathbf{p}$ and $\mathbf{q}$, there exists a transport map $\mathrm{t} : \mathbb{R}^d \to \mathbb{R}^d$, so that $(\mathrm{t} \times \mathrm{id})_{\#}\boldsymbol{\nu} \in \Gamma_{\mathrm{opt}}(\mathbf{p}, \mathbf{q})$, where the push-forward operator $\#$ is defined as $\mathrm{t}_{\#}\boldsymbol{\nu}(\theta) = \boldsymbol{\nu}(\mathrm{t}(\theta))$. With this notion, we can make use of the underlying $L^2$ Hilbert space to define strong subdifferentials.

Letting $\mathcal{L} : \mathcal{P}_2 \to \mathbb{R}$ be a proper functional, define $\xi \in \partial\mathcal{L}$ as the strong subdifferential of $\mathcal{L}$ (taken at density $\mathbf{p}$ associated with measure $\boldsymbol{\mu}$) if, for any transport map t, we have:

$$\mathcal{L}(\mathrm{t}_{\#}\boldsymbol{\mu}) - \mathcal{L}(\boldsymbol{\mu}) \geq \int_{\mathbb{R}^d} \langle \xi(\theta), \mathrm{t}(\theta) - \theta \rangle \, \mathrm{d}\boldsymbol{\mu}(\theta) + o\left( \int_{\mathbb{R}^d} \|\mathrm{t}(\theta) - \theta\|_2 \, \mathrm{d}\boldsymbol{\mu}(\theta) \right).$$

See [23, Definition 10.1.1] for more details. This strong subdifferential provides us the proper notion of "gradient." In particular, for functionals with enough regularity, the strong subdifferential of $\mathcal{L}$ taken at $\mathbf{p}$ can be expressed as $\nabla_\theta \frac{\delta \mathcal{L}}{\delta \mathbf{p}}$, where $\frac{\delta}{\delta \mathbf{p}}$ is the functional derivative taken at $\mathbf{p}$ and $\nabla_\theta$ is the ordinary gradient operator in the space of $\theta$ [23, Lemma 10.4.1].

# 3 Underdamped Langevin Algorithm as Accelerated Gradient Descent

A recent trend in optimization theory involves casting the analysis of algorithms into a continuous dynamical systems framework [41, 45, 47, 40]. This approach involves two steps: (1) a continuous-time system is specified and a convergence rate is obtained for the continuous dynamics; (2) the continuous dynamics is discretized, yielding a discrete-time algorithm, and the discretization error is analyzed, yielding an overall convergence rate. Our work follows in this vein. We first study a continuous-time stochastic dynamical system that can be interpreted as an accelerated gradient flow with respect to the KL divergence $\mathrm{KL}(\mathbf{p}_t \| \mathbf{p}^*)$. We then derive the underdamped Langevin algorithm as a discretization of the accelerated gradient flow. We show that this discretization is precisely accelerated gradient descent with respect to $\mathrm{KL}(\mathbf{p}_t \| \mathbf{p}^*)$.

## 3.1 Gradient descent dynamics with respect to KL divergence

We start by defining the dynamics of gradient descent via a consideration of the gradient flow associated with the KL divergence $\mathrm{KL}(\mathbf{p}_t \| \mathbf{p}^*)$. We first formulate the "vector flow" associated with the following stochastic differential equation with Lipschitz continuous drift $b : \mathbb{R}^d \to \mathbb{R}^d$:

$$\mathrm{d}\theta_t = b(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{3}$$

where $B_t$ is a standard Brownian motion. The evolution of the probability density function $\mathbf{p}_t$ of the random variable $\theta_t$ follows the transport of probability mass along a vector flow $v_t$ in the state space:

$$\frac{\partial}{\partial t}\mathbf{p}_t(\theta) + \nabla^{\mathrm{T}} \left( \mathbf{p}_t(\theta) v_t(\theta) \right) = 0, \tag{4}$$

where the vector flow can be calculated as: $v_t(\theta) = b(\theta) - \nabla \ln \mathbf{p}_t(\theta)$. This can be compared with the following Liouville equation:

$$\frac{\partial}{\partial t}\bar{\mathbf{p}}_t(\theta) + \nabla^{\mathrm{T}} \left( \bar{\mathbf{p}}_t(\theta) b(\theta) \right) = 0,$$

which describes the evolution of the probability along a deterministic vector field, $\frac{\mathrm{d}}{\mathrm{d}t}\bar{\theta}_t = b(\bar{\theta}_t)$.

On the other hand, we formulate the "gradient" of the KL divergence corresponding to the vector flow point of view. For the objective functional $\mathcal{F}[\mathbf{p}_t]$, its time change when $\theta_t$ follows Eq. (3) is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\mathbf{p}_t] = \mathbb{E}_{\theta \sim \mathbf{p}_t} \left[ \left\langle \nabla \frac{\delta \mathcal{F}[\mathbf{p}_t]}{\delta \mathbf{p}_t}(\theta), b(\theta) - \nabla \ln \mathbf{p}_t \right\rangle \right],$$

where $\nabla \frac{\delta \mathcal{F}[\mathbf{p}_t]}{\delta \mathbf{p}_t}(\theta)$ is the strong subdifferential of $\mathcal{F}[\mathbf{p}_t]$ associated with the 2-Wasserstein metric (See Sec. 2.3). Therefore, we can consider the gradient-descent dynamics with respect to the functional $\mathcal{F}[\mathbf{p}_t]$ as taking the vector flow $v_t$ in Eq. (4) as $v_t(\theta) = -\nabla \frac{\delta \mathcal{F}[\mathbf{p}_t]}{\delta \mathbf{p}_t}(\theta)$. When the functional is the KL divergence, $\mathcal{F}[\mathbf{p}_t] = \mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right)$, the gradient descent flow $v_t^{GD}$ involves taking

$$v_t^{GD}(\theta) = -\nabla \frac{\delta \mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right)}{\delta \mathbf{p}_t}(\theta) = -\nabla \ln \frac{\mathbf{p}_t(\theta)}{\mathbf{p}^*(\theta)},$$

or, equivalently, $b^{GD}(\theta) = -\nabla U(\theta)$ in Eq. (3).

Along this gradient descent flow, $v_t^{GD}$, the time evolution of the KL divergence is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right) = -\mathbb{E}_{\theta \sim \mathbf{p}_t}\left[\left\|\nabla \frac{\delta \mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right)}{\delta \mathbf{p}_t}(\theta)\right\|^2\right] = -\mathbb{E}_{\theta \sim \mathbf{p}_t}\left[\left\|\nabla \ln \frac{\mathbf{p}_t(\theta)}{\mathbf{p}^*(\theta)}\right\|^2\right].$$

If $\mathbf{p}^*(\theta)$ satisfies Assumption **A1** then taking $g = \frac{\mathbf{p}_t}{\mathbf{p}^*}$ in the log-Sobolev inequality yields:

$$\mathbb{E}_{\theta \sim \mathbf{p}_t}\left[\ln\left(\frac{\mathbf{p}_t(\theta)}{\mathbf{p}^*(\theta)}\right)\right] \leq \frac{1}{2\rho}\mathbb{E}_{\theta \sim \mathbf{p}_t}\left[\left\|\nabla \ln\left(\frac{\mathbf{p}_t(\theta)}{\mathbf{p}^*(\theta)}\right)\right\|^2\right]. \tag{5}$$

Note the resemblance of this bound to the Polyak-Łojasiewicz condition [37] used in optimization theory for studying the convergence of gradient methods—in both cases the difference in objective value from the current iterate to the optimum is upper bounded by the squared norm of the gradient of the objective. With the log-Sobolev inequality, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right) = -\mathbb{E}_{\theta \sim \mathbf{p}_t}\left[\left\|\nabla \ln \frac{\mathbf{p}_t(\theta)}{\mathbf{p}^*(\theta)}\right\|^2\right] \leq -2\rho \mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right),$$

which implies the linear convergence of $\mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right)$ along the gradient descent flow.

## 3.2 Accelerated gradient descent in KL divergence: A continuous perspective

We now introduce an accelerated dynamics in the space of probabilities via the incorporation of a momentum variable $r \in \mathbb{R}^d$. Denote $x = (\theta, r)$ and let the joint target distribution be $\mathbf{p}^*(x) = \mathbf{p}^*(\theta)\mathbf{p}^*(r) = \exp\left(-U(\theta) - \frac{\xi}{2}||r||_2^2\right)$.[2] To design the accelerated gradient descent dynamics with respect to the KL divergence, we leverage the acceleration phenomenon in optimization, which uses the gradient of the expanded objective function to guide the algorithm (see the discussion in Sec. 3.2.2). We expand the KL divergence (in both the $\theta$ and $r$ coordinates) to obtain:

$$\begin{aligned}\mathrm{KL}\left(\mathbf{p}_t(\theta, r) \| \mathbf{p}^*(\theta)\mathbf{p}^*(r)\right) &= \int\int \mathbf{p}_t(\theta, r) \ln \frac{\mathbf{p}_t(\theta, r)}{\mathbf{p}^*(\theta)\mathbf{p}^*(r)}\mathrm{d}\theta \mathrm{d}r \\ &= \mathrm{KL}\left(\mathbf{p}_t(\theta) \| \mathbf{p}^*(\theta)\right) + \mathbb{E}_{\theta \sim \mathbf{p}_t(\theta)}\left[\mathrm{KL}\left(\mathbf{p}_t(r|\theta) \| \mathbf{p}^*(r)\right)\right],\end{aligned}$$

and form the vector field:

$$v_t^{AGD}(x) = -\begin{pmatrix} 0 & -\mathrm{I} \\ \mathrm{I} & \gamma \mathrm{I} \end{pmatrix}\begin{pmatrix} \nabla_\theta \frac{\delta \mathrm{KL}(\mathbf{p}_t \| \mathbf{p}^*)}{\delta \mathbf{p}_t} \\ \nabla_r \frac{\delta \mathrm{KL}(\mathbf{p}_t \| \mathbf{p}^*)}{\delta \mathbf{p}_t} \end{pmatrix} \tag{6}$$

$$= \begin{pmatrix} \nabla_r \ln \mathbf{p}_t(\theta, r) + \xi r \\ -\nabla_\theta \ln \mathbf{p}_t(\theta, r) - \nabla U(\theta) - \gamma \nabla_r \ln \frac{\mathbf{p}_t(\theta, r)}{\mathbf{p}^*(r)} \end{pmatrix}. \tag{7}$$

[2] We will use $\mathbf{p}^*(\theta)$ and $\mathbf{p}_t(\theta)$ to denote marginal distributions of $\mathbf{p}^*(\theta, r)$ and $\mathbf{p}_t(\theta, r)$, respectively, after integration over $r$.

The corresponding continuity equation defined by this vector field is

$$\begin{aligned}0 &= \frac{\partial}{\partial t}\mathbf{p}_t(\theta,r) + \nabla^{\mathrm{T}}\left(\mathbf{p}_t(\theta,r)v_t^{AGD}(\theta,r)\right)\\ &= \frac{\partial}{\partial t}\mathbf{p}_t(\theta,r) + \left(\nabla_\theta^{\mathrm{T}},\nabla_r^{\mathrm{T}}\right)\left[\mathbf{p}_t(\theta,r)\begin{pmatrix}\xi r\\ -\nabla U(\theta) - \gamma\nabla_r \ln\frac{\mathbf{p}_t(\theta,r)}{\mathbf{p}^*(r)}\end{pmatrix}\right].\end{aligned}$$

This implies that the vector field can be implemented via the following stochastic differential equation

$$\begin{cases} d\theta_t = \xi r_t \mathrm{d}t\\ dr_t = -\nabla U(\theta_t)\mathrm{d}t - \gamma\xi r_t\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}B_t,\end{cases} \tag{8}$$

which is the underdamped Langevin dynamics [20].

### 3.2.1 Convergence of the accelerated gradient-descent dynamics

If we consider the time derivative of the KL divergence, we have: $\mathrm{KL}\left(\mathbf{p}_t\|\mathbf{p}^*\right)$,

$$\begin{aligned}\frac{d}{\mathrm{d}t}\mathrm{KL}\left(\mathbf{p}_t\|\mathbf{p}^*\right) &= \int\left\langle \nabla_x\frac{\delta\mathrm{KL}\left(\mathbf{p}_t\|\mathbf{p}^*\right)}{\delta\mathbf{p}_t}, v_t^{AGD}(\theta,r)\right\rangle \mathbf{p}_t\ \mathrm{d}x\\ &= \int\left\langle \nabla_x\frac{\delta\mathrm{KL}\left(\mathbf{p}_t\|\mathbf{p}^*\right)}{\delta\mathbf{p}_t}, -\begin{pmatrix}0 & -\mathrm{I}\\ \mathrm{I} & \gamma\mathrm{I}\end{pmatrix}\nabla_x\ln\frac{\mathbf{p}_t}{\mathbf{p}^*}\right\rangle \mathbf{p}_t\ \mathrm{d}x\\ &= -\gamma\mathbb{E}_{\mathbf{p}_t}\left[\left\|\nabla_r\ln\frac{\mathbf{p}_t}{\mathbf{p}^*}\right\|^2\right].\end{aligned} \tag{9}$$

This only demonstrates the contractive property in the $r$ coordinates (note that the gradient is only in $r$ in Line (9)) and does not directly provide a linear convergence rate over time. To quantify the convergence rate for this accelerated gradient descent dynamics with respect to the KL divergence objective, we need to couple the convergence in $\theta$ coordinates to that in $r$. To this end, we follow recent work in the optimization literature [45] and design a Lyapunov functional which makes use of a quadratic form of the gradient of the distance $\mathcal{D}$ between the current iteration $\mathbf{p}_t$ and the stationary solution $\mathbf{p}^*$:

$$\begin{aligned}\mathcal{L}[\mathbf{p}_t] &= \mathrm{KL}\left(\mathbf{p}_t\|\mathbf{p}^*\right) + \mathbb{E}_{\mathbf{p}_t}\left[\left\langle\nabla_x\frac{\delta\mathcal{D}[\mathbf{p}_t,\mathbf{p}^*]}{\delta\mathbf{p}_t}, S\nabla_x\frac{\delta\mathcal{D}[\mathbf{p}_t,\mathbf{p}^*]}{\delta\mathbf{p}_t}\right\rangle\right]\\ &= \mathbb{E}_{\mathbf{p}_t}\left[\ln\frac{\mathbf{p}_t}{\mathbf{p}^*} + \left\langle\nabla_x\ln\frac{\mathbf{p}_t}{\mathbf{p}^*}, S\nabla_x\ln\frac{\mathbf{p}_t}{\mathbf{p}^*}\right\rangle\right],\end{aligned} \tag{10}$$

where we take the distance measure between $\mathbf{p}_t$ and $\mathbf{p}^*$ as the KL divergence itself: $\mathcal{D}[\mathbf{p}_t,\mathbf{p}^*] = \mathrm{KL}\left(\mathbf{p}_t\|\mathbf{p}^*\right) = \mathbb{E}_{\mathbf{p}_t}\left[\ln\frac{\mathbf{p}_t}{\mathbf{p}^*}\right]$. Here we set the positive definite matrix in the quadratic form to be

$$S = \frac{1}{L_G}\begin{pmatrix}1/4\ \mathrm{I}_{d\times d} & 1/2\ \mathrm{I}_{d\times d}\\ 1/2\ \mathrm{I}_{d\times d} & 2\ \mathrm{I}_{d\times d}\end{pmatrix}. \tag{11}$$

Interestingly, similar forms appear in the analyses of both accelerated gradient descent dynamics [31, 45] and hypocoercive diffusion operators [42, 4].

We then make use of this Lyapunov functional to obtain a linear convergence rate for the accelerated gradient descent dynamics with respect to the KL divergence.

**Proposition 1.** *Under Assumptions **A1**–**A3**, the time evolution of the Lyapunov functional $\mathcal{L}$ with respect to the continuous time vector flow $v_t^{AGD}$ in Eq. (7) with $\gamma = 2$ and $\xi = 2L_G$ is upper bounded as:*

$$\frac{d}{\mathrm{d}t}\mathcal{L}[\mathbf{p}_t] \leq -\frac{\rho}{10}\mathcal{L}[\mathbf{p}_t].$$

This establishes linear convergence of the continuous process with a rate of $\frac{\rho}{10}$.

### 3.2.2 Accelerated gradient descent dynamics for optimization

It is worth noting that the derivation in the previous subsection has a close correspondence to recent analyses of the accelerated gradient descent dynamics in convex optimization [41, 45]. Indeed, when optimizing a strongly convex function $U(\theta)$ on a Euclidean space with the accelerated gradient descent dynamics, the continuous limit of the algorithm is expressed as an ordinary differential equation [45]:

$$\frac{\mathrm{d}^2\theta_t}{\mathrm{d}t^2} + \gamma\xi\frac{\mathrm{d}\theta_t}{\mathrm{d}t} + \xi\nabla U(\theta_t) = 0.$$

We can expand the space of interest via introducing a "momentum" variable, $r_t = \frac{1}{\xi}\frac{\mathrm{d}\theta_t}{\mathrm{d}t}$, to obtain a vector field point of view on the joint space of $x_t = (\theta_t, r_t)$:

$$\begin{cases} \frac{\mathrm{d}\theta_t}{\mathrm{d}t} = \xi r_t \\ \frac{\mathrm{d}r_t}{\mathrm{d}t} = -\nabla U(\theta_t) - \gamma\xi r_t. \end{cases}$$

We also extend the original objective function $U(\theta)$ to $H(x) = U(\theta) + \frac{\xi}{2}\|r\|_2^2$ to capture the overall dynamical behavior in the space of $x$. With the definition of this extended objective function $H$, we can simplify the expression of the dynamics:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\begin{pmatrix} 0 & -\mathrm{I} \\ \mathrm{I} & \gamma\mathrm{I} \end{pmatrix}\begin{pmatrix} \nabla_\theta H(x) \\ \nabla_r H(x) \end{pmatrix}. \tag{12}$$

To quantify convergence for the strongly convex objective $U$, [45] considers a Lyapunov function of the form $l(x) = H(\theta) + \left\langle \nabla_x^T D_h(x), S\nabla_x D_h(x) \right\rangle$, where $D_h(x) = \frac{1}{2}\left\|\theta - \theta^*\right\|^2 + \frac{1}{2}\left\|r\right\|^2$ is the squared distance from $(\theta, r)$ to the optimum of $H$, $(\theta^*, 0)$.

Comparing the dynamics of Eq. (12) versus Eq. (6) and the convergence analyses for them, we observe that the underdamped Langevin diffusion defined in Eq. (8) is precisely accelerated gradient descent with respect to the KL divergence.

## 3.3 Underdamped Langevin via second-order discretization

While the continuous-time perspective yields insight into the convergence rates achievable by acceleration, for these insights to apply to discrete-time algorithms it is necessary to understand the effects of discretization. In optimization, an emerging literature has begun to show how to design discretization procedures that retain accelerated rates from continuous time [45, 47, 40]. The literature in MCMC has not yet formalized lower bounds on convergence rates that allow characterizations of acceleration, in either continuous time or discrete time, but there are results that exhibit the importance of discretization for convergence. In particular, higher order (and more accurate) discretization schemes are found to accelerate convergence [29, 21, 8, 11, 27, 28].

In this section we show how to design a discretization for the an underdamped Langevin algorithm that yields accelerated rates. Following [8], we discretize the time dimension underlying Eq. (8) into intervals of equal length $h$ (at the end of the $k$-th iteration, we have $t = kh$). Then in the $(k+1)$-th step, we define a continuous dynamics in the interval of $\tau \in [kh, (k+1)h]$ by conditioning on the initial value of $x_{kh}$:

$$\begin{cases} \mathrm{d}\theta_\tau = \xi r_\tau \mathrm{d}\tau \\ \mathrm{d}r_\tau = -\gamma\xi r_\tau \mathrm{d}\tau - \nabla U(\theta_{kh})\mathrm{d}\tau + \sqrt{2\gamma}\mathrm{d}B_\tau. \end{cases} \tag{13}$$

In Appendix B we derive explicit formulas for $x_\tau$ given $x_{kh}$. These are used to generate the $(k+1)$-th iterate. In particular, define the hyperparameters $\gamma = 2$, $\xi = 2L_G$, and set the step size as follows:

$$h = \frac{1}{56}\frac{1}{\sqrt{L_G}}\min\left\{\frac{1}{24}\frac{\rho}{L_G}, \frac{\sqrt{L_G}\rho}{L_H}\right\} \cdot \min\left\{\left(\widetilde{C_N} + 2\right)^{-1/2}\sqrt{\frac{\epsilon}{d}}, \sqrt{\frac{\epsilon}{C_M}}\right\}, \tag{14}$$

**Algorithm 1:** Underdamped Langevin Algorithm

```
Let $x_0 = (\theta_0, r_0)$, where $\theta_0, r_0 \sim \mathcal{N}\left(0, \frac{1}{L_G}\mathrm{I}\right)$.
for $k = 0, \cdots, K-1$ do
    Sample $x_{(k+1)h} \sim \mathcal{N}\left(\mu\left(x_{kh}\right), \Sigma\right)$, where $\mu\left(x_{kh}\right)$ and $\Sigma$ are defined in Eq. (35) and (36).
end for
```

where $\widetilde{C_N} = C_N + \frac{1}{2}\ln\frac{L_G}{2\pi}$. The discretized vector field is

$$\hat{v}_\tau^{AGD} = \left(\begin{array}{l} \xi r_\tau \\ -\nabla U(\theta_{kh}) - \gamma \nabla_r \ln \frac{\mathbf{p}(\theta_\tau, r_\tau)}{\mathbf{p}^*(r_\tau)} \end{array}\right) = \left(\begin{array}{l} \xi r_\tau \\ -\nabla U(\theta_{kh}) - \gamma\xi r_\tau - \gamma \nabla_r \ln \mathbf{p}(\theta_\tau, r_\tau) \end{array}\right). \tag{15}$$

This leads to a high-order discretization scheme that is defined explicitly in Appendix B and summarized in Algorithm 1.

By way of comparison, the Euler-Maruyama discretization scheme corresponds to:

$$\hat{v}_\tau^{E-M} = \left(\begin{array}{l} \xi r_{kh} \\ -\nabla U(\theta_{kh}) - \gamma\xi r_{kh} - \gamma \nabla_r \ln \mathbf{p}(\theta_\tau, r_\tau) \end{array}\right).$$

After integration, we obtain that for $\tau \in [kh, (k+1)h]$:

$$\left\{\begin{array}{l} \theta_\tau = \theta_{kh} + (\tau - kh)\xi r_{kh} \\ r_\tau = \left(1 - (\tau - kh)\gamma\xi\right) r_{kh} - (\tau - kh)\nabla U(\theta_{kh}) + \sqrt{2\gamma} B_{\tau - kh}, \end{array}\right.$$

where the Brownian motion is defined as $B_{\tau-kh} \sim \mathcal{N}\left(0, (\tau - kh)\mathrm{I}_{d\times d}\right)$. This low-order integration scheme does not grant accelerated convergence guarantees.

There are other higher-order discretization schemes that can be considered in addition to our scheme in Eq. (15). In particular, note that $v_t^{AGD}$ decomposes into two parts:

$$v_t^{AGD} = \left(\begin{array}{l} \xi r_t \\ -\nabla U(\theta_t) \end{array}\right) + \left(\begin{array}{l} 0 \\ -\gamma \nabla_r \ln \frac{\mathbf{p}(\theta_t, r_t)}{\mathbf{p}^*(r_t)} \end{array}\right),$$

where each part preserves $p^*$ as the invariant distribution. This inspires a splitting scheme for integrating $v_t^{AGD}$. The first part is a Hamiltonian vector flow, which can be integrated via symplectic integration schemes such as the leapfrog method. The second part can be explicitly integrated to yield $r_{\tau-kh} \sim \mathcal{N}\left(e^{-\gamma\xi(\tau-kh)} r_{kh}, \frac{1}{\xi}\left(1 - e^{-2\gamma\xi(\tau-kh)}\right)\mathrm{I}\right)$.

Taking $(\tau - kh) \to \infty$, $r$ is resampled as: $r \sim \mathcal{N}\left(0, \frac{1}{\xi}\mathrm{I}\right)$ according to the stationary distribution $\mathbf{p}^*(r)$. This recovers the Hamiltonian Monte Carlo (HMC) method [29]. Relating to concepts in optimization, this "momentum resampling" step corresponds to a "momentum restart" method in optimization: one periodically restarts the momentum from the stationary point [33]. In optimization this has a theoretical justification in terms of increasing convergence rate; for HMC it has been observed empirically that *not* taking $(\tau - kh) \to \infty$ at every step increases mixing [35].

# 4 Convergence of the Underdamped Langevin Algorithm

From Fig. 1, we see that the underdamped Langevin algorithm, Eq. (34), seems to have a similar profile to accelerated gradient descent; it uses oscillatory behavior to increase the convergence rate. In this section, we

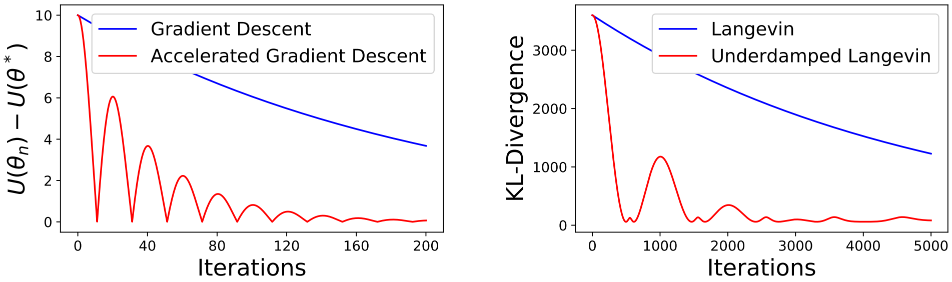

Figure 1: Acceleration phenomenon in optimization and sampling. Left: The (accelerated) gradient descent algorithms minimize the objective function value $|U(\theta_t) - U(\theta^*)|$. Right: The (underdamped) Langevin algorithms minimize the KL divergence $\mathrm{KL}\left(\mathbf{p}_t(\theta)\|\mathbf{p}^*(\theta)\right)$, where $\mathbf{p}^*(\theta) \propto e^{-U(\theta)}$. In both cases, $U$ is a quadratic function in 100 dimensions with condition number $L/m = 100$.

rigorously establish acceleration, by proving that the convergence of the underdamped Langevin algorithm is of order $\mathcal{O}\left(\sqrt{d/\epsilon}\right)$ in terms of KL divergence.

Let the KL divergence from $\mathbf{p}_t(\theta)$ to $\mathbf{p}^*(\theta)$ be the target functional to minimize:

$$\mathrm{KL}\left(\mathbf{p}_t(\theta)\|\mathbf{p}^*(\theta)\right) \leq \mathrm{KL}\left(\mathbf{p}_t(\theta, r)\|\mathbf{p}^*(\theta)\mathbf{p}^*(r)\right).$$

We have the following theorem.

**Theorem 1.** *Assume $\mathbf{p}^*(\theta) \propto e^{-U(\theta)}$ satisfies Assumptions **A1**–**A3**. We use $\rho$ to denote the minimum of the log-Sobolev constant and* 1*. Then if we iterate the underdamped Langevin algorithm* (34) *with initial condition $\theta_0 \sim \mathcal{N}\left(0, \frac{1}{L_G}\mathrm{I}\right)$ for*

$$k \geq \mathcal{O}\left(\sqrt{\frac{d}{\epsilon}}\ln\left(\frac{d}{\epsilon}\right)\right)$$

*steps, we have* $\mathrm{KL}\left(\mathbf{p}_{kh}(\theta)\|\mathbf{p}^*(\theta)\right) \leq \mathrm{KL}\left(\mathbf{p}_{kh}(\theta, r)\|\mathbf{p}^*(\theta)\mathbf{p}^*(r)\right) < \epsilon$, $\forall \epsilon \leq 2d$.

*If we further assume that the function $U$ is locally nonconvex with radius $R$ and has global strong convexity $m$ (Assumption (a)–(c)), we obtain an explicit dependence of the convergence time $K$ on other constants:*

$$K = \mathcal{O}\left(\max\left\{\frac{L_G^{3/2}}{\rho^2}, \frac{L_H}{\rho^2}\right\}\sqrt{\frac{d}{\epsilon}}\ln\frac{d}{\epsilon}\right),$$

*where* $\rho = \min\left\{\frac{m}{2}e^{-16L_G R^2}, 1\right\}$.

We devote the remainder of Section 4 to the proof of Theorem 1. As advertised, the proof decomposes into a continuous-time analysis and a discretization analysis. We first establish the convergence rate of the continuous underdamped Langevin dynamics in Proposition 1 to quantify the instantaneous contraction provided by the dynamics. We then study the discretization error of the underdamped Langevin algorithm in each step. Combining these two results and integrating over the time steps leads us to the final conclusion.

We begin by formulating the instantaneous change of the probability density $\mathbf{p}(x_\tau)$ within each step of the underdamped Langevin algorithm. The time evolution of $\mathbf{p}(x_\tau|x_{kh})$ following the discretized vector flow

$\hat{v}_\tau^{AGD}$ for $\tau \in [kh, (k+1)h]$ is as follows:

$$\begin{aligned}\frac{\partial \mathbf{p}(x_\tau|x_{kh})}{\partial \tau} &= -\nabla_x^{\mathrm{T}}\left(\mathbf{p}(x_\tau|x_{kh}) \cdot \hat{v}_\tau^{AGD}\right) \\ &= -\nabla_x^{\mathrm{T}}\left(\mathbf{p}(x_\tau|x_{kh}) \cdot v_\tau^{AGD}\right) - \nabla_x^{\mathrm{T}}\left(\mathbf{p}(x_\tau|x_{kh}) \cdot (\hat{v}_\tau^{AGD} - v_\tau^{AGD})\right).\end{aligned}$$

Therefore, for the unconditioned probability density $\mathbf{p}(x_\tau) = \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\mathbf{p}(x_\tau|x_{kh})\right]$,

$$\frac{\partial \mathbf{p}(x_\tau)}{\partial \tau} = -\nabla_x^{\mathrm{T}}\left(\mathbf{p}(x_\tau) \cdot v_\tau^{AGD}\right) - \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\nabla_x^{\mathrm{T}}\left((\hat{v}_\tau^{AGD} - v_\tau^{AGD})\mathbf{p}(x_\tau|x_{kh})\right)\right]. \tag{16}$$

We have thus separated the time evolution of $\mathbf{p}(x_\tau)$ into two parts: the continuous component and the discretization error component.

Recall the Lyapunov functional, $\mathcal{L}(\mathbf{p}_t) = \mathbb{E}_{\mathbf{p}_t}\left[\ln\frac{\mathbf{p}_t}{\mathbf{p}^*} + \left\langle \nabla_x \ln\frac{\mathbf{p}_t}{\mathbf{p}^*}, S\nabla_x \ln\frac{\mathbf{p}_t}{\mathbf{p}^*}\right\rangle\right]$, that we defined in Sec. 3.2). We use this Lyapunov functional to analyze the convergence of the underdamped Langevin algorithm. Note that the instantaneous change of the Lyapunov functional $\mathcal{L}$ follows the overall vector flow $\hat{v}_t^{AGD}$, and derives from the continuous vector flow $v_t^{AGD}$ and the discretization error $\hat{v}_t^{AGD} - v_t^{AGD}$:

$$\begin{aligned}\frac{d}{dt}\mathcal{L}[\mathbf{p}(x_\tau)] &= \int \frac{\delta\mathcal{L}}{\delta\mathbf{p}(x_\tau)}\frac{\partial\mathbf{p}(x_\tau)}{\partial t}\,\mathrm{d}x_\tau = \int\left\langle \nabla_x\frac{\delta\mathcal{L}}{\delta\mathbf{p}(x_\tau)}, \hat{v}_\tau^{AGD}\right\rangle \mathbf{p}(x_\tau)\,\mathrm{d}x_\tau \\ &= \int\left\langle \nabla_x\frac{\delta\mathcal{L}}{\delta\mathbf{p}(x_\tau)}, v_\tau^{AGD}\right\rangle \mathbf{p}(x_\tau)\,\mathrm{d}x_\tau \qquad (17a) \\ &+ \int\left\langle \nabla_x\frac{\delta\mathcal{L}}{\delta\mathbf{p}(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\hat{v}_\tau^{AGD} - v_\tau^{AGD}\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle\,\mathrm{d}x_\tau. \qquad (17b)\end{aligned}$$

We now analyze term (17a) and term (17b) separately, returning later to combine the analyses and obtain the overall convergence rate.

We use Lemma 7 in the Appendix to expand term (17a) and quantify the convergence of $\mathcal{L}$ with respect to the continuous vector flow $v_\tau^{AGD}$:

$$\begin{aligned}\int\left\langle \nabla_x\frac{\delta\mathcal{L}}{\delta\mathbf{p}_t}(x), v_\tau^{AGD}(x)\right\rangle \mathbf{p}_\tau(x)\,\mathrm{d}x = &-4\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x\nabla_r \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), S\nabla_x\nabla_r\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right] \\ &- \mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), M_C\nabla_x\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle\right],\end{aligned} \tag{18}$$

where $M_C$ is defined in Eq. (39). The two terms on the right-hand side of Eq. (18) are both less than or equal to zero. We will use the first term to cancel similar terms in the discretization error and use the second term to drive the convergence of the process (by way of the log-Sobolev inequality).

## 4.1 Discretization error

For term (17b) capturing the discretization error, we provide an upper bound in the following proposition.

**Proposition 2.** *Under Assumption **A2**, when $\tau - kh \le \frac{1}{8L_G}$, $\gamma = 2$, and $\xi = 2L_G$, term* (17b) *is upper* *bounded as:*

$$
\begin{aligned}
&\int \left\langle \nabla_x \frac{\delta \mathcal{L}}{\delta \mathbf{p}(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\hat{v}_\tau^{AGD} - v_\tau^{AGD}\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \,\mathrm{d}x_\tau \\
&\le 4\mathbb{E}_{\mathbf{p}_\tau(x_\tau)}\left[\left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\rangle_F\right] \\
&+ \frac{1}{32}\mathbb{E}_{\mathbf{p}_\tau}\left[\left\|\nabla_\theta \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\|^2\right] + \frac{9}{16}\mathbb{E}_{\mathbf{p}_\tau}\left[\left\|\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\|^2\right] \\
&+ \left(68L_G^2 + \frac{1}{8}\frac{L_H^2}{L_G}\right)\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau - \theta_{kh}\|^2\right] + 18eL_G d \max\left\{L_G^4(\tau-kh)^4, L_G^2(\tau-kh)^2\right\}.
\end{aligned}
$$

Roughly speaking, Proposition 2 upper bounds the instantaneous contribution of the discretization error by the terms appearing in Eq. (18) (the contraction of the continuous process), the variance of $\theta_\tau - \theta_{kh}$ (the progress of $\theta$ within one step), and constant terms that depend on the step size. After combining Proposition 2 with Proposition 1, the only nonnegative terms that remain are the variance of $\theta_\tau - \theta_{kh}$ and other constant terms.

We devote the rest of this subsection to the proof of Proposition 2. We first expand term (17b) using the definitions of the functional $\mathcal{L}$ as well as the discrete and continuous vector flows $\hat{v}_\tau^{AGD}$ and $v_\tau^{AGD}$.

**Lemma 3.** *For $\tau - kh \le \frac{1}{8L_G}$, the time evolution of the Lyapunov functional $\mathcal{L}$ with respect to the discretization error $\hat{v}_\tau^{AGD} - v_\tau^{AGD}$ is:*

$$
\begin{aligned}
&\int \left\langle \nabla_x \frac{\delta \mathcal{L}}{\delta \mathbf{p}(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\hat{v}_\tau^{AGD} - v_\tau^{AGD}\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \,\mathrm{d}x_\tau \\
&= 2\int \left\langle \nabla_\theta \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle_F \,\mathrm{d}x_\tau \qquad (19a) \\
&+ 9\int \left\langle \nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \,\mathrm{d}x_\tau \qquad (19b) \\
&+ 2\int \left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right]\right\rangle_F \mathbf{p}_\tau(x_\tau)\,\mathrm{d}x_\tau. \qquad (19c)
\end{aligned}
$$

It can be observed that of the three terms (19a)–(19c) in Lemma 3, there are two types of term: Terms (19a) and (19b) only involve first-order derivatives, $\nabla_\# \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}$ (for # labeling $\theta$ or $r$); while term (19c) involves a second-order derivative, $\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}$.

For terms (19a) and (19b), we make use of Young's inequality to obtain upper bounds:

$$
\begin{aligned}
&\int \left\langle \nabla_\theta \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \,\mathrm{d}x_\tau \\
&\le \frac{1}{64}\int \left\|\nabla_\theta \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\|^2 \mathbf{p}_\tau(x_\tau)\,\mathrm{d}x_\tau + 16L_G^2\mathbb{E}_{\mathbf{p}(x_\tau,x_{kh})}\left[\|\theta_\tau - \theta_{kh}\|^2\right]. \qquad (20a)
\end{aligned}
$$

$$
\begin{aligned}
&\int \left\langle \nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\left(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \,\mathrm{d}x_\tau \\
&\le \frac{1}{16}\int \left\|\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\|^2 \mathbf{p}_\tau(x_\tau)\,\mathrm{d}x_\tau + 4L_G^2\mathbb{E}_{\mathbf{p}(x_\tau,x_{kh})}\left[\|\theta_\tau - \theta_{kh}\|^2\right]. \qquad (20b)
\end{aligned}
$$

The main difficulty is in bounding term (19c), which is the object of the following lemma.

**Lemma 4.** *Under Assumption **A2**, we provide an explicit bound for term* (19c)*. When* $\tau - kh \leq \frac{1}{8L_G}$*,* $\gamma = 2$*, and* $\xi = 2L_G$*,*

$$\begin{aligned}
&\int \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S \nabla_{x_\tau} \mathbb{E}_{x_{kh} \sim \mathbf{p}(x_{kh}|x_\tau)} \left[\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right] \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&\leq 2\mathbb{E}_{\mathbf{p}_\tau(x_\tau)} \left[ \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)} \right\rangle_F \right] \\
&+ 9eL_G d \max\left\{L_G^4(\tau - kh)^4, L_G^2(\tau - kh)^2\right\} + \frac{1}{16}\frac{L_H^2}{L_G} \mathbb{E}_{\mathbf{p}(x_{kh}|x_\tau)\mathbf{p}_\tau(x_\tau)} \left[\|\theta_\tau - \theta_{kh}\|^2\right].
\end{aligned}$$

In the proof of Lemma 4, we first upper bound the Frobenius inner product in term (19c) by the (weighted) Frobenius norms of $\nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}$ and $\nabla_{x_\tau} \mathbb{E}_{x_{kh} \sim \mathbf{p}(x_{kh}|x_\tau)} [\nabla U(\theta_\tau) - \nabla U(\theta_{kh})]$. We then use a synchronous coupling technique to calculate $\nabla_{x_\tau} \mathbb{E}_{x_{kh} \sim \mathbf{p}(x_{kh}|x_\tau)} [\nabla U(\theta_\tau) - \nabla U(\theta_{kh})]$ and provide an upper bound of its Frobenius norm. We defer the complete proof to Appendix D.

Applying Eq. (20a)–(20b) and Lemma 4 to Eq. (19a)–(19c), we bound the overall discretization error and finish the proof of Proposition 2 as follows:

$$\begin{aligned}
&\int \left\langle \nabla_x \frac{\delta \mathcal{L}}{\delta \mathbf{p}(x_\tau)}, \mathbb{E}_{x_{kh} \sim \mathbf{p}(x_{kh})} \left[\left(\hat{v}_\tau^{AGD} - v_\tau^{AGD}\right) \mathbf{p}(x_\tau | x_{kh})\right] \right\rangle \, \mathrm{d}x_\tau \\
&\leq 4\mathbb{E}_{\mathbf{p}_\tau(x_\tau)} \left[ \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)} \right\rangle_F \right] \\
&+ \frac{1}{32} \mathbb{E}_{\mathbf{p}_\tau} \left[ \left\| \nabla_\theta \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)} \right\|^2 \right] + \frac{9}{16} \mathbb{E}_{\mathbf{p}_\tau} \left[ \left\| \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)} \right\|^2 \right] \\
&+ \left(68L_G^2 + \frac{1}{8}\frac{L_H^2}{L_G}\right) \mathbb{E}_{\mathbf{p}(x_{kh}, x_\tau)} \left[\|\theta_\tau - \theta_{kh}\|^2\right] + 18eL_G d \max\left\{L_G^4(\tau - kh)^4, L_G^2(\tau - kh)^2\right\}.
\end{aligned}$$

## 4.2 Convergence of the underdamped Langevin algorithm

Combining Propositions 1 and 2, which establish the convergence rates of the continuous underdamped Langevin dynamics and the discretization error, we find that the overall time evolution of the Lyapunov functional $\mathcal{L}$ within each step of the underdamped Langevin algorithm can be upper bounded as follows:

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}(\mathbf{p}_t)}{\mathrm{d}t} &= \int \left\langle \nabla_x \frac{\delta \mathcal{L}}{\delta \mathbf{p}_t}, v_t^{AGD} \right\rangle \mathbf{p}_t \, \mathrm{d}x \\
&+ \int \left\langle \nabla_x \frac{\delta \mathcal{L}}{\delta \mathbf{p}_t}, \mathbb{E}_{x_{kh} \sim \mathbf{p}(x_{kh})} \left[\left(\hat{v}_\tau^{AGD} - v_\tau^{AGD}\right) \mathbf{p}(x_\tau | x_{kh})\right] \right\rangle \mathbf{p}_t \, \mathrm{d}x \\
&\leq -\mathbb{E}_{\mathbf{p}_t} \left[ \left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), M \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right) \right\rangle_F \right] \qquad (21a) \\
&+ \left(68L_G^2 + \frac{1}{8}\frac{L_H^2}{L_G}\right) \mathbb{E}_{\mathbf{p}(x_{kh}, x_\tau)} \left[\|\theta_\tau - \theta_{kh}\|^2\right] \qquad (21b) \\
&+ 18eL_G d \max\left\{L_G^4(\tau - kh)^4, L_G^2(\tau - kh)^2\right\}, \qquad (21c)
\end{aligned}$$

where

$$M = \begin{pmatrix} \frac{31}{32}\mathrm{I}_{d\times d} & 4 \cdot \mathrm{I}_{d\times d} - \frac{1}{8}\frac{\nabla^2 U(\theta)}{L_G} \\ 4 \cdot \mathrm{I}_{d\times d} - \frac{1}{8}\frac{\nabla^2 U(\theta)}{L_G} & \frac{279}{16}\mathrm{I}_{d\times d} - \frac{1}{2}\frac{\nabla^2 U(\theta)}{L_G} \end{pmatrix}.$$

In this section, we will further analyze terms (21a)–(21c) to obtain the overall convergence rate of the underdamped Langevin algorithm. We will need to quantify the convergence contributed by term (21a) and upper bound the extra discretization error in terms (21b)–(21c) as the algorithm progresses. After these two steps, choosing a suitable step size will finish the proof of Theorem 1.

We begin by using the log-Sobolev inequality to relate term (21a) to the Lyapunov functional $\mathcal{L}(\mathbf{p}_t)$. A key step is lower bounding matrix $M$ which is done in the following Lemma 5 (the proof of which is deferred to Appendix E).

**Lemma 5.** *Under Assumption **A2**, for any* $L_G \geq 2\rho$, $M \succeq \frac{\rho}{30}\left(S + \frac{1}{2\rho}\mathrm{I}_{2d\times 2d}\right)$.

We can thus upper bound term (21a) using this lower bound on $M$ in conjunction with the log-Sobolev inequality, Eq. (5):

$$
\begin{aligned}
&-\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), M\nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right] \\
&\leq -\frac{\rho}{30}\left(\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right), S\nabla_x \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right)\right\rangle\right] + \frac{1}{2\rho}\mathbb{E}_{\mathbf{p}_t}\left[\left\|\nabla_x \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right)\right\|^2\right]\right) \\
&\leq -\frac{\rho}{30}\left(\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right), S\nabla_x \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right)\right\rangle\right] + \mathbb{E}_{\mathbf{p}_t}\left[\ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right)\right]\right) \\
&\leq -\frac{\rho}{30}\cdot\mathcal{L}[\mathbf{p}_t]. \qquad (22)
\end{aligned}
$$

Consequently, Eq. (21a)–(21c) simplify to:

$$
\begin{aligned}
\frac{\mathrm{d}\mathcal{L}(\mathbf{p}_t)}{\mathrm{d}t} &\leq -\frac{\rho}{30}\mathcal{L}(\mathbf{p}_t) && (23a)\\
&\quad + \left(68L_G^2 + \frac{1}{8}\frac{L_H^2}{L_G}\right)\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau - \theta_{kh}\|^2\right] && (23b)\\
&\quad + 18eL_G d\max\left\{L_G^4(\tau-kh)^4, L_G^2(\tau-kh)^2\right\}. && (23c)
\end{aligned}
$$

This implies that without the extra discretization error of terms (23b)–(23c), the Markov process converges exponentially (similarly as for the continuous dynamics) with a rate of $\rho/30$, proportional to the log-Sobolev constant.

We now focus on the second task of upper bounding terms (23b)–(23c). The crux of the argument is to upper bound the variance of $\theta_\tau - \theta_{kh}$ as the algorithm progresses. In the following lemma we show that for a suitable choice of step size, $\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau - \theta_{kh}\|^2\right]$ is uniformly upper bounded by a term that scales as $\mathcal{O}(h^2 d)$.

**Lemma 6.** *Assume that function $U$ satisfies Assumption **A1**–**A3**, where $\rho$ denotes the minimum of the log-Sobolev constant and* 1. *Assume that we take $\gamma = 2$, $\xi = 2L_G$, and*

$$
h = \frac{1}{56}\frac{1}{\sqrt{L_G}}\min\left\{\frac{1}{24}\frac{\rho}{L_G}, \frac{\sqrt{L_G}\rho}{L_H}\right\}\cdot\min\left\{\left(\widetilde{C_N}+2\right)^{-1/2}\sqrt{\frac{\epsilon}{d}}, \sqrt{\frac{\epsilon}{C_M}}\right\},
$$

*where $\epsilon \leq 2d$. Then for $\theta_\tau$ following Eq.* (13), *$\forall n \in \mathbb{N}^+$ and $\forall \tau \in [kh,(k+1)h]$,*

$$
\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau - \theta_{kh}\|^2\right] \leq \left(\left(24\widetilde{C_N}+26\right)\frac{L_G}{\rho}\cdot d + 24C_M\frac{L_G}{\rho}\right)h^2 = \mathcal{O}\left(\frac{L_G}{\rho}d\cdot h^2\right).
$$

To establish this uniform upper bound, we use an inductive argument—we prove that if the above bound holds for $t \leq kh$, then, given the effect of contraction and the discretization error in $[kh,\tau]$, the bound will still hold for any $\tau \in [kh,(k+1)h]$. We defer the complete proof of Lemma 6 to Appendix E.

Given this uniform bound for $\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau-\theta_{kh}\|^2\right]$ across the entire interval, we can upper bound term (23b) using our choice of the parameters $\gamma = 2$, $\xi = 2L_G$, and the step size $h$:

$$\begin{aligned}
&\left(68L_G^2+\frac{1}{8}\frac{L_H^2}{L_G}\right)\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau-\theta_{kh}\|^2\right]\\
&=\left(68L_G^2+\frac{1}{8}\frac{L_H^2}{L_G}\right)\left(\left(24\widetilde{C_N}+26\right)\frac{L_G}{\rho}\cdot d+24C_M\frac{L_G}{\rho}\right)h^2\\
&\le \rho\cdot L_G\max\left\{136\frac{L_G}{\rho},\frac{1}{4}\frac{L_H^2}{L_G^2\rho}\right\}\cdot\max\left\{\left(48\widetilde{C_N}+52\right)\frac{L_G}{\rho}d,48C_M\frac{L_G}{\rho}\right\}h^2\\
&\le \frac{49}{4}\rho\cdot L_G\max\left\{24^2\frac{L_G^2}{\rho^2},\frac{L_H^2}{L_G\rho^2}\right\}\cdot\max\left\{\left(\widetilde{C_N}+2\right)d,C_M\right\}h^2\\
&\le \frac{\rho}{30}\cdot\frac{\epsilon}{4}. \qquad (24a)
\end{aligned}$$

For term (23c), we obtain that

$$\begin{aligned}
18eL_Gd\max\left\{L_G^4(\tau-kh)^4,L_G^2(\tau-kh)^2\right\}&\le\frac{\rho}{30}\cdot540e\frac{L_G}{\rho}d\max\{L_G^4h^4,L_G^2h^2\}\\
&\le\frac{\rho}{30}\cdot\frac{\epsilon}{4}. \qquad (24b)
\end{aligned}$$

Plugging Eqs. (24a)–(24b) into Eqs. (23b)–(23c), we obtain the following upper bound for $\frac{\mathrm{d}\mathcal{L}(\mathbf{p}_t)}{\mathrm{d}t}$:

$$\frac{\mathrm{d}\mathcal{L}(\mathbf{p}_t)}{\mathrm{d}t}\le-\frac{\rho}{30}\cdot\left(\mathcal{L}(\mathbf{p}_t)-\frac{\epsilon}{2}\right).$$

Applying Grönwall's lemma, we arrive at a bound for the Lyapunov functional at every step:

$$\mathcal{L}[\mathbf{p}_{kh}]-\frac{\epsilon}{2}\le e^{-\frac{\rho}{30}h}\left(\mathcal{L}[\mathbf{p}_{(k-1)h}]-\frac{\epsilon}{2}\right)\le e^{-\frac{\rho}{30}hk}\left(\mathcal{L}[\mathbf{p}_0]-\frac{\epsilon}{2}\right)<e^{-\frac{\rho}{30}hk}\mathcal{L}[\mathbf{p}_0].$$

Therefore, for any $k\ge K=\frac{30}{\rho h}\ln\left(\frac{2\mathcal{L}[\mathbf{p}_0]}{\epsilon}\right)$, we have $\mathrm{KL}\left(\mathbf{p}_{kh}\|\mathbf{p}^*\right)\le\mathcal{L}[\mathbf{p}_{kh}]\le\epsilon$.

We now use the definition of the step size $h$ and the upper bound on the initial value $\mathcal{L}[\mathbf{p}_0]$ from Lemma 12 to obtain the number of iterations for Algorithm 1 to converge to within $\epsilon$ of the target distribution $\mathbf{p}^*$:

$$\begin{aligned}
K&=1680\max\left\{24\frac{L_G^{3/2}}{\rho^2},\frac{L_H}{\rho^2}\right\}\cdot\max\left\{\sqrt{\widetilde{C_N}+2}\sqrt{\frac{d}{\epsilon}},\sqrt{\frac{C_M}{\epsilon}}\right\}\cdot\ln\left(4\max\left\{\left(\widetilde{C_N}+1\right)\frac{d}{\epsilon},\frac{C_M}{\epsilon}\right\}\right)\\
&=\mathcal{O}\left(\sqrt{\frac{d}{\epsilon}}\ln\frac{d}{\epsilon}\right).
\end{aligned}$$

If the function $U$ further satisfies assumptions **A1**—**A3** (that $U$ is nonconvex inside a region of radius $R$ and $m$-strongly convex outside of it), we can instantiate the constants $\rho\ge\frac{m}{2}e^{-16L_GR^2}$, $\widetilde{C_N}=C_N+\frac{1}{2}\ln\frac{L_G}{2\pi}\le\frac{1}{2}\ln\frac{2L_G}{m}$, and $C_M\le32\frac{L_G^2}{m^2}L_GR^2$, and study the computational complexity in more detail. The number of iterations required becomes:

$$\begin{aligned}
K=4800e^{32L_GR^2}&\max\left\{24\frac{L_G^{3/2}}{m^2},\frac{L_H}{m^2}\right\}\cdot\max\left\{\sqrt{\ln\frac{L_G}{m}+5}\sqrt{\frac{d}{\epsilon}},8R\frac{L_G}{m}\sqrt{\frac{L_G}{\epsilon}}\right\}\\
&\cdot\ln\left(2\max\left\{\left(\ln\frac{L_G}{m}+4\right)\frac{d}{\epsilon},64R^2\frac{L_G^2}{m^2}\frac{L_G}{\epsilon}\right\}\right).
\end{aligned}$$

Emphasizing the dimension dependency, we have:

$$K = \mathcal{O}\left(\max\left\{\frac{L_G^{3/2}}{\rho^2}, \frac{L_H}{\rho^2}\right\}\sqrt{\frac{d}{\epsilon}}\ln\frac{d}{\epsilon}\right).$$

# 5 Discussion

We have shown that there is an analog of Nesterov accelerated gradient method for MCMC—it is the underdamped Langevin algorithm. We demonstrated this by adopting a view of sampling algorithms as optimizing over the space of probability measures, with KL divergence as the objective functional. By constructing an appropriate Lyapunov functional, we were able to prove that the underdamped Langevin algorithm has an accelerated convergence rate compared to the classical overdamped Langevin algorithm.

A line of recent results leverage richer stochastic dynamics to obtain better pre-conditioning and employ higher-order discretization schemes [17, 22, 21]. They observe that in practice such dynamics increase stability and in turn results in faster convergence of the algorithm.

Our particular approach involves multiplying the strong sub-differential of the KL divergence by a symplectic matrix and a positive semidefinite matrix. An interesting direction for future research would be to consider other, more general choices. Indeed, a general construction of underdamped stochastic processes would involve taking a vector field $v_t$ to have the following form:

$$v_t = -(D(x) + Q(x))\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right), \tag{25}$$

where $D(x)$ is a positive semidefinite *diffusion* matrix, and $Q(x)$ is a skew-symmetric *curl* matrix. This has the form of a generic dynamics for smooth optimization. It can be checked that when $\mathbf{p}_t(x) = \mathbf{p}^*(x)$, $v_t = 0$. Therefore, $\mathbf{p}^*$ is a stationary distribution when $\mathbf{p}_t$ follows the vector flow $v_t$:

$$\begin{aligned}\frac{\partial \mathbf{p}_t(x)}{\partial t} =& - \nabla \cdot (\mathbf{p}_t(x) \cdot v_t)\\ =& \nabla \cdot \left(\mathbf{p}_t(x) \cdot (D(x) + Q(x))\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right)\right).\end{aligned} \tag{26}$$

It has been previously proved [24, 26] that any continuous Markov process with the stationary distribution $\mathbf{p}^*$ which satisfies an integrability condition can be represented in the form of Eq. (26).

To simulate the dynamics of $v_t$ on the state space of $x$, we can realize it as a stochastic process with an Itô diffusion:

$$\begin{aligned}\frac{\partial \mathbf{p}_t(x)}{\partial t} =& \nabla \cdot \left(\mathbf{p}_t(x)\big(D(x) + Q(x)\big)\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right)\right)\\ =& \sum_i\sum_j \frac{\partial^2}{\partial x_i \partial x_j}\big(\mathbf{p}_t(x) D_{i,j}(x)\big) - \nabla \cdot \Big(\mathbf{p}_t(x)\Big(\big(D(x) + Q(x)\big)\nabla \ln \mathbf{p}^*(x) + \Gamma(x)\Big)\Big),\end{aligned} \tag{27}$$

where $\Gamma_i(x) = \sum_j \frac{\partial}{\partial x_j}\left[D(x) + Q(x)\right]_{i,j}$. Eq. (27) corresponds to the probability density of $x_t$ following a stochastic differential equation:

$$\mathrm{d}x_t = \left((D(x) + Q(x))\nabla \ln\left(\mathbf{p}^*(x)\right) + \Gamma(x)\right)\mathrm{d}t + \sqrt{2D(x)}\mathrm{d}B_t. \tag{28}$$

To study convergence of this process, denote the first variation of a functional $\mathcal{G}[\mathbf{p}_t]$ as $\frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t] : \mathbb{R}^d \to \mathbb{R}$. If the vector flow $v_t$ satisfies the continuity equation for $\mathbf{p}_t$, Eq. (26), then

$$\begin{aligned}
\frac{d}{dt}\mathcal{G}[\mathbf{p}_t] &= \int \frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t](x) \frac{d}{dt}\mathbf{p}_t(x)\mathrm{d}x \\
&= \int \frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t](x)(-\nabla \cdot (\mathbf{p}_t(x) v_t(x)))\mathrm{d}x \\
&= \int \left\langle \nabla \frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t](x), v_t(x) \right\rangle \mathbf{p}_t(x)\mathrm{d}x \\
&= \mathbb{E}_{\mathbf{p}_t}\left[ \left\langle \nabla \frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t](x), v_t(x) \right\rangle \right] \\
&= -\mathbb{E}_{\mathbf{p}_t}\left[ \left\langle \nabla \frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t](x), (D(x)+Q(x))\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right) \right\rangle \right]. \qquad (29)
\end{aligned}$$

Using notation from statistical mechanics, we can represent Eq. (29) in a more compact form using a (Ginzburg-Landau) dissipative bracket and a generalized Poisson bracket to generate the stochastic process $\frac{d}{dt}\mathbf{p}_t(x)$ with $\nabla \frac{\delta \mathcal{F}}{\delta \mathbf{p}_t}$. Define the dissipative bracket $\{\cdot, \cdot\}$ as

$$\{\mathcal{G}[\mathbf{p}_t], \mathcal{F}[\mathbf{p}_t]\} = \mathbb{E}_{\mathbf{p}_t(x)}\left[ \left\langle \nabla \frac{\delta \mathcal{G}[\mathbf{p}_t](x)}{\delta \mathbf{p}_t(y)}, D(x)\nabla \frac{\delta \mathcal{F}[\mathbf{p}_t](x)}{\delta \mathbf{p}_t(y)} \right\rangle \right]; \qquad (30)$$

and the generalized Poisson bracket $[\cdot, \cdot]$ as

$$[\mathcal{G}[\mathbf{p}_t], \mathcal{F}[\mathbf{p}_t]] = \mathbb{E}_{\mathbf{p}_t(x)}\left[ \left\langle \nabla \frac{\delta \mathcal{G}[\mathbf{p}_t](x)}{\delta \mathbf{p}_t(y)}, Q(x)\nabla \frac{\delta \mathcal{F}[\mathbf{p}_t](x)}{\delta \mathbf{p}_t(y)} \right\rangle \right]. \qquad (31)$$

Then

$$\begin{aligned}
\frac{d}{dt}\mathcal{G}[\mathbf{p}_t] &= -\mathbb{E}_{\mathbf{p}_t}\left[ \left\langle \nabla \frac{\delta \mathcal{G}}{\delta \mathbf{p}_t}[\mathbf{p}_t](x), (D(x)+Q(x))\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right) \right\rangle \right] \\
&= -\{\mathcal{G}[\mathbf{p}_t], \mathcal{F}[\mathbf{p}_t]\} - [\mathcal{G}[\mathbf{p}_t], \mathcal{F}[\mathbf{p}_t]]. \qquad (32)
\end{aligned}$$

By taking $\mathcal{G} = \mathcal{F}$ as the KL-divergence, we can calculate its time derivative as:

$$\begin{aligned}
\frac{d}{dt}\mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right) &= \mathbb{E}_{\mathbf{p}_t}\left[ \left\langle \nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right), -(D(x)+Q(x))\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right) \right\rangle \right] \\
&= -\mathbb{E}_{\mathbf{p}_t}\left[ \left\langle \nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right), D(x)\nabla \ln\left(\frac{\mathbf{p}_t(x)}{\mathbf{p}^*(x)}\right) \right\rangle \right] \leq 0, \qquad (33)
\end{aligned}$$

where we know from the positive semidefiniteness of $D(x)$ that $\mathrm{KL}\left(\mathbf{p}_t \| \mathbf{p}^*\right)$ is monotonically non-increasing. If $D(x)$ were to be positive definite, we can directly obtain a linear convergence rate for the *continuous process* using the log-Sobolev inequality. However if $D(x)$ is just positive semidefinite (as is the case for the diffusion matrix that we encountered while analyzing the underdamped Langevin algorithm) we need to choose a well-designed Lyapunov functional to prove convergence (if the process indeed converges).

Some attempts have been made in this direction in the stochastic optimization literature for a class of constant $D$ and $Q$ matrices [16]. For the generic case, [15] explores an approach based on Stein factors; this seems like a particularly promising avenue to explore further.

## 6 Acknowledgements

We would like to thank Jianfeng Lu, Chi Jin, and Nilesh Tripuraneni for many helpful discussions and insights. This work was partially supported by Army Research Office grant W911NF-17-1-0304, and National Science Foundation Grant NSF-IIS-1740855, NSF-IIS-1909365, and NSF-IIS-1619362.

# References

[1] J. Bierkens, P. Fearnhead, and G. Roberts. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320, 2019.

[2] N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. arXiv:1805.00452, 2018.

[3] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Am. Stat. Assoc.*, 113(522):855–867, 2018.

[4] S. Calogero. Exponential convergence to equilibrium for kinetic Fokker-Planck equations. *Comm. Part. Differ. Equat.*, 37(8):1357–1390, 2012.

[5] N. Chatterji, N. Flammarion, Y.-A. Ma, P. Bartlett, and M. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 764–773, 2018.

[6] X. Cheng and P. L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, pages 186–211, 2018.

[7] 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:1805.01648, 2018.

[8] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 300–323, 2018.

[9] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. Royal Stat. Soc. B*, 79(3):651–676, 2017.

[10] A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. arXiv:1710.00095, 2017.

[11] A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. arXiv:1807.09382, 2018.

[12] A. Durmus and E. Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. arXiv:1605.01559, 2016.

[13] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017.

[14] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! arXiv:1801.02309, 2018.

[15] M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems 31*, pages 9671–9680. 2018.

[16] X. Gao, M. Gurbuzbalaban, and L. Zhu. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. arXiv:1812.07725, 2019.

[17] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Royal Stat. Soc. B*, 73(2):123–214, 2011.

[18] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math*, 97(4):1061–1083, 1975.

[19] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, January 1998.

[20] P. Langevin. On the theory of Brownian motion (sur la théorie du mouvement brownien). *C. R. Acad. Sci. (Paris)*, 146:530–533, 1908.

[21] B. Leimkuhler and X. Shang. Adaptive thermostats for noisy gradient systems. *SIAM J. Sci. Comput.*, 38(2):A712–A736, 2016.

[22] C. Liu, J. Zhu, and Y. Song. Stochastic gradient geodesic MCMC methods. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 642–651. 2016.

[23] A. Luigi, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Springer Science & Business Media, 2nd edition, 2008.

[24] Y.-A Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems (NIPS) 28*, pages 2899–2907. 2015.

[25] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proc. Natl. Acad. Sci. U.S.A.*, 116:20881–20885, 2019.

[26] Y.-A Ma, E. B. Fox, T. Chen, and L. Wu. Irreversible samplers from jump and continuous Markov processes. *Stat. Comput.*, pages 1–26, 2018.

[27] O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. arXiv:1708.07114, 2017.

[28] O. Mangoubi and N. K. Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. arXiv:1802.08898, 2018.

[29] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[30] A. Nemirovskii and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

[31] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[32] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer, Boston, 2004.

[33] B. O'donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.*, 15(3):715–732, 2015.

[34] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

[35] M. Ottobre, N. S. Pillai, F. J. Pinski, and A. M. Stuart. A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, 22(1):60–106, 02 2016.

[36] H. J. M. Peters and P. P. Wakker. Convex functions on non-convex domains. *Econ. Lett.*, 22(2):251–255, 1986.

[37] B. T. Polyak. Gradient methods for minimizing functionals. *Zh. Vychisl. Mat. Mat. Fiz.*, 3(4):643–653, 1963.

[38] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4:337–357, 2002.

[39] P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys.*, 69(10):4628, 1978.

[40] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via Symplectic Discretization of High-Resolution Differential Equations. arXiv:1902.03694, 2019.

[41] W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 2510–2518. 2014.

[42] C. Villani. Hypocoercivity. *Mem. Am. Math. Soc.*, 202(950), 2009.

[43] C. Villani. *Optimal Transport: Old and New.* Wissenschaften. Springer, Berlin, 2009.

[44] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. arXiv:1802.08089, 2018.

[45] A. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. arXiv:1611.02635, 2016.

[46] M. Yan. Extension of convex function. *J. Convex. Anal.*, 21(4):965–987, 2014.

[47] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems (NeuIPS) 31*, pages 3900–3909. 2018.

# A Local Nonconvexity Assumption

For $\mathbf{p}^*(\theta) \propto e^{-U(\theta)}$, we call a function $U : \mathbb{R}^d \to \mathbb{R}$ *locally nonconvex* with radius $R$ and global strong convexity $m$ if it satisfies the following assumptions:

(a) $U(\theta)$ is $m$-strongly convex for $\|\theta\| > R$.

That is: $V(\theta) = U(\theta) - \frac{m}{2}\|\theta\|_2^2$ is convex on $\Omega = \mathbb{R}^d \backslash \mathbb{B}(0, R)^3$. We then follow the definition of convexity on nonconvex domains [36, 46] to require that $\forall \theta \in \Omega$, any convex combination of $\theta = \lambda_1\theta_1 + \cdots + \lambda_k\theta_{kh}$ with $\theta_1, \cdots, \theta_{kh} \in \Omega$ satisfies:

$$V(\theta) \leq \lambda_1 V(\theta_1) + \cdots + \lambda_k V(\theta_{kh}).$$

(b) $U(\theta)$ is $L_G$-Lipschitz smooth and Hessian $L_H$-Lipschitz.

That is: $U \in C^2(\mathbb{R}^d)$; $\forall \theta, \vartheta \in \mathbb{R}^d$, $\|\nabla U(\theta) - \nabla U(\vartheta)\| \leq L_G \|\theta - \vartheta\|$ and $\left\|\nabla^2 U(\theta) - \nabla^2 U(\vartheta)\right\|_F \leq L_H \|\theta - \vartheta\|$.

(c) For convenience, let $\nabla U(0) = 0$ (i.e., zero is a local extremum).

From [25], we know that $\rho \geq \frac{m}{2} e^{-16 L_G R^2}$. We prove that the constants in Assumption **A3** are also upper bounded by functions of $m$, $L_G$, and $R$.

**Fact 1.** *If $\mathbf{p}^*(\theta) \propto e^{-U(\theta)}$ satisfy Assumptions (a)–(c), then the normalization constant $\int \exp(-U(\theta))\mathrm{d}\theta$ is upper bounded as follows:*

$$\ln \int \exp\left(-U(\theta)\right) \mathrm{d}\theta = \frac{d}{2} \ln \frac{4\pi}{m} + 32 \frac{L_G^2}{m^2} L_G R^2.$$

*In other words, constants in Assumption **A3** are bounded as: $C_N \leq \frac{1}{2} \ln \frac{4\pi}{m}$, and $C_M \leq 32 \frac{L_G^2}{m^2} L_G R^2$.*

# B Explicit Iteration Rule for Algorithm 1

We provide an explicit iteration formula for $x_\tau$ given $x_{kh}$ in Eq. (13). Given $x_{kh}$ at the previous iteration, $x_\tau$ can be calculated as:

$$\begin{cases} \theta_\tau = \theta_{kh} + \dfrac{1 - e^{-\gamma\xi(\tau - kh)}}{\gamma} r_{kh} - \dfrac{1}{\gamma}\left((\tau - kh) - \dfrac{1 - e^{-\gamma\xi(\tau - kh)}}{\gamma\xi}\right) \nabla U(\theta_{kh}) + W_\theta \\ r_\tau = e^{-\gamma\xi(\tau - kh)} r_{kh} - \dfrac{1 - e^{-\gamma\xi(\tau - kh)}}{\gamma\xi} \nabla U(\theta_{kh}) + W_r \end{cases}, \tag{34}$$

where

$$\begin{pmatrix} W_\theta \\ W_r \end{pmatrix} \sim \mathcal{N}\left(0, \Sigma_\tau\right).$$

The covariance matrix $\Sigma \in \mathbb{R}^{2d \times 2d}$ is

$$\Sigma_\tau = \begin{pmatrix} \Sigma_{1,1}(\tau)\ \mathrm{I}_{d\times d} & \Sigma_{1,2}(\tau)\ \mathrm{I}_{d\times d} \\ \Sigma_{1,2}(\tau)\ \mathrm{I}_{d\times d} & \Sigma_{2,2}(\tau)\ \mathrm{I}_{d\times d} \end{pmatrix},$$

[3] Here we let $\mathbb{B}(0, R)$ denote the closed ball of radius $R$ centered at 0.

where

$$\begin{aligned}
\Sigma_{1,1}(\tau) &= \frac{1}{\gamma}\left(2(\tau-kh)-\frac{3}{\gamma\xi}+\frac{4}{\gamma\xi}e^{-\gamma\xi(\tau-kh)}-\frac{1}{\gamma\xi}e^{-2\gamma\xi(\tau-kh)}\right);\\
\Sigma_{1,2}(\tau) &= \frac{1+e^{-2\gamma\xi(\tau-kh)}-2e^{-\gamma\xi(\tau-kh)}}{\gamma\xi};\\
\Sigma_{2,2}(\tau) &= \frac{1-e^{-2\gamma\xi(\tau-kh)}}{\xi}.
\end{aligned}$$

Therefore, the update rule in Algorithm 1 can be expressed as:

$$x_{(k+1)h}\sim\mathcal{N}\left(\mu\left(x_{kh}\right),\Sigma\right),$$

where

$$\mu\left(x_{kh}\right)=\left(\begin{array}{l}\theta_{kh}+\dfrac{1-e^{-\gamma\xi h}}{\gamma}r_{kh}-\dfrac{1}{\gamma}\left(h-\dfrac{1-e^{-\gamma\xi h}}{\gamma\xi}\right)\nabla U(\theta_{kh})\\ e^{-\gamma\xi h}r_{kh}-\dfrac{1-e^{-\gamma\xi h}}{\gamma\xi}\nabla U(\theta_{kh})\end{array}\right),\tag{35}$$

and

$$\Sigma=\left(\begin{array}{cc}\dfrac{1}{\gamma}\left(2h-\dfrac{3}{\gamma\xi}+\dfrac{4}{\gamma\xi}e^{-\gamma\xi h}-\dfrac{1}{\gamma\xi}e^{-2\gamma\xi h}\right)\mathrm{I}_{d\times d} & \dfrac{1+e^{-2\gamma\xi h}-2e^{-\gamma\xi h}}{\gamma\xi}\mathrm{I}_{d\times d}\\ \dfrac{1+e^{-2\gamma\xi h}-2e^{-\gamma\xi h}}{\gamma\xi}\mathrm{I}_{d\times d} & \dfrac{1-e^{-2\gamma\xi h}}{\xi}\mathrm{I}_{d\times d}\end{array}\right).\tag{36}$$

In Algorithm 1, the hyperparameters are set to be: $\gamma=2$, $\xi=2L_G$, and

$$h=\frac{1}{56}\frac{1}{\sqrt{L_G}}\min\left\{\frac{1}{24}\frac{\rho}{L_G},\frac{\sqrt{L_G}\rho}{L_H}\right\}\cdot\min\left\{\left(\widetilde{C_N}+2\right)^{-1/2}\sqrt{\frac{\epsilon}{d}},\sqrt{\frac{\epsilon}{C_M}}\right\},\tag{37}$$

where $\widetilde{C_N}=C_N+\dfrac{1}{2}\ln\dfrac{L_G}{2\pi}$.

# C Convergence of the Continuous Process

To simplify the notations in the proofs, we let $a=\dfrac{1}{L_G}$, $b=\dfrac{1}{4L_G}$, and $c=\dfrac{2}{L_G}$, so that

$$S=\frac{1}{L_G}\left(\begin{array}{cc}1/4\ \mathrm{I}_{d\times d} & 1/2\ \mathrm{I}_{d\times d}\\ 1/2\ \mathrm{I}_{d\times d} & 2\ \mathrm{I}_{d\times d}\end{array}\right)=\left(\begin{array}{cc}b\ \mathrm{I}_{d\times d} & a/2\ \mathrm{I}_{d\times d}\\ a/2\ \mathrm{I}_{d\times d} & c\ \mathrm{I}_{d\times d}\end{array}\right).$$

**Proof of Proposition 1** We first compute the time evolution of the Lyapunov function $\mathcal{L}$ with respect to the continuous time vector flow $v_t^{AGD}$ in Eq. (7).

**Lemma 7.** *The time derivative of the Lyapunov functional $\mathcal{L}$ with respect to the continuous time vector flow $v_t^{AGD}$ in Eq. (7) with $\gamma=2$ and $\xi=2L_G$ is:*

$$\begin{aligned}
\frac{d}{\mathrm{d}t}\mathcal{L}[\mathbf{p}_t] &= \int\left\langle\nabla_x\frac{\delta\mathcal{L}}{\delta\mathbf{p}_t},v_t^{AGD}\right\rangle\mathbf{p}_t\ \mathrm{d}x\\
&= -\mathbb{E}_{\mathbf{p}_t}\left[\left\langle\nabla_x\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right),M_C\nabla_x\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right]\\
&\quad-4\mathbb{E}_{\mathbf{p}_t}\left[\left\langle\nabla_x\nabla_r\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right),S\nabla_x\nabla_r\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right],
\end{aligned}$$

*where*

$$M_C = \begin{pmatrix} \frac{a}{2}\xi \cdot \mathrm{I} & \frac{c+a\gamma}{2}\xi \cdot \mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) \\ \frac{c+a\gamma}{2}\xi \cdot \mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) & \gamma\left(2c\xi+1\right)\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta) \end{pmatrix}$$
$$= \begin{pmatrix} \mathrm{I}_{d\times d} & 4\cdot \mathrm{I}_{d\times d} - \frac{1}{8}\frac{\nabla^2 U(\theta)}{L_G} \\ 4\cdot \mathrm{I}_{d\times d} - \frac{1}{8}\frac{\nabla^2 U(\theta)}{L_G} & 18\cdot \mathrm{I}_{d\times d} - \frac{1}{2}\frac{\nabla^2 U(\theta)}{L_G} \end{pmatrix}. \tag{38}$$

We then upper bound the time derivative of $\mathcal{L}$ by a negative factor times itself to obtain linear convergence rate.

**Lemma 8.** *For $L_G$-Lipschitz smooth $U$, matrix $M_C$ defined in Eq.* (38) *satisfy:*

$$M_C \succeq \frac{\rho}{10}\left(S + \frac{1}{2\rho}\mathrm{I}_{2d\times 2d}\right). \tag{39}$$

Since the matrix $S$ is positive definite, we can directly bound the evolution of the Lyapunov functional $\mathcal{L}$ as

$$\begin{aligned}\frac{d}{\mathrm{d}t}\mathcal{L}[\mathbf{p}_t] &\le -\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), M_C \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right] \\ &\le -\frac{\rho}{10}\left(\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), S \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right] + \frac{1}{2\rho}\mathbb{E}_{\mathbf{p}_t}\left[\left\|\nabla_x \ln \frac{\mathbf{p}_t}{\mathbf{p}^*}\right\|^2\right]\right).\end{aligned}$$

Using the log-Sobolev inequality in Assumption **A1**, we directly obtain:

$$\begin{aligned}\frac{d}{\mathrm{d}t}\mathcal{L}[\mathbf{p}_t] &\le -\frac{\rho}{10}\left(\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), S \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right] + \frac{1}{2\rho}\mathbb{E}_{\mathbf{p}_t}\left[\left\|\nabla_x \ln \frac{\mathbf{p}_t}{\mathbf{p}^*}\right\|^2\right]\right) \\ &\le -\frac{\rho}{10}\left(\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), S \nabla_x \ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right] + \mathbb{E}_{\mathbf{p}_t}\left[\ln \frac{\mathbf{p}_t}{\mathbf{p}^*}\right]\right) \\ &= -\frac{\rho}{10}\mathcal{L}[\mathbf{p}_t],\end{aligned}$$

which implies the linear convergence of the continuous process with a rate of $\frac{\rho}{10}$. ■

**Proof of Lemma 7** Denote $h(\mathbf{p}_t) = \sqrt{\frac{\mathbf{p}_t}{\mathbf{p}^*}}$. Then

$$\mathcal{L}[\mathbf{p}_t] = \mathbb{E}_{\mathbf{p}_t}\left[2\ln h + 4\left\langle \nabla_x \ln h, S\nabla_x \ln h\right\rangle\right] = 2\mathbb{E}_{\mathbf{p}_t}\left[\ln h\right] + 4\mathbb{E}_{\mathbf{p}^*}\left[\left\langle \nabla_x h, S\nabla_x h\right\rangle\right].$$

The variational derivative of $\mathcal{L}[\mathbf{p}_t]$ can be thus calculated as:

$$\frac{\delta \mathcal{L}[\mathbf{p}_t]}{\delta \mathbf{p}_t} = 2\ln h + 1 + \frac{4}{h}(\nabla_x)^* S \nabla_x h,$$

where the adjoint operator of $\nabla_x$ is with respect to the inner product: $\mathbb{E}_{\mathbf{p}^*}\left[\langle\cdot,\cdot\rangle\right]$. Since:

$$\mathbb{E}_{p^*}\left[\langle \nabla_x f, \overrightarrow{v}\rangle\right] = \mathbb{E}_{p^*}\left[\left(-\nabla_x^{\mathrm{T}}\overrightarrow{v} - \nabla_x^{\mathrm{T}}\ln \mathbf{p}^*(x)\overrightarrow{v}\right) f\right]^{4},$$

[4] Here we define the $\nabla_x^{\mathrm{T}}$ operator over a vector field $\overrightarrow{v}(x)$ as its divergence: $\nabla_x^{\mathrm{T}}\overrightarrow{v}(x) = \sum_i \frac{\partial v_i(x)}{\partial x_i}$.

the adjoint operator can be expressed as:

$$(\nabla_x)^* = -\nabla_x^{\mathrm{T}} - \nabla_x^{\mathrm{T}} \ln \mathbf{p}^*(x) = \left(-\nabla_\theta^{\mathrm{T}} + \nabla^{\mathrm{T}} U(\theta), -\nabla_r^{\mathrm{T}} + \xi r^{\mathrm{T}}\right).$$

The vector flow $v_t$ can also be expressed in terms of $h(\mathbf{p}_t)$ as:

$$v_t = -2(D(x) + Q(x))\nabla_x \ln h = -\frac{2}{h}(D(x) + Q(x))\nabla_x h.$$

Therefore,

$$\begin{align}
&\mathbb{E}_{p_t}\left[\left\langle \nabla_x \frac{\delta \mathcal{L}}{\delta \mathbf{p}_t}, v_t \right\rangle\right] \notag \\
&= -4\mathbb{E}_{p^*}\left[\langle \nabla_x h, (D(x) + Q(x))\nabla_x h\rangle\right] \tag{40} \\
&\quad - 8\mathbb{E}_{p^*}\left[\langle \nabla_x (\nabla_x)^* S \nabla_x h, (D(x) + Q(x))\nabla_x h\rangle\right] \tag{41} \\
&\quad + 8\mathbb{E}_{p^*}\left[\langle \nabla_x h, (D(x) + Q(x))\nabla_x h\rangle \frac{(\nabla_x)^* S \nabla_x h}{h}\right]. \tag{42}
\end{align}$$

For Line (40),

$$-4\mathbb{E}_{p^*}\left[\langle \nabla_x h, (D(x) + Q(x))\nabla_x h\rangle\right] = -4\gamma \mathbb{E}_{p^*}\left[\|\nabla_r h\|^2\right] = -\gamma \mathbb{E}_{\mathbf{p}_t}\left[\left\|\nabla_r \ln \frac{\mathbf{p}_t}{\mathbf{p}^*}\right\|^2\right],$$

same as in Eq. (9).

For Line (42),

$$\begin{align}
&8\mathbb{E}_{p^*}\left[\langle \nabla_x h, (D(x) + Q(x))\nabla_x h\rangle \frac{(\nabla_x)^* S \nabla_x h}{h}\right] \notag \\
&= 8\gamma \mathbb{E}_{p^*}\left[\frac{1}{h}\langle \nabla_r h, \nabla_r h\rangle (\nabla_x)^* S \nabla_x h\right] \notag \\
&= 8\gamma \mathbb{E}_{p^*}\left[\left\langle \frac{1}{h}\nabla_x \|\nabla_r h\|^2 - \frac{1}{h^2}\|\nabla_r h\|^2 \nabla_x h, S\nabla_x h\right\rangle\right] \notag \\
&= 16\gamma \mathbb{E}_{p^*}\left[\left\langle \frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h, S\nabla_x \nabla_r^{\mathrm{T}} h\right\rangle_F\right] - 8\gamma \mathbb{E}_{p^*}\left[\left\langle \frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h, S\frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h\right\rangle_F\right]. \tag{43}
\end{align}$$

Next we focus on Line (41).

**Lemma 9.**

$$\begin{align}
&-8\mathbb{E}_{p^*}\left[\langle \nabla_x (\nabla_x)^* S \nabla_x h, (D(x) + Q(x))\nabla_x h\rangle\right] \notag \\
&= -8\gamma \mathbb{E}_{p^*}\left[\langle \nabla_x \nabla_r h, S\nabla_x \nabla_r h\rangle_F\right] \tag{44} \\
&\quad - 4a\xi \mathbb{E}_{p^*}\left[||\nabla_\theta h||^2\right] \notag \\
&\quad - 4\mathbb{E}_{p^*}\left[\left\langle \nabla_r h, \left(2c\gamma\xi \mathrm{I} - a\nabla^2 U(\theta)\right)\nabla_r h\right\rangle\right] \notag \\
&\quad - 4\mathbb{E}_{p^*}\left[\left\langle \nabla_\theta h, \left((c\xi - a\gamma\xi)\mathrm{I} - 2b\nabla^2 U(\theta)\right)\nabla_r h\right\rangle\right]. \tag{45}
\end{align}$$

Then Line (44) combines with Eq. (43):

$$\begin{aligned}
&-8\gamma\mathbb{E}_{p^*}\left[\langle \nabla_x\nabla_r h, S\nabla_x\nabla_r h\rangle_F\right]\\
&+16\gamma\mathbb{E}_{p^*}\left[\left\langle \frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h, S\nabla_x\nabla_r h\right\rangle_F\right]\\
&-8\gamma\mathbb{E}_{p^*}\left[\left\langle \frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h, S\frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h\right\rangle_F\right]\\
&=-8\gamma\mathbb{E}_{p^*}\left[\left\langle \left(\nabla_x\nabla_r^{\mathrm{T}} h-\frac{\nabla_x h}{h}\nabla_r h\right), S\left(\nabla_x\nabla_r^{\mathrm{T}} h-\frac{\nabla_x h}{h}\nabla_r h\right)\right\rangle_F\right]
\end{aligned}$$

Therefore, Lines (40)–(42) sum up to be:

$$\begin{aligned}
&\mathbb{E}_{p_t}\left[\left\langle \nabla_x\frac{\delta L}{\delta \mathbf{p}_t}, v_t\right\rangle\right]\\
&=-8\gamma\mathbb{E}_{p^*}\left[\left\langle \left(\nabla_x\nabla_r h-\frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h\right), S\left(\nabla_x\nabla_r h-\frac{\nabla_x h}{h}\nabla_r^{\mathrm{T}} h\right)\right\rangle_F\right]\\
&-4\mathbb{E}_{p^*}\left\langle \begin{pmatrix}\nabla_\theta h\\ \nabla_r h\end{pmatrix}, M_C\begin{pmatrix}\nabla_\theta h\\ \nabla_r h\end{pmatrix}\right\rangle\\
&=-8\gamma\mathbb{E}_{\mathbf{p}_t}\left[\langle \nabla_x\nabla_r\ln h, S\nabla_x\nabla_r\ln h\rangle_F\right]-4\mathbb{E}_{\mathbf{p}_t}\left[\langle \nabla_x\ln h, M_C\nabla_x\ln h\rangle_F\right]\\
&=-2\gamma\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x\nabla_r\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), S\nabla_x\nabla_r\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right]\\
&-\mathbb{E}_{\mathbf{p}_t}\left[\left\langle \nabla_x\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right), M_C\nabla_x\ln\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)\right\rangle_F\right],
\end{aligned}$$

where

$$\begin{aligned}
M_C&=\begin{pmatrix}\frac{a}{2}\xi\cdot\mathrm{I} & \frac{c+a\gamma}{2}\xi\cdot\mathrm{I}-\frac{b}{2}\nabla^2U(\theta)\\ \frac{c+a\gamma}{2}\xi\cdot\mathrm{I}-\frac{b}{2}\nabla^2U(\theta) & \gamma\left(2c\xi+1\right)\mathrm{I}-\frac{a}{2}\nabla^2U(\theta)\end{pmatrix}\\
&=\begin{pmatrix}\mathrm{I} & 4\cdot\mathrm{I}-\frac{1}{8}\frac{\nabla^2U(\theta)}{L_G}\\ 4\cdot\mathrm{I}-\frac{1}{8}\frac{\nabla^2U(\theta)}{L_G} & 18\cdot\mathrm{I}-\frac{1}{2}\frac{\nabla^2U(\theta)}{L_G}\end{pmatrix}.
\end{aligned}\tag{46}$$

■

**Proof of Lemma 8** We aim to prove that

$$\begin{aligned}
M_C&=\begin{pmatrix}\frac{a}{2}\xi\cdot\mathrm{I} & \frac{c+a\gamma}{2}\xi\cdot\mathrm{I}-\frac{b}{2}\nabla^2U(\theta)\\ \frac{c+a\gamma}{2}\xi\cdot\mathrm{I}-\frac{b}{2}\nabla^2U(\theta) & \gamma\left(2c\xi+1\right)\mathrm{I}-\frac{a}{2}\nabla^2U(\theta)\end{pmatrix}\\
&\succeq\lambda\left(S+\frac{1}{2\rho}\mathrm{I}\right)=\lambda\begin{pmatrix}\left(b+\frac{1}{2\rho}\right)\mathrm{I} & \frac{a}{2}\mathrm{I}\\ \frac{a}{2}\mathrm{I} & \left(c+\frac{1}{2\rho}\right)\mathrm{I}\end{pmatrix},
\end{aligned}$$

for $a = \dfrac{1}{L_G}$, $b = \dfrac{1}{4L_G}$, $c = \dfrac{2}{L_G}$, $\gamma = 2$, $\xi = 2L_G$, and $\lambda = \dfrac{\rho}{10}$. That is equivalent to having:

$$\widehat{M_C} = \begin{pmatrix} \left(\frac{a}{2}\xi - \left(b + \frac{1}{2\rho}\right)\lambda\right)\mathrm{I} & \left(\frac{c+a\gamma}{2}\xi - \frac{a}{2}\lambda\right)\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) \\ \left(\frac{c+a\gamma}{2}\xi - \frac{a}{2}\lambda\right)\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) & \left(\gamma\left(2c\xi + 1\right) - \left(c + \frac{1}{2\rho}\right)\lambda\right)\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta) \end{pmatrix}$$

to be positive semidefinite.

Denote $\alpha = \dfrac{a}{2}\xi - \left(b + \dfrac{1}{2\rho}\right)\lambda$, $\beta = \dfrac{c+a\gamma}{2}\xi - \dfrac{a}{2}\lambda$, and $\sigma = \gamma\left(2c\xi+1\right) - \left(c + \dfrac{1}{2\rho}\right)\lambda$. We analyze the eigenvalues of $\widehat{M_C} = \begin{pmatrix} \alpha\mathrm{I} & \beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) \\ \beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) & \sigma\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta) \end{pmatrix}$ and ask when they will all be nonnegative. We write the characteristic equation for $\widehat{M}$:

$$\begin{aligned} \det\left[\widehat{M_C} - l\cdot\mathrm{I}\right] &= \det\left[\begin{pmatrix} (\alpha - l)\mathrm{I} & \beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) \\ \beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) & (\sigma - l)\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta) \end{pmatrix}\right] \\ &= \det\left[(\alpha - l)(\sigma - l)\mathrm{I} - \frac{a}{2}(\alpha - l)\nabla^2 U(\theta) - \left(\beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta)\right)^2\right] = 0, \end{aligned}$$

since $\beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta)$ and $(\sigma - l)\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta)$ commute. Diagonalizing $\nabla^2 U(\theta) = V^{-1}\Lambda V$, we obtain a set of independent equations based on each eigenvalue $\Lambda_j$ of $\nabla^2 U(\theta)$:

$$l^2 + \left(\frac{a}{2}\Lambda_j - \alpha - \sigma\right) l - \left(\frac{b^2}{4}\Lambda_j^2 + \left(\frac{a}{2}\alpha - b\beta\right)\Lambda_j + \beta^2 - \alpha\sigma\right) = 0.$$

To guarantee that $l \geq 0$, we need that $\forall \Lambda_j \in [-L_G, L_G]$,

$$\begin{cases} \frac{a}{2}\Lambda_j - \alpha - \sigma \leq 0 \\ \frac{b^2}{4}\Lambda_j^2 + \left(\frac{a}{2}\alpha - b\beta\right)\Lambda_j + \beta^2 - \alpha\sigma \leq 0 \end{cases}.$$

Since the linear function $\frac{a}{2}\Lambda_j - \alpha - \sigma$ of $\Lambda_j$ is increasing; the quadratic function $\frac{b^2}{4}\Lambda_j^2 + \left(\frac{a}{2}\alpha - b\beta\right)\Lambda_j + \beta^2 - \alpha\sigma$ of $\Lambda_j$ is convex, we simply need the inequality to be satisfied at the end points:

$$\begin{cases} \frac{a}{2}L_G - \alpha - \sigma \leq 0 \\ \frac{b^2}{4}L_G^2 - \left(\frac{a}{2}\alpha - b\beta\right)L_G + \beta^2 - \alpha\sigma \leq 0 \\ \frac{b^2}{4}L_G^2 + \left(\frac{a}{2}\alpha - b\beta\right)L_G + \beta^2 - \alpha\sigma \leq 0 \end{cases}.$$

We verify these inequalities by plugging in the setting of $a = \dfrac{1}{L_G}$, $b = \dfrac{1}{4L_G}$, $c = \dfrac{2}{L_G}$, $\gamma = 2$, $\xi = 2L_G$, and

$\lambda = \dfrac{\rho}{10}$ in the definition of $\alpha$, $\beta$, and $\sigma$. We obtain:

$$\begin{cases} \dfrac{a}{2}L_G - \alpha - \sigma = -\dfrac{92}{5} + \dfrac{9\rho}{40L_G} \leq 0 \\ \dfrac{b^2}{4}L_G^2 - \left(\dfrac{a}{2}\alpha - b\beta\right) L_G + \beta^2 - \alpha\sigma = -\dfrac{819}{1600} + \dfrac{191\rho}{800L_G} - \dfrac{\rho^2}{400L_G^2} \leq 0 \\ \dfrac{b^2}{4}L_G^2 + \left(\dfrac{a}{2}\alpha - b\beta\right) L_G + \beta^2 - \alpha\sigma = -\dfrac{2499}{1600} + \dfrac{191\rho}{800L_G} - \dfrac{\rho^2}{400L_G^2} \leq 0 \end{cases} .$$

Therefore, $M_C \succeq \lambda \left(S + \dfrac{1}{2\rho}\mathrm{I}_{2d\times 2d}\right)$ for $a = \dfrac{1}{L_G}$, $b = \dfrac{1}{4L_G}$, $c = \dfrac{2}{L_G}$, $\gamma = 2$, $\xi = 2L_G$, and $\lambda = \dfrac{L_G}{10}$. ■

## C.1 Supporting Proof for Lemma 7

**Proof of Lemma 9** First note that $-8\mathbb{E}_{p^*}\left[\langle \nabla_x(\nabla_x)^* S\nabla_x h, (D(x)+Q(x))\nabla_x h\rangle\right]$ separates into three terms:

$$\begin{aligned} &-8\mathbb{E}_{p^*}\left[\langle \nabla_x(\nabla_x)^* S\nabla_x h, (D(x)+Q(x))\nabla_x h\rangle\right] \\ &= -4a\mathbb{E}_{p^*}\left[\left\langle \nabla_x\big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), (D+Q)\nabla_x h\right\rangle\right] && (47) \\ &- 8b\mathbb{E}_{p^*}\left[\langle \nabla_x(\nabla_\theta)^*\nabla_\theta h, (D+Q)\nabla_x h\rangle\right] && (48) \\ &- 8c\mathbb{E}_{p^*}\left[\langle \nabla_x(\nabla_r)^*\nabla_r h, (D+Q)\nabla_x h\rangle\right]. && (49) \end{aligned}$$

We then deal with the three terms one by one.

1. For the cross term $-4a\mathbb{E}_{p^*}\left[\left\langle \nabla_x\big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), (D+Q)\nabla_x h\right\rangle\right]$ in Line 47,

$$\begin{aligned} &-\mathbb{E}_{p^*}\left[\left\langle \nabla_x\big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), (D+Q)\nabla_x h\right\rangle\right] \\ &= -\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} \big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), (D+Q)\begin{pmatrix} \nabla_\theta h \\ \nabla_r h \end{pmatrix} \right\rangle\right] \\ &= -\gamma\mathbb{E}_{p^*}\left[\left\langle \nabla_r\big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), \nabla_r h\right\rangle\right] && (50) \\ &- \mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} \big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), Q\begin{pmatrix} \nabla_\theta h \\ \nabla_r h \end{pmatrix} \right\rangle\right]. && (51) \end{aligned}$$

Here, $\nabla_\theta$ commutes with $\nabla_r$ and $(\nabla_r)^*$.

- Hence Line (50) equals:

$$\begin{aligned} &-\gamma\mathbb{E}_{p^*}\left[\langle \nabla_r(\nabla_\theta)^*\nabla_r h, \nabla_r h\rangle + \langle \nabla_r(\nabla_r)^*\nabla_\theta h, \nabla_r h\rangle\right] \\ &= -\gamma\mathbb{E}_{p^*}\left[\langle \nabla_r h, \nabla_\theta(\nabla_r)^*\nabla_r h\rangle + \langle \nabla_\theta h, \nabla_r(\nabla_r)^*\nabla_r h\rangle\right] \\ &= -\gamma\mathbb{E}_{p^*}\left[\langle \nabla_r h, (\nabla_r)^*\nabla_r\nabla_\theta h\rangle + \langle \nabla_\theta h, \nabla_r(\nabla_r)^*\nabla_r h\rangle\right] \text{ [5]} \\ &= -\gamma\mathbb{E}_{p^*}\left[\left\langle \nabla_\theta h, \big((\nabla_r)^*\nabla_r + \nabla_r(\nabla_r)^*\big)\nabla_r h\right\rangle\right]. \end{aligned}$$

We make use of the commutator of $\nabla_r$ and $(\nabla_r)^*$, $[\nabla_r, (\nabla_r)^*]\overrightarrow{v} = \nabla_r(\nabla_r)^*\overrightarrow{v}(x) - (\nabla_r)^*\nabla_r\overrightarrow{v}(x) = -\nabla_r\nabla_r^{\mathrm{T}}\overrightarrow{v} + \xi\overrightarrow{v} + \nabla_r^{\mathrm{T}}\nabla_r\overrightarrow{v}$, and simplify Line (50):

$$\begin{aligned} &-\gamma\mathbb{E}_{p^*}\left[\langle \nabla_r(\nabla_\theta)^*\nabla_r h, \nabla_r h\rangle + \langle \nabla_r(\nabla_r)^*\nabla_\theta h, \nabla_r h\rangle\right] \\ &= -\gamma\mathbb{E}_{p^*}\left[\left\langle \nabla_\theta h, \big(2(\nabla_r)^*\nabla_r + [\nabla_r, (\nabla_r)^*]\big)\nabla_r h\right\rangle\right] \\ &= -\gamma\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, 2(\nabla_r)^*\nabla_r\nabla_r h + \xi\nabla_r h\rangle\right] \\ &= -2\gamma\mathbb{E}_{p^*}\left[\langle \nabla_r\nabla_\theta h, \nabla_r\nabla_r h\rangle_F\right] - \gamma\xi\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, \nabla_r h\rangle\right], \end{aligned}$$

[5] Here $(\nabla_r)^*\nabla_r\nabla_\theta h$ is a column vector with its elements defined as: $\big((\nabla_r)^*\nabla_r\nabla_\theta h\big)_i = \sum_j \left(\frac{\partial}{\partial r_j}\right)^* \frac{\partial}{\partial r_j}\frac{\partial}{\partial \theta_i} h$.

where we have used $\langle\cdot,\cdot\rangle_F$ to also denote Frobenius inner product between matrices.

- Line (51) can be simplified by using the representation of the vector flow in Eq. (7):

$$\begin{aligned}
&-\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} \big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), Q \begin{pmatrix} \nabla_\theta h \\ \nabla_r h \end{pmatrix} \right\rangle\right] \\
&= -\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} \big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), Q \begin{pmatrix} \nabla U(\theta) \\ \xi r \end{pmatrix} \frac{h}{2} \right\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} h, \begin{pmatrix} \nabla_r \\ \nabla_\theta \end{pmatrix} \big(\xi r^{\mathrm{T}}\nabla_\theta h - \nabla^{\mathrm{T}} U(\theta)\nabla_r h\big) \right\rangle\right]. \qquad (52)
\end{aligned}$$

Denote $B[h] = \xi r^{\mathrm{T}}\nabla_\theta h - \nabla^{\mathrm{T}} U(\theta)\nabla_r h$, then $B$ is an anti-symmetric operator: $B^*[h] = -B[h]$. Then Eq. (52) can be further simplified:

$$\begin{aligned}
&-\frac{1}{2}\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} h, \begin{pmatrix} \nabla_r \\ \nabla_\theta \end{pmatrix} \big(\xi r^{\mathrm{T}}\nabla_\theta h - \nabla^{\mathrm{T}} U(\theta)\nabla_r h\big) \right\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} h, \begin{pmatrix} \nabla_r \\ \nabla_\theta \end{pmatrix} B[h] \right\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, \nabla_r B[h]\rangle + \langle \nabla_r h, \nabla_\theta B[h]\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, \nabla_r B[h]\rangle + \langle \nabla_r h, B\nabla_\theta [h]\rangle + \langle \nabla_r h, [\nabla_\theta, B][h]\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, \nabla_r B[h]\rangle - \langle B\nabla_r h, \nabla_\theta [h]\rangle + \langle \nabla_r h, [\nabla_\theta, B][h]\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, [\nabla_r, B][h]\rangle + \langle \nabla_r h, [\nabla_\theta, B][h]\rangle\right]. \qquad (53)
\end{aligned}$$

Since $[\nabla_r, B][h] = \xi\nabla_\theta h$ and $[\nabla_\theta, B][h] = -\nabla^2 U(\theta)\nabla_r h$, Eq. (53) becomes

$$\begin{aligned}
&-\frac{1}{2}\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, [\nabla_r, B][h]\rangle + \langle \nabla_r h, [\nabla_\theta, B][h]\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\xi\,\langle \nabla_\theta h, \nabla_\theta h\rangle - \left\langle \nabla_r h, \nabla^2 U(\theta)\nabla_r h\right\rangle\right].
\end{aligned}$$

Therefore, Line (51) is

$$\begin{aligned}
&-\mathbb{E}_{p^*}\left[\left\langle \begin{pmatrix} \nabla_\theta \\ \nabla_r \end{pmatrix} \big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), Q \begin{pmatrix} \nabla_\theta h \\ \nabla_r h \end{pmatrix} \right\rangle\right] \\
&= -\frac{1}{2}\mathbb{E}_{p^*}\left[\xi\,\langle \nabla_\theta h, \nabla_\theta h\rangle - \left\langle \nabla_r h, \nabla^2 U(\theta)\nabla_r h\right\rangle\right].
\end{aligned}$$

Summing up Lines (50) and (51),

$$\begin{aligned}
&-\mathbb{E}_{p^*}\left[\left\langle \nabla_x\big((\nabla_\theta)^*\nabla_r h + (\nabla_r)^*\nabla_\theta h\big), (D+Q)\nabla_x h\right\rangle\right] \\
&= -2\gamma\mathbb{E}_{p^*}\left[\langle \nabla_\theta\nabla_r h, \nabla_r\nabla_r h\rangle_F\right] \\
&- \gamma\xi\mathbb{E}_{p^*}\left[\langle \nabla_\theta h, \nabla_r h\rangle\right] - \frac{\xi}{2}\mathbb{E}_{p^*}\left[||\nabla_\theta h||^2\right] + \frac{1}{2}\mathbb{E}_{p^*}\left[\left\langle \nabla_r h, \nabla^2 U(\theta)\nabla_r h\right\rangle\right].
\end{aligned}$$

2. For $-8b\mathbb{E}_{p^*}\left[\langle\nabla_x(\nabla_\theta)^*\nabla_\theta h,(D+Q)\nabla_x h\rangle\right]$ in Line 48,

$$
\begin{aligned}
&-2\mathbb{E}_{p^*}\left[\langle\nabla_x(\nabla_\theta)^*\nabla_\theta h,(D+Q)\nabla_x h\rangle\right]\\
&=-2\mathbb{E}_{p^*}\left[\left\langle\begin{pmatrix}\nabla_\theta\\ \nabla_r\end{pmatrix}(\nabla_\theta)^*\nabla_\theta h,(D+Q)\begin{pmatrix}\nabla_\theta h\\ \nabla_r h\end{pmatrix}\right\rangle\right]\\
&=-2\gamma\mathbb{E}_{p^*}\left[\langle\nabla_r(\nabla_\theta)^*\nabla_\theta h,\nabla_r h\rangle\right]-\mathbb{E}_{p^*}\left[\langle\nabla_\theta h,\nabla_\theta B[h]\rangle\right]\\
&=-2\gamma\mathbb{E}_{p^*}\left[\langle\nabla_r(\nabla_\theta)^*\nabla_\theta h,\nabla_r h\rangle\right]-\mathbb{E}_{p^*}\left[\langle\nabla_\theta h,B\nabla_\theta h+[\nabla_\theta,B][h]\rangle\right]\\
&=-2\gamma\mathbb{E}_{p^*}\left[\langle\nabla_\theta\nabla_r h,\nabla_\theta\nabla_r h\rangle_F\right]+\mathbb{E}_{p^*}\left[\left\langle\nabla_\theta h,\nabla^2U(\theta)\nabla_r h\right\rangle\right].
\end{aligned}
$$

3. For $-8c\mathbb{E}_{p^*}\left[\langle\nabla_x(\nabla_r)^*\nabla_r h,(D+Q)\nabla_x h\rangle\right]$ in Line 49,

$$
\begin{aligned}
&-2\mathbb{E}_{p^*}\left[\langle\nabla_x(\nabla_r)^*\nabla_r h,(D+Q)\nabla_x h\rangle\right]\\
&=-2\mathbb{E}_{p^*}\left[\left\langle\begin{pmatrix}\nabla_\theta\\ \nabla_r\end{pmatrix}(\nabla_r)^*\nabla_r h,(D+Q)\begin{pmatrix}\nabla_\theta h\\ \nabla_r h\end{pmatrix}\right\rangle\right]\\
&=-2\gamma\mathbb{E}_{p^*}\left[\langle\nabla_r(\nabla_r)^*\nabla_r h,\nabla_r h\rangle\right]-\mathbb{E}_{p^*}\left[\langle\nabla_r h,\nabla_r B[h]\rangle\right]\\
&=-2\gamma\mathbb{E}_{p^*}\left[\langle\left((\nabla_r)^*\nabla_r+[\nabla_r,(\nabla_r)^*]\right)\nabla_r h,\nabla_r h\rangle\right]\\
&-\mathbb{E}_{p^*}\left[\langle\nabla_r h,B\nabla_r h+[\nabla_r,B][h]\rangle\right]\\
&=-2\gamma\mathbb{E}_{p^*}\left[\langle\nabla_r\nabla_r h,\nabla_r\nabla_r h\rangle_F\right]\\
&-2\gamma\xi\mathbb{E}_{p^*}\left[\langle\nabla_r h,\nabla_r h\rangle\right]-\xi\mathbb{E}_{p^*}\left[\langle\nabla_\theta h,\nabla_r h\rangle\right].
\end{aligned}
$$

Summing everything up,

$$
\begin{aligned}
&-8\mathbb{E}_{p^*}\left[\langle\nabla_x(\nabla_x)^*S\nabla_x h,(D(x)+Q(x))\nabla_x h\rangle\right]\\
&=-8a\gamma\mathbb{E}_{p^*}\left[\langle\nabla_\theta\nabla_r h,\nabla_r\nabla_r h\rangle_F\right] &(54)\\
&-8b\gamma\mathbb{E}_{p^*}\left[\langle\nabla_\theta\nabla_r h,\nabla_\theta\nabla_r h\rangle_F\right] &(55)\\
&-8c\gamma\mathbb{E}_{p^*}\left[\langle\nabla_r\nabla_r h,\nabla_r\nabla_r h\rangle_F\right] &(56)\\
&-2a\xi\mathbb{E}_{p^*}\left[||\nabla_\theta h||^2\right]\\
&-4\mathbb{E}_{p^*}\left[\left\langle\nabla_r h,\left(2c\gamma\xi\mathrm{I}-\frac{a}{2}\nabla^2U(\theta)\right)\nabla_r h\right\rangle\right]\\
&-4\mathbb{E}_{p^*}\left[\left\langle\nabla_\theta h,\left((c\xi+a\gamma\xi)\mathrm{I}-b\nabla^2U(\theta)\right)\nabla_r h\right\rangle\right].
\end{aligned}
$$

For Lines (54)–(56),

$$
\begin{aligned}
&-a\mathbb{E}_{p^*}\left[\langle\nabla_\theta\nabla_r h,\nabla_r\nabla_r h\rangle_F\right]\\
&-b\mathbb{E}_{p^*}\left[\langle\nabla_\theta\nabla_r h,\nabla_\theta\nabla_r h\rangle_F\right]\\
&-c\mathbb{E}_{p^*}\left[\langle\nabla_r\nabla_r h,\nabla_r\nabla_r h\rangle_F\right]\\
&=-\gamma\mathbb{E}_{p^*}\left[\langle\nabla_x\nabla_r h,S\nabla_x\nabla_r h\rangle_F\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&-8\mathbb{E}_{p^*}\left[\langle\nabla_x(\nabla_x)^*S\nabla_x h,(D(x)+Q(x))\nabla_x h\rangle\right]\\
&=-8\gamma\mathbb{E}_{p^*}\left[\langle\nabla_x\nabla_r h,S\nabla_x\nabla_r h\rangle_F\right] &(57)\\
&-2a\xi\mathbb{E}_{p^*}\left[||\nabla_\theta h||^2\right]\\
&-4\mathbb{E}_{p^*}\left[\left\langle\nabla_r h,\left(2c\gamma\xi\mathrm{I}-\frac{a}{2}\nabla^2U(\theta)\right)\nabla_r h\right\rangle\right]\\
&-4\mathbb{E}_{p^*}\left[\left\langle\nabla_\theta h,\left((c\xi+a\gamma\xi)\mathrm{I}-b\nabla^2U(\theta)\right)\nabla_r h\right\rangle\right]. &(58)
\end{aligned}
$$

■

# D Discretization Error

**Proof of Lemma 3** As in the continuous case, define $h = \sqrt{\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}}$, and denote $a = \frac{1}{L_G}$, $b = \frac{1}{4L_G}$, $c = \frac{2}{L_G}$. First note that

$$\begin{aligned}
&\int \left\langle \nabla_r \frac{\delta L}{\delta \mathbf{p}_t}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \left( 2 \ln h + 4 \frac{\nabla_x^* S \nabla_x h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau .
\end{aligned}$$

We prove in the following that

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_x^* S \nabla_x h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \qquad (59) \\
&= \int \left\langle \nabla_x \nabla_r \ln h, S \nabla_x \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&+ \xi \int \left\langle \frac{a}{2} \nabla_\theta \ln h + c \nabla_r \ln h, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \left(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau .
\end{aligned}$$

Similar to the continuous case, the term in Line (59) separates into four terms:

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_x^* S \nabla_x h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= b \int \left\langle \nabla_r \left( \frac{\nabla_\theta^* \nabla_\theta h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \qquad (60) \\
&+ \frac{a}{2} \int \left\langle \nabla_r \left( \frac{\nabla_\theta^* \nabla_r h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \qquad (61) \\
&+ \frac{a}{2} \int \left\langle \nabla_r \left( \frac{\nabla_r^* \nabla_\theta h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \qquad (62) \\
&+ c \int \left\langle \nabla_r \left( \frac{\nabla_r^* \nabla_r h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau . \qquad (63)
\end{aligned}$$

We first simplify Lines (60) and (61) and then deal with Lines (62) and (63).

1. For Lines (60) and (61):

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_\theta^* \nabla_\# h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right] \right\rangle \,\mathrm{d}x_\tau \\
&= \int \left\langle h\nabla_r \nabla_\theta^* \nabla_\# h - \nabla_r h \nabla_\theta^* \nabla_\# h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \right\rangle \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \nabla_\# h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \nabla_\theta^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&- \int \left\langle \nabla_\theta \nabla_r h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \nabla_\#^T h \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&+ \int \left\langle h\nabla_r \nabla_\# h - \nabla_r h \nabla_\#^T h, \nabla_\theta \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \nabla_\# h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \nabla_\theta^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&- \int \left\langle \nabla_\theta \nabla_r h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \nabla_\#^T h \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r \nabla_\# \ln h, \nabla_\theta \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau)\,\mathrm{d}x_\tau.
\end{aligned}$$

When $\# = \theta$,

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_\theta^* \nabla_\theta h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right] \right\rangle \,\mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \nabla_\theta \ln h, \nabla_\theta \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau)\,\mathrm{d}x_\tau.
\end{aligned}$$

When $\# = r$,

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_\theta^* \nabla_r h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right] \right\rangle \,\mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r^2 h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \nabla_\theta^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&- \int \left\langle \nabla_\theta \nabla_r h, \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \nabla_r^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau)\,\mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r^2 \ln h, \nabla_\theta \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau)\,\mathrm{d}x_\tau.
\end{aligned}$$

2. For Lines (62) and (63):

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_r^* \nabla_\# h}{h} \right), \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= \int \left\langle h \nabla_r \nabla_r^* \nabla_\# h - \nabla_r h \nabla_r^* \nabla_\# h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right\rangle \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&= \xi \int \left\langle \frac{\nabla_\# h}{h}, \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r \nabla_\# h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \nabla_r^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&- \int \left\langle \nabla_r^2 h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \nabla_\#^T h \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&+ \int \left\langle h \nabla_r \nabla_\# h - \nabla_r h \nabla_\#^T h, \nabla_r \left( \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&= \xi \int \left\langle \nabla_\# \ln h, \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r \nabla_\# h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \nabla_r^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&- \int \left\langle \nabla_r^2 h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \nabla_\#^T h \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r \nabla_\# \ln h, \nabla_r \left( \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau .
\end{aligned}$$

When $\# = \theta$,

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_r^* \nabla_\theta h}{h} \right), \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= \xi \int \left\langle \nabla_\theta \ln h, \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r \nabla_\theta h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \nabla_r^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&- \int \left\langle \nabla_r^2 h, \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \nabla_\theta^{\mathrm{T}} h \right\rangle_F \mathbf{p}^*(x_\tau) \, \mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r \nabla_\theta \ln h, \nabla_r \left( \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau .
\end{aligned}$$

When $\# = r$,

$$\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_r^* \nabla_r h}{h} \right), \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= \xi \int \left\langle \nabla_r \ln h, \mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau \\
&+ \int \left\langle \nabla_r^2 \ln h, \nabla_r \left( \frac{\mathbb{E}_{x_{kh}\sim \mathbf{p}(x_{kh})} \left[ \big( \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \big) \mathbf{p}(x_\tau | x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau .
\end{aligned}$$

Therefore, Lines (60)–(63) combines to be:

$$
\begin{aligned}
&\int \left\langle \nabla_r \left( \frac{\nabla_x^* S \nabla_x h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= b \int \left\langle \nabla_r \nabla_\theta \ln h, \nabla_\theta \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&+ \frac{a}{2} \int \left\langle \nabla_r^2 \ln h, \nabla_\theta \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&+ \frac{a}{2} \int \left\langle \nabla_r \nabla_\theta \ln h, \nabla_r \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&+ c \int \left\langle \nabla_r^2 \ln h, \nabla_r \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&+ \xi \int \left\langle \frac{a}{2} \nabla_\theta \ln h + c \nabla_r \ln h, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau \\
&= \int \left\langle \nabla_x \nabla_r \ln h, S \nabla_x \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&+ \xi \int \left\langle \frac{a}{2} \nabla_\theta \ln h + c \nabla_r \ln h, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau .
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\int \left\langle \nabla_r \frac{\delta L}{\delta \mathbf{p}_t}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \left( 2 \ln h + 4 \frac{\nabla_x^* S \nabla_x h}{h} \right), \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle \, \mathrm{d}x_\tau \\
&+ 2 \int \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S \nabla_x \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \qquad (64) \\
&+ 2\xi \int \left\langle \frac{a}{2} \nabla_\theta \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)} + c \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right] \right\rangle_F \, \mathrm{d}x_\tau .
\end{aligned}
$$

It can be seen that the expectation in Line (64) can be rewritten as $x_{kh}$ conditioning on $x_\tau$:

$$
\begin{aligned}
&\int \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S \nabla_x \left( \frac{\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})} \left[ \big(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\big) \mathbf{p}(x_\tau|x_{kh}) \right]}{\mathbf{p}_\tau(x_\tau)} \right) \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau \\
&= \int \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S \nabla_{x_\tau} \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)} \left[ \nabla U(\theta_\tau) - \nabla U(\theta_{kh}) \right] \right\rangle_F \mathbf{p}_\tau(x_\tau) \, \mathrm{d}x_\tau .
\end{aligned}
$$

Therefore,

$$\begin{aligned}
&\int \left\langle \nabla_r \frac{\delta L}{\delta \mathbf{p}_t}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \, \mathrm{d}x_\tau \\
&= \int \left\langle \nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \, \mathrm{d}x_\tau \\
&+ 2\int \left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right]\right\rangle_F \mathbf{p}_\tau(x_\tau)\, \mathrm{d}x_\tau \\
&+ 2\xi\int \left\langle \frac{a}{2}\nabla_\theta \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)} + c\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle_F \, \mathrm{d}x_\tau \\
&= a\xi\int \left\langle \nabla_\theta \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle_F \, \mathrm{d}x_\tau \\
&+ (2c\xi+1)\int \left\langle \nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh})}\left[\big(\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\big)\mathbf{p}(x_\tau|x_{kh})\right]\right\rangle \, \mathrm{d}x_\tau \\
&+ 2\int \left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right]\right\rangle_F \mathbf{p}_\tau(x_\tau)\, \mathrm{d}x_\tau .
\end{aligned}$$

■

**Proof of Lemma 4** We first explicitly calculate $\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right]$ in the following Lemma 10. To obtain the expression, we use synchronous coupling of the trajectories of underdamped Langevin algorithm with infinitesimally different initial conditions.

**Lemma 10.** *Denote $\nu = \tau - kh \le h$ and*

$$\eta = \frac{1}{\gamma}\left(\frac{e^{\gamma\xi(\tau-kh)}\left(1-e^{-\gamma\xi(\tau-kh)}\right)^2}{\gamma\xi} - \left((\tau-kh) - \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma\xi}\right)\right) \sim \mathcal{O}\left(\xi\nu^2\right).$$

*Then for $\nu \le \dfrac{1}{8L_G}$ (and $\gamma = 2$, and $\xi = 2L_G$),*

$$\begin{aligned}
&\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right] \\
&= \mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\begin{pmatrix} \left(\nabla^2 U(\theta_\tau)-\nabla^2 U(\theta_{kh})\right) + \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}.
\end{aligned} \tag{65}$$

Taking Lemma 10 as given, we can separate Term (19c) into two:

$$\begin{aligned}
&\int \left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right]\right\rangle_F \mathbf{p}_\tau(x_\tau)\, \mathrm{d}x_\tau \\
&= \int\int \left\langle S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \begin{pmatrix} \nabla^2 U(\theta_\tau)-\nabla^2 U(\theta_{kh}) \\ 0 \end{pmatrix}\right\rangle_F \mathbf{p}(x_{kh}|x_\tau)\mathbf{p}_\tau(x_\tau)\, \mathrm{d}x_{kh}\mathrm{d}x_\tau && \text{(66a)} \\
&+ \int\int \left\langle S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}\right\rangle_F \\
&\qquad \cdot\, \mathbf{p}(x_{kh}|x_\tau)\mathbf{p}_\tau(x_\tau)\, \mathrm{d}x_{kh}\mathrm{d}x_\tau . && \text{(66b)}
\end{aligned}$$

We then make use of the properties of Frobenius inner product to upper bound Terms (66a) and (66b) by the Frobenius norms:

$$\begin{aligned}
&\left\langle S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, A_{2d\times d}\right\rangle_F \\
&= \left\langle \sqrt{S}\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \sqrt{S}A_{2d\times d}\right\rangle_F \\
&\leq \alpha \left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\rangle_F + \frac{1}{4\alpha}\left\langle A_{2d\times d}, SA_{2d\times d}\right\rangle_F .
\end{aligned}$$

As a result, we obtain that for Term (66a),

$$\begin{aligned}
&\left\langle S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \begin{pmatrix} \nabla^2 U(\theta_\tau) - \nabla^2 U(\theta_{kh}) \\ 0 \end{pmatrix}\right\rangle_F \\
&\leq \frac{\gamma}{2}\left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\rangle_F + \frac{b}{2\gamma}\left\|\nabla^2 U(\theta_\tau) - \nabla^2 U(\theta_{kh})\right\|_F^2 ;
\end{aligned}$$

and for Term (66b),

$$\begin{aligned}
&\left\langle S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, \begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}\right\rangle_F \\
&\leq \frac{\gamma}{2}\left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\rangle_F \\
&+ \frac{1}{2\gamma}\Bigg\langle \begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}, \\
&\qquad S\begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}\Bigg\rangle_F \\
&\leq \frac{\gamma}{2}\left\langle \nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_x\nabla_r \ln\frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\rangle_F \\
&+ \frac{(b+c)d}{2\gamma}\left\|\begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}\right\|_2^2 . \qquad (67)
\end{aligned}$$

To obtain the final bound, we simplify Eq. (67) by demonstrating the following fact.

**Fact 2.** *For* $0 \leq \nu \leq \min\left\{\dfrac{1}{\gamma\xi}, \dfrac{1}{\sqrt{2eL_G\xi}}\right\}$,

$$\left\|\begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}\right\|_2 \leq 4e\max\{L_G^2\xi\nu^2, L_G\xi\nu\}.$$

Since $\nu \leq \dfrac{1}{8L_G} \leq \min\left\{\dfrac{1}{\gamma\xi}, \dfrac{1}{\sqrt{2eL_G\xi}}\right\}$, and $\left\|\nabla^2 U(\theta_\tau) - \nabla^2 U(\theta_{kh})\right\|_F \leq L_H\left\|\theta_\tau - \theta_{kh}\right\|$, we plug the above inequalities into Terms (66a) and (66b) and arrive at our conclusion:

$$
\begin{aligned}
&\int \left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right]\right\rangle_F \mathbf{p}_\tau(x_\tau)\ \mathrm{d}x_\tau \\
&\leq \gamma \mathbb{E}_{\mathbf{p}_\tau(x_\tau)}\left[\left\langle \nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}, S\nabla_x \nabla_r \ln \frac{\mathbf{p}_\tau(x_\tau)}{\mathbf{p}^*(x_\tau)}\right\rangle_F\right] \\
&+ \frac{2e(b+c)d}{\gamma}\max\{L_G^4\xi^2\nu^4, L_G^2\xi^2\nu^2\} + \frac{bL_H^2}{2\gamma}\mathbb{E}_{\mathbf{p}(x_{kh}|x_\tau)\mathbf{p}_\tau(x_\tau)}\left[\|\theta_\tau - \theta_{kh}\|^2\right].
\end{aligned}
$$

■

**Proof of Lemma 10** We study the following term with an arbitrary vector $v \in \mathbb{R}^{2d}$ (and denote $\hat{x}_n = \left(\hat{\theta}_n, \hat{r}_n\right) \in \mathbb{R}^{2d}$):

$$
\begin{aligned}
&v^{\mathrm{T}}\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right] \\
&= \lim_{h\to 0}\frac{1}{h}\mathbb{E}_{\substack{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau) \\ \hat{x}_n\sim\mathbf{p}(\hat{x}_n|x_\tau+hv)}}\left[\left(\nabla U(\theta_\tau + hv) - \nabla U(\hat{\theta}_n)\right) - \left(\nabla U(\theta_\tau) - \nabla U(\theta_{kh})\right)\right] \\
&= \lim_{h\to 0}\frac{1}{h}\mathbb{E}_{(x_{kh},\hat{x}_n)\sim\Gamma(\mathbf{p}(x_{kh}|x_\tau),\mathbf{p}(\hat{x}_n|x_\tau+hv))}\left[\left(\nabla U(\theta_\tau + hv) - \nabla U(\theta_\tau)\right) - \left(\nabla U(\hat{\theta}_n) - \nabla U(\theta_{kh})\right)\right],
\end{aligned}
$$

where $\Gamma\left(\mathbf{p}(x_{kh}|x_\tau), \mathbf{p}(\hat{x}_n|x_\tau + hv)\right)$ is any joint distribution of $x_{kh}$ and $\hat{x}_n$ with marginal distributions being $\mathbf{p}(x_{kh}|x_\tau)$ and $\mathbf{p}(\hat{x}_n|x_\tau + hv)$ – any coupling between the two random variables.

Recall from (34) that the relation between $x_\tau$ and $x_{kh}$ is:

$$
\left\{
\begin{aligned}
\theta_\tau &= \theta_{kh} + \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma}r_{kh} - \frac{1}{\gamma}\left((\tau - kh) - \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma\xi}\right)\nabla U(\theta_{kh}) + W_\theta \\
r_\tau &= r_{kh} - \left(1-e^{-\gamma\xi(\tau-kh)}\right)r_{kh} - \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma\xi}\nabla U(\theta_{kh}) + W_r
\end{aligned}
\right. , \tag{68}
$$

where $W_x^{\mathrm{T}} = \left(W_\theta^{\mathrm{T}}, W_r^{\mathrm{T}}\right)$ is the Gaussian random variable. It can be proven that for step size $\nu \leq h \leq \dfrac{1}{8L_G}$, $x_{kh}$ is uniquely determined given $x_\tau$ and $W_x$. Here we take the parallel coupling between $x_{kh}$ and $\hat{x}_n$. Namely, we take:

$$
\left\{
\begin{aligned}
\theta_\tau + hv_\theta &= \hat{\theta}_n + \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma}\hat{r}_n - \frac{1}{\gamma}\left((\tau - kh) - \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma\xi}\right)\nabla U(\hat{\theta}_n) + W_\theta \\
r_\tau + hv_r &= \hat{r}_n - \left(1-e^{-\gamma\xi(\tau-kh)}\right)\hat{r}_n - \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma\xi}\nabla U(\hat{\theta}_n) + W_r
\end{aligned}
\right. ,
$$

where the Gaussian random variable $W_x$ takes the same value as that in Eq. (68). Then we get that for any pair of $(x_{kh}, \hat{x}_n)$ following this joint law,

$$
\hat{\theta}_n - \theta_{kh} = hv_\theta + h\Delta(\bar{\theta}),
$$

where we define

$$
\Delta(\bar{\theta}) = \left(\left(\mathrm{I} + \eta\nabla^2 U(\bar{\theta})\right)^{-1} - \mathrm{I}\right)v_\theta - \frac{e^{\gamma\xi(\tau-kh)} - 1}{\gamma}\left(\mathrm{I} + \eta\nabla^2 U(\bar{\theta})\right)^{-1}v_r,
$$

$\bar{\theta}$ a convex combination of $\theta_{kh}$ and $\hat{\theta}_n$, and

$$
\eta = \frac{1}{\gamma}\left(\frac{e^{\gamma\xi(\tau-kh)}\left(1-e^{-\gamma\xi(\tau-kh)}\right)^2}{\gamma\xi} - \left((\tau - kh) - \frac{1-e^{-\gamma\xi(\tau-kh)}}{\gamma\xi}\right)\right) \sim \mathcal{O}(\xi\nu^2).
$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{(x_{kh},\hat{x}_n)\sim\Gamma(\mathbf{p}(x_{kh}|x_\tau),\mathbf{p}(\hat{x}_n|x_\tau+hv))}\left[\left(\nabla U(\theta_\tau+hv)-\nabla U(\theta_\tau)\right)-\left(\nabla U(\hat{\theta}_n)-\nabla U(\theta_{kh})\right)\right]\\
&=\mathbb{E}_{(x_{kh},\hat{x}_n)\sim\Gamma(\mathbf{p}(x_{kh}|x_\tau),\mathbf{p}(\hat{x}_n|x_\tau+hv))}\left[\nabla^2 U(\tilde{\theta})hv_\theta-\nabla^2 U(\bar{\theta})\left(\hat{\theta}_n-\theta_{kh}\right)\right]\\
&=\mathbb{E}_{\Gamma}\left[\left(\nabla^2 U(\tilde{\theta})-\nabla^2 U(\bar{\theta})\right)hv_\theta+h\nabla^2 U(\bar{\theta})\Delta(\bar{\theta})\right],
\end{aligned}$$

where $\tilde{\theta}$ is a convex combination of $\theta_\tau$ and $\theta_\tau+hv$. Taking the limit $h\to 0$, we have:

$$\begin{aligned}
&v^{\mathrm{T}}\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right]\\
&=\lim_{h\to 0}\frac{1}{h}\mathbb{E}_{\substack{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)\\ \hat{x}_n\sim\mathbf{p}(\hat{x}_n|x_\tau+hv)}}\left[\left(\nabla U(\theta_\tau+hv)-\nabla U(\hat{\theta}_n)\right)-\left(\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right)\right]\\
&=\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\left(\nabla^2 U(\theta_\tau)-\nabla^2 U(\theta_{kh})\right)v_\theta+\nabla^2 U(\theta_{kh})\Delta(\theta_{kh})\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\nabla_{x_\tau}\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left[\nabla U(\theta_\tau)-\nabla U(\theta_{kh})\right]\\
&=\mathbb{E}_{x_{kh}\sim\mathbf{p}(x_{kh}|x_\tau)}\left(\begin{array}{c}\left(\nabla^2 U(\theta_\tau)-\nabla^2 U(\theta_{kh})\right)+\nabla^2 U(\theta_{kh})\left(\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1}-\mathrm{I}\right)\\ -\dfrac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I}+\eta\nabla^2 U(\theta_{kh})\right)^{-1}\end{array}\right).
\end{aligned}$$

■

# E Overall Convergence of the Underdamped Langevin Algorithm

**Proof of Lemma 5** We aim to prove that

$$\begin{aligned}
M&=\begin{pmatrix}\dfrac{31}{64}a\xi\cdot\mathrm{I} & \dfrac{c+a\gamma}{2}\xi\cdot\mathrm{I}-\dfrac{b}{2}\nabla^2 U(\theta)\\ \dfrac{c+a\gamma}{2}\xi\cdot\mathrm{I}-\dfrac{b}{2}\nabla^2 U(\theta) & \dfrac{31}{32}\gamma\left(2c\xi+1\right)\mathrm{I}-\dfrac{a}{2}\nabla^2 U(\theta)\end{pmatrix}\\
&\succeq\lambda\left(S+\frac{1}{2\rho}\mathrm{I}\right)=\lambda\begin{pmatrix}\left(b+\dfrac{1}{2\rho}\right)\mathrm{I} & \dfrac{a}{2}\mathrm{I}\\ \dfrac{a}{2}\mathrm{I} & \left(c+\dfrac{1}{2\rho}\right)\mathrm{I}\end{pmatrix},
\end{aligned}$$

for $a=\dfrac{1}{L_G}$, $b=\dfrac{1}{4L_G}$, $c=\dfrac{2}{L_G}$, $\gamma=2$, $\xi=2L_G$, and $\lambda=\dfrac{\rho}{30}$. That is equivalent to having:

$$\widehat{M}=\begin{pmatrix}\left(\dfrac{31}{64}a\xi-\left(b+\dfrac{1}{2\rho}\right)\lambda\right)\mathrm{I} & \left(\dfrac{c+a\gamma}{2}\xi-\dfrac{a}{2}\lambda\right)\mathrm{I}-\dfrac{b}{2}\nabla^2 U(\theta)\\ \left(\dfrac{c+a\gamma}{2}\xi-\dfrac{a}{2}\lambda\right)\mathrm{I}-\dfrac{b}{2}\nabla^2 U(\theta) & \left(\dfrac{31}{32}\gamma\left(2c\xi+1\right)-\left(c+\dfrac{1}{2\rho}\right)\lambda\right)\mathrm{I}-\dfrac{a}{2}\nabla^2 U(\theta)\end{pmatrix}$$

to be positive semidefinite.

Denote $\alpha=\dfrac{31}{64}a\xi-\left(b+\dfrac{1}{2\rho}\right)\lambda$, $\beta=\dfrac{c+a\gamma}{2}\xi-\dfrac{a}{2}\lambda$, and $\sigma=\dfrac{31}{32}\gamma\left(2c\xi+1\right)-\left(c+\dfrac{1}{2\rho}\right)\lambda$. Then we analyze the eigenvalues of $\widehat{M}=\begin{pmatrix}\alpha\mathrm{I} & \beta\mathrm{I}-\dfrac{b}{2}\nabla^2 U(\theta)\\ \beta\mathrm{I}-\dfrac{b}{2}\nabla^2 U(\theta) & \sigma\mathrm{I}-\dfrac{a}{2}\nabla^2 U(\theta)\end{pmatrix}$ and ask when they will all be nonnegative. We write the characteristic equation for $\widehat{M}$:

$$\det\left[\widehat{M} - l\cdot \mathrm{I}\right] = \det\left[\begin{pmatrix} (\alpha - l)\mathrm{I} & \beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) \\ \beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta) & (\sigma - l)\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta) \end{pmatrix}\right]$$

$$= \det\left[(\alpha - l)(\sigma - l)\mathrm{I} - \frac{a}{2}(\alpha - l)\nabla^2 U(\theta) - \left(\beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta)\right)^2\right] = 0,$$

since $\beta\mathrm{I} - \frac{b}{2}\nabla^2 U(\theta)$ and $(\sigma - l)\mathrm{I} - \frac{a}{2}\nabla^2 U(\theta)$ commute. Diagonalizing $\nabla^2 U(\theta) = V^{-1}\Lambda V$, we obtain a set of independent equations based on each eigenvalue $\Lambda_j$ of $\nabla^2 U(\theta)$:

$$l^2 + \left(\frac{a}{2}\Lambda_j - \alpha - \sigma\right) l - \left(\frac{b^2}{4}\Lambda_j^2 + \left(\frac{a}{2}\alpha - b\beta\right)\Lambda_j + \beta^2 - \alpha\sigma\right) = 0.$$

To guarantee that $l \geq 0$, we need that $\forall \Lambda_j \in [-L_G, L_G]$,

$$\begin{cases} \frac{a}{2}\Lambda_j - \alpha - \sigma \leq 0 \\ \frac{b^2}{4}\Lambda_j^2 + \left(\frac{a}{2}\alpha - b\beta\right)\Lambda_j + \beta^2 - \alpha\sigma \leq 0 \end{cases}.$$

Since the linear function $\frac{a}{2}\Lambda_j - \alpha - \sigma$ of $\Lambda_j$ is increasing; the quadratic function $\frac{b^2}{4}\Lambda_j^2 + \left(\frac{a}{2}\alpha - b\beta\right)\Lambda_j + \beta^2 - \alpha\sigma$ of $\Lambda_j$ is convex, we simply need the inequality to satisfy at the end points:

$$\begin{cases} \frac{a}{2}L_G - \alpha - \sigma \leq 0 \\ \frac{b^2}{4}L_G^2 - \left(\frac{a}{2}\alpha - b\beta\right)L_G + \beta^2 - \alpha\sigma \leq 0 \\ \frac{b^2}{4}L_G^2 + \left(\frac{a}{2}\alpha - b\beta\right)L_G + \beta^2 - \alpha\sigma \leq 0 \end{cases}.$$

We verify these inequalities by plugging in the setting of $a = \frac{1}{L_G}$, $b = \frac{1}{4L_G}$, $c = \frac{2}{L_G}$, $\gamma = 2$, $\xi = 2L_G$, and $\lambda = \frac{\rho}{30}$, in the definition of $\alpha$, $\beta$, and $\sigma$. Then for $L_G \geq 2\rho$, we obtain that

$$\begin{cases} \frac{a}{2}L_G - \alpha - \sigma = -\frac{8579}{480} + \frac{3\rho}{40L_G} \leq 0 \\ \frac{b^2}{4}L_G^2 - \left(\frac{a}{2}\alpha - b\beta\right)L_G + \beta^2 - \alpha\sigma = -\frac{5357}{115200} + \frac{241\rho}{3200L_G} - \frac{\rho^2}{3600L_G^2} \leq 0 \\ \frac{b^2}{4}L_G^2 + \left(\frac{a}{2}\alpha - b\beta\right)L_G + \beta^2 - \alpha\sigma = -\frac{126077}{115200} + \frac{241\rho}{3200L_G} - \frac{\rho^2}{3600L_G^2} \leq 0 \end{cases}.$$

Therefore, $M \succeq \lambda\left(S + \frac{1}{2\rho}\mathrm{I}_{2d\times 2d}\right)$ when we take $a = \frac{1}{L_G}$, $b = \frac{1}{4L_G}$, $c = \frac{2}{L_G}$, $\gamma = 2$, and $\xi = 2L_G$, where the contraction rate $\lambda$ is $\lambda = \frac{\rho}{30}$.

■

**Proof of Lemma 6** For the expectation of $\left\|\theta_\tau - \theta_{kh}\right\|^2$ taken over the joint distribution of $(x_\tau, x_{kh})$, we use the definition of $x_\tau$ in our Equation (13) to expand it (by way of Jensen's inequality):

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\left\|\theta_\tau - \theta_{kh}\right\|^2\right] &= \xi\mathbb{E}\left[\left\|\int_{kh}^{\tau} r_s \mathrm{d}s\right\|^2\right] \\
&\leq \xi h \int_{kh}^{\tau} \mathbb{E}\left[\left\|r_s\right\|^2\right] \mathrm{d}s \\
&\leq \xi h^2 \sup_{s\in[kh,(k+1)h]} \mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\left\|r_s\right\|^2\right] \\
&= 2L_G h^2 \sup_{s\in[kh,(k+1)h]} \mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\left\|r_s\right\|^2\right]. \qquad (69)
\end{aligned}$$

In the following Lemma 11, we uniformly upper bound $\mathbb{E}\left[\left\|r_s\right\|^2\right]$ by $\mathcal{O}\left(\frac{d}{\rho}\right)$.

**Lemma 11.** *Assume that function $U$ satisfies Assumption **A1**–**A3**, where $\rho$ denotes the minimum of the log-Sobolev constant and* 1. *If we take $\gamma = 2$, $\xi = 2L_G$, and*

$$h = \frac{1}{56}\frac{1}{\sqrt{L_G}}\min\left\{\frac{1}{24}\frac{\rho}{L_G}, \frac{\sqrt{L_G}\rho}{L_H}\right\}\cdot\min\left\{\left(\widetilde{C_N}+2\right)^{-1/2}\sqrt{\frac{\epsilon}{d}}, \sqrt{\frac{\epsilon}{C_M}}\right\},$$

*where $\epsilon \leq d\frac{L_G}{\rho}$. Then for $r_s$ following Equation* (13), *$\forall s \geq 0$,*

$$\mathbb{E}\left[\left\|x_s\right\|^2\right] \leq \left(12\widetilde{C_N}+13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho} = \mathcal{O}\left(\frac{d}{\rho}\right).$$

We defer the proof of Lemma 11 to Sec. E.1.

Taking Lemma 11 as given, we can find that $\mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\left\|r_s\right\|^2\right]$ in Eq. (69) is upper bounded as:

$$\sup_{s\in[kh,(k+1)h]} \mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\left\|r_s\right\|^2\right] \leq \sup_{s\in[kh,(k+1)h]} \mathbb{E}_{x_s\sim\mathbf{p}_s}\left[\left\|x_s\right\|^2\right] \leq \left(12\widetilde{C_N}+13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho},$$

resulting in the final bound for $\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\left\|\theta_\tau - \theta_{kh}\right\|^2\right]$ to be:

$$\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\left\|\theta_\tau - \theta_{kh}\right\|^2\right] \leq \left(\left(24\widetilde{C_N}+26\right)\frac{L_G}{\rho}\cdot d + 24C_M\frac{L_G}{\rho}\right)h^2 = \mathcal{O}\left(\frac{L_G}{\rho}d\cdot h^2\right).$$

■

**Lemma 12.** *Let $\mathbf{p}_0(x) = \mathbf{p}_0(\theta)\mathbf{p}_0(r)$, where*

$$\mathbf{p}_0(\theta) = \left(\frac{L_G}{2\pi}\right)^{d/2}\exp\left(-\frac{L_G}{2}\left\|\theta\right\|^2\right),$$

*and*

$$\mathbf{p}_0(r) = \left(\frac{\xi}{2\pi}\right)^{d/2}\exp\left(-\frac{\xi}{2}\left\|r\right\|^2\right).$$

*For $\mathbf{p}^*(x) \propto \left(-U(\theta) - \frac{\xi}{2}\left\|r\right\|^2\right)$, if $U(\theta)$ follows Assumptions **A1**–**A3**, then we can define $\widetilde{C_N} = C_N + \frac{1}{2}\ln\frac{L_G}{2\pi}$ and obtain that*

$$\mathrm{KL}\left(\mathbf{p}_0\|\mathbf{p}^*\right) = \int \mathbf{p}_0(x)\ln\left(\frac{\mathbf{p}_0(x)}{\mathbf{p}^*(x)}\right)\mathrm{d}x \leq \widetilde{C_N}\cdot d + C_M, \qquad (70)$$

*and*

$$\begin{aligned}\mathcal{L}[\mathbf{p}_0] &= \mathrm{KL}\left(\mathbf{p}_0 \| \mathbf{p}^*\right) + \mathbb{E}_{\mathbf{p}_0}\left[\left\langle \nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}, S\nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}\right\rangle\right] \\ &\leq \left(\widetilde{C_N} + 1\right) d + C_M. \end{aligned} \tag{71}$$

*With the setting of $\xi = 2L_G$, we can also obtain that*

$$\mathbb{E}_{x\sim\mathbf{p}^*}\left[\|x\|^2\right] \leq \left(4\frac{\widetilde{C_N}}{\rho} + \frac{5}{2}\frac{1}{L_G}\right) \cdot d + 4\frac{C_M}{\rho}. \tag{72}$$

**Proof of Lemma 12** We want to bound $\mathrm{KL}\left(\mathbf{p}_0 \| \mathbf{p}^*\right) = \int \mathbf{p}_0(x) \ln\left(\frac{\mathbf{p}_0(x)}{\mathbf{p}^*(x)}\right) \mathrm{d}x = \int \mathbf{p}_0(\theta) \ln\left(\frac{\mathbf{p}_0(\theta)}{\mathbf{p}^*(\theta)}\right) \mathrm{d}\theta$, where $\mathbf{p}^*(\theta) \propto e^{-U(\theta)}$ and $\mathbf{p}_0(\theta) = \left(\frac{L_G}{2\pi}\right)^{d/2} \exp\left(-\frac{L_G}{2}\|\theta\|^2\right)$. First note that

$$\mathbf{p}^*(\theta) = \exp\left(-U(\theta)\right) \Big/ \int \exp\left(-U(\theta)\right) \mathrm{d}\theta.$$

By Assumptions **A2** and **A3**, $U(\theta) \leq \frac{L_G}{2}\|\theta\|^2$, $\forall \theta \in \mathbb{R}^d$. We also know that: $\ln\int \exp\left(-U(\theta)\right) \mathrm{d}\theta \leq C_N \cdot d + C_M$.

Therefore,

$$\begin{aligned} -\ln p^*(\theta) &= U(\theta) + \ln \int \exp\left(-U(\theta)\right) \mathrm{d}\theta \\ &\leq \frac{L_G}{2}\|\theta\|^2 + C_N \cdot d + C_M. \end{aligned} \tag{73}$$

Hence

$$-\int \mathbf{p}_0(\theta) \ln p^*(\theta) \mathrm{d}\theta \leq \frac{d}{2} + C_N \cdot d + C_M.$$

We can also calculate that

$$\int \mathbf{p}_0(\theta) \ln p_0(\theta) \mathrm{d}\theta = -\frac{d}{2} - \frac{d}{2} \ln \frac{2\pi}{L_G}.$$

Therefore,

$$\begin{aligned} \mathrm{KL}\left(\mathbf{p}_0 \| \mathbf{p}^*\right) &= \int \mathbf{p}_0(\theta) \ln p_0(\theta) \mathrm{d}\theta - \int \mathbf{p}_0(\theta) \ln p^*(\theta) \mathrm{d}\theta \\ &\leq \left(C_N + \frac{1}{2} \ln \frac{L_G}{2\pi}\right) \cdot d + C_M \\ &= \widetilde{C_N} \cdot d + C_M. \end{aligned}$$

For $\mathbb{E}_{\mathbf{p}_0}\left[\left\langle \nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}, S\nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}\right\rangle\right]$, since $U$ is $L_G$-Lipschitz smooth, $\left\|\nabla_\theta \ln p^*(x)\right\|^2 \le L_G^2 \left\|\theta\right\|^2$, and thus

$$\begin{aligned}
&\mathbb{E}_{\mathbf{p}_0}\left[\left\langle \nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}, S\nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}\right\rangle\right] \\
&= \frac{1}{4L_G}\mathbb{E}_{\mathbf{p}_0}\left[\left\|\nabla_\theta \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}\right\|^2\right] \\
&\le \frac{1}{2L_G}\mathbb{E}_{\mathbf{p}_0}\left[\left\|\nabla_\theta \ln \mathbf{p}_0\right\|^2 + \left\|\nabla_\theta \ln \mathbf{p}^*\right\|^2\right] \\
&\le L_G \mathbb{E}_{\mathbf{p}_0}\left[\left\|\theta\right\|^2\right] \\
&= d.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathcal{L}[\mathbf{p}_0] &= \mathrm{KL}\left(\mathbf{p}_0 \| \mathbf{p}^*\right) + \mathbb{E}_{\mathbf{p}_0}\left[\left\langle \nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}, S\nabla_x \ln \frac{\mathbf{p}_0}{\mathbf{p}^*}\right\rangle\right] \\
&\le \left(\widetilde{C_N} + 1\right) d + C_M. \qquad (74)
\end{aligned}$$

For $\mathbb{E}_{x^* \sim \mathbf{p}^*}\left[\left\|x^*\right\|^2\right]$, we bound it using $W_2(\mathbf{p}^*, \mathbf{p}_0)$. We choose an auxiliary random variable $\theta_0$ following the law of $\mathbf{p}_0(\theta)$ and couples optimally with $\theta^* \sim \mathbf{p}^*(\theta)$: $(\theta^*, \theta_0) \sim \gamma \in \Gamma_{opt}(\mathbf{p}^*, \mathbf{p}_0)$. We then have

$$\begin{aligned}
\mathbb{E}_{x^* \sim \mathbf{p}^*}\left[\left\|x^*\right\|^2\right] &= \mathbb{E}_{r^* \sim \mathbf{p}^*(r)}\left[\left\|r^*\right\|^2\right] + \mathbb{E}_{\theta^* \sim \mathbf{p}^*(\theta)}\left[\left\|\theta^*\right\|^2\right] \\
&= \frac{d}{\xi} + \mathbb{E}_{(\theta^*, \theta_0) \sim \gamma}\left[\left\|\theta_0 + (\theta^* - \theta_0)\right\|^2\right] \\
&\le \frac{d}{\xi} + 2\mathbb{E}_{\theta_0 \sim \mathbf{p}_0}\left[\left\|\theta_0\right\|^2\right] + 2\mathbb{E}_{(\theta^*, \theta_0) \sim \gamma}\left[\left\|\theta^* - \theta_0\right\|^2\right] \\
&= \frac{d}{\xi} + \frac{2d}{L_G} + 2W_2^2(\mathbf{p}^*, \mathbf{p}_0).
\end{aligned}$$

We further expand this inequality by using the extended Talagrand inequality, Eq. (1), which applies to the joint density function $\mathbf{p}^*(\theta, r) \propto \exp\left(-U(\theta) - \frac{\xi}{2}\left\|r\right\|^2\right)$ with log-Sobolev constant greater than or equal to $\rho$ and Lipschitz smoothness of $U + \frac{\xi}{2}\left\|r\right\|^2$ less than or equal to $4L_G$:

$$W_2^2(\mathbf{p}_s, \mathbf{p}^*) \le \frac{2}{\rho}\mathrm{KL}\left(\mathbf{p}_s \| \mathbf{p}^*\right).$$

Therefore, for $\xi = 2L_G$,

$$\begin{aligned}
\mathbb{E}_{x^* \sim \mathbf{p}^*}\left[\left\|x^*\right\|^2\right] &\le \frac{d}{\xi} + \frac{2d}{L_G} + \frac{4}{\rho}\mathrm{KL}\left(\mathbf{p}_0 \| \mathbf{p}^*\right) \\
&\le \left(\frac{4\widetilde{C_N}}{\rho} + \frac{1}{\xi} + \frac{2}{L_G}\right) \cdot d + \frac{4C_M}{\rho} \\
&= \left(4\frac{\widetilde{C_N}}{\rho} + \frac{5}{2}\frac{1}{L_G}\right) \cdot d + 4\frac{C_M}{\rho}.
\end{aligned}$$

It is worth noting that the choice of the initial condition $\mathbf{p}_0$ can be flexible. For example, if we choose $x_0 \sim \mathcal{N}(0, \mathrm{I})$, then $\mathrm{KL}(\mathbf{p}_0 \| \mathbf{p}^*) \leq \left(C_N + \frac{L_G}{2} - \frac{1}{2} - \frac{1}{2}\ln(2\pi)\right) \cdot d + C_M$ (resulting in merely an extra $\ln L_G$ term in the overall computation complexity). $\blacksquare$

### E.1 Supporting Proof for Lemma 6

**Proof of Lemma 11** In what follows, we will prove that:

1. $\mathbb{E}\left[\|x_0\|^2\right] \leq \left(12\widetilde{C_N} + 13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho}$.
2. If $\forall s \leq kh$, $\mathbb{E}\left[\|x_s\|^2\right] \leq \left(12\widetilde{C_N} + 13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho}$, then $\forall s \in [kh, (k+1)h]$,
$$\mathbb{E}\left[\|x_s\|^2\right] \leq \left(12\widetilde{C_N} + 13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho}.$$

By induction, this will prove Lemma 11.

For claim 1, we can calculate that $\mathbb{E}_{x_0 \sim \mathbf{p}_0}\left[\|x_0\|^2\right] = \frac{3}{2} \cdot \frac{d}{L_G} \leq \left(12\widetilde{C_N} + 13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho}$.

We prove claim 2 in a two step procedure: we first prove in the following Lemma 13 that if $\mathbb{E}\left[\|x_{kh}\|^2\right]$ is bounded, then $\mathbb{E}\left[\|x_s\|^2\right]$ remains bounded for $s \in [kh, (k+1)h]$. We then provide a specific bound of it.

**Lemma 13.** *Assume the step size $h \leq \frac{1}{8L_G}$ and let $\gamma = 2$ and $\xi = 2L_G$. Then $\forall s \in [kh, (k+1)h]$, $\mathbb{E}\left[\|x_s\|^2\right] \leq 2\mathbb{E}\left[\|x_{kh}\|^2\right] + \frac{d}{L_G}$.*

It can be verified that for $\epsilon \leq 2d$ and $\rho \leq 1$, $h$ is indeed smaller than $\frac{1}{8L_G}$. Thus Lemma 13, in conjunction with the induction hypothesis, gives us a rough bound that $\forall s \in [kh, (k+1)h]$,

$$\begin{aligned}\mathbb{E}\left[\|x_s\|^2\right] &\leq 2\mathbb{E}\left[\|x_{kh}\|^2\right] + \frac{d}{L_G} \leq \left(24\widetilde{C_N} + 26\right)\frac{d}{\rho} + 24\frac{C_M}{\rho} + \frac{d}{L_G} \\ &\leq \left(24\widetilde{C_N} + 27\right)\frac{d}{\rho} + 24\frac{C_M}{\rho}. \end{aligned} \tag{75}$$

Then to accurately bound $\mathbb{E}\left[\|x_s\|^2\right]$, we use $\mathbb{E}_{x^* \sim \mathbf{p}^*}\left[\|x^*\|^2\right]$ as an anchor point and bound the Wasserstein-2 distance between $p_s$ and $p^*$. To this end, we choose an auxiliary random variable $x^*$ following the law of $\mathbf{p}^*$ and couples optimally with $\mathbf{p}(x_s)$: $(x_s, x^*) \sim \zeta \in \Gamma_{opt}(\mathbf{p}(x_s), \mathbf{p}^*(x^*))$. Then using Young's inequality and Eq. (72) in Lemma 12,

$$\begin{aligned}\mathbb{E}\left[\|x_s\|^2\right] &= \mathbb{E}_{(x_s, x^*) \sim \zeta}\left[\|x^* + (x_s - x^*)\|^2\right] \\ &\leq 2\mathbb{E}_{\mathbf{p}^*}\left[\|x^*\|^2\right] + 2\mathbb{E}_{(x_s, x^*) \sim \zeta}\left[\|x_s - x^*\|^2\right] \\ &\leq \left(8\frac{\widetilde{C_N}}{\rho} + 5\frac{1}{L_G}\right) \cdot d + 8\frac{C_M}{\rho} + 2W_2^2(\mathbf{p}_s, \mathbf{p}^*).\end{aligned}$$

Applying the extended Talagrand inequality, Eq. (1), we obtain that

$$\mathbb{E}\left[\|x_s\|^2\right] \leq \left(8\frac{\widetilde{C_N}}{\rho} + 5\frac{1}{L_G}\right) \cdot d + 8\frac{C_M}{\rho} + \frac{4}{\rho}\mathrm{KL}\left(\mathbf{p}_s\|\mathbf{p}^*\right). \tag{76}$$

On the other hand, we can use dissipation of the Lyapunov functional to bound the growth of the KL-divergence, and in turn the growth of $\mathbb{E}\left[\|x_s\|^2\right]$ in Eq. (76). This is the thesis of the following Lemma 14.

**Lemma 14.** *Let $x_s$ follow the underdamped Langevin algorithm 1 with parameters $\xi = 2L_G$, $\gamma = 2$, and the step size $h = (k+1)h - kh$ given in Eq. (37). Also let $p_s$ be the probability distribution of $x_s$. Assume that Eq. (75) (given by the induction hypothesis in conjunction with Lemma 13) holds for any $s \in [kh, (k+1)h]$. Then for $\epsilon \leq 2d$ and $\rho \leq 1$, $\forall s \in [kh, (k+1)h]$,*

$$\frac{\mathrm{d}\mathcal{L}[\mathbf{p}_s]}{\mathrm{d}s} \leq -\frac{\rho}{30} \cdot \left(\mathcal{L}[\mathbf{p}_s] - \frac{\epsilon}{2}\right). \tag{77}$$

Applying Grönwall's Lemma in Eq. (77), we obtain that the objective functional $\mathcal{L}$ will not increase by more than $\epsilon/2$ throughout the progress of the algorithm:

$$\mathcal{L}[\mathbf{p}_s] - \frac{\epsilon}{2} \leq e^{-\frac{\rho}{30}(s-kh)}\left(\mathcal{L}[\mathbf{p}_{kh}] - \frac{\epsilon}{2}\right) \leq e^{-\frac{\rho}{30}kh - \frac{\rho}{30}(s-kh)}\left(\mathcal{L}[\mathbf{p}_0] - \frac{\epsilon}{2}\right) \leq \mathcal{L}[\mathbf{p}_0],$$

where $\mathcal{L}[\mathbf{p}_s] = \mathrm{KL}\left(\mathbf{p}_s\|\mathbf{p}^*\right) + \mathbb{E}_{\mathbf{p}_s}\left[\left\langle \nabla_x \ln\frac{\mathbf{p}_s}{\mathbf{p}^*}, S\nabla_x \ln\frac{\mathbf{p}_s}{\mathbf{p}^*}\right\rangle\right]$. Therefore, we can bound $\mathrm{KL}\left(\mathbf{p}_s\|\mathbf{p}^*\right)$ using initial conditions

$$\mathrm{KL}\left(\mathbf{p}_s\|\mathbf{p}^*\right) \leq \mathcal{L}[\mathbf{p}_s] \leq \mathcal{L}[\mathbf{p}_0] + \frac{\epsilon}{2}.$$

From Lemma 12, we know that $\mathcal{L}[\mathbf{p}_0] \leq \left(\widetilde{C_N} + 1\right) d + C_M$. Therefore, for $\epsilon \leq 2d$,

$$\begin{aligned}\mathrm{KL}\left(\mathbf{p}_s\|\mathbf{p}^*\right) &\leq \left(\widetilde{C_N} + 1\right) d + C_M + \frac{\epsilon}{2} \\ &\leq \left(\widetilde{C_N} + 2\right) d + C_M.\end{aligned} \tag{78}$$

Plugging Eq. (78) into Eq. (76), we obtain our final result that

$$\begin{aligned}\mathbb{E}\left[\|x_s\|^2\right] &\leq \left(8\frac{\widetilde{C_N}}{\rho} + 5\frac{1}{L_G}\right) d + 8\frac{C_M}{\rho} + \frac{4}{\rho}\mathrm{KL}\left(\mathbf{p}_s\|\mathbf{p}^*\right) \\ &\leq \left(12\frac{\widetilde{C_N}}{\rho} + 8\frac{1}{\rho} + 5\frac{1}{L_G}\right) d + 12\frac{C_M}{\rho} \\ &\leq \left(12\widetilde{C_N} + 13\right)\frac{d}{\rho} + 12\frac{C_M}{\rho},\end{aligned}$$

since $\rho \leq L_G$. ■

**Proof of Lemma 13** We begin from the discretized dynamics of underdamped Langevin diffusion Eq. (15)

to calculate that $\forall s \in [kh,(k+1)h]$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[\|x_s\|^2\right] &= \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[\|\theta_s\|^2+\|r_s\|^2\right]\\
&= 2\mathbb{E}\left[\left\langle \begin{pmatrix}\theta_s\\ r_s\end{pmatrix}, \begin{pmatrix}\xi r_s\\ -\nabla U(\theta_{kh})-\gamma\xi r_s-\gamma\nabla_r \ln \mathbf{p}_s\end{pmatrix}\right\rangle\right]\\
&\le 2\mathbb{E}\left[\xi\langle\theta_s,r_s\rangle-\langle r_s,\theta_{kh}\rangle-\gamma\xi\|r_s\|^2\right]-2\gamma\int_{\mathbb{R}^d}\langle r_s,\nabla_r\ln\mathbf{p}_s\rangle\mathbf{p}_s\mathrm{d}x_s\\
&\le 2\mathbb{E}\left[\xi\|\theta_s\|\|r_s\|+L_G\|\theta_{kh}\|\|r_s\|-\gamma\xi\|r_s\|^2\right]+2\gamma d\\
&\le 2L_G\mathbb{E}\left[\|\theta_s\|^2+\|r_s\|^2\right]+2L_G\mathbb{E}\left[\|\theta_{kh}\|^2+\|r_{kh}\|^2\right]+2\gamma d,
\end{aligned}\tag{79}$$

where the last step follows from plugging in the setting of $\gamma=2$ and $\xi=2L_G$ and using Young's inequality. Multiplying $e^{-2L_G s}>0$ on both ends of Eq. (79), we obtain that $\forall s$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(e^{-2L_G s}\mathbb{E}\left[\|x_s\|^2\right]\right)\le e^{-2L_G s}\left(2L_G\mathbb{E}\left[\|x_{kh}\|^2\right]+2\gamma d\right).\tag{80}$$

Applying the fundamental theorem of calculus and multiplying $e^{2L_G\tau}>0$ on both sides, we have that

$$\begin{aligned}
\mathbb{E}\left[\|x_\tau\|^2\right] &\le e^{2L_G\tau}\int_{kh}^{\tau}e^{-2L_G s}\left(2L_G\mathbb{E}\left[\|x_{kh}\|^2\right]+2\gamma d\right)\mathrm{d}s+e^{2L_G(\tau-kh)}\mathbb{E}\left[\|x_{kh}\|^2\right]\\
&= \frac{1}{2L_G}\left(e^{2L_G(\tau-kh)}-1\right)\left(2L_G\mathbb{E}\left[\|x_{kh}\|^2\right]+2\gamma d\right)+e^{2L_G(\tau-kh)}\mathbb{E}\left[\|x_{kh}\|^2\right].
\end{aligned}$$

It can then be checked that when $\tau-kh\le h\le\frac{1}{8L_G}$, the factor $\left(e^{2L_G(\tau-kh)}-1\right)\le\frac{1}{2}$, and that

$$\mathbb{E}\left[\|x_\tau\|^2\right]\le 2\mathbb{E}\left[\|x_{kh}\|^2\right]+\frac{d}{L_G},\quad \forall\tau\in[kh,(k+1)h].$$

■

**Proof of Lemma 14** Applying the result of Eq. (69) that:

$$\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau-\theta_{kh}\|^2\right]\le 2L_G h^2\sup_{s\in[kh,(k+1)h]}\mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\|r_s\|^2\right]$$

to Eq. (23a)–(23c), we obtain that for $\xi=2L_G$, $\gamma=2$, and $\forall\tau\in[kh,(k+1)h]$,

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}(\mathbf{p}_\tau)}{\mathrm{d}\tau} &\le -\frac{\rho}{30}\mathcal{L}(\mathbf{p}_\tau)\\
&\quad+\left(68L_G^2+\frac{1}{8}\frac{L_H^2}{L_G}\right)\mathbb{E}_{\mathbf{p}(x_{kh},x_\tau)}\left[\|\theta_\tau-\theta_{kh}\|^2\right]+18eL_G d\max\left\{L_G^4(\tau-kh)^4,L_G^2(\tau-kh)^2\right\}\\
&\le -\frac{\rho}{30}\Bigg(\mathcal{L}(\mathbf{p}_\tau)-60\frac{L_G}{\rho}\left(68L_G^2+\frac{1}{8}\frac{L_H^2}{L_G}\right)h^2\sup_{s\in[kh,(k+1)h]}\mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\|r_s\|^2\right]\\
&\qquad\qquad-540e\frac{L_G}{\rho}d\max\left\{L_G^4h^4,L_G^2h^2\right\}\Bigg)\\
&\le -\frac{\rho}{30}\Bigg(\mathcal{L}(\mathbf{p}_\tau)-60\frac{L_G}{\rho}\max\left\{136L_G^2,\frac{1}{4}\frac{L_H^2}{L_G}\right\}h^2\sup_{s\in[kh,(k+1)h]}\mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\|r_s\|^2\right]\\
&\qquad\qquad-1500\frac{L_G}{\rho}d\max\left\{L_G^4h^4,L_G^2h^2\right\}\Bigg).
\end{aligned}\tag{81}$$

Using the definition of $h = \frac{1}{56}\frac{1}{\sqrt{L_G}} \min\left\{\frac{1}{24}\frac{\rho}{L_G}, \frac{\sqrt{L_G}\rho}{L_H}\right\} \cdot \min\left\{\left(\widetilde{C_N}+2\right)^{-1/2}\sqrt{\frac{\epsilon}{d}}, \sqrt{\frac{\epsilon}{C_M}}\right\}$ in Eq. (37), we know that

$$L_G^2 h^2 \leq \frac{1}{6000}\frac{1}{\widetilde{C_N}+2} \cdot \frac{\rho^2}{L_G}\frac{\epsilon}{d}.$$

Plugging this setting into the last term of Eq. (81), we obtain that for $\epsilon \leq 2d$ and $\rho \leq 1$,

$$1500\frac{L_G}{\rho} d \max\left\{L_G^4 h^4, L_G^2 h^2\right\} \leq \frac{\epsilon}{4}.$$

We can similarly combine this setting of the step size $h$ with the premise of this Lemma, Eq. (75), that $\sup_{s\in[kh,(k+1)h]} \mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\|r_s\|^2\right] \leq \left(24\widetilde{C_N}+27\right)\frac{d}{\rho} + 24\frac{C_M}{\rho}$, and obtain:

$$\begin{aligned}
&60\frac{L_G}{\rho}\max\left\{136L_G^2, \frac{1}{4}\frac{L_H^2}{L_G}\right\} h^2 \cdot \sup_{s\in[kh,(k+1)h]} \mathbb{E}_{r_s\sim\mathbf{p}_s}\left[\|r_s\|^2\right] \\
&\leq 60\frac{L_G}{\rho}\max\left\{144L_G^2, \frac{1}{4}\frac{L_H^2}{L_G}\right\} \cdot \left(\left(24\widetilde{C_N}+27\right)\frac{d}{\rho} + 24\frac{C_M}{\rho}\right) h^2 \\
&\leq 28^2 L_G \max\left\{24^2\frac{L_G^2}{\rho^2}, \frac{L_H^2}{L_G\rho^2}\right\} \cdot \max\left\{\left(\widetilde{C_N}+2\right)d, C_M\right\} h^2 \leq \frac{\epsilon}{4}.
\end{aligned}$$

Consequently, the time derivative of the Lyapunov functional $\mathcal{L}$ is bounded as:

$$\frac{\mathrm{d}\mathcal{L}[\mathbf{p}_s]}{\mathrm{d}s} \leq -\rho \cdot \left(\mathcal{L}[\mathbf{p}_s] - \frac{\epsilon}{2}\right). \tag{82}$$

■

# F Proofs for Auxiliary Facts

**Proof of Fact 1** By Assumptions (b) and (c), $U(\theta) \leq \frac{L_G}{2}\|\theta\|^2$, $\forall\theta \in \mathbb{R}^d$. We also prove in the following that

- $U(\theta) \geq \frac{m}{4}\|\theta\|^2$, $\forall\theta \in \mathbb{R}^d \setminus \mathbb{B}\left(0, \frac{8L_G}{m}R\right)$;
- $U(\theta) \geq -\frac{L_G}{2}\|\theta\|^2$, $\forall\theta \in \mathbb{B}\left(0, \frac{8L_G}{m}R\right)$.

The latter case follows directly from Assumptions (b) and (c). For the former case where $\|\theta\| \geq \frac{8L_G}{m}R$, define $\vartheta = \frac{R}{\|\theta\|}\theta$. Since $\|\vartheta\| = R$,

$$\langle\nabla U(\vartheta), \vartheta\rangle \geq -L_G R^2.$$

Because any convex combination of $\theta$ and $\vartheta$ belongs to the set $\mathbb{R}^d \setminus \mathbb{B}(0, R)$, where $U$ is $m$-strongly convex,

$$\begin{aligned} U(\theta) - U(\vartheta) &\geq \langle \nabla U(\vartheta), \theta - \vartheta \rangle + \frac{m}{2} \|\theta - \vartheta\|^2 \\ &= \left(\frac{\|\theta\|}{R} - 1\right) \langle \nabla U(\vartheta), \vartheta \rangle + \frac{m}{2} \left(\frac{\|\theta\|}{R} - 1\right)^2 \\ &\geq -\left(\frac{\|\theta\|}{R} - 1\right) L_G R^2 + \frac{m}{2} \left(\frac{\|\theta\|}{R} - 1\right)^2 \\ &\geq \frac{m}{4} \|\theta\|^2 + L_G R^2, \end{aligned}$$

since $\|\theta\| \geq \frac{8 L_G}{m} R$. Again, using Assumptions (b) and (c), $U(\vartheta) \geq -\frac{L_G}{2} R^2$, which leads to the result that $U(\theta) \geq \frac{m}{4} \|\theta\|^2$.

Therefore, $U(\theta) \geq \frac{m}{4} \|\theta\|^2 - 32 \frac{L_G^2}{m^2} L_G R^2$ and

$$\begin{aligned} \ln \int \exp\left(-U(\theta)\right) \mathrm{d}\theta &\leq \ln \int \exp\left(-\frac{m}{4} \|\theta\|^2 + 32 \frac{L_G^2}{m^2} L_G R^2\right) \mathrm{d}\theta \\ &= \frac{d}{2} \ln \frac{4\pi}{m} + 32 \frac{L_G^2}{m^2} L_G R^2. \end{aligned}$$

Hence $C_N \leq \frac{1}{2} \ln \frac{4\pi}{m}$ and $C_M \leq 32 \frac{L_G^2}{m^2} L_G R^2$. ■

**Proof of Fact 2** We begin with the definition of

$$\eta = \frac{1}{\gamma} \left( \frac{e^{\gamma\xi(\tau - kh)} \left(1 - e^{-\gamma\xi(\tau - kh)}\right)^2}{\gamma\xi} - \left( (\tau - kh) - \frac{1 - e^{-\gamma\xi(\tau - kh)}}{\gamma\xi} \right) \right),$$

and provide bound for it when $0 \leq (\tau - kh) \leq \min\left\{\frac{1}{\gamma\xi}, \frac{1}{\sqrt{2eL_G\xi}}\right\}$.

First note that for $0 \leq (\tau - kh) \leq \frac{1}{\gamma\xi}$,

$$1 - \gamma\xi\nu \leq e^{-\gamma\xi(\tau - kh)} \leq 1 + \gamma\xi(\tau - kh).$$

Then we obtain that

$$\begin{aligned} \eta &= \frac{1}{\gamma} \left( \frac{e^{\gamma\xi(\tau - kh)}}{\gamma\xi} \left(\gamma\xi(\tau - kh)\right)^2 - (\tau - kh) + \frac{-\gamma\xi(\tau - kh)}{\gamma\xi} \right) \\ &= \gamma(\tau - kh)^2 e^{\gamma\xi(\tau - kh)} \\ &\leq e\xi(\tau - kh)^2. \end{aligned}$$

We then prove Fact 2 by separating the following term:

$$\begin{aligned} &\left\| \begin{pmatrix} \nabla^2 U(\theta_{kh}) \left( \left(\mathrm{I} + \eta \nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I} \right) \\ -\frac{e^{\gamma\xi(\tau - kh)} - 1}{\gamma} \nabla^2 U(\theta_{kh}) \left(\mathrm{I} + \eta \nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix} \right\|_2 \\ &\leq 2 \max \Bigg\{ \left\| \nabla^2 U(\theta_{kh}) \left( \left(\mathrm{I} + \eta \nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I} \right) \right\|_2, \\ &\qquad\qquad \left\| \frac{1 - e^{\gamma\xi(\tau - kh)}}{\gamma} \nabla^2 U(\theta_{kh}) \left(\mathrm{I} + \eta \nabla^2 U(\theta_{kh})\right)^{-1} \right\|_2 \Bigg\}. \end{aligned}$$

Since $\left\|\eta\nabla^2 U(\theta_{kh})\right\| \le e\xi\nu^2 \left\|\nabla^2 U(\theta_{kh})\right\| \le eL_G\xi\nu^2 < 1$ for $\nu \le \min\left\{\frac{1}{\gamma\xi}, \frac{1}{\sqrt{2eL_G\xi}}\right\}$, $\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1}$ admits the following series expansion:

$$\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1} = \sum_{n=0}^{\infty} \left(-\eta\nabla^2 U(\theta_{kh})\right)^n.$$

Consequently,

$$\left\|\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1}\right\|_2 \le \sum_{n=0}^{\infty} \left(\eta L_G\right)^n = \frac{1}{1-\eta L_G} \le 2,$$

and

$$\left\|\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right\|_2 \le \sum_{n=1}^{\infty} \left(\eta L_G\right)^n = \frac{\eta L_G}{1-\eta L_G} \le 2\eta L_G = 2eL_G\xi\nu^2.$$

Therefore, for the first term,

$$\begin{aligned}
&\left\|\nabla^2 U(\theta_{kh})\left(\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right)\right\|_2 \\
&\le \left\|\nabla^2 U(\theta_{kh})\right\|_2 \left\|\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right\|_2 \\
&\le 2eL_G^2\xi\nu^2.
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\left\|\frac{1-e^{\gamma\xi(\tau-kh)}}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1}\right\|_2 \\
&\le \xi\nu\left\|\nabla^2 U(\theta_{kh})\right\| \left\|\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1}\right\|_2 \\
&\le 2L_G\xi\nu.
\end{aligned}$$

Therefore,

$$\left\|\begin{pmatrix} \nabla^2 U(\theta_{kh})\left(\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1} - \mathrm{I}\right) \\ -\frac{e^{\gamma\xi(\tau-kh)}-1}{\gamma}\nabla^2 U(\theta_{kh})\left(\mathrm{I} + \eta\nabla^2 U(\theta_{kh})\right)^{-1} \end{pmatrix}\right\|_2 \le 4e\max\{L_G^2\xi\nu^2, L_G\xi\nu\}.$$

■