Journal of Machine Learning Research 22 (2021) 1-41 Submitted 6/20; Revised 10/20; Published 1/21

High-Order Langevin Diffusion Yields an Accelerated MCMC Algorithm

Wenlong Mou* WMOU@BERKELEY.EDU
Department of EECS
University of California, Berkeley
Berkeley, CA, 94720, USA

Yi-An Ma* YIANMA.UCSD@GMAIL.COM
Halicioğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92023, USA

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Department of EECS and Department of Statistics
University of California, Berkeley
Berkeley, CA, 94720, USA

Peter L. Bartlett PETER@BERKELEY.EDU
Department of EECS and Department of Statistics
University of California, Berkeley
Berkeley, CA, 94720, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of EECS and Department of Statistics
University of California, Berkeley
Berkeley, CA, 94720, USA

Editor: Arnak Dalalyan

Abstract

We propose a Markov chain Monte Carlo (MCMC) algorithm based on third-order Langevin dynamics for sampling from distributions with smooth, log-concave densities. The higher-order dynamics allow for more flexible discretization schemes, and we develop a specific method that combines splitting with more accurate integration. For a broad class of d-dimensional distributions arising from generalized linear models, we prove that the resulting third-order algorithm produces samples from a distribution that is at most $\varepsilon > 0$ in Wasserstein distance from the target distribution in $O\left(\frac{d^{1/4}}{\varepsilon^{1/2}}\right)$ steps. This result requires only Lipschitz conditions on the gradient. For general strongly convex potentials with α-th order smoothness, we prove that the mixing time scales as $O\left(\frac{d^{1/4}}{\varepsilon^{1/2}} + \frac{d^{1/2}}{\varepsilon^{1/(\alpha-1)}}\right)$.

Keywords: Markov Chain Monte Carlo; mixing time; Langevin diffusion; Bayesian inference; discretization of SDEs

. * Wenlong Mou and Yi-An Ma contributed equally to this work.

1. Introduction

Recent years have witnessed substantial progress in the theoretical analysis of algorithms for large-scale statistical inference. There are lines of work on both the optimization algorithms used to compute frequentist point estimates as well as the sampling algorithms that underpin Bayesian inference. In both cases, nonasymptotic rates of convergence have been obtained and, increasingly, those rates include dimension dependence (see, e.g., Dalalyan, 2017; Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019; Cheng et al., 2018b; Cheng and Bartlett, 2018; Dwivedi et al., 2019; Ma et al., 2019b; Chatterji et al., 2018). In particular, for the gradient-based algorithms that have become the state-of-the-art in many applications, the dimension dependence is generally linear or sublinear, providing strong theoretical support for the deployment of these algorithms in large-scale problems.

Although progress has been made in both optimization and sampling, the latter has lagged the former, arguably because the stochasticity inherent to sampling involves mathematical challenges. Indeed, much of the recent progress in both paradigms has involved taking a continuous-time point of view, whereby a discrete-time algorithm is seen as arising from discretizing some underlying continuous dynamical system, and this line of attack turns out to be more challenging for sampling methods. In the setting of optimization, continuous dynamics can be represented in terms of ordinary differential equations (ODEs) (Brown and Bartholomew-Biggs, 1989; Su et al., 2014; Wilson et al., 2016; Shi et al., 2018), whereas for sampling, the underlying dynamics are characterized as stochastic differential equations (SDEs) (Roberts and Rosenthal, 2001; Dalalyan, 2017; Durmus and Moulines, 2019). The non-smooth nature of the Brownian motion underlying these SDEs presents challenges in performing the discretization required to translate continuous-time results to discrete time.

In this paper, we focus on densities over $\mathbb{R}^d$ of the form $p^*(\theta) \propto \exp(-U(\theta))$, where the potential function $U : \mathbb{R}^d \to \mathbb{R}$ is strongly convex and Lipschitz smooth. There is a substantial body of past work on sampling problems of this type; among other results, it has been shown (Cheng et al., 2018b,a; Dalalyan and Riou-Durand, 2020; Ma et al., 2019a) that a discretization of the second-order Langevin diffusion has mixing time that scales as $O(\sqrt{d})$; notably, this scaling of the mixing time compares favorably to the best known $O(d)$ scaling of the first-order Langevin diffusion (Dalalyan, 2017; Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019). The results were further improved by Shen and Lee (2019), who used a uniform random time to construct a low-bias estimator for an integral, leading to an improved discretization scheme with $O(d^{1/3}/\varepsilon^{2/3})$ mixing time. Furthermore, Cao et al. (2020) showed for second-order Langevin diffusions that this rate cannot be further improved.

However, if additional and relatively strong assumptions are imposed on the density, then even faster rates of convergence can be obtained (Mangoubi and Smith, 2017; Mangoubi and Vishnoi, 2018; Lee et al., 2018). Here we ask whether it is possible to accelerate convergence of sampling algorithms beyond the $O(d^{1/3})$ barrier for a large class of machine learning models *without* imposing additional assumptions beyond strong convexity and Lipschitz smoothness. The main contribution of our paper is an affirmative answer to this question.

Let us provide some context for our line of attack and our contributions. As is well known from past work (Dalalyan, 2017; Bou-Rabee et al., 2020), the continuous-time Langevin dynamics converge to the target distribution at an exponential rate, with no dependence on dimension. However, in any implementation with digital computation, the continuous-time dynamics must be approximated with a discrete-time scheme, leading to numerical error that does scale with the dimension and the conditioning of the problem. Direct application of higher-order schemes to the Langevin diffusion is hindered by the non-smoothness of the Brownian motion. One way to circumvent this problem is to augment the traditional Langevin diffusion by moving to higher-order continuous dynamics. In particular, recent work has studied the second-order (or underdamped) Langevin algorithm, which lifts the original d-dimensional space to a $2d$-dimensional space consisting of vectors of the form $x = (\theta, r) \in \mathbb{R}^d \times \mathbb{R}^d$, and considers SDEs in these variables (Cheng et al., 2018b,a; Dalalyan and Riou-Durand, 2020; Ma et al., 2019a). There is a natural hierarchy of such lifted schemes, and this paper is based on proposing and analyzing a carefully designed *third-order lifting*, to be described in Section 2.4. We provide a thorough analysis of a particular discretization of this third-order scheme, establishing non-asymptotic bounds on mixing time for particular classes of potential functions.

Our presentation begins with an analysis of potential functions that have the following ridge-separable form:

$$U(\theta) = \sum_{i=1}^{n} u_i\big(a_i^{\mathrm{T}}\theta\big), \tag{1}$$

where $\{u_i\}_{i=1}^n$ are a collection of univariate functions, and $\{a_i\}_{i=1}^n$ are a given collection of vectors in $\mathbb{R}^d$. Many log-concave sampling problems that arise in statistics and machine learning involve potential functions of this form. In particular, posterior sampling in Bayesian generalized linear models, including Bayesian logistic regression and one-layer neural networks, involve densities of the form (1). It is worth noticing that we do not impose any additional assumptions on the vectors $\{a_i\}_{i=1}^n$. The ridge-separable form is needed only to make sure that certain one-dimensional integrals have closed-form solutions.

Given a distribution of the form (1), with the potential function U being strongly convex and smooth, we prove that $O(d^{\frac{1}{4}}/\varepsilon^{\frac{1}{2}})$ steps suffice to make the Wasserstein distance between the sample and target distributions less than ε. This is the first time that the $O(d^{1/3})$ barrier for the log-concave sampling problem has been overcome without additional structural assumptions on the data, even for the special case of Bayesian logistic regression. The dependency on ε is also improved relative to the current state of the art. It is worth noting that our analysis allows for arbitrary vectors $\{a_i\}_{i=1}^n$ and functions $\{u_i\}_{i=1}^n$, as long as the smoothness and strong convexity of U are guaranteed. This is in contrast to previous work (Mangoubi and Smith, 2017; Mangoubi and Vishnoi, 2018; Lee et al., 2018) that requires incoherence conditions on the data vectors $\{a_i\}_{i=1}^n$ and/or high-order smoothness conditions on the component functions $\{u_i\}_{i=1}^n$.

We then tackle the more general setting in which the potential function U need not be ridge-separable (1). Assuming only access to a black-box gradient oracle, we show that the dimension dependency of our algorithm is $O(\sqrt{d})$, but the dependency on the final accuracy ε associated with this term can be adaptive to the smoothness assumptions satisfied by U. In particular, when the potential function U satisfies an α-th order smoothness condition for

$\alpha \geq 2$, we establish an upper bound on the mixing time of $O\left(d^{\frac{1}{4}}/\varepsilon^{\frac{1}{2}} + d^{\frac{1}{2}}/\varepsilon^{\frac{1}{\alpha-1}}\right)$. This bound compares favorably to the existing $O(\sqrt{d}/\varepsilon)$ rates when a high-accuracy solution is desired.

The remainder of the paper is organized as follows. Section 2 is devoted to background on both the Langevin diffusion and various higher-order variants of it. In Section 3, we describe the third-order Langevin scheme analyzed in this paper and state our two main results: Theorem 1 for the case of ridge-separable functions, and Theorem 2 for general functions under black-box gradient access with additional smoothness. Section 4 is devoted to the proofs of our main results, with more technical aspects of the arguments deferred to the appendices. We conclude with a discussion in Section 5.

2. Background and problem formulation

In this section, we first introduce the class of sampling problems that are our focus, before turning to specific sampling algorithms that are based on discretizations of diffusion processes. We begin with the classical first-order discretization of Langevin diffusion, and then introduce the higher-order discretization that is our principal object of study.

2.1 A class of sampling problems

We consider the problem of drawing samples from a distribution with density written in the form $p^*(\theta) \propto \exp(-U(\theta))$. The *potential function* $U : \mathbb{R}^d \to \mathbb{R}$ is assumed to be strongly convex and smooth in the following sense:

Assumption 1 (Strong convexity and smoothness) *The function U is differentiable, m-strongly convex and L-smooth, meaning that*

$$\frac{m}{2}\|\theta' - \theta\|_2^2 \;\leq\; U(\theta') - U(\theta) - \langle \nabla U(\theta),\, \theta' - \theta\rangle \;\leq\; \frac{L}{2}\|\theta' - \theta\|_2^2 \qquad \text{for all } \theta, \theta' \in \mathbb{R}^d, \tag{2}$$

where $0 < m \leq L$ are constants.

We say that the potential is (m, L)-*convex-smooth* when this sandwich relation holds. The *condition number* of the problem is given by the ratio $\kappa := \frac{L}{m} \in [1, \infty)$.

Given an iterative algorithm that generates a random vector $\theta^{(k)}$ at each step k, we use $\pi^{(k)}$ to denote the law of $\theta^{(k)}$. We are interested in the convergence $\pi^{(k)}$ to the measure π^* defined by the target density p^*. In order to quantify closeness of the measures $\pi^{(k)}$ and π^*, we use the Wasserstein-2 distance (see the book by Villani (2009) for background). Given a pair of distributions p and q on $\mathbb{R}^d$, a *coupling* γ is a joint distribution over the product space $\mathbb{R}^d \times \mathbb{R}^d$ that has p and q as its marginal distributions. We let $\Gamma(p, q)$ denote the space of all possible couplings of p and q. With this notation, the *Wasserstein-2 distance* is given by

$$\mathcal{W}_2^2(p, q) := \inf_{\gamma \in \Gamma(p,q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y). \tag{3}$$

Our goal is to analyze the ε-mixing time in terms of this distance: the Wasserstein mixing time is the minimum number of steps the algorithm takes to converge to within ε of the target measure π^* in $\mathcal{W}_2$ distance:

$$T_{\text{mix}}(\varepsilon) := \min\left\{k \;\mid\; \mathcal{W}_2(\pi^{(k)}, \pi^*) \le \varepsilon\right\}. \tag{4}$$

2.2 First-order Langevin algorithm

We refer to the stochastic process represented by the following stochastic differential equation as the *continuous-time Langevin dynamics*:

$$\mathrm{d}\theta_t = -\frac{1}{L}\nabla U(\theta_t)\ \mathrm{d}t + \sqrt{2/L}\ \mathrm{d}B_t. \tag{5}$$

Here the reader should recall that L is the smoothness parameter from Assumption 1.

It is a classical result that the distribution of the continuous-time Langevin dynamics converges to the distribution with density p^* at an exponential rate; moreover, under Assumption 1, it is known (Ledoux, 2000) that the convergence rate is independent of the dimension d. However, after discretization using the Euler scheme, the resulting algorithm—a discrete-time stochastic process—has a mixing rate that scales as $\mathcal{O}(d)$; see Dalalyan (2017). As this result makes clear, the principal difficulty in high-dimensional sampling problems based on Langevin diffusion is the numerical error that arises from the integration of the continuous-time dynamics. A standard approach to this problem is to introduce higher-order forms of discretization, but in the setting of Langevin diffusion a major challenge arises—namely, the non-smoothness of the Brownian motion makes it difficult to control the higher-order numerical error.

2.3 Underdamped (second-order) Langevin dynamics

One way to proceed is to augment the dynamics to yield smoother trajectories that are more readily discretized. For example, the *second-order Langevin algorithm*, also known as the underdamped Langevin algorithm, lifts the original d-dimensional space to a $2d$-dimensional space consisting of vectors of the form $x = (\theta, r) \in \mathbb{R}^d \times \mathbb{R}^d$, and defines the following SDE:

$$\begin{cases} \mathrm{d}\theta_t &= r_t\, \mathrm{d}t \\ \mathrm{d}r_t &= -\frac{1}{L}\nabla U(\theta_t)\, \mathrm{d}t - \xi r_t\, \mathrm{d}t + \sqrt{2\xi/L}\ \mathrm{d}B_t^r, \end{cases} \tag{6}$$

where $\xi > 0$ is an algorithmic parameter.

In the second-order Langevin dynamics determined by the system (6), the trajectory θ_t has one additional order of smoothness compared to the Brownian motion B_t^r. As a result, it is possible to introduce higher-order discretizations for the augmented dynamics. Examples of such discretizations include Hamiltonian Monte Carlo (Neal, 2010; Chen et al., 2020) and underdamped Langevin algorithms (Cheng et al., 2018b), both of which are derived from equation (6). These methods can provably accelerate convergence; in particular, the underdamped Langevin algorithm provides a convergence rate of $\mathcal{O}(\sqrt{d})$ when the potential function U is strongly convex and Lipschitz smooth.

2.4 A third-order scheme

It is natural to ask whether one can further accelerate the convergence of Langevin algorithms via higher-order dynamics, where we expand the ambient space and drive the variable of interest via higher-order integration of an SDE. In order to pursue this question, let us recall a generic recipe for constructing Markov dynamics with a desired stationary distribution. Consider the family of SDEs of the form

$$\mathrm{d}x_t = (\mathrm{D}+\mathrm{Q})\nabla H(x_t)\ \mathrm{d}t + \sqrt{2\mathrm{D}}\ \mathrm{d}B_t, \tag{7}$$

where D is a positive semidefinite matrix, and Q is a constant skew-symmetric matrix. It can be shown (Ma et al., 2015, 2018) that for any choice of the matrices (D, Q) respecting these constraints, the SDE in equation (7) has $p^*(x) \propto \exp(-H(x))$ as its invariant distribution.

Within this general framework, note that the second-order Langevin dynamics (6) are obtained by setting $x_t = (\theta_t, r_t)$, and choosing

$$H(x_t) = H(\theta_t, r_t) = U(\theta_t) + \frac{L}{2}\left\|r_t\right\|^2, \quad \mathrm{D} = \frac{1}{L}\begin{bmatrix} 0 & 0 \\ 0 & \xi \mathrm{I} \end{bmatrix}, \quad \text{and} \quad \mathrm{Q} = \frac{1}{L}\begin{bmatrix} 0 & \mathrm{I} \\ -\mathrm{I} & 0 \end{bmatrix}.$$

Note that the positive semidefinite matrix D has a zero top-left block matrix (corresponding to the θ coordinates), which means that θ_t is not directly coupled with the Brownian motion.

This observation motivates us to choose an even more singular D matrix. Beginning from the general equation (7), let $x_t = (\theta_t, p_t, r_t)$, and define the function $H(x_t) = U(\theta_t) + \frac{L}{2}\left\|p_t\right\|^2 + \frac{L}{2}\left\|r_t\right\|^2$, along with the matrices

$$\mathrm{D} = \frac{1}{L}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \xi\mathrm{I} \end{bmatrix}, \quad \text{and} \quad \mathrm{Q} = \frac{1}{L}\begin{bmatrix} 0 & \mathrm{I} & 0 \\ -\mathrm{I} & 0 & \gamma\mathrm{I} \\ 0 & -\gamma\mathrm{I} & 0 \end{bmatrix}.$$

Given these definitions, we set up a third-order form of Langevin dynamics as follows:

$$\begin{cases} \mathrm{d}\theta_t = p_t\ \mathrm{d}t \\ \mathrm{d}p_t = -\frac{1}{L}\nabla U(\theta_t)\ \mathrm{d}t + \gamma r_t\ \mathrm{d}t \\ \mathrm{d}r_t = -\gamma p_t\ \mathrm{d}t - \xi r_t\ \mathrm{d}t + \sqrt{2\xi/L}\ \mathrm{d}B_t^r. \end{cases} \tag{8}$$

The trajectory of θ_t under these third-order dynamics is even smoother than the corresponding trajectory under the underdamped Langevin dynamics; this increased smoothness affords more control over discretization errors. In particular, in our numerical analysis of equation (8), we exploit the fact that the Brownian motion B_t^r and ∇U are passed into the time derivative of two different variables, r_t and p_t, respectively. This allows us to adopt a splitting scheme that takes advantage of the structure of U and thereby provides an improvement in convergence rate relative to past work. In Section 3, we describe the integration scheme that yields this faster convergence rate.

3. Main results

In this section, we describe our higher-order Langevin algorithm, and state two theorems that characterize its convergence rate.

3.1 Discretized third-order algorithm

At a high level, we propose an algorithm, akin to the Langevin or underdamped Langevin algorithm, that at every iteration generates a normal random variable centered according to the previous iterate (see Algorithm 1). The algorithm is constructed via a three-stage discretization of the continuous-time Markov dynamics (8). See Section 3.4 for a detailed discussion of the discretization scheme.

Recalling the (m, L)-strong-convexity-smoothness condition given in Assumption 1, we see that the potential function U has a unique global minimizer $\theta^* \in \mathbb{R}^d$ such that $\nabla U(\theta^*) = 0$. We initialize our algorithm at a point θ_0 satisfying $\|\theta_0 - \theta^*\|_2 \leq \frac{1}{L}$. Such a point can be found in $O(\sqrt{\kappa}\log(d\kappa))$ gradient evaluations using accelerated gradient methods (Nesterov, 2004). Our algorithm generates a sequence of vector triples $x^{(k)} = (\theta^{(k)}, p^{(k)}, r^{(k)})$ for $k = 1, 2, \ldots$ in a recursive manner. Any instance of the algorithm is specified by a stepsize parameter $\eta > 0$, two positive auxiliary parameters γ and ξ, and a function $\Delta U : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We provide specific choices of these parameters and the function ΔU in the theory to follow.

Algorithm 1: Discretized Third-Order Langevin Algorithm

Let $x^{(0)} = (\theta^{(0)}, p^{(0)}, r^{(0)}) = (\theta^*, 0, 0)$.
for $k = 0, \cdots, N-1$ **do**
  Sample $x^{(k+1)} \sim \mathcal{N}\left(\boldsymbol{\mu}\left(x^{(k)}\right), \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are defined in (9a) and (9b).
end for

Given the iterate $x^{(k)}$ at step k, the next iterate $x^{(k+1)}$ is obtained by drawing from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}(x^{(k)})$, where

$$\boldsymbol{\mu}(x) := \begin{pmatrix} \theta - \frac{\eta}{2L}\Delta U(\theta, p) + \mu_{12}p + \mu_{13}r \\ -\frac{1}{L}\Delta U(\theta, p) + \mu_{22}p + \mu_{23}r \\ \frac{\mu_{31}}{L}\Delta U(\theta, p) + \mu_{32}p + \mu_{33}r \end{pmatrix}, \tag{9a}$$

and covariance

$$\boldsymbol{\Sigma} := \frac{1}{L}\begin{pmatrix} \sigma_{11}\cdot \mathrm{I}_{d\times d} & \sigma_{12}\cdot \mathrm{I}_{d\times d} & \sigma_{13}\cdot \mathrm{I}_{d\times d} \\ \sigma_{12}\cdot \mathrm{I}_{d\times d} & \sigma_{22}\cdot \mathrm{I}_{d\times d} & \sigma_{23}\cdot \mathrm{I}_{d\times d} \\ \sigma_{13}\cdot \mathrm{I}_{d\times d} & \sigma_{23}\cdot \mathrm{I}_{d\times d} & \sigma_{33}\cdot \mathrm{I}_{d\times d} \end{pmatrix}. \tag{9b}$$

The constants μ_{12}–μ_{33}, as well as σ_{11}–σ_{33} above are entirely determined by the triple (γ, ξ, η); see Appendix C for their explicit definitions.

We make a few remarks about the algorithm:

- The vector $\Delta U(\theta, p)$ is chosen to be either an exact or approximate value of the integral $\int_0^\eta \nabla U(\theta + tp)dt$. As we discuss in the two versions of the main theorem, different choices of $\Delta U(\theta, p)$ are available depending on the starting assumptions, and each such choice leads to a different mixing time bound.

- In each iteration, we only need to compute $\Delta U(\theta, p)$ once. Below we provide choices of the function ΔU for which this step has equivalent computational cost with a gradient evaluation.
- As we will see from the derivation in Section 3.4 and the closed-form expressions in Appendix C, when the stepsize η is small, the leading terms in $(\mu(x^{(k)}), \Sigma)$ are the same as the dynamics in equation (8). However, the high-order correction terms are essential for achieving accelerated rates. This high-order scheme allows us to separate ∇U, the only nonlinear part of the equation, and carry out a direct integration on a deterministic path.
- While our description allows for different choices of the parameters γ and ξ, in our analysis, we adopt the choices $\gamma = \kappa$ and $\xi = 2\kappa$.

3.2 Guarantees for ridge-separable potentials

We begin by describing and analyzing a version of our algorithm applicable when the potential function U is of ridge-separable form (1). In this case, the integral $\int_0^\eta \nabla U(\theta + tp)dt$ can be computed exactly using the Newton-Leibniz formula. This fact allows us to run Algorithm 1 with the choice

$$\Delta U(\theta, p) := \int_0^\eta \left(\sum_{i=1}^n u_i'(a_i^{\mathrm{T}}(\theta + tp)) a_i \right) dt \; = \; \sum_i \left(u_i \left(a_i^{\mathrm{T}}(\theta + \eta p) \right) - u_i \left(a_i^{\mathrm{T}} \theta \right) \right) \frac{a_i}{a_i^{\mathrm{T}} p}. \tag{10}$$

Note that for ridge-separable functions, the total computational cost needed for computing $\Delta U(\theta, p)$ is of the same order as the cost for computing the gradient itself.

We claim that an $O(d^{\frac{1}{4}})$ mixing time can be achieved in this way. More precisely, we have:

Theorem 1 *Let U be an (m, L)-convex-smooth potential of the form* (1). *Given a desired Wasserstein accuracy $\varepsilon \in (0, \sqrt{d/L})$, suppose that we run Algorithm 1 with stepsize $\eta = c\kappa^{-\frac{11}{4}} d^{-\frac{1}{4}} L^{\frac{1}{4}} \varepsilon^{\frac{1}{2}}$, using the function $\Delta U(\theta, p)$ defined in equation* (10). *Then there is a universal constant C such that the mixing time is bounded as*

$$T_{mix}(\varepsilon) \leq C \cdot \kappa^{\frac{19}{4}} \, (d/L)^{\frac{1}{4}} \left(\frac{1}{\sqrt{\varepsilon}} \right) \log \left(\frac{d\kappa}{\varepsilon} \right).$$

Note that the result holds true for any potential function of the form of equation (1), regardless of the distribution of the data points. In particular, we do not assume any form of incoherence, as in the work of Lee et al. (2018) and Mangoubi and Smith (2017), nor do we impose any conditions on the norm of vectors a_i. Furthermore, only strong convexity and smoothness assumptions are used, without requiring high-order smoothness assumptions. Many log-concave sampling problems of practical interest in statistical applications arise from a Gibbs measure defined by generalized linear potential functions. Under this setup, our result significantly improves the previous best known $O(d^{1/3}/\varepsilon^{2/3})$ rate (Shen and Lee, 2019) in the dependency on both ε and d. Finally, it is also worth noticing that the ridge-separable form is needed only to ensure the close-form expression (10). Theorem 1 also applies to a function that does not satisfy equation (1), but allows the closed-form integration of $\Delta U(\theta, p)$.

As a caveat, we note that the stepsize used in running Algorithm 1 depends on knowledge of κ and L; these problem-specific parameters may be unknown in practice. An important direction for future work is to provide an automated mechanism for stepsize selection with similar guarantees. We note that such automated procedures exist in the context of optimization algorithms (e.g., backtracking line search).

3.3 Guarantees under black-box gradient access

We now turn to the more general setting, in which the potential function U is no longer ridge-separable. Suppose moreover that we have access to U only via a black-box gradient oracle, meaning that we can compute the gradient $\nabla U(\theta)$ at any point θ of our choice. Under these assumptions, the closed-form integrator described in Algorithm 1 is no longer available. However, by using Lagrange interpolation as an approximation, we can still derive a practical high-order algorithm that yields a faster mixing time. In particular, while the mixing time scales as $O(\sqrt{d})$, as with lower-order methods, we show that the ε-dependency term can be adaptive to the degree of smoothness of the function U.

Lagrange interpolation: When the objective U does not take the form of a generalized linear function, we use Lagrange interpolating polynomials with Chebyshev nodes (Stoer and Bulirsch, 2002) to approximate the function:

$$s \mapsto \nabla U\left(\theta^{(k)} + (s - k\eta)p^{(k)}\right) \quad \text{over the interval } s \in [k\eta, (k+1)\eta].$$

In particular, for a given smoothness parameter $\alpha \in \mathbb{N}^+$, the Chebyshev nodes are given by:

$$s_i = k\eta + \frac{\eta}{2}\left(1 + \cos\frac{2i-1}{2\alpha}\pi\right) \quad \text{for } i = 1, 2, \ldots, \alpha.$$

The Chebyshev polynomial interpolation operator Φ takes as inputs a scalar t, and a function $z : [0, \eta] \to \mathbb{R}^d$, and returns the scalar $\Phi(t; z) := \sum_{i=1}^{\alpha} z(s_i) \prod_{j \neq i} \frac{t - s_i}{s_j - s_i}$. Note that the integral of this function over t can be computed in closed form.

For each pair (θ, p) define the mapping $\psi_{\theta,p}(s) = \nabla U(\theta + sp)$ from $\mathbb{R}$ to $\mathbb{R}^d$. Applying the interpolation polynomial to this mapping, we define

$$\Delta U(\theta, p) := \int_0^{\eta} \Phi(t; \psi_{\theta,p}) dt. \tag{11}$$

Note that $t \mapsto \Phi(t; \psi_{\theta,p})$ is a polynomial function, and hence the integral over t can be computed exactly. Computing this integral requires α gradient evaluations in total, along with additional computational costs to calculate the linear combination of the gradient. Thus, when the smoothness α is viewed as a constant, the computational complexity is order-equivalent to a gradient evaluation.

As is well known from the numerical analysis literature, higher-order polynomial approximations are suitable to approximate functions that satisfy higher-order smoothness condition. In our analysis, we impose a higher-order smoothness condition on U in the following way. Note that the gradient $\nabla U(\theta)$ at any given θ is simply a vector, or equivalently a first-order tensor. For a first-order tensor T, its tensor norm is given by $|\!|\!| T |\!|\!|_{\text{tsr}}^{(1)} = \|T\|_2$, corresponding to the ordinary Euclidean norm. For a k-th order tensor T, we recursively define its tensor norm as $|\!|\!| T |\!|\!|_{\text{tsr}}^{(k)} := \sup_{v \in \mathbb{S}^{d-1}} |\!|\!| Tv |\!|\!|_{\text{tsr}}^{(k-1)}$, where $\mathbb{S}^{d-1}$ denotes the Euclidean sphere in d-dimensions. With this definition, the second-order tensor $|\!|\!| \cdot |\!|\!|_{\text{tsr}}^{(2)}$ norm for a matrix is equivalent to its ℓ_2-operator norm.

Assumption 2 (High-order smoothness) *For some $\alpha \geq 2$, the potential function U is α-th order differentiable, and the associated tensor of derivatives satisfies the bound*

$$|\!|\!| \nabla^\alpha U(x) |\!|\!|_{tsr}^{(\alpha)} \leq L_\alpha^{\alpha-1},$$

for some $L_\alpha > 0$.

Note that in the special case $\alpha = 3$, Assumption 2 corresponds to a Lipschitz condition on the Hessian function, as has been used in past analysis of sampling algorithms.

Under this smoothness assumption, we have the following guarantee:

Theorem 2 *Consider a potential U satisfying Assumptions 1 and 2 for some $\alpha \geq 2$. Given a desired Wasserstein accuracy $\varepsilon \in (0, \sqrt{d/L})$, suppose that we run Algorithm 1 with stepsize*

$$\eta = c \cdot \min \left(\kappa^{-\frac{11}{4}} d^{-\frac{1}{4}} L^{\frac{1}{4}} \varepsilon^{\frac{1}{2}}, L_\alpha^{-1} L^{\frac{1}{2}} \kappa^{-4} d^{-\frac{1}{2}} \varepsilon^{\frac{1}{\alpha-1}}, \kappa^{-5} \right), \tag{12a}$$

using $\Delta U(\theta, p) = \int_0^\eta \Phi(t; \psi_{\theta,p}) dt$ where $\psi_{\theta,p}(s) = \nabla U(\theta + sp)$. Then there is a universal constant C such that

$$T_{mix}(\varepsilon) \leq C \cdot \max \left(L_\alpha \kappa^6 \sqrt{d/L} \varepsilon^{-\frac{1}{\alpha-1}}, \kappa^{\frac{19}{4}} (d/L)^{\frac{1}{4}} \varepsilon^{-\frac{1}{2}}, \kappa^7 \right). \tag{12b}$$

We observe that the dimension dependence becomes $d^{1/2}$, but the corresponding ϵ dependence is reduced to $\epsilon^{-1/(\alpha-1)}$, so higher-order smoothness leads to better accuracy dependence.

3.4 Derivation of the discretization

In this section, we show how Algorithm 1 arises from a particular discretization scheme applied to the continuous-time process (8). Our overall approach involves a combination of a splitting method and a high-order integration scheme. More precisely, we reduce the problem of one-step simulation of equation (8) to (approximately) computing the integral of ∇U along a straight line, using three stages of discretization. Letting $\widehat{g}_s$ be an approximation for $\nabla U \left(\theta^{(k)} + (s - k\eta) p^{(k)} \right)$ and letting $\Delta U(\theta^{(k)}, p^{(k)}) := \int_{k\eta}^{(k+1)\eta} \widehat{g}_s ds$, the discretization error bound depends on the accuracy with which $\widehat{g}_s$ approximates ∇U. Depending on the assumptions imposed on the target distribution, various choices of $\widehat{g}$ can be used, the exact integration of which leads to different choices of ΔU. Theorems 1 and 2 correspond to two instances of this general approach.

Our three-stage discretization scheme is similar to classical splitting schemes for Langevin dynamics (Abdulle et al., 2015; Leimkuhler and Shang, 2016). We construct two continuous-time processes: (a) an interpolation process $(\widetilde{\theta}, \widetilde{p}, \widetilde{r})$; and (b) an auxiliary process $(\widehat{\theta}, \widehat{p}, \widehat{r})$

defined over the time interval $[k\eta, (k+1)\eta)$. At time $k\eta$, both processes are started from point $(\theta^{(k)}, p^{(k)}, r^{(k)})$, and after one step of Algorithm 1, we are guaranteed to have the equivalence $(\theta^{(k+1)}, p^{(k+1)}, r^{(k+1)}) = (\widetilde{\theta}_{(k+1)\eta}, \widetilde{p}_{(k+1)\eta}, \widetilde{r}_{(k+1)\eta})$. In other words, the process $(\widetilde{\theta}_t, \widetilde{p}_t, \widetilde{r}_t)_{k\eta \le t \le (k+1)\eta}$ forms a continuous interpolation between discrete-time steps of Algorithm 1. It is worth noting that we only need to calculate the process $(\widetilde{\theta}, \widetilde{p}, \widetilde{r})$ at time points that are integer multiples of η in an algorithmic manner, which is explicitly derived in Appendix C.

At a high level, the procedure summarized in equations (13)—(15) first computes coarse estimates $(\widehat{\theta}_t, \widehat{p}_t, \widehat{r}_t)_{k\eta \le t \le (k+1)\eta}$ for the path taken by the third-order dynamics over the time interval $[k\eta, (k+1)\eta]$. Then, a finer approximation $(\widetilde{\theta}_t, \widetilde{p}_t, \widetilde{r}_t)_{k\eta \le t \le (k+1)\eta}$ is constructed based on the coarse estimates. The key idea for the construction of the discretization scheme is that the integration for the p-part of the dynamics is performed separately over a deterministically chosen line (c.f. equation (14)). This choice makes the use of direct integration methods possible.

First stage: We begin by constructing estimators $\widehat{\theta}_{(k+1)\eta}$ and $\widehat{r}_{(k+1)\eta}$ following the Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930):

$$\begin{cases} \mathrm{d}\widehat{\theta}_t = \widetilde{p}_{k\eta}\mathrm{d}t, \\ \mathrm{d}\widehat{r}_t = -\gamma\widetilde{p}_{k\eta}\mathrm{d}t - \xi\widehat{r}_t\mathrm{d}t + \sqrt{2\xi/L}\mathrm{d}B_t^r, \end{cases} \qquad \text{for all } t \in [k\eta, (k+1)\eta]. \tag{13}$$

The values $(\widehat{\theta}_{(k+1)\eta}, \widehat{r}_{(k+1)\eta})$ are then used to calculate a high-accuracy result by adding a correction term. We use the function $\widehat{g}_s$ as an approximation of the gradient

$$\nabla U(\widehat{\theta}_s) = \nabla U\left(\theta^{(k)} + (s - k\eta)p^{(k)}\right).$$

It is worth noting that $\widehat{g}_s$ approximates a function along a fixed curve determined by $x_{k\eta}$, and has no interaction with other variables nor the noise. This makes it possible to obtain high-accuracy solutions to the equation by integration of a deterministic and known function.

Second stage: In the second stage, we solve the system of differential equations

$$\begin{cases} \mathrm{d}\widetilde{p}_t = -\frac{1}{L}\widehat{g}_t\mathrm{d}t + \gamma\widehat{r}_t dt, \\ \mathrm{d}\widehat{p}_t = -\frac{1}{L\eta}\left(\int_{k\eta}^{(k+1)\eta} \widehat{g}_s\mathrm{d}s\right)\mathrm{d}t + \gamma\widehat{r}_t dt, \end{cases} \qquad \forall t \in [k\eta, (k+1)\eta]. \tag{14}$$

Solving these equations amounts to integrating $\widehat{g}_t$ and $\widehat{r}_t$, whereas the quantity $\widehat{p}_t$ corresponds to a linear approximation of the integral component of $\widetilde{p}_t$. This choice ensures that calculations for $\widetilde{\theta}$ in our third stage are straightforward. From equation (14), we always have $\widetilde{p}_{(k+1)\eta} = \widehat{p}_{(k+1)\eta}$, which can be used as the value of $p^{(k+1)}$ in Algorithm 1.

Third stage: In the third stage, we use the estimator $\widehat{p}_t$ constructed from equation (14) in order to construct approximate solutions to the following systems of SDEs:

$$\begin{cases} \mathrm{d}\widetilde{\theta}_t = \widehat{p}_t\mathrm{d}t \\ \mathrm{d}\widetilde{r}_t = -\gamma\widehat{p}_t\mathrm{d}t - \xi\widetilde{r}_t\mathrm{d}t + \sqrt{2\xi/L}\ \mathrm{d}B_t^r, \end{cases} \qquad t \in [k\eta, (k+1)\eta]. \tag{15}$$

Note that the Brownian motion $(B_t^r : t \geq 0)$ used in process (15) must be the same as that used in process (13), so that the two processes must be solved jointly. As shown in Appendix C, we can carry out the integrations in closed form, so as to obtain the explicit quantities required to implement Algorithm 1.

Choices of approximation $\widehat{g}_t$: The three-stage discretization scheme that we have described is a general framework, where we are free to make different choices of $\Delta(\theta, p) = \int_0^\eta \widehat{g}_{k\eta+s} \mathrm{d}s$. The choice of $\widehat{g}_t$ is constrained by the need to make the integration exactly solvable, and it has to serve as a good approximation for $\nabla U(\widehat{\theta}_t)$. If the potential function is of the form defined in equation (1), we can simply take $\widehat{g}_t = \nabla U(\theta + tp)$, and the integration can be carried out in closed form by the Newton-Leibniz formula. Alternatively, if U is given by a black-box gradient oracle and satisfies appropriate higher-order smoothness conditions, we can use the Chebyshev node-interpolation method, and approximate $\nabla U(\theta + tp)$ using a polynomial in t. In such a case, the time integral of $\widehat{g}_t$ can also be computed exactly.

4. Proofs

In this section, we provide the proofs of our main results. We begin in Section 4.1 by stating and proving a result (Proposition 3) on the exponential convergence of the third-order dynamics in continuous time. Section 4.2 is devoted to our proofs of our main results (Theorems 1 and 2) on the behavior of the discrete-time algorithm. In all cases, we defer the proofs of more technical results to the appendices.

4.1 Exponential convergence in continuous time

We begin by studying the process $\{x_t\}_{t\geq 0}$ defined by the continuous-time third-order dynamics (8), with the particular goal of showing convergence in the Wasserstein-2 distance. In all of our analysis—in this section as well as others—we make the choices $\gamma = \kappa$ and $\xi = 2\kappa$ in defining the dynamics. Throughout this section, we use $\mathrm{eig}_i(A)$ to denote the i-th largest eigenvalue for a real symmetric matrix A.

It is known (Ma et al., 2015, 2018) that the limiting stationary distribution of the process $x_t = (\theta_t, p_t)$ has a product form:

$$p^*(x) \propto e^{-H(x)} = e^{-U(\theta) - \frac{L}{2}\|p\|^2 - \frac{L}{2}\|r\|^2}.$$

Our goal is to show that the distribution of $x_t = (\theta_t, p_t)$ converges at an exponential rate in the Wasserstein-2 distance, as previously defined in equation (3), to this expanded target distribution.

In order to do so, we consider two processes following the third-order dynamics (8), where the process $\{x_t\}_{t\geq 0}$ and $\{x_t^*\}_{t\geq 0}$ are started, respectively, from the initial distributions p_0 and p^*. We then couple these two processes via a synchronous coupling. In order to establish a convergence rate, we make use of the following Lyapunov function:

$$t \mapsto \mathcal{L}_t = \inf_{\zeta_t \in \Gamma(p_t, p^*)} \mathbb{E}_{(x_t, x^*) \sim \zeta_t} \left[(x_t - x^*)^{\mathrm{T}} S (x_t - x^*) \right], \tag{16a}$$

where the symmetric matrix S is given by

$$S := \begin{pmatrix} \frac{\kappa^7+3\kappa^4+5\kappa^3+\kappa+1}{4\kappa^5}\mathrm{I} & \frac{\kappa}{2}\mathrm{I} & \frac{1}{4}\left(1-\frac{1}{\kappa^3}-\frac{1}{\kappa^4}\right)\kappa\mathrm{I} \\ \frac{\kappa}{2}\mathrm{I} & \frac{4\kappa^4+6\kappa^3+\kappa+1}{4\kappa^4}\mathrm{I} & \frac{\kappa+1}{2\kappa}\mathrm{I} \\ \frac{1}{4}\left(1-\frac{1}{\kappa^3}-\frac{1}{\kappa^4}\right)\kappa\mathrm{I} & \frac{\kappa+1}{2\kappa}\mathrm{I} & \frac{\kappa+2}{4\kappa}\mathrm{I} \end{pmatrix}. \tag{16b}$$

Here, the matrix S is constructed in order to guarantee $S \succ 0$ and that $SM + MS \prec 0$ for any block matrix M of the form:

$$M = \frac{1}{L}\begin{bmatrix} 0 & \mathrm{I} & 0 \\ -H & 0 & \gamma\mathrm{I} \\ 0 & -\gamma\mathrm{I} & 0 \end{bmatrix}, \quad \text{for some symmetric } H \text{ such that } m\mathrm{I} \preceq H \preceq L\mathrm{I}.$$

This condition is sufficient for the exponential convergence in $\mathcal{W}_2$ distance. The form of this condition resembles the celebrated Lyapunov stability condition for linear systems. In the 3×3 block matrix case, the problem of constructing S reduces to the construction of cubic equation problems. See Section 4.1 for the proof of the desired properties. With this setup, our main result on the continuous-time dynamics is the following:

Proposition 3 *Let the random initializations x_0 and x_0^* follow the laws of p_0 and p^*, respectively. Then the process $\{x_t\}_{t\geq 0}$ defined by the dynamics* (8) *satisfies the bound*

$$\inf_{\zeta_t \in \Gamma(p_t, p^*)} \mathbb{E}_{(x_t, x^*)\sim\zeta_t}\left[(x_t - x^*)^{\mathrm{T}} S (x_t - x^*)\right] \leq e^{-\frac{t}{5\kappa^2+50}} \inf_{\zeta_0 \in \Gamma(p_0, p^*)} \mathbb{E}_{(x_0, x^*)\sim\zeta_0}\left[(x_0 - x^*)^{\mathrm{T}} S (x_0 - x^*)\right].$$

As shown in Lemma 6, to be stated in the next section, the eigenvalues of S lie in the interval $[1/(5\kappa),\ \kappa^2 + 10]$. Thus, Proposition 3 implies convergence in the Wasserstein-2 distance at an exponential rate.

The remainder of this section is devoted to the proof of Proposition 3. The first step in the proof involves establishing a differential inequality for the Lyapunov function $\mathcal{L}_t$ via the coupling technique.

Lemma 4 *Let the processes $\{x_t\}_{t\geq 0}$ and $\{x_t^*\}_{t\geq 0}$ follow the third-order dynamics* (8) *with initial conditions x_0 and $x_0^* \in \mathbb{R}^{3d}$. Then there exists a coupling $\bar{\zeta} \in \Gamma\left(\mathcal{L}(x_t : t \geq 0), \mathcal{L}(x_t^* : t \geq 0)\right)$ of the laws of the processes $\{x_t\}_{t\geq 0}$ and $\{x_t^*\}_{t\geq 0}$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}t}(x_t - x_t^*)^{\mathrm{T}} S (x_t - x_t^*) \leq -\frac{1}{5\kappa^2 + 50}(x_t - x_t^*)^{\mathrm{T}} S (x_t - x_t^*) \qquad \textit{for all } t \geq 0. \tag{17}$$

The proof of Lemma 4, given in Section 4.1.1, is based on the synchronous coupling technique, in which two processes are coupled based on the same underlying Brownian motion.

Taking Lemma 4 as given, we can now complete the proof of Proposition 3. Applying Grönwall's lemma to equation (17) yields

$$(x_t - x_t^*)^{\mathrm{T}} S (x_t - x_t^*) \leq e^{-t/(5\kappa^2+50)}\left((x_0 - x_0^*)^{\mathrm{T}} S (x_0 - x_0^*)\right).$$

Let ζ_0^* denote the optimal coupling between the initial distributions p_0 and p^*. Noting that $\widehat{\zeta} = \mathbb{E}_{(x_0,x_0^*)\sim\zeta_0^*}\left[\cdot|x_0, x_0^*)\right]$ is a coupling, and let $\widehat{\zeta}_t$ be its marginal at time point t. We find that:

$$\begin{aligned}
&\inf_{\zeta_t\in\Gamma(p_t,p_t^*)} \mathbb{E}_{(x_t,x_t^*)\sim\zeta_t}\left[(x_t - x_t^*)^{\mathrm{T}} S(x_t - x_t^*)\right] \\
&\leq \mathbb{E}_{(x_t,x^*)\sim\widehat{\zeta}_t}\left[(x_t - x_t^*)^{\mathrm{T}} S(x_t - x_t^*)\right] \\
&= \mathbb{E}_{(x_0,x_0^*)\sim\zeta_0^*}\left[\mathbb{E}_{(x_t,x^*)\sim\bar{\zeta}(x_t,x_t^*|x_0,x_0^*)}\left[(x_t - x_t^*)^{\mathrm{T}} S(x_t - x_t^*)\right]\right] \\
&\leq \mathbb{E}_{(x_0,x_0^*)\sim\zeta_0^*}\left[e^{-t/(5\kappa^2+50)}\left((x_0 - x_0^*)^{\mathrm{T}} S(x_0 - x_0^*)\right)\right] \\
&= e^{-t/(5\kappa^2+50)} \inf_{\zeta_0\in\Gamma(p_0,p^*)} \mathbb{E}_{(x_0,x^*)\sim\zeta_0}\left[(x_0 - x^*)^{\mathrm{T}} S(x_0 - x^*)\right],
\end{aligned}$$

which establishes the bound in Proposition 3.

4.1.1 Proof of Lemma 4

We prove Lemma 4 by choosing a synchronous coupling $\bar{\zeta} \in \Gamma\left(\mathcal{L}(x_t)_{t\geq 0}, \mathcal{L}(x_t^*)_{t\geq 0}\right)$ for the laws of x_t and x_t^*. (A synchronous coupling simply means that we use the same Brownian motion B_t^r in defining both x_t and x_t^*.) In this way, for any pair $(x_t, x_t^*) \sim \bar{\zeta}$, we have the equivalence

$$\begin{pmatrix} \mathrm{d}(\theta_t - \theta_t^*) \\ \mathrm{d}(p_t - p_t^*) \\ \mathrm{d}(r_t - r_t^*) \end{pmatrix} = (\mathrm{D} + \mathrm{Q}) \begin{pmatrix} \nabla U(\theta_t) - \nabla U(\theta_t^*) \\ L(p_t - p_t^*) \\ L(r_t - r_t^*) \end{pmatrix} \mathrm{d}t. \tag{18}$$

Since U is a Lipschitz-smooth function defined on $\mathbb{R}^d$, an open convex domain, we can use the mean-value theorem for vector-valued functions and write $\nabla U(\theta_t) - \nabla U(\theta_t^*) = H_t(\theta_t - \theta_t^*)$, where

$$H_t = \int_0^1 \nabla^2 U\left(\theta_t^* + \lambda\left(\theta_t - \theta_t^*\right)\right) \mathrm{d}\lambda. \tag{19}$$

We obtain that $\mathrm{d}(x_t - x_t^*) = M_t(x_t - x_t^*)\mathrm{d}t$, where the matrix M_t takes the form

$$M_t := \begin{pmatrix} 0 & \mathrm{I} & 0 \\ -\frac{H_t}{L}\mathrm{I} & 0 & \gamma\mathrm{I} \\ 0 & -\gamma\mathrm{I} & -\xi\mathrm{I} \end{pmatrix}.$$

Consequently, the derivative of the function $t \mapsto (x_t - x_t^*)^{\mathrm{T}} S(x_t - x_t^*)$ is given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(x_t - x_t^*)^{\mathrm{T}} S(x_t - x_t^*) &= 2(x_t - x_t^*)^{\mathrm{T}} S M_t(x_t - x_t^*) \\
&= (x_t - x_t^*)^{\mathrm{T}}(S M_t + M_t^{\mathrm{T}} S)(x_t - x_t^*).
\end{aligned}$$

In order to proceed, we need to relate the eigenvalues of the matrix $SM_t + M_t^{\mathrm{T}} S$ to those of S. The following lemmas allow us to carry out this conversion:

Lemma 5 *For any $\kappa = L/m \geq 1$ and matrix H_t of the form (19) such that $m\mathrm{I} \preceq H_t \preceq L\mathrm{I}$, the eigenvalues of $SM_t + M_t^{\mathrm{T}} S$ are smaller than $-1/5$.*

Lemma 6 *For any $\kappa = L/m \geq 1$, the eigenvalues of the matrix S lie in the interval $[\frac{1}{5\kappa},\ \kappa^2 + 10]$.*

Using these two lemmas, it follows that for any pair of random variables $(x_t, x_t^*) \sim \bar{\zeta}$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(x_t - x_t^*)^{\mathrm{T}} S (x_t - x_t^*) &= (x_t - x_t^*)^{\mathrm{T}}(SM_t + M_t^{\mathrm{T}} S)(x_t - x_t^*) \\ &\overset{(i)}{\leq} -\frac{1}{5}\left\|x_t - x_t^*\right\|^2 \\ &\overset{(ii)}{\leq} -\frac{1}{5\kappa^2 + 50}(x_t - x_t^*)^{\mathrm{T}} S (x_t - x_t^*),\end{aligned}$$

where inequality (i) follows from Lemma 5 and inequality (ii) follows from the upper bound on the eigenvalues of S in Lemma 6. This completes the proof of Lemma 4.

Finally, we turn to the proofs of Lemmas 5 and 6.

4.1.2 Proof of Lemma 5

Let $\{l_k\}_{k=1}^d$ denote the eigenvalues of the rescaled matrix H_t/L. Since H_t is the Hessian of the potential function U—which is m-strongly convex and L-Lipschitz smooth—all of the eigenvalues of H_t belong to the interval $[1/\kappa, 1]$.

Letting $\{\lambda_k\}$ denote the eigenvalues of $SM_t + M_t^{\mathrm{T}} S$, the following lemma provides control on them:

Lemma 7 *The eigenvalues $\{\lambda_k\}$ take the form $\lambda_0 = -1$ and*

$$\lambda_k^{\pm} := -\frac{2(l_k + 1/\kappa) \pm (l_k - 1/\kappa)\sqrt{g_1(\kappa)}}{4/\kappa},$$

for a function g_1 (see equation (28a) *in Appendix B.1) such that*

$$2 - \sqrt{g_1(\kappa)} \leq 0 \qquad \text{for all } \kappa \geq 1, \text{ and} \tag{20a}$$

$$-\frac{\kappa}{4}\left(\left(2 - \sqrt{g_1(\kappa)}\right) + 2/\kappa + 1/\kappa\sqrt{g_1(\kappa)}\right) \leq -1/5. \tag{20b}$$

See Appendix B.2 for the proof of this claim.

We now use Lemma 7 to establish the claimed upper bound on the eigenvalues. We first upper bound λ_k^+. Since $(l_k - 1/\kappa)\sqrt{g_1(\kappa)} \geq 0$, we find that

$$\lambda_k^+ \leq -\frac{2(l_k + 1/\kappa)}{4/\kappa} \leq -1 \ \leq\ -\tfrac{1}{5},$$

as claimed.

Turning to the bound on λ_k^-, we can rewrite it as:

$$\lambda_k^- = -\frac{\kappa}{4}\left(\left(2 - \sqrt{g_1(\kappa)}\right) l_k + 2/\kappa + 1/\kappa\sqrt{g_1(\kappa)}\right).$$

Using the bound (20a) and the fact that $l_k \leq 1$, we can upper bound λ_k^- as

$$\lambda_k^- \leq -\frac{\kappa}{4}\left(\left(2 - \sqrt{g_1(\kappa)}\right) + 2/\kappa + 1/\kappa\sqrt{g_1(\kappa)}\right).$$

Finally, applying the bound (20b), we conclude that $\mathrm{eig}_i(SM_t + M_t^{\mathrm{T}} S) \leq -\frac{1}{5}$ for all $i = 1, \cdots, 3d$, which completes the proof of Lemma 5.

4.1.3 Proof of Lemma 6

We show that the eigenvalues $\mathrm{eig}_i(S)$ satisfy a certain third-order equation. For each $i = 1, \ldots, 3d$, we claim that the variable $x = \frac{4}{\kappa^5} \cdot \mathrm{eig}_i(S)$, satisfies the following cubic equation:

$$f(x) = x^3 - g_2(\kappa) \cdot x^2 + (1/\kappa^9) g_3(\kappa) \cdot x - (1/\kappa^{15}) g_3(\kappa) = 0, \tag{21}$$

where the coefficients $g_2(\kappa)$ and $g_3(\kappa)$ are defined in Appendix B.1. Since S is a symmetric matrix, all the roots of equation (21) are real.

In order for the eigenvalues $\mathrm{eig}_i(S)$ to lie in the interval $[\frac{1}{5\kappa}, \kappa^2 + 10]$, it suffices to show that the roots of the function f from equation (21) all lie in the interval $[\frac{4}{5\kappa^6}, \ \frac{4}{\kappa^3} + \frac{40}{\kappa^5}]$.

Lemma 8 *For any $\kappa \geq 1$, the function f in equation (21) satisfies the bounds*

$$\begin{aligned} f(x) < 0 \quad &\text{for all } x \leq \tfrac{4}{5\kappa^6}, \text{ and} \\ f(x) > 0 \quad &\text{for all } x \geq \tfrac{4}{\kappa^3} + \tfrac{40}{\kappa^5}. \end{aligned}$$

We defer the proof of this lemma to Appendix B.3. Note that Lemma 8 implies that all real roots of equation (21) lie in the range of $\left[\frac{4}{5\kappa^6}, \frac{4}{\kappa^3} + \frac{40}{\kappa^5}\right]$, which completes the proof.

4.2 Proofs of discrete-time results

We now turn to the proofs of our two main results—namely, Theorems 1 and Theorem 2—that establish the behavior of the discrete-time algorithm. We begin with a general roadmap for the proofs, along with a key auxiliary result (Proposition 9) common to both arguments.

4.2.1 Roadmap and a key auxiliary result

Let $(\widetilde{x}_t = (\widetilde{\theta}_t, \widetilde{p}_t, \widetilde{r}_t) : t \geq 0)$ be the process defined in Section 3.4. For each $k \in \mathbb{N}$, the random vector $\widetilde{x}_{k\eta}$ has the same distribution as $x^{(k)}$ defined by Algorithm 1. We construct a coupling between the process $\widetilde{x}$ and the stationary diffusion process $(x_t : t \geq 0)$ with $x_0 \sim e^{-U(\theta) - \frac{L}{2}\|p\|_2^2 - \frac{L}{2}\|r\|_2^2}$. (It is obvious that x_t is also following the stationary distribution for any $t > 0$). Given a d-dimensional Brownian motion, $(B_t : t \geq 0)$, we use it to drive both processes. For any $N \in \mathbb{N}$, we have:

$$\mathcal{W}_2(\pi^{(N)}, \pi) \leq \left(\mathbb{E}\left\|\widetilde{x}_{N\eta} - x_{N\eta}\right\|_2^2\right) \leq \|S^{-1}\|_{op}\left(\mathbb{E}\left\|\widetilde{x}_{N\eta} - x_{N\eta}\right\|_S^2\right).$$

In the last step we transform the $\|\cdot\|_2$ norm into the $\|\cdot\|_S$ norm, which is possible because the process $(x_t : t \geq 0)$ is contractive under the $\|\cdot\|_S$ norm.

To analyze the one-step discretization error, we have the following key lemma, which holds in general for any approximation $\widehat{g}_t\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)$ of

$$\nabla U(\widehat{\theta}_{k\eta+t}) = \nabla U\left(\frac{t-k\eta}{\eta}\widehat{\theta}_{(k+1)\eta} + \frac{(k+1)\eta - t}{\eta}\widetilde{\theta}_{k\eta}\right),$$

for $t \in [k\eta, (k+1)\eta]$. Note that Proposition 9 is used in the proof of both theorems.

Proposition 9 *Given an (m, L)-convex-smooth potential, let the process $(\widehat{\theta}_t, \widetilde{\theta}_t, \widetilde{p}_t, \widetilde{r}_t)$ be defined by equations (13)–(15), define $\widetilde{x}_t = (\widetilde{\theta}_t, \widetilde{p}_t, \widetilde{r}_t)$, and let $x_t = (\theta_t, p_t, r_t)$ be generated from the continuous-time dynamics (8) initialized with the stationary distribution. Suppose that we use the same Brownian motion for both processes, and assume that, for $k\eta \leq t \leq (k+1)\eta$, $\widehat{g}_t\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)$ belongs to* $\mathrm{conv}\left(\left\{\nabla U(\lambda\widetilde{\theta}_{k\eta} + (1-\lambda)\widehat{\theta}_{(k+1)\eta}) : \lambda \in [0,1]\right\}\right)$*. Then there is a universal constant C such that*

$$\begin{aligned}\mathbb{E}\left\|\widetilde{x}_{(k+1)\eta} - x_{(k+1)\eta}\right\|_S^2 &\leq \left(1 - \frac{\eta}{20\kappa^2 + 200}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta} - x_{k\eta}\right\|_S^2 + C\frac{\kappa^8\eta^5 d}{L} \\ &\quad + C\frac{\kappa^5\eta\mathbb{E}\Delta_k(g)^2}{L^2} + C\kappa^6\frac{\eta^5}{L^2}\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\frac{\mathrm{d}}{\mathrm{d}s}\widehat{g}_s\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)\right\|_2^2, \end{aligned} \tag{22a}$$

where

$$\Delta_k(g) := \sup_{t\in[k\eta,(k+1)\eta]}\left\|\widehat{g}_t\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right) - \nabla U\left(\frac{t-k\eta}{\eta}\widehat{\theta}_{(k+1)\eta} + \frac{(k+1)\eta-t}{\eta}\widetilde{\theta}_{k\eta}\right)\right\|_2. \tag{22b}$$

See Appendix A for the proof of the proposition. In the following, we give an overview of the proof strategy, highlighting the reason why an $O(d^{1/4})$ mixing time is available.

Note that the process $\widetilde{x}_t$ also satisfies a system of SDEs, with drift terms defined by $\widehat{g}, \widehat{p}$ and $\widehat{r}$. Those drift terms are dependent on the future moves of the Brownian motion, so the interpolated process $(\widetilde{x}_t)_{k\eta\leq t\leq(k+1)\eta}$ is not a Markov diffusion. Nevertheless, we can still compare it with the process (8) using a synchronous coupling, and compute the evolution of $\|\widetilde{x}_t - x_t\|_S^2$ along the path:

$$\begin{aligned}\|\widetilde{x}_t - x_t\|_S^2 &\leq \|\widetilde{x}_{k\eta} - x_{k\eta}\|_S^2 - \frac{1}{20\kappa^2+200}\int_{k\eta}^t \|\widetilde{x}_s - x_s\|_S^2\, ds \\ &\quad + (20\kappa^2+200)\|S\|_{op}\int_{k\eta}^t\left(\left\|\widehat{g}_s - \nabla U(\widetilde{\theta}_s)\right\|_2^2 + \|\widehat{r}_s - \widetilde{r}_s\|_2^2 + (1+\gamma^2)\|\widehat{p}_s - \widetilde{p}_s\|_2^2\right)\mathrm{d}s. \end{aligned} \tag{23}$$

In the right-hand-side of equation (23), the contraction term follows from the continuous-time contraction result in Lemma 4, and the additional error terms are the differences between the drift terms in the dynamics defined by equation (13)–(15), and the true drift terms in the continuous-time dynamics. The discretization error analysis relies on bounding the three terms respectively.

Comparing the constructions in equations (13)–(15), we note that:

$$\begin{aligned}\widehat{r}_t-\widetilde{r}_t =&\gamma\int_{k\eta}^t\int_{k\eta}^s(-\widehat{g}_\ell/L-\gamma\widehat{r}_\ell)\mathrm{d}\ell\mathrm{d}s-\xi\int_{k\eta}^t(\widehat{r}_s-\widetilde{r}_s)\mathrm{d}s,\\ \widehat{p}_t-\widetilde{p}_t =&\frac{1}{L}\int_{k\eta}^t\frac{1}{\eta}\int_{k\eta}^{(k+1)\eta}(\widehat{g}_s-\widehat{g}_\ell)d\ell ds.\end{aligned}$$

The quantity $(-\widehat{g}_\ell/L-\gamma\widehat{r}_\ell)$ scales at the order of $O(\sqrt{d})$, which is integrated twice and leads to the bound $\|\widehat{r}_s-\widetilde{r}_s\|_2=O(\eta^2\sqrt{d})$. This bound furthermore makes an $O(\eta^2\sqrt{d})$ contribution to the final bound on the discretization error. For the term $\widehat{p}_t-\widetilde{p}_t$, we note that $\|\widehat{g}_s-\widehat{g}_\ell\|_2=O(\eta\sqrt{d})$. Integrating it along the path leads to an $O(\eta^2\sqrt{d})$ bound on $\|\widehat{p}_t-\widetilde{p}_t\|_2$. Therefore, the latter two terms in the error parts of equation (23) can both be bounded by $O(\eta^2\sqrt{d})$. When the function U is ridge separable, taking exact integral leads to the final $O(d^{1/4}/\sqrt{\varepsilon})$ mixing time bound in Theorem 1; when the function U is given by a black-box gradient oracle, the discretization error bound relies on the magnitude of $\Delta_k(g)$, the error bounds in Chebyshev polynomials, which eventually leads to the guarantees in Theorem 2.

4.2.2 Proof of Theorem 1

We now turn to the proof of Theorem 1. Let $\widehat{g}_t(\theta_1,\theta_2)=\nabla U\left(\frac{(k+1)\eta-t}{\eta}\theta_1+\frac{t-k\eta}{\eta}\theta_2\right)$. Since the function U is in the form of equation (1), by Section 3.4, the one-step update can be explicitly solved in closed form. We thus have $\Delta_k(g)=0$ for any $k\in\mathbb{N}_+$.

By Proposition 9, for the synchronous coupling between the process $\widetilde{x}_t$ defined by the algorithm and the process (8), we have

$$\begin{aligned}\mathbb{E}\left\|\widetilde{x}_{(k+1)\eta}-x_{(k+1)\eta}\right\|_S^2\leq&\left(1-\frac{\eta}{20\kappa^2+200}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta}-x_{k\eta}\right\|_S^2+C\kappa^8\eta^5 d/L\\&+C\kappa^5\eta\mathbb{E}\Delta_k(g)^2/L^2+C\kappa^6\frac{\eta^5}{L^2}\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\frac{\mathrm{d}}{\mathrm{d}s}\widehat{g}_s\left(\widetilde{\theta}_{k\eta},\widehat{\theta}_{(k+1)\eta}\right)\right\|_2^2\\ \leq&\left(1-\frac{\eta}{20\kappa^2+200}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta}-x_{k\eta}\right\|_S^2+C\kappa^8\eta^5 d/L\\&+C\kappa^6\frac{\eta^5}{L^2}\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\frac{\mathrm{d}}{\mathrm{d}s}\nabla U\left(\widetilde{\theta}_{k\eta}+(s-k\eta)\widetilde{p}_{k\eta}\right)\right\|_2^2.\end{aligned}$$

Note that by the L-smoothness condition (Assumption 1), we have

$$\begin{aligned}\left\|\frac{\mathrm{d}}{\mathrm{d}s}\nabla U\left(\widetilde{\theta}_{k\eta}+(s-k\eta)\widetilde{p}_{k\eta}\right)\right\|_2&=\left\|\nabla^2 U\left(\widetilde{\theta}_{k\eta}+(s-k\eta)\widetilde{p}_{k\eta}\right)\widetilde{p}_{k\eta}\right\|_2\\&\leq\left\|\nabla^2 U\left(\widetilde{\theta}_{k\eta}+(s-k\eta)\widetilde{p}_{k\eta}\right)\right\|_{op}\cdot\left\|\widetilde{p}_{k\eta}\right\|_2\leq L\left\|\widetilde{p}_{k\eta}\right\|_2.\end{aligned}$$

The moments can further be controlled as:

$$\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2\leq 2\mathbb{E}\left\|p_{k\eta}\right\|_2^2+2\mathbb{E}\left\|p_{k\eta}-\widetilde{p}_{k\eta}\right\|_2^2\leq\frac{2d}{L}+2\|S^{-1}\|_{op}\mathbb{E}\left\|x_{k\eta}-\widetilde{x}_{k\eta}\right\|_S^2.$$

Therefore, with $\eta < c\kappa^{-11/4}$, we have $2C\kappa^6\eta^5 \|S^{-1}\|_{op} \leq \frac{\eta}{40\kappa^2+400}$, and consequently:

$$\begin{aligned}\mathbb{E}\left\|\widetilde{x}_{(k+1)\eta} - x_{(k+1)\eta}\right\|_S^2 &\leq \left(1 - \frac{\eta}{20\kappa^2+200}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta} - x_{k\eta}\right\|_S^2 + C\kappa^8\eta^5 d/L. \\ &\quad + C\kappa^6\eta^5\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 \\ &\leq \left(1 - \frac{\eta}{40\kappa^2+400}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta} - x_{k\eta}\right\|_S^2 + 2C\kappa^8\eta^5 d/L.\end{aligned}$$

Solving the recursion, we obtain:

$$\mathbb{E}\left\|\widetilde{x}_{(k+1)\eta} - x_{(k+1)\eta}\right\|_S^2 \leq \left(1 - \frac{\eta}{20\kappa^2+200}\right)^k \mathbb{E}\left\|\widetilde{x}_0 - x_0\right\|_S^2 + C'\kappa^{10}\eta^4 d/L. \tag{24}$$

Starting the algorithm from $\widetilde{x}_0 := (\theta_0, 0, 0)$, we have $\mathbb{E}\left\|x_0 - \widetilde{x}_0\right\|_2^2 \leq \frac{8d}{L}$.

For a given $\varepsilon > 0$, to make the desired bound hold, we use the synchronous coupling we constructed, and needing $\mathbb{E}\left\|\widetilde{x}_{N\eta} - x_{N\eta}\right\|_2^2 \leq \varepsilon^2$, we let both terms in equation (24) scale as at most $\frac{1}{2}\|S^{-1}\|_{op}^{-1}\varepsilon^2$, which leads to:

$$\begin{cases}\eta < c\kappa^{-\frac{11}{4}} d^{-\frac{1}{4}} \varepsilon^{\frac{1}{2}}, \\ N \geq C \cdot \frac{\kappa^2}{\eta} \log \frac{3d}{L\varepsilon}.\end{cases}$$

Choosing the parameters accordingly completes the proof.

4.2.3 Proof of Theorem 2

Now we describe the proof of Theorem 2, a general result for functions with high-order smoothness. The proof is also based on synchronous coupling. In addition to exploiting Proposition 9, we also use the following standard result for Chebyshev node interpolation (Stewart, 1998):

Lemma 10 *For a curve $(x_t \mid t \in [0, \ell])$ in $\mathbb{R}^d$ with up to α-th order derivatives, let $(\Phi(t; x) \mid t \in [0, \ell])$ be the $(\alpha - 1)$-order Lagrange polynomial defined at the α-th order Chebyshev nodes. Then the interpolation error is bounded as*

$$\sup_{t\in[0,\ell]} \left\|x_t - \Phi(t; x)\right\|_2 \leq \frac{1}{2^{\alpha-1}\alpha!}\ell^\alpha \sup_{t\in[0,\ell]} \left\|\frac{\mathrm{d}^\alpha}{\mathrm{d}t^\alpha} x_t\right\|_2. \tag{25}$$

Now suppose that we use $(\alpha - 1)$-th order Chebyshev nodes in order to construct a polynomial $\widehat{g}_t$ that approximates $\nabla U(\widetilde{\theta}_{k\eta} + (t - k\eta)\widetilde{p}_{k\eta})$. By Lemma 10, the approximation error can be controlled by:

$$\begin{aligned}\Delta_k(g) &= \sup_{t\in[k\eta,(k+1)\eta]} \left\|\widehat{g}_t - \nabla U(\widetilde{\theta}_{k\eta} + (t - k\eta)\widetilde{p}_{k\eta})\right\|_2 \\ &\leq \eta^{\alpha-1} \sup_{t\in[k\eta,(k+1)\eta]} \left\|\frac{\mathrm{d}^{\alpha-1}}{\mathrm{d}s^{\alpha-1}} \nabla U(\widetilde{\theta}_{k\eta} + (s - k\eta)\widetilde{p}_{k\eta}\right\|_2 \\ &\leq \eta^{\alpha-1} \sup_{t\in[k\eta,(k+1)\eta]} \left\|\nabla^\alpha U(\widetilde{\theta}_{k\eta} + (t - k\eta)\widetilde{p}_{k\eta}\right\|_{\mathrm{tsr}}^{(\alpha)} \cdot \left\|\widetilde{p}_{k\eta}\right\|_2^{\alpha-1} \\ &\leq \eta^{\alpha-1} L_\alpha^{\alpha-1} \cdot \left\|\widetilde{p}_{k\eta}\right\|_2^{\alpha-1}.\end{aligned}$$

Thus, by Lemma 13, we have:

$$\mathbb{E}\Delta_k(g)^2 \leq \left(C\eta L_\alpha \kappa^3 \sqrt{\frac{d+2\alpha}{L}}\right)^{2(\alpha-1)}.$$

For Lagrange interpolating polynomials, the time derivative is the finite difference between interpolation points. These differences can be further bounded by the time derivative of the original process $\nabla U(\theta + tp)$:

$$\sup_{s\in[0,\eta]} \left\|\frac{\mathrm{d}}{\mathrm{d}s}\widehat{g}_{k\eta+s}\right\|_2 \leq L \left\|\widetilde{p}_{k\eta}\right\|_2.$$

Similar to the proof of Theorem 1, since the weights in Lagrangian interpolation at Chebyshev nodes are non-negative, using Proposition 9, we obtain the bound:

$$\begin{aligned}\left\|\widetilde{x}_{(k+1)\eta} - x_{(k+1)\eta}\right\|_2^2 \leq \left(1 - \frac{\eta}{40\kappa^2 + 400}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta} - x_{k\eta}\right\|_S^2 \\ + C\kappa^8\eta^5 d/L + \eta\left(C\eta L_\alpha \kappa^3 \sqrt{\frac{d+2\alpha}{L}}\right)^{2(\alpha-1)}.\end{aligned}$$

In order to ensure that $\mathbb{E}\left\|\widetilde{x}_{N\eta} - x_{N\eta}\right\|_2^2 \leq \varepsilon^2$, we need each of the three error terms to be bounded by $\frac{\varepsilon}{3}$. Solving the resulting equations leads to

$$\begin{cases}\eta < c\cdot\min\left(\kappa^{-\frac{11}{4}}d^{-\frac{1}{4}}L^{-\frac{1}{4}}\varepsilon^{-\frac{1}{2}}, L_\alpha^{-1}\kappa^{-3-\frac{1}{\alpha-1}}L^{\frac{1}{2}}d^{-\frac{1}{2}}\varepsilon^{-\frac{1}{\alpha-1}}, \kappa^{-5}\right) \\ N \geq C\cdot\frac{\kappa^2}{\eta}\log\frac{3d}{L\varepsilon},\end{cases}$$

which completes the proof.

5. Discussion

In this paper, we focused on accelerating the convergence of gradient-based MCMC algorithms for high-dimensional problems. Our analysis was based on breaking the problem into two steps: (1) design of a splitting scheme that reduces the problem of SDE discretization to that of integration along a fixed straight line; and (2) a third-order Langevin dynamics that allows for a fine-grained discretization analysis while satisfying exponentially fast convergence.

For the second step, we constructed a third-order Langevin dynamics that has smoother trajectories, and for which the integration of ∇U can be separated from the Brownian motion part. We then utilized this dynamics to design MCMC algorithms adaptive to underlying structure of the problem. We proved a mixing time of order $\mathcal{O}\left(d^{1/4}/\varepsilon^{1/2}\right)$ for ridge-separable potentials, which cover a large class of machine learning models. Under a more general black-box oracle model, we established a rate of order $O(d^{1/4}/\varepsilon^{1/2} + d^{1/2}/\varepsilon^{1/(\alpha-1)})$ for densities defined by a potential function that satisfies an α-th order smoothness condition,

An important future direction is to further investigate higher-order splitting schemes with the use of higher-order dynamics so as to further reduce the dimension dependence of the mixing time. We conjecture that the exponent on d can be further reduced, with a trade-off between the dependency on the dimension and the condition number.

Acknowledgements

This work was partially supported by the Office of Naval Research Grant ONR-N00014-18-1-2640 to MJW, Army Research Office grant W911NF-17-1-0304, and National Science Foundation Grants NSF-IIS-1740855, NSF-IIS-1909365, and NSF-IIS-1619362. We thank Santosh Vempala and Chris Junchi Li for helpful discussions.

References

A. Abdulle, G. Vilmart, and K. C. Zygalakis. Long time accuracy of Lie-Trotter splitting methods for Langevin dynamics. *SIAM J. Numer. Anal.*, 53(1):1–16, 2015.

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.*, 30(3):1209–1250, 06 2020. doi: 10.1214/19-AAP1528.

A. A. Brown and M. C. Bartholomew-Biggs. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. *J. Optimiz. Theory App.*, 62(2):211–224, 1989.

Y. Cao, J. Lu, and L. Wang. Complexity of randomized algorithms for underdamped langevin dynamics. *arXiv preprint arXiv:2003.09906*, 2020.

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 (ICML)*, volume 80, pages 764–773, 2018.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, page To appear, September 2020. Originally posted as arXiv:1905.12247.

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.

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, 2018a.

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, 2018b.

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

A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Process. Appl.*, 2019.

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

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

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

M. Ledoux. The geometry of Markov diffusion generators. *Ann Fac Sci Toulouse Math*, 9(6): 305–366, 2000.

Y.-T. Lee, Z. Song, and S. S. Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities. arXiv:1812.06243, 2018.

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

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

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.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for MCMC? arXiv:1902.00996, 2019a.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019b.

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

O. Mangoubi and N. K. Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 6030–6040. 2018.

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

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

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019.

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

G. W. Stewart. *Afternotes Goes to Graduate School: Lectures on Advanced Numerical Analysis*, volume 58. Siam, 1998.

J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer, New York, 3rd edition, 2002.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2510–2518. 2014.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. *Phys. Rev.*, 36: 823–841, 1930.

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

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

Appendix A. Proof of Proposition 9

In this appendix, we prove Proposition 9, as previously stated in Section 4.2.1. Recall that this result provides a bound on the discretization error with certain choice of $\widehat{g}_t$ used in equation (14). Our proof of this bound is based on direct coupling estimates.

A.1 A decomposition into three terms

Comparing the two processes along the path, we obtain:

$$\mathrm{d}\begin{bmatrix}\theta_s-\widetilde{\theta}_s\\ p_s-\widetilde{p}_s\\ r_s-\widetilde{r}_s\end{bmatrix} = (D+Q)\begin{bmatrix}\nabla U(\theta_s)-\nabla U(\widetilde{\theta}_s)\\ L(p_s-\widetilde{p}_s)\\ L(r_s-\widetilde{r}_s)\end{bmatrix}\mathrm{d}s + \frac{1}{L}\begin{bmatrix}0\\ \nabla U(\widetilde{\theta}_s)-\widehat{g}_s\left(\widetilde{\theta}_{k\eta},\widehat{\theta}_{(k+1)\eta}\right)\\ 0\end{bmatrix}\mathrm{d}s$$
$$+\begin{bmatrix}0\\ \widehat{r}_s-\widetilde{r}_s\\ 0\end{bmatrix}\mathrm{d}s+\begin{bmatrix}\widetilde{p}_s-\widehat{p}_s\\ 0\\ \gamma(\widehat{p}_s-\widetilde{p}_s)\end{bmatrix}\mathrm{d}s.$$

Introducing the function $J(u)=u^\top Su$, for the one-step analysis with $t\in[k\eta,(k+1)\eta]$, we have:

$$J(\widetilde{x}_t-x_t)=J(\widetilde{x}_{k\eta}-x_{k\eta})+\int_{k\eta}^t\begin{bmatrix}\theta_s-\widetilde{\theta}_s\\ p_s-\widetilde{p}_s\\ r_s-\widetilde{r}_s\end{bmatrix}^{\mathrm{T}} S(D+Q)\begin{bmatrix}\nabla U(\theta_s)-\nabla U(\widetilde{\theta}_s)\\ L(p_s-\widetilde{p}_s)\\ L(r_s-\widetilde{r}_s)\end{bmatrix}\mathrm{d}s$$
$$+\frac{1}{L}\int_{k\eta}^t(\widetilde{x}_s-x_s)^{\mathrm{T}}S\begin{bmatrix}0\\ \nabla U(\widetilde{\theta}_s)-\widehat{g}_s\\ 0\end{bmatrix}\mathrm{d}s+\int_{k\eta}^t(\widetilde{x}_s-x_s)^{\mathrm{T}}S\begin{bmatrix}0\\ \widehat{r}_t-\widetilde{r}_t\\ 0\end{bmatrix}\mathrm{d}s.$$

Consequently, we have

$$J(\widetilde{x}_t - x_t) \leq \|\widetilde{x}_{k\eta} - x_{k\eta}\|_S^2 - \frac{1}{5\kappa^2+50}\int_{k\eta}^t \|\widetilde{x}_s - x_s\|_S^2 \,\mathrm{d}s + \frac{3}{20\kappa^2+200}\int_{k\eta}^t \|\widetilde{x}_s - x_s\|_S^2 \,\mathrm{d}s$$
$$+ (20\kappa^2+200)\|\!|S|\!\|_{op}\int_{k\eta}^t \left(\frac{1}{L^2}\underbrace{\left\|\nabla U(\widetilde{\theta}_s) - \widehat{g}_s\right\|_2^2}_{I_1(s)} + \underbrace{\|\widehat{r}_t - \widetilde{r}_t\|_2^2}_{I_2(s)} + (1+\gamma^2)\underbrace{\|\widehat{p}_s - \widetilde{p}_s\|_2^2}_{I_3(s)}\right)\mathrm{d}s.$$

Simplifying further, we have shown that

$$J(\widetilde{x}_t - x_t) \leq \|\widetilde{x}_{k\eta} - x_{k\eta}\|_S^2 - \frac{1}{20\kappa^2+200}\int_{k\eta}^t \|\widetilde{x}_s - x_s\|_S^2 \,\mathrm{d}s$$
$$+ (20\kappa^2+200)\|\!|S|\!\|_{op}\int_{k\eta}^t (I_1(s)/L^2 + I_2(s) + (1+\gamma^2)I_3(s))\mathrm{d}s. \quad (26)$$

The remainder of the proof is devoted to bounding the terms $\{I_j\}_{j=1}^3$.

A.2 Some auxiliary results

In order to bound the terms $\{I_j\}_{j=1}^3$, we require a number of auxiliary results, stated here. Our first result bounds the error in using the line $\left(\left(1-\frac{t}{\ell}\right)x_0 + \frac{t}{\ell}x_\ell \ : \ t \in [0,\ell]\right)$ to approximate a curve $(x_t : t \in [0,\ell])$ in $\mathbb{R}^d$.

Lemma 11 *The straight-line approximation error of the curve is uniformly bounded as*

$$\sup_{t\in[0,\ell]}\left\|\left(1-\frac{t}{\ell}\right)x_0 + \frac{t}{\ell}x_\ell - x_t\right\|_2 \leq \ell^2\left\|\ddot{x}_t\right\|_2.$$

Our second auxiliary lemma relates the squared Euclidean norm of the interpolation process (c.f. equations (13)–(15)) with the squared norm at discrete time steps.

Lemma 12 *For the process $(\widetilde{p}_t, \widetilde{r}_t, \widetilde{\theta}_t)$ defined by equations (13)–(15), we have*

$$\sup_{s\in[k\eta,(k+1)\eta]}\mathbb{E}\left(\|\widetilde{r}_s\|_2^2 + \|\widetilde{p}_s\|_2^2\right) \leq C\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + d/L\right),$$

for some universal constant $C > 0$.

Our third auxiliary lemma upper bounds the higher order moments of the stochastic process generated by Algorithm 1. These bounds are useful for controlling certain higher order derivatives along the path.

Lemma 13 *Assuming that $\widehat{g}$ satisfies $\widehat{g}_t(\theta_1,\theta_2) \in \mathrm{conv}\left(\{\nabla U(\lambda\theta_1 + (1-\lambda)\theta_2) : \lambda \in [0,1]\}\right)$ for any θ_1, θ_2, consider the process $x^{(k)} = (\theta^{(k)}, p^{(k)}, r^{(k)})$ defined by Algorithm 1, for any $\alpha \in \mathbb{N}_+$ and $\eta < c\kappa^{-5}$, we have*

$$\left(\mathbb{E}\left\|x^{(k)} - (\theta^*, 0, 0)\right\|_2^{2\alpha}\right)^{\frac{1}{2\alpha}} \leq C\kappa^3\sqrt{\frac{d+2\alpha}{L}},$$

for some universal constants $c, C > 0$.

We return to prove all of these claims in Appendix A.5.

A.3 Bounding the three terms

Taking the auxiliary lemmas as given for now, we now bound each of the terms $\{I_j\}_{j=1}^3$ in succession.

A.3.1 Upper bound for I_1

Note that:

$$\begin{aligned}\mathbb{E}I_1(s) &\le 2\mathbb{E}\left\|\nabla U(\widetilde{\theta}_t) - \nabla U\left(\frac{t-k\eta}{\eta}\widehat{\theta}_{(k+1)\eta} + \frac{(k+1)\eta - t}{\eta}\widetilde{\theta}_{k\eta}\right)\right\|_2^2 + 2\mathbb{E}\Delta_k(g)^2 \\ &\le 2L^2\mathbb{E}\left\|\widetilde{\theta}_t - \frac{t-k\eta}{\eta}\widehat{\theta}_{(k+1)\eta} - \frac{(k+1)\eta - t}{\eta}\widetilde{\theta}_{k\eta}\right\|_2^2 + 2\mathbb{E}\Delta_k(g)^2.\end{aligned}$$

For the first term, note that:

$$\begin{aligned}\left\|\widetilde{\theta}_t - \left(\frac{t-k\eta}{\eta}\widehat{\theta}_\eta + \frac{(k+1)\eta - t}{\eta}\widetilde{\theta}_{k\eta}\right)\right\|_2 &= \left\|\int_{k\eta}^t \widetilde{p}_s \mathrm{d}s - \int_{k\eta}^t \frac{\widehat{\theta}_{(k+1)\eta} - \widetilde{\theta}_{k\eta}}{\eta}\mathrm{d}t\right\|_2 \\ &= \left\|\int_{k\eta}^t \widetilde{p}_s \mathrm{d}s - \int_{k\eta}^t \frac{1}{\eta}\int_{k\eta}^{(k+1)\eta}\widetilde{p}_{k\eta}\mathrm{d}\ell\mathrm{d}s\right\|_2 \\ &\le \int_{k\eta}^t \left\|\widetilde{p}_s - \widetilde{p}_{k\eta}\right\|_2 \mathrm{d}s.\end{aligned}$$

Moreover, by definition, we have $\widetilde{p}_s - \widetilde{p}_{k\eta} = \int_{k\eta}^s (-\widehat{g}_\ell/L - \gamma\widehat{r}_\ell)\mathrm{d}\ell$. Applying the Cauchy-Schwartz inequality, we obtain:

$$\begin{aligned}\mathbb{E}\left\|\widetilde{p}_s - \widetilde{p}_{k\eta}\right\|_2^2 &\le 2\eta\int_{k\eta}^s \left(\mathbb{E}\left\|\widehat{g}_\ell\right\|_2^2/L^2 + \gamma^2\mathbb{E}\left\|\widehat{r}_\ell\right\|_2^2\right)\mathrm{d}\ell \\ &\le \frac{2\eta^2}{L^2}\mathbb{E}\sup_{s\in[k\eta,\le(k+1)\eta]}\left\|\nabla U(\widehat{\theta}_s)\right\|_2^2 + 2\eta^2\gamma^2\sup_{\ell\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widehat{r}_\ell\right\|_2^2.\end{aligned}$$

Therefore, applying Cauchy-Schwartz to the integral, we arrive at:

$$\begin{aligned}\mathbb{E}I_1(t) &\le 2L^2(t-k\eta)\int_{k\eta}^t \mathbb{E}\left\|\widetilde{p}_s - \widetilde{p}_{k\eta}\right\|_2^2 \mathrm{d}s + 2\mathbb{E}\Delta_k(g)^2 \\ &\le 2\eta^4\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\nabla U(\widehat{\theta}_s)\right\|_2^2 + 2\eta^4\gamma^2L^2\sup_{\ell\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widehat{r}_\ell\right\|_2^2 + 2\mathbb{E}\Delta_k(g)^2.\end{aligned}$$

The trajectory of $(\widehat{r}_t : k\eta \le t \le (k+1)\eta)$ follows an Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930), which admits a closed-form expression:

$$\widehat{r}_t = e^{-\xi(t-k\eta)}\widehat{r}_{k\eta} + \left(1 - e^{-\xi(t-k\eta)}\right)\frac{\gamma}{\xi}\widetilde{p}_{k\eta} + \int_{k\eta}^t e^{-\xi(t-s)}\sqrt{2\xi/L}\mathrm{d}B_s.$$

Taking the second moment and applying Young's inequality, we obtain:

$$\mathbb{E}\left\|\widehat{r}_t\right\|_2^2 \leq 3\mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + 6\eta^2\gamma^2\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \frac{6\xi\eta d}{L}.$$

For the gradient-norm term, by Assumption 1, we can relate it to moments of $\widetilde{\theta}$ and $\widetilde{p}$:

$$\begin{aligned}\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\nabla U(\widehat{\theta}_s)\right\|_2^2 &\leq L^2\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\widehat{\theta}_s-\theta^*\right\|_2^2\\ &\leq 2L^2\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta}-\theta^*\right\|_2^2+\eta^2\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2\right).\end{aligned}$$

The bound on $\mathbb{E}\Delta_k(g)^2$ depends on specific choice of the gradient approximation g, which will be discussed in the main proof of Theorem 1 and Theorem 2.

A.3.2 Upper bound for I_2

Recalling the dynamics of $\widetilde{r}$ and $\widehat{r}$, we have

$$\widehat{r}_t-\widetilde{r}_t=-\gamma\int_{k\eta}^t(\widetilde{p}_{k\eta}-\widetilde{p}_s)\mathrm{d}s-\xi\int_{k\eta}^t(\widehat{r}_s-\widetilde{r}_s)\mathrm{d}s.$$

Solving it as an ODE for $\widehat{r}_t-\widetilde{r}_t$ within time interval $[k\eta,(k+1)\eta]$, we obtain:

$$\widehat{r}_t-\widetilde{r}_t=-\gamma\int_{k\eta}^t(\widetilde{p}_{k\eta}-\widetilde{p}_s)e^{-\xi(t-s)}\mathrm{d}s.$$

By equation (14), we have:

$$\widetilde{p}_s-\widetilde{p}_{k\eta}=-\int_{k\eta}^s\widehat{g}_\ell/Ld\ell+\gamma\int_{k\eta}^s\widehat{r}_\ell d\ell.$$

Note that we have:

$$\mathbb{E}\left\|-\widehat{g}_\ell/L-\gamma\widehat{r}_\ell\right\|_2^2\leq 2\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\nabla U(\widehat{\theta}_s)\right\|_2^2/L^2+2\gamma^2\sup_{\ell\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widehat{r}_\ell\right\|_2^2.$$

Using Cauchy-Schwartz twice, we obtain:

$$\begin{aligned}\mathbb{E}I_2(t)=\mathbb{E}\left\|\widehat{r}_t-\widetilde{r}_t\right\|_2^2&=\gamma^2\mathbb{E}\left\|\int_{k\eta}^t(\widetilde{p}_{k\eta}-\widetilde{p}_s)e^{-\xi(t-s)}\mathrm{d}s\right\|_2^2\leq\gamma^2\eta\int_{k\eta}^t\mathbb{E}\left\|\widetilde{p}_{k\eta}-\widetilde{p}_s\right\|_2^2\mathrm{d}s\\ &\leq\gamma^2\eta\int_{k\eta}^t\left\|-\int_{k\eta}^s\widehat{g}_\ell/Ld\ell+\gamma\int_{k\eta}^s\widehat{r}_\ell d\ell\right\|_2^2ds\leq\gamma^2\eta^2\int_{k\eta}^t\int_{k\eta}^s\mathbb{E}\left\|-\widehat{g}_\ell/L-\gamma\widehat{r}_\ell\right\|_2^2d\ell ds\\ &\leq 2\gamma^2\eta^4\left(\mathbb{E}\sup_{s\in[k\eta(k+1)\eta]}\left\|\nabla U(\widehat{\theta}_s)\right\|_2^2/L^2+\gamma^2\sup_{\ell\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widehat{r}_\ell\right\|_2^2\right).\end{aligned}$$

The upper bound involves the expected supremum of the squared gradient norm along the path of $\widehat{\theta}_s$, which was obtained in Section A.3.1:

$$\mathbb{E} \sup_{s\in[k\eta,(k+1)\eta]} \left\|\nabla U(\widehat{\theta}_s)\right\|_2^2 \leq 2L^2 \left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \eta^2 \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2\right).$$

Putting these results together, for $t \in [k\eta, (k+1)\eta]$ we have:

$$\mathbb{E}I_2(t) \leq 4\gamma^2\eta^4 \left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \eta^2 \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \gamma^2 \sup_{\ell\in[k\eta,(k+1)\eta]} \mathbb{E}\left\|\widehat{r}_\ell\right\|_2^2\right).$$

A.3.3 Upper Bound for I_3

Note that both the process $\widetilde{p}_s$ and the process $\widehat{p}_s$ can be written in terms of integrals as follows:

$$\begin{aligned}
\widetilde{p}_t &= \widetilde{p}_{k\eta} - \int_{k\eta}^t \frac{1}{L}\widehat{g}_s \mathrm{d}s + \int_{k\eta}^t \gamma \widehat{r}_s \mathrm{d}s, \\
\widehat{p}_t &= \widetilde{p}_{k\eta} - \frac{t-k\eta}{L\eta}\left(\int_{k\eta}^{(k+1)\eta} \widehat{g}_s \mathrm{d}s\right)\mathrm{d}s + \int_{k\eta}^t \gamma \widehat{r}_s \mathrm{d}s.
\end{aligned}$$

By Lemma 11, for the process $\iota(s) := \int_{k\eta}^{k\eta+s} \widehat{g}_\ell d\ell$, we have:

$$\begin{aligned}
\sup_{s\in[0,\eta]} \left\|\left(1-\frac{s}{\eta}\right)\iota(0) + \frac{s}{\eta}\iota(\eta) - \iota(s)\right\|_2 &\leq \eta^2 \sup_{s\in[0,\eta]} \left\|\ddot{\iota}_s\right\|_2 \\
&= \eta^2 \sup_{s\in[k\eta,(k+1)\eta]} \left\|\frac{\mathrm{d}}{\mathrm{d}s}\widehat{g}_{k\eta+s}\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)\right\|_2.
\end{aligned}$$

In this way, we obtain

$$\left\|\widetilde{p}_t - \widehat{p}_t\right\|_2 = \frac{1}{L}\left\|\left(1-\frac{s}{\eta}\right)\iota(0) + \frac{s}{\eta}\iota(\eta) - \iota(s)\right\|_2 \leq \frac{\eta^2}{L} \sup_{s\in[0,\eta]} \left\|\frac{\mathrm{d}}{\mathrm{d}s}\widehat{g}_s\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)\right\|_2,$$

as claimed.

A.4 Obtaining the final bound

If we have $\left\|\widetilde{\theta}_t - \theta^*\right\|_2, \left\|\widetilde{p}_t\right\|_2, \left\|\widetilde{r}_t\right\|_2 = O(\sqrt{d})$, the above upper bound for $I_1(t)$ is of order $O(\eta^4 d)$ and the upper bound for $I_2(t)$ is of order $O(\eta^4 d)$, making it possible to achieve a final discretization error of order $O(\eta^2 d^{1/2})$.

In order to make this intuition precise, we use Lemma 12, which shows that the supremum of second moments of $\widetilde{r}$ and $\widetilde{p}$ along the path can be related to the second moment at time $k\eta$:

$$\sup_{s\in[k\eta,(k+1)\eta]} \mathbb{E}\left(\left\|\widetilde{r}_s\right\|_2^2 + \left\|\widetilde{p}_s\right\|_2^2\right) \leq C\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + d/L\right). \quad (27)$$

We substitute the moment upper bounds (27) into the estimates for I_1, I_2 and I_3 to find that for any stepsize $\eta < \min(1/\gamma, 1/\xi)$, we have

$$\begin{aligned}
\mathbb{E}I_1(s) &\leq C\kappa^2\eta^4 L^2 \left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + d/L \right) + 2\mathbb{E}\Delta_k(g)^2, \\
\mathbb{E}I_2(s) &\leq C\kappa^4\eta^4 \left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + d/L \right), \\
\mathbb{E}I_3(s) &\leq \frac{\eta^4}{L^2}\mathbb{E} \sup_{s \in [k\eta, (k+1)\eta]} \left\| \frac{\mathrm{d}}{\mathrm{d}s} \widehat{g}_s\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)\right\|_2^2.
\end{aligned}$$

So we have:

$$\begin{aligned}
&\mathbb{E}(I_1(s)/L^2 + I_2(s)) \\
&\leq C\kappa^4\eta^4 \left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta} - \theta^*\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + d/L^2 \right) + 2\mathbb{E}\Delta_k(g)^2/L^2 \\
&\leq 2C\kappa^4\eta^4 \left(\mathbb{E}\left\|\theta_{k\eta} - \theta^*\right\|_2^2 + \mathbb{E}\left\|p_{k\eta}\right\|_2^2 + \mathbb{E}\left\|r_{k\eta}\right\|_2^2 + \mathbb{E}\left\|x_{k\eta} - \widetilde{x}_{k\eta}\right\|_2^2 + d/L \right) + \frac{2\mathbb{E}\Delta_k(g)^2}{L^2} \\
&\leq 2C\kappa^4\eta^4\mathbb{E}\left\|x_{k\eta} - \widetilde{x}_{k\eta}\right\|_2^2 + 2C\kappa^4\eta^4 d/L + 2\mathbb{E}\Delta_k(g)^2/L^2.
\end{aligned}$$

Given a stepsize $\eta < c\kappa^{-\frac{11}{4}}$ with sufficiently small $c > 0$, we are guaranteed that $2(15\kappa^2 + 150)C\|S\|_{op}\|S^{-1}\|_{op}\kappa^4\eta^4 < \frac{1}{10\kappa^2+100}$. Plugging back to the upper bound for $\|\widetilde{x}_t - x_t\|_S^2$ along the path, for $t \in [k\eta, (k+1)\eta]$, we obtain:

$$\begin{aligned}
\mathbb{E}\left\|\widetilde{x}_t - x_t\right\|_S^2 \leq{}& \mathbb{E}\left\|\widetilde{x}_{k\eta} - x_{k\eta}\right\|_S^2 - \frac{1}{20\kappa^2 + 200}\int_{k\eta}^t \left\|\widetilde{x}_s - x_s\right\|_S^2 \mathrm{d}s \\
&+ C\kappa^4(t - k\eta)\left(\kappa^4\eta^4 d/L + \mathbb{E}\Delta_k(g)^2/L^2\right) \\
&+ C\frac{\eta^5}{L^2}\kappa^6\mathbb{E}\sup_{s \in [k\eta, (k+1)\eta]} \left\| \frac{\mathrm{d}}{\mathrm{d}s} \widehat{g}_s\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)\right\|_2^2.
\end{aligned}$$

Applying Grönwall's inequality yields

$$\begin{aligned}
\mathbb{E}\left\|\widetilde{x}_{(k+1)\eta} - x_{(k+1)\eta}\right\|_S^2 \leq{}& \left(1 - \frac{\eta}{10\kappa^2 + 100}\right)\mathbb{E}\left\|\widetilde{x}_{k\eta} - x_{k\eta}\right\|_S^2 \\
&+ C\left(\frac{\kappa^8\eta^5 d}{L} + \frac{\kappa^4\eta}{L^2}\mathbb{E}\Delta_k(g)^2 + \frac{\kappa^6\eta^5}{L^2}\mathbb{E}\sup_{s \in [k\eta, (k+1)\eta]} \left\| \frac{\mathrm{d}}{\mathrm{d}s} \widehat{g}_s\left(\widetilde{\theta}_{k\eta}, \widehat{\theta}_{(k+1)\eta}\right)\right\|_2^2 \right).
\end{aligned}$$

A.5 Proof of auxiliary lemmas

We now return to prove the auxiliary results that were stated and used in the previous sections—namely, Lemmas 11, 12 and 13.

A.5.1 Proof of Lemma 11

Introducing the shorthand $\lambda := t/\ell$, a direct calculation with the mean value theorem yields:

$$\begin{aligned}(1-\lambda)x_0 + \lambda x_\ell - x_{\lambda\ell} &= (1-\lambda)\ell \int_0^\lambda \dot{x}(\tau\ell)\mathrm{d}\tau - \lambda\ell \int_0^{1-\lambda} \dot{x}((1-\tau)\ell)\mathrm{d}\tau \\ &= (1-\lambda)\ell\lambda \left(\frac{1}{\lambda}\int_0^\lambda \dot{x}(\tau\ell)\mathrm{d}\tau - \frac{1}{1-\lambda}\int_0^{1-\lambda} \dot{x}((1-\tau)\ell)\mathrm{d}\tau \right) \\ &= \lambda(1-\lambda)\ell \int_0^1 \left(\dot{x}(\tau\ell\lambda) - \dot{x}((1-(1-\lambda)\tau)\ell)\right)\mathrm{d}\tau.\end{aligned}$$

Taking the Euclidean norm yields

$$\left\|(1-\lambda)x_0 + \lambda x_\ell - x_{\lambda\ell}\right\|_2 \leq \ell^2\lambda(1-\lambda)\sup_{t\in[0,\ell]} \left\|\ddot{x}_t\right\|_2.$$

A.5.2 Proof of Lemma 12

For $\widetilde{p}_t$, note that:

$$\mathbb{E}\left\|\widetilde{p}_t\right\|_2^2 \leq 3\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + 3\eta\int_{k\eta}^t \mathbb{E}\left\|\widehat{g}_s\right\|_2^2/L^2\mathrm{d}s + 3\gamma^2\eta\int_{k\eta}^t \mathbb{E}\left\|\widehat{r}_s\right\|_2^2\mathrm{d}s.$$

The two terms appearing in the above upper bound are both easy to control:

$$\mathbb{E}\left\|\widehat{g}_s\right\|_2^2 \leq \mathbb{E}\sup_{s\in[k\eta,(k+1)\eta]}\left\|\nabla U(\widehat{\theta}_s)\right\|_2^2 \leq 2L^2\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta}-\theta^*\right\|_2^2 + \eta^2\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2\right).$$

$$\begin{aligned}\mathbb{E}\left\|\widehat{r}_s\right\|_2^2 &\leq 3\mathbb{E}\left\|\widehat{r}_{k\eta}\right\|_2^2 + 3\gamma^2\eta\int_{k\eta}^t\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 dt + 3\mathbb{E}\left\|\int_{k\eta}^t\sqrt{2\xi/L}\mathrm{d}B_t^r\right\|_2^2 \\ &\leq 3\mathbb{E}\left\|\widehat{r}_{k\eta}\right\|_2^2 + 3\gamma^2\eta^2\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + 3\eta\xi d/L.\end{aligned}$$

Putting them together, with $\eta < \min(1/\gamma, 1/\xi)$, we have:

$$\begin{aligned}\sup_{t\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widetilde{p}_t\right\|_2^2 &\leq 3\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + 6\eta^2\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta}-\theta^*\right\|_2^2 + \eta^2\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2\right) \\ &\quad + 3\gamma^2\eta^2\left(\eta^2\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + 4\eta\xi d/L\right) \\ &\leq 12\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta}-\theta^*\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2 + d/L\right).\end{aligned}$$

For $\widetilde{r}_t$, the argument is a bit more involved, since we need to relate back to the integral of its own moments along the path. But since the time interval is short, we can control it easily.

$$\begin{aligned}\mathbb{E}\left\|\widetilde{r}_s - \widetilde{r}_{k\eta}\right\|_2^2 &= \int_{k\eta}^s\left(-\mathbb{E}\langle\widetilde{r}_\ell-\widetilde{r}_{k\eta},\,\widetilde{p}_\ell\rangle - \mathbb{E}\langle\widetilde{r}_\ell-\widetilde{r}_{k\eta},\,\widetilde{r}_\ell\rangle + 2\xi d/L\right)\mathrm{d}\ell \\ &\leq \frac{1}{2}\int_{k\eta}^s\left(\mathbb{E}\left\|\widetilde{r}_\ell-\widetilde{r}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{p}_\ell\right\|_2^2\right)\mathrm{d}\ell + \frac{1}{2}\int_{k\eta}^s\left(3\mathbb{E}\left\|\widetilde{r}_\ell-\widetilde{r}_{k\eta}\right\|_2^2 + \mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2\right)\mathrm{d}\ell \\ &\quad + 2(s-k\eta)\xi d/L.\end{aligned}$$

Applying Grönwall's inequality yields

$$\begin{aligned}\mathbb{E}\left\|\widetilde{r}_s-\widetilde{r}_{k\eta}\right\|_2^2 &\leq (e^{2(s-k\eta)}-1)\left(\sup_{\ell\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widetilde{p}_\ell\right\|_2^2+\mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2+2\xi d/L\right)\\ &\leq 14\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta}-\theta^*\right\|_2^2+\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2+\mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2+d/L\right).\end{aligned}$$

In the second inequality, we plug in the bound for $\mathbb{E}\left\|\widetilde{p}_\ell\right\|_2^2$. Consequently,

$$\sup_{s\in[k\eta,(k+1)\eta]}\mathbb{E}\left\|\widetilde{r}_s\right\|_2^2\leq 30\left(\mathbb{E}\left\|\widetilde{\theta}_{k\eta}-\theta^*\right\|_2^2+\mathbb{E}\left\|\widetilde{p}_{k\eta}\right\|_2^2+\mathbb{E}\left\|\widetilde{r}_{k\eta}\right\|_2^2+d/L\right),$$

which completes the proof.

A.5.3 Proof of Lemma 13

For notational convenience, we assume $\theta^*=0$ in the proof of this lemma. This assumption can be made without loss of generality. Conditional on $x^{(k)}$, we have, by the definition in Algorithm 1 (we omit the superscripts and denote $(\theta^{(k)},p^{(k)},r^{(k)})$ by (θ,p,r).):

$$\mu(x^{(k)})-x^{(k)}=\begin{pmatrix}-\frac{\eta}{2L}\int_0^\eta\widehat{g}_{t+k\eta}\left(\theta,\theta+\eta p\right)dt+\mu_{12}p+\mu_{13}r\\ -\frac{1}{L}\int_0^\eta\widehat{g}_{t+k\eta}\left(\theta,\theta+\eta p\right)dt+(\mu_{22}-1)p+\mu_{23}r\\ \frac{\mu_{31}}{L}\int_0^\eta\widehat{g}_{t+k\eta}\left(\theta,\theta+\eta p\right)dt+\mu_{32}p+(\mu_{33}-1)r\end{pmatrix}.$$

Since $\widehat{g}$ belongs to the convex hull of the curve $\nabla U(\theta+tp)$, as assumed in the statement of the lemma, we have:

$$\left\|\frac{1}{\eta}\int_0^\eta\widehat{g}_{t+k\eta}\left(\theta,\theta+\eta p\right)dt-\nabla U(\theta)\right\|_2\leq\sup_{t\in[0,\eta]}\left\|\nabla U(\theta+tp)-\nabla U(\theta)\right\|_2\leq L\eta\left\|p\right\|_2,$$

and using the smoothness of U, we can easily see that

$$\left\|\nabla U(\theta)\right\|_2\leq L\left\|\theta-\theta^*\right\|_2,\forall\theta\in\mathbb{R}^d.$$

Collecting the main terms and bound the rest of terms directly using the norm of $x^{(k)}$, we obtain:

$$\begin{aligned}\mu(x^{(k)})-x^{(k)}&=\eta\begin{pmatrix}p^{(k)}\\ -\frac{1}{L}\nabla U(\theta^{(k)})+\gamma r^{(k)}\\ -\gamma p^{(k)}-\xi r^{(k)}\end{pmatrix}+\zeta_k\\ &=\eta\begin{pmatrix}0&I&0\\ -\frac{\nabla^2U(x^{(k)})}{L}I&0&\gamma I\\ 0&-\gamma I&-\xi I\end{pmatrix}x^{(k)}+\zeta_k=\eta M_kx^{(k)}+\zeta_k.\end{aligned}$$

By Appendix C, it is easy to see that $\|\zeta_k\|_S \leq C\eta^2\kappa^2 \left\|x^{(k)}\right\|_2 \leq C'\eta^2\kappa^3 \left\|x^{(k)}\right\|_S$.

Therefore, we have:

$$\begin{aligned}
\left\|\mu(x^{(k)})\right\|_S^2 &= \left((I+\eta M_k)x^{(k)} + \zeta_k\right)^\top S \left((I+\eta M_k)x^{(k)} + \zeta_k\right) \\
&\leq \left\|x^{(k)}\right\|_S^2 + 2\eta(x^{(k)})^\top M_k^\top S x^{(k)} + \eta^2 \|\!|M_k^\top S M_k\|\!|_{\mathrm{op}} \cdot \|\!|S^{-1}\|\!|_{\mathrm{op}} \cdot \left\|x^{(k)}\right\|_S^2 \\
&\quad + 2\|\zeta_k\|_S (1+\eta\|\!|M_k\|\!|_{op}) \left\|x^{(k)}\right\|_S + \|\zeta_k\|_S^2 \\
&\leq \left(1 - \frac{\eta}{5\kappa^2+50}\right)\left\|x^{(k)}\right\|_S^2 + C(\eta^2\kappa^3 + \eta^3\kappa^4 + \eta^4\kappa^6)\left\|x^{(k)}\right\|_S^2.
\end{aligned}$$

For $\eta < c\kappa^{-5}$ with some universal constant $c > 0$, we have:

$$\left\|\mu(x^{(k)})\right\|_S^2 \leq \left(1 - \frac{\eta}{10\kappa^2+100}\right)\left\|x^{(k)}\right\|_S^2.$$

Now we turn to deal with the stochastic part. By our construction, it is easy to verify that $\|\!|\Sigma\|\!|_{op} \leq C\kappa^2\eta$ for some universal constant $C > 0$.

Letting $w_k \sim \mathcal{N}(0, \frac{1}{L}I)$ be independent from $x^{(k)}$, we have:

$$\begin{aligned}
\mathbb{E}\left\|x^{(k+1)}\right\|_S^{2\alpha} &= \mathbb{E}\left\|S^{\frac{1}{2}}\mu(x^{(k)}) + (S\Sigma)^{\frac{1}{2}}w_k\right\|_2^{2\alpha} \\
&\leq \sum_{j=0}^{2\alpha}\binom{2\alpha}{j}\mathbb{E}\left\|S^{\frac{1}{2}}\mu(x^{(k)})\right\|_2^j \mathbb{E}\left\|(S\Sigma)^{\frac{1}{2}}w_k\right\|_2^{2\alpha-j} \\
&\leq \sum_{j=0}^{2\alpha}\binom{2\alpha}{j}\left(\mathbb{E}\left\|S^{\frac{1}{2}}\mu(x^{(k)})\right\|_2^{2\alpha}\right)^{\frac{j}{2\alpha}}\left(\mathbb{E}\left\|(S\Sigma)^{\frac{1}{2}}w_k\right\|_2^{2\alpha}\right)^{1-\frac{j}{2\alpha}} \\
&= \left(\left(\mathbb{E}\left\|\mu(x^{(k)})\right\|_S^{2\alpha}\right)^{\frac{1}{2\alpha}} + \left(\mathbb{E}\left\|(S\Sigma)^{\frac{1}{2}}w_k\right\|_2^{2\alpha}\right)^{\frac{1}{2\alpha}}\right)^{2\alpha}.
\end{aligned}$$

So we obtain:

$$\begin{aligned}
\left(\mathbb{E}\left\|x^{(k+1)}\right\|_S^{2\alpha}\right)^{\frac{1}{2\alpha}} &\leq \left(\mathbb{E}\left\|\mu(x^{(k)})\right\|_S^{2\alpha}\right)^{\frac{1}{2\alpha}} + \left(\mathbb{E}\left\|(S\Sigma)^{\frac{1}{2}}w_k\right\|_2^{2\alpha}\right)^{\frac{1}{2\alpha}} \\
&\leq \left(1 - \frac{\eta}{10\kappa^2+100}\right)\left(\mathbb{E}\left\|x^{(k)}\right\|_S^{2\alpha}\right)^{\frac{1}{2\alpha}} + C\kappa\sqrt{\frac{\eta}{L}(d+2\alpha)}.
\end{aligned}$$

Noting that $\left\|x^{(0)}\right\|_2 \leq \frac{1}{L}$, by solving the recursion inequalities, we obtain:

$$\left(\mathbb{E}\left\|x^{(k)}\right\|_2^{2\alpha}\right)^{\frac{1}{2\alpha}} \leq C\kappa^3\sqrt{\frac{d+2\alpha}{L}}.$$

Appendix B. Auxiliary results for Lemmas 5 and 6

This appendix is dedicated to proofs of two auxiliary results—namely, Lemmas 7 and 8—that underlie the proofs of Lemmas 5 and 6.

B.1 Definitions of the functions g_1, g_2 and g_3

The functions g_1, g_2 and g_3 are given by:

$$g_1(\kappa) = 4 + \tfrac{20}{\kappa^2} + \tfrac{56}{\kappa^3} + \tfrac{40}{\kappa^4} + \tfrac{8}{\kappa^5} + \tfrac{20}{\kappa^6} + \tfrac{12}{\kappa^7} + \tfrac{1}{\kappa^8} + \tfrac{2}{\kappa^9} + \tfrac{1}{\kappa^{10}} \tag{28a}$$

$$g_2(\kappa) = \tfrac{1}{\kappa^3} + \tfrac{5}{\kappa^5} + \tfrac{11}{\kappa^6} + \tfrac{5}{\kappa^7} + \tfrac{1}{\kappa^8} + \tfrac{2}{\kappa^9} + \tfrac{1}{\kappa^{10}}, \quad \text{and} \tag{28b}$$

$$g_3(\kappa) = 8 + \tfrac{24}{\kappa^2} + \tfrac{60}{\kappa^3} + \tfrac{41}{\kappa^4} + \tfrac{10}{\kappa^5} + \tfrac{21}{\kappa^6} + \tfrac{12}{\kappa^7} + \tfrac{1}{\kappa^8} + \tfrac{2}{\kappa^9} + \tfrac{1}{\kappa^{10}}. \tag{28c}$$

B.2 Proof of Lemma 7

By inspecting the definition of g_1 in equation (28a), we see that $\forall \kappa \geq 1$, $g_1(\kappa)$ always lies in the interval $[4, +\infty)$, which implies the inequality $2 - \sqrt{g_1(\kappa)} \leq 0$.

As for the second inequality, we first group the $\sqrt{g_1(\kappa)}$ term and rewrite the inequality (20b) into the following equivalent form:

$$2\kappa + \frac{6}{5} \geq (\kappa - 1)\sqrt{g_1(\kappa)}.$$

Since both the left and right sides are non-negative, we can square both sides, thereby obtaining

$$\left(2\kappa + \tfrac{6}{5}\right)^2 - (\kappa - 1)^2 g_1(\kappa) \geq 0.$$

Expanding the left hand side, we obtain the equivalent form of equation (20b):

$$\tfrac{64}{5}\kappa - \tfrac{564}{25} - \tfrac{16}{\kappa} + \tfrac{52}{\kappa^2} + \tfrac{16}{\kappa^3} - \tfrac{44}{\kappa^4} + \tfrac{20}{\kappa^5} + \tfrac{3}{\kappa^6} - \tfrac{12}{\kappa^7} + \tfrac{2}{\kappa^8} - \tfrac{1}{\kappa^{10}} \geq 0.$$

Since $\kappa \geq 1$, we can divide by κ on both sides and obtain an equivalent inequality with a polynomial function of $1/\kappa$:

$$g_6(1/\kappa) = \tfrac{64}{5} - \tfrac{564}{25\kappa} - \tfrac{16}{\kappa^2} + \tfrac{52}{\kappa^3} + \tfrac{16}{\kappa^4} - \tfrac{44}{\kappa^5} + \tfrac{20}{\kappa^6} + \tfrac{3}{\kappa^7} - \tfrac{12}{\kappa^8} + \tfrac{2}{\kappa^9} - \tfrac{1}{\kappa^{11}} \geq 0.$$

We can rewrite $g_6(1/\kappa)$ as a polynomial of $\left(\frac{1}{\kappa} - \frac{1}{2}\right)$:

$$\begin{aligned} g_6(1/\kappa) = \tfrac{201599}{51200} &- \tfrac{49211}{25600}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right) + \tfrac{24025}{512}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^2 + \tfrac{2955}{256}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^3 \\ &- \tfrac{3397}{64}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^4 - \tfrac{1399}{32}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^5 - \tfrac{751}{16}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^6 - \tfrac{381}{8}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^7 \\ &- \tfrac{189}{8}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^8 - \tfrac{47}{4}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^9 - \tfrac{11}{2}\left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^{10} - \left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^{11}. \end{aligned}$$

For all κ such that $1/\kappa \in (0, 1]$, we have

$$\begin{aligned} g_6(1/\kappa) \geq \tfrac{201599}{51200} - \tfrac{49211}{25600}\tfrac{1}{2} + \left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^2 \Big(&\tfrac{24025}{512} - \tfrac{2955}{256}\tfrac{1}{2} - \tfrac{3397}{64}\left(\tfrac{1}{2}\right)^2 - \tfrac{1399}{32}\left(\tfrac{1}{2}\right)^3 - \tfrac{751}{16}\left(\tfrac{1}{2}\right)^4 \\ &- \tfrac{381}{8}\left(\tfrac{1}{2}\right)^5 - \tfrac{189}{8}\left(\tfrac{1}{2}\right)^6 - \tfrac{47}{4}\left(\tfrac{1}{2}\right)^7 - \frac{11}{2}\left(\tfrac{1}{2}\right)^8 - \left(\tfrac{1}{2}\right)^9 \Big) \\ &= \tfrac{38097}{12800} + \left(\tfrac{1}{\kappa} - \tfrac{1}{2}\right)^2 \left(\tfrac{5523}{128}\right), \end{aligned}$$

which is strictly positive. This completes the proof of Lemma 7.

B.3 Proof of Lemma 8

In order to prove the bounds in this lemma, we first examine the monotonicity properties of the cubic function

$$f(x) := x^3 - g_2(\kappa) \cdot x^2 + \frac{1}{\kappa^9} g_3(\kappa) \cdot x - \frac{1}{\kappa^{15}} g_3(\kappa). \tag{29}$$

Lemma 14 *For any $\kappa \geq 1$, the function f from equation (29) has the following properties:*

(a) It is monotonically increasing over the interval $(\frac{4}{\kappa^3} + \frac{40}{\kappa^5}, \infty)$.

(b) It is monotonically increasing over the interval $(-\infty, \frac{4}{5\kappa^6})$.

Therefore, $\max_{x \leq \frac{4}{5\kappa^6}} f(x) = f\left(\frac{4}{5\kappa^6}\right)$ and $\min_{x \geq \frac{4}{\kappa^3} + \frac{40}{\kappa^5}} f(x) = f\left(\frac{4}{\kappa^3} + \frac{40}{\kappa^5}\right)$. Then we simply need to prove that $f\left(\frac{4}{5\kappa^6}\right) < 0$ and $f\left(\frac{4}{\kappa^3} + \frac{40}{\kappa^5}\right) > 0$ to obtain the result.

Now observe that

$$\begin{aligned} f\left(\tfrac{4}{\kappa^3} + \tfrac{40}{\kappa^5}\right) = \tfrac{48}{\kappa^9} + \tfrac{1520}{\kappa^{11}} - \tfrac{144}{\kappa^{12}} + \tfrac{15920}{\kappa^{13}} - \tfrac{3120}{\kappa^{14}} + \tfrac{54600}{\kappa^{15}} \\ - \tfrac{16812}{\kappa^{16}} - \tfrac{6224}{\kappa^{17}} - \tfrac{256}{\kappa^{18}} - \tfrac{2793}{\kappa^{19}} - \tfrac{766}{\kappa^{20}} + \tfrac{467}{\kappa^{21}} + \tfrac{32}{\kappa^{22}} + \tfrac{79}{\kappa^{23}} + \tfrac{38}{\kappa^{24}} - \tfrac{1}{\kappa^{25}}, \end{aligned}$$

from which we can see that

$$f\left(\tfrac{4}{\kappa^3} + \tfrac{40}{\kappa^5}\right) \geq \tfrac{48}{\kappa^9} + \tfrac{1376}{\kappa^{11}} + \tfrac{12800}{\kappa^{13}} + \tfrac{27749}{\kappa^{15}} + \tfrac{467}{\kappa^{21}} > 0,$$

using the fact that $1/\kappa \in (0, 1]$. Moreover, we also have

$$f\left(\tfrac{4}{5\kappa^6}\right) = -\tfrac{56}{25\kappa^{15}} - \tfrac{8}{\kappa^{17}} - \tfrac{2316}{125\kappa^{18}} - \tfrac{57}{5\kappa^{19}} - \tfrac{66}{25\kappa^{20}} - \tfrac{137}{25\kappa^{21}} - \tfrac{76}{25\kappa^{22}} - \tfrac{1}{5\kappa^{23}} - \tfrac{2}{5\kappa^{24}} - \tfrac{1}{5\kappa^{25}},$$

which is negative. Therefore, we conclude that for any $\kappa > 1$, the cubic function f satisfies the inequalities

$$f(x) < 0 \quad \text{if } x \leq \tfrac{4}{5\kappa^6}, \text{ and} \qquad f(x) > 0 \quad \text{if } x \geq \tfrac{4}{\kappa^3} + \tfrac{40}{\kappa^5}.$$

B.3.1 Proof of Lemma 14

We divide our proof into separate parts, corresponding to claims (a) and (b) in the lemma statement. For both parts, we establish monotonicity of the function f from equation (29) by studying its derivative

$$f'(x) = 3x^2 - 2g_2(\kappa) \cdot x + \frac{1}{\kappa^9} g_3(\kappa). \tag{30}$$

Proof of part (a): By inspection, for large enough x, the quadratic function f' is positive, and hence the function f is monotonically increasing in this range. Concretely, we claim that f' remains positive for all $x > \frac{4}{\kappa^3} + \frac{40}{\kappa^5}$.

Our strategy is to compute the solutions $x^*_{\pm}$ to the quadratic equation $f'(x) = 0$, and prove that the larger one x^*_+ satisfies the lower bound

$$x^*_+ \leq \frac{4}{\kappa^3} + \frac{40}{\kappa^5} \quad \text{for any } \kappa \geq 1. \tag{31}$$

In detail, the two solutions to the quadratic equation $f'(x) = 0$ are given by the pair $x^*_{\pm} = \frac{1}{3}\left(g_4(\kappa) \pm \sqrt{g_5(\kappa)}\right)$, where we define

$$g_4(\kappa) = \tfrac{1}{\kappa^3} + \tfrac{5}{\kappa^5} + \tfrac{11}{\kappa^6} + \tfrac{5}{\kappa^7} + \tfrac{1}{\kappa^8} + \tfrac{2}{\kappa^9} + \tfrac{1}{\kappa^{10}} \tag{32a}$$

and

$$g_5(\kappa) = \tfrac{1}{\kappa^6} + \tfrac{10}{\kappa^8} - \tfrac{2}{\kappa^9} + \tfrac{35}{\kappa^{10}} + \tfrac{40}{\kappa^{11}} - \tfrac{5}{\kappa^{12}} - \tfrac{1}{\kappa^{13}} + \tfrac{37}{\kappa^{14}} + \tfrac{1}{\kappa^{15}} + \tfrac{7}{\kappa^{16}} + \tfrac{11}{\kappa^{17}} + \tfrac{1}{\kappa^{19}} + \tfrac{1}{\kappa^{20}}. \tag{32b}$$

From the fact that $\frac{1}{\kappa} \in (0, 1]$, it follows that

$$g_4(\kappa) \leq \frac{1}{\kappa^3} + \frac{25}{\kappa^5}, \quad \text{and} \quad g_5(\kappa) \leq \frac{11}{\kappa^6} + \frac{125}{\kappa^{10}}. \tag{33a}$$

Hence

$$\sqrt{g_5(\kappa)} \leq \sqrt{\frac{11}{\kappa^6}} + \sqrt{\frac{125}{\kappa^{10}}} \leq \frac{4}{\kappa^3} + \frac{12}{\kappa^5}. \tag{33b}$$

Combining equations (33a) and (33b) yields the bound (31), and hence completes the proof of part (a).

Proof of part (b): Recall the solutions to the quadratic equation $f'(x) = 0$ that we computed in the previous section. For this part, it suffices to show that the smaller solution $x^*_- = \frac{1}{3}\left(g_4(\kappa) - \sqrt{g_5(\kappa)}\right)$ satisfies the bound

$$x^*_- \geq \frac{4}{5\kappa^6} \quad \text{for any } \kappa \geq 1. \tag{34}$$

We begin by noting that

$$\begin{aligned}
\frac{1}{3}\left(g_4(\kappa) - \sqrt{g_5(\kappa)}\right) &= \frac{1}{3}\frac{g_4(\kappa)^2 - g_5(\kappa)}{g_4(\kappa) + \sqrt{g_5(\kappa)}} \\
&= \frac{1}{\kappa^9}\frac{\left(8 + \frac{24}{\kappa^2} + \frac{60}{\kappa^3} + \frac{41}{\kappa^4} + \frac{10}{\kappa^5} + \frac{21}{\kappa^6} + \frac{12}{\kappa^7} + \frac{1}{\kappa^8} + \frac{2}{\kappa^9} + \frac{1}{\kappa^{10}}\right)}{g_4(\kappa) + \sqrt{g_5(\kappa)}}.
\end{aligned} \tag{35}$$

Equation (33b) states that $\sqrt{g_5(\kappa)} \leq \frac{4}{\kappa^3} + \frac{12}{\kappa^5}$, and hence for any $1/\kappa \in (0, 1]$,

$$\begin{aligned}
g_4(\kappa) + \sqrt{g_5(\kappa)} &\leq \tfrac{5}{\kappa^3} + \tfrac{17}{\kappa^5} + \tfrac{11}{\kappa^6} + \tfrac{5}{\kappa^7} + \tfrac{1}{\kappa^8} + \tfrac{2}{\kappa^9} + \tfrac{1}{\kappa^{10}} \\
&\leq \tfrac{5}{\kappa^3} + \tfrac{17}{\kappa^5} + \tfrac{20}{\kappa^6} \\
&\leq \frac{8 + \frac{24}{\kappa^2} + \frac{60}{\kappa^3} + \frac{41}{\kappa^4} + \frac{10}{\kappa^5} + \frac{21}{\kappa^6} + \frac{12}{\kappa^7} + \frac{1}{\kappa^8} + \frac{2}{\kappa^9} + \frac{1}{\kappa^{10}}}{\kappa^3}.
\end{aligned} \tag{36}$$

Since $g_4(\kappa) + \sqrt{g_5(\kappa)}$ is positive for $\kappa \geq 1$, we can substitute the bound (36) into equation (35), thereby finding that

$$
\begin{aligned}
x_-^* &= \frac{1}{3}\left(g_4(\kappa) - \sqrt{g_5(\kappa)}\right) \\
&= \frac{1}{\kappa^9} \frac{8 + \frac{24}{\kappa^2} + \frac{60}{\kappa^3} + \frac{41}{\kappa^4} + \frac{10}{\kappa^5} + \frac{21}{\kappa^6} + \frac{12}{\kappa^7} + \frac{1}{\kappa^8} + \frac{2}{\kappa^9} + \frac{1}{\kappa^{10}}}{g_4(\kappa) + \sqrt{g_5(\kappa)}} \\
&\geq \frac{1}{\kappa^6} \geq \frac{4}{5\kappa^6},
\end{aligned}
$$

which completes the proof of the bound (34).

Appendix C. Details of Algorithm 1

This section is devoted to explicit definition of all constants involved in Algorithm 1, as well as a derivation of how the algorithm's updates follows from integrating equations (13)–(15). We begin by defining the constants precisely:

$$
\begin{aligned}
\mu_{12} &= \left(1 + \frac{\gamma^2}{\xi^2}\right)\eta - \frac{\gamma^2}{2\xi}\eta^2 - \frac{\gamma^2}{\xi^3}\left(1 - e^{-\xi\eta}\right) \\
\mu_{13} &= \frac{\gamma}{\xi}\eta + \frac{\gamma}{\xi^2}\left(e^{-\xi\eta} - 1\right) \\
\mu_{22} &= 1 + \frac{\gamma^2}{\xi^2}\left(1 - \xi\eta - e^{-\xi\eta}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\mu_{23} &= \frac{\gamma}{\xi}\left(1 - e^{-\xi\eta}\right) \\
\mu_{31} &= \frac{\gamma}{\xi} - \frac{\gamma}{\xi^2}\frac{1 - e^{-\xi\eta}}{\eta} \\
\mu_{32} &= \frac{\gamma^3}{\xi^2}\eta + \frac{\gamma^3}{\xi^2}\eta e^{-\xi\eta} - \left(\frac{2\gamma^3}{\xi^3} + \frac{\gamma}{\xi}\right)\left(1 - e^{-\xi\eta}\right) \\
\mu_{33} &= e^{-\xi\eta} + \frac{\gamma^2}{\xi}\eta e^{-\xi\eta} - \frac{\gamma^2}{\xi^2}\left(1 - e^{-\xi\eta}\right),
\end{aligned}
$$

as well as

$$\sigma_{11} = \frac{2\gamma^2}{\xi^3}\eta - \frac{2\gamma^2}{\xi^2}\eta^2 + \frac{2\gamma^2}{3\xi}\eta^3 - \frac{4\gamma^2}{\xi^3}\eta e^{-\eta\xi} + \frac{\gamma^2}{\xi^4}\left(1 - e^{-2\xi\eta}\right)$$

$$\sigma_{12} = \frac{\gamma^2}{\xi^3 L}\left(\xi\eta - \left(1 - e^{-\xi\eta}\right)\right)^2$$

$$\sigma_{22} = \frac{2\gamma^2}{\xi}\eta - \frac{4\gamma^2}{\xi^2}\left(1 - e^{-\xi\eta}\right) + \frac{\gamma^2}{\xi^2}\left(1 - e^{-2\xi\eta}\right)$$

$$\sigma_{13} = -\frac{\gamma^3}{\xi^2}\eta^2\left(2e^{-\xi\eta} + 1\right) + \left(\frac{2\gamma^3}{\xi^3} - \frac{\gamma^3}{\xi^3}e^{-2\xi\eta} - \frac{4\gamma^3}{\xi^3}e^{-\eta\xi} - \frac{2\gamma}{\xi}e^{-\eta\xi}\right)\eta + \left(\frac{3\gamma^3}{2\xi^4} + \frac{\gamma}{\xi^2}\right)\left(1 - e^{-2\xi\eta}\right)$$

$$\sigma_{23} = \frac{\gamma^3}{\xi^2}\left(e^{-2\xi\eta} - 2e^{-\xi\eta} - 2\right)\eta + \frac{3\gamma^3}{2\xi^3}\left(e^{-2\xi\eta} - 4e^{-\xi\eta} + 3\right) + \frac{\gamma}{\xi}\left(1 - e^{-\xi\eta}\right)^2$$

$$\sigma_{33} = -\frac{\gamma^4}{\xi^2}\eta^2 e^{-2\xi\eta} + \left(-\frac{2\gamma^2}{\xi}e^{-2\xi\eta} + \frac{\gamma^4}{\xi^3}\left(-3e^{-2\xi\eta} + 4e^{-\xi\eta} + 2\right)\right)\eta$$
$$+ \frac{\gamma^4}{2\xi^4}\left(-5e^{-2\eta\xi} + 16e^{-\eta\xi} - 11\right) + \frac{\gamma^2}{\xi^2}\left(-3e^{-2\eta\xi} + 4e^{-\eta\xi} - 1\right) + \left(1 - e^{-2\eta\xi}\right).$$

Given these definitions, we now demonstrate how to obtain Algorithm 1 via integrating equations (13) through (15).

First step: The first step involving the Ornstein-Uhlenbeck process can be explicitly solved as (let $B_0 = 0$):

$$\begin{cases} \widehat{\theta}_t = \widetilde{\theta}_{k\eta} + (t - k\eta)\widetilde{p}_{k\eta}, \\ \widehat{r}_t = e^{-\xi(t-k\eta)}\widetilde{r}_{k\eta} - \frac{\gamma}{\xi}\left(1 - e^{-\xi(t-k\eta)}\right)\widetilde{p}_{k\eta} + \sqrt{\frac{2\xi}{L}}\int_{k\eta}^t e^{-\xi(t-s)}\mathrm{d}B_s^r. \end{cases} \tag{37}$$

It is worth noting that we have set at the beginning of every step $\widehat{\theta}_{k\eta} = \widetilde{\theta}_{k\eta}$ and $\widehat{r}_{k\eta} = \widetilde{r}_{k\eta}$.

Second step: Next we integrate the right-hand side of equation (14) to obtain $\widetilde{p}$:

$$\widehat{p}_t = \widetilde{p}_{k\eta} - \frac{1}{L\eta}\int_{k\eta}^t \left(\int_{k\eta}^{(k+1)\eta} \widehat{g}_s \mathrm{d}s\right)\mathrm{d}\varsigma + \gamma \int_{k\eta}^t \widehat{r}_s \mathrm{d}s \tag{38}$$

For a given gradient approximation $\widehat{g}_s$ at $s \in [k\eta, (k+1)\eta)$, define

$$\Delta U(\theta, p) := \int_{k\eta}^{(k+1)\eta} \widehat{g}_s(\theta, \theta + \eta p) ds.$$

Plugging into the equation, we have:

$$\int_{k\eta}^t \left(\int_{k\eta}^{(k+1)\eta} \widehat{g}_s \mathrm{d}s\right)\mathrm{d}\varsigma = (t - k\eta)\Delta U(\widetilde{\theta}_{k\eta}, \widetilde{p}_{k\eta}).$$

For $\int_{k\eta}^{t} \widehat{r}_s \mathrm{d}s$, we use Fubini's theorem and obtain that

$$\begin{aligned}
\int_{k\eta}^{t} \widehat{r}_s \mathrm{d}s &= \int_{k\eta}^{t} \left(e^{-\xi(s-k\eta)} \widetilde{r}_{k\eta} - \frac{\gamma}{\xi} \left(1 - e^{-\xi(s-k\eta)}\right) \widetilde{p}_{k\eta} + \sqrt{\frac{2\xi}{L}} \int_{k\eta}^{s} e^{-\xi(s-\widehat{s})} \mathrm{d}B^r_{\widehat{s}} \right) \mathrm{d}s \\
&= \frac{1}{\xi} \left(1 - e^{-\xi(t-k\eta)}\right) \widetilde{r}_{k\eta} + \frac{\gamma}{\xi^2} \left(1 - \xi\,(t - k\eta) - e^{-\xi(t-k\eta)}\right) \widetilde{p}_{k\eta} \\
&+ \sqrt{\frac{2\xi}{L}} \int_{k\eta}^{t} \left(\int_{k\eta}^{s} e^{-\xi(s-\widehat{s})} \mathrm{d}B^r_{\widehat{s}} \right) \mathrm{d}s \\
&= \frac{\gamma}{\xi^2} \left(1 - \xi\,(t - k\eta) - e^{-\xi(t-k\eta)}\right) \widetilde{p}_{k\eta} + \frac{1}{\xi} \left(1 - e^{-\xi(t-k\eta)}\right) \widetilde{r}_{k\eta} \\
&+ \sqrt{\frac{2\xi}{L}} \int_{k\eta}^{t} \left(\int_{\widehat{s}}^{t} e^{-\xi(s-\widehat{s})} \mathrm{d}s \right) \mathrm{d}B^r_{\widehat{s}} \\
&= \frac{\gamma}{\xi^2} \left(1 - \xi\,(t - k\eta) - e^{-\xi(t-k\eta)}\right) \widetilde{p}_{k\eta} + \frac{1}{\xi} \left(1 - e^{-\xi(t-k\eta)}\right) \widetilde{r}_{k\eta} \\
&+ \sqrt{\frac{2}{\xi L}} \int_{k\eta}^{t} \left(1 - e^{-\xi(t-\widehat{s})}\right) \mathrm{d}B^r_{\widehat{s}}.
\end{aligned}$$

Therefore, we obtain $\widehat{p}_t$ from explicit integration:

$$\begin{aligned}
\widehat{p}_t &= \underbrace{-\frac{t - k\eta}{L\eta} \Delta U(\widetilde{\theta}_{k\eta}, \widetilde{p}_{k\eta})}_{T_1(t)} \\
&+ \underbrace{\left(1 + \frac{\gamma^2}{\xi^2} \left(1 - \xi\,(t - k\eta) - e^{-\xi(t-k\eta)}\right)\right) \widetilde{p}_{k\eta} + \frac{\gamma}{\xi} \left(1 - e^{-\xi(t-k\eta)}\right) \widetilde{r}_{k\eta}}_{T_2(t)} \\
&+ \underbrace{\sqrt{\frac{2\gamma^2}{\xi L}} \int_{k\eta}^{t} \left(1 - e^{-\xi(t-\widehat{s})}\right) \mathrm{d}B^r_{\widehat{s}}}_{T_3(t)}.
\end{aligned} \tag{39}$$

Third step: Next observe that

$$\widetilde{\theta}_{(k+1)\eta} = \widetilde{\theta}_{k\eta} + \int_{k\eta}^{(k+1)\eta} \widehat{p}_t \mathrm{d}t \; = \; \widetilde{\theta}_{k\eta} + \int_{k\eta}^{(k+1)\eta} \left(T_1(t) + T_2(t) + T_3(t)\right) \mathrm{d}t,$$

where

$$\begin{aligned}
\int_{k\eta}^{(k+1)\eta} T_1(t) \mathrm{d}t &= -\frac{1}{L\eta} \int_{k\eta}^{(k+1)\eta} (t - k\eta) \mathrm{d}t \cdot \int_{k\eta}^{(k+1)\eta} \widehat{g}_s \mathrm{d}s \\
&= -\frac{\eta}{2L} \Delta U(\widetilde{\theta}_{k\eta}, \widetilde{p}_{k\eta}),
\end{aligned}$$

$$\int_{k\eta}^{(k+1)\eta} T_2(t)\mathrm{d}t = \left(\left(1+\frac{\gamma^2}{\xi^2}\right)\eta - \frac{\gamma^2}{2\xi}\eta^2 - \frac{\gamma^2}{\xi^3}\left(1-e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta} + \left(\frac{\gamma}{\xi}\eta + \frac{\gamma}{\xi^2}\left(e^{-\xi\eta}-1\right)\right)\widetilde{r}_{k\eta},$$

and

$$\begin{aligned}
\int_{k\eta}^{(k+1)\eta} T_3(t)\mathrm{d}t &= \sqrt{\frac{2\gamma^2}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(\int_{k\eta}^{t}\left(1-e^{-\xi(t-s)}\right)\mathrm{d}B_s^r\right)\mathrm{d}t\\
&= \sqrt{\frac{2\gamma^2}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(\int_{s}^{(k+1)\eta}\left(1-e^{-\xi(t-s)}\right)\mathrm{d}t\right)\mathrm{d}B_s^r\\
&= \sqrt{\frac{2\gamma^2}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(((k+1)\eta - s) + \frac{1}{\xi}\left(e^{-\xi((k+1)\eta-s)}-1\right)\right)\mathrm{d}B_s^r.
\end{aligned}$$

Putting together the pieces, we find that

$$\begin{aligned}
\widetilde{\theta}_{(k+1)\eta} &= \widetilde{\theta}_{k\eta} - \frac{\eta}{2L}\Delta U(\widetilde{\theta}_{k\eta},\widetilde{p}_{k\eta})\\
&\quad + \left(\left(1+\frac{\gamma^2}{\xi^2}\right)\eta - \frac{\gamma^2}{2\xi}\eta^2 - \frac{\gamma^2}{\xi^3}\left(1-e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta} + \left(\frac{\gamma}{\xi}\eta + \frac{\gamma}{\xi^2}\left(e^{-\xi\eta}-1\right)\right)\widetilde{r}_{k\eta}\\
&\qquad + \sqrt{\frac{2\gamma^2}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(((k+1)\eta - s) + \frac{1}{\xi}\left(e^{-\xi((k+1)\eta-s)}-1\right)\right)\mathrm{d}B_s^r.
\end{aligned}$$

We then calculate $\widetilde{r}_{(k+1)\eta}$:

$$\begin{aligned}
e^{\xi\eta}\widetilde{r}_{(k+1)\eta} - \widetilde{r}_{k\eta} &= \int_{k\eta}^{(k+1)\eta}\mathrm{d}\left(e^{\xi(t-k\eta)}\widetilde{r}_t\right)\\
&= \int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}\left(-\gamma\widehat{p}_t\mathrm{d}t + \sqrt{2\xi/L}\,\mathrm{d}B_t^r\right)\\
&= -\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}\widehat{p}_t\mathrm{d}t + \sqrt{\frac{2\xi}{L}}\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}\,\mathrm{d}B_t^r\\
&= -\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}\left(T_1(t)+T_2(t)+T_3(t)\right)\mathrm{d}t + \sqrt{\frac{2\xi}{L}}\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}\,\mathrm{d}B_t^r.
\end{aligned}$$

Term $-\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}T_1(t)\mathrm{d}t$ equals:

$$\begin{aligned}
&-\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}T_1(t)\mathrm{d}t\\
&= \frac{\gamma}{L\eta}\int_{k\eta}^{(k+1)\eta}(t-k\eta)e^{\xi(t-k\eta)}\mathrm{d}t\cdot\int_{k\eta}^{(k+1)\eta}\widehat{g}_s\mathrm{d}s\\
&= e^{\xi\eta}\left(\frac{\gamma}{L\xi} - \frac{\gamma}{L\xi^2}\frac{1-e^{-\xi\eta}}{\eta}\right)\cdot\Delta U(\widetilde{\theta}_{k\eta},\widetilde{p}_{k\eta}).
\end{aligned}$$

Term $-\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}T_2(t)\mathrm{d}t$ is equal to:

$$\begin{aligned}
&-\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}T_2(t)\mathrm{d}t\\
&= e^{\xi\eta}\left(\frac{\gamma^3}{\xi^2}\eta+\frac{\gamma^3}{\xi^2}\eta e^{-\xi\eta}-\left(\frac{2\gamma^3}{\xi^3}+\frac{\gamma}{\xi}\right)\left(1-e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta}\\
&+e^{\xi\eta}\left(\frac{\gamma^2}{\xi}\eta e^{-\xi\eta}-\frac{\gamma^2}{\xi^2}\left(1-e^{-\xi\eta}\right)\right)\widetilde{r}_{k\eta}.
\end{aligned}$$

Term $-\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}T_3(t)\mathrm{d}t$ can be calculated to be:

$$\begin{aligned}
&-\gamma\int_{k\eta}^{(k+1)\eta} e^{\xi(t-k\eta)}T_3(t)\mathrm{d}t\\
&=-\sqrt{\frac{2\gamma^4}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(\int_{k\eta}^{t} e^{\xi(t-k\eta)}\left(1-e^{-\xi(t-s)}\right)\mathrm{d}B_s^r\right)\mathrm{d}t\\
&=-\sqrt{\frac{2\gamma^4}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(\int_{s}^{(k+1)\eta}\left(e^{\xi(t-k\eta)}-e^{\xi(s-k\eta)}\right)\mathrm{d}t\right)\mathrm{d}B_s^r\\
&=\sqrt{\frac{2\gamma^4}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(e^{\xi(s-k\eta)}\left((k+1)\eta-s\right)-\frac{1}{\xi}\left(e^{\xi\eta}-e^{\xi(s-k\eta)}\right)\right)\mathrm{d}B_s^r.
\end{aligned}$$

Summing these terms, we obtain that

$$\begin{aligned}
\widetilde{r}_{(k+1)\eta} &= e^{-\xi\eta}\widetilde{r}_{k\eta}\\
&+\left(\frac{\gamma}{L\xi}-\frac{\gamma}{L\xi^2}\frac{1-e^{-\xi\eta}}{\eta}\right)\cdot\Delta U(\widetilde{\theta}_{k\eta},\widetilde{p}_{k\eta})\\
&+\left(\frac{\gamma^3}{\xi^2}\eta+\frac{\gamma^3}{\xi^2}\eta e^{-\xi\eta}-\left(\frac{2\gamma^3}{\xi^3}+\frac{\gamma}{\xi}\right)\left(1-e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta}+\left(\frac{\gamma^2}{\xi}\eta e^{-\xi\eta}-\frac{\gamma^2}{\xi^2}\left(1-e^{-\xi\eta}\right)\right)\widetilde{r}_{k\eta}\\
&+\sqrt{\frac{2\xi}{L}}\int_{k\eta}^{(k+1)\eta}\left(\frac{\gamma^2}{\xi}\left((k+1)\eta-s\right)e^{-\xi((k+1)\eta-s)}+\left(1+\frac{\gamma^2}{\xi^2}\right)e^{-\xi((k+1)\eta-s)}-\frac{\gamma^2}{\xi^2}\right)\mathrm{d}B_s^r.
\end{aligned}$$

For $\widetilde{p}_{(k+1)\eta}$, we directly know from equation (39) that

$$\begin{aligned}
\widetilde{p}_{(k+1)\eta} = -\frac{1}{L}\Delta U(\widetilde{\theta}_{k\eta},\widetilde{p}_{k\eta})+\left(1+\frac{\gamma^2}{\xi^2}\left(1-\xi\eta-e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta}+\frac{\gamma}{\xi}\left(1-e^{-\xi\eta}\right)\widetilde{r}_{k\eta}\\
+\sqrt{\frac{2\gamma^2}{\xi L}}\int_{k\eta}^{(k+1)\eta}\left(1-e^{-\xi((k+1)\eta-s)}\right)\mathrm{d}B_s^r.
\end{aligned}$$

Therefore, $\widetilde{x}_{(k+1)}=(\widetilde{\theta}_{(k+1)},\widetilde{p}_{(k+1)},\widetilde{r}_{(k+1)})$ conditioning on $\widetilde{x}_{(k)}$ follows a normal distribution. We calculate its mean and covariance below. We first find $\mathbb{E}\left[\widetilde{x}_{(k+1)}\right]$ using properties of the

Itô integral:

$$\mathbb{E}\left[\widetilde{\theta}_{(k+1)\eta}\right] = \widetilde{\theta}_{k\eta} - \frac{\eta}{2L}\Delta U(\widetilde{\theta}_{k\eta}, \widetilde{p}_{k\eta}) \\ + \left(\left(1 + \frac{\gamma^2}{\xi^2}\right)\eta - \frac{\gamma^2}{2\xi}\eta^2 - \frac{\gamma^2}{\xi^3}\left(1 - e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta} + \left(\frac{\gamma}{\xi}\eta + \frac{\gamma}{\xi^2}\left(e^{-\xi\eta} - 1\right)\right)\widetilde{r}_{k\eta},$$

$$\mathbb{E}\left[\widetilde{p}_{(k+1)\eta}\right] = -\frac{1}{L}\Delta U(\widetilde{\theta}_{k\eta}, \widetilde{p}_{k\eta}) + \left(1 + \frac{\gamma^2}{\xi^2}\left(1 - \xi\eta - e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta} + \frac{\gamma}{\xi}\left(1 - e^{-\xi\eta}\right)\widetilde{r}_{k\eta}.$$

$$\mathbb{E}\left[\widetilde{r}_{(k+1)\eta}\right] = \left(\frac{\gamma}{L\xi} - \frac{\gamma}{L\xi^2}\frac{1 - e^{-\xi\eta}}{\eta}\right)\cdot\Delta U(\widetilde{\theta}_{k\eta}, \widetilde{p}_{k\eta}) + \left(\frac{\gamma^3}{\xi^2}\eta + \frac{\gamma^3}{\xi^2}\eta e^{-\xi\eta} - \left(\frac{2\gamma^3}{\xi^3} + \frac{\gamma}{\xi}\right)\left(1 - e^{-\xi\eta}\right)\right)\widetilde{p}_{k\eta} \\ + \left(e^{-\xi\eta} + \frac{\gamma^2}{\xi}\eta e^{-\xi\eta} - \frac{\gamma^2}{\xi^2}\left(1 - e^{-\xi\eta}\right)\right)\widetilde{r}_{k\eta}.$$

Since all three processes $\widetilde{\theta}$, $\widetilde{p}$ and $\widetilde{r}$ share the same Brownian motion, we can use Itô isometry to calculate their covariance. For example, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\theta}_{(k+1)\eta} - \mathbb{E}\left[\widetilde{\theta}_{(k+1)\eta}\right]\right)\left(\widetilde{p}_{(k+1)\eta} - \mathbb{E}\left[\widetilde{p}_{(k+1)\eta}\right]\right)^{\mathrm{T}}\right] \\
&= \frac{2\gamma^2}{\xi L}\mathbb{E}\left[\left(\int_{k\eta}^{(k+1)\eta}\left(((k+1)\eta - s) + \frac{1}{\xi}\left(e^{-\xi((k+1)\eta - s)} - 1\right)\right)\mathrm{d}B_s^r\right)\left(\int_{k\eta}^{(k+1)\eta}\left(1 - e^{-\xi((k+1)\eta - s)}\right)\mathrm{d}B_s^r\right)^{\mathrm{T}}\right] \\
&= \frac{2\gamma^2}{\xi L}\int_{k\eta}^{(k+1)\eta}\left(((k+1)\eta - s) + \frac{1}{\xi}\left(e^{-\xi((k+1)\eta - s)} - 1\right)\right)\left(1 - e^{-\xi((k+1)\eta - s)}\right)\mathrm{d}s \cdot \mathrm{I}_{d\times d} \\
&= \frac{\gamma^2}{\xi^3 L}\left(\xi\eta - \left(1 - e^{-\xi\eta}\right)\right)^2 \cdot \mathrm{I}_{d\times d}.
\end{aligned}$$

Similarly, we obtain the entire covariance matrix for the tuple $\widetilde{x}_{(k+1)} = (\widetilde{\theta}_{(k+1)}, \widetilde{p}_{(k+1)}, \widetilde{r}_{(k+1)})$:

$$\mathbb{E}\left[\left(\widetilde{x}_{(k+1)\eta} - \mathbb{E}\left[\widetilde{x}_{(k+1)\eta}\right]\right)\left(\widetilde{x}_{(k+1)\eta} - \mathbb{E}\left[\widetilde{x}_{(k+1)\eta}\right]\right)^{\mathrm{T}}\right] \\ = \frac{1}{L}\begin{pmatrix} \sigma_{11}\cdot\mathrm{I}_{d\times d} & \sigma_{12}\cdot\mathrm{I}_{d\times d} & \sigma_{13}\cdot\mathrm{I}_{d\times d} \\ \sigma_{12}\cdot\mathrm{I}_{d\times d} & \sigma_{22}\cdot\mathrm{I}_{d\times d} & \sigma_{23}\cdot\mathrm{I}_{d\times d} \\ \sigma_{13}\cdot\mathrm{I}_{d\times d} & \sigma_{23}\cdot\mathrm{I}_{d\times d} & \sigma_{33}\cdot\mathrm{I}_{d\times d} \end{pmatrix}. \quad (40)$$

where

$$\sigma_{11} = \frac{2\gamma^2}{\xi^3}\eta - \frac{2\gamma^2}{\xi^2}\eta^2 + \frac{2\gamma^2}{3\xi}\eta^3 - \frac{4\gamma^2}{\xi^3}\eta e^{-\eta\xi} + \frac{\gamma^2}{\xi^4}\left(1 - e^{-2\xi\eta}\right),$$

$$\sigma_{12} = \frac{\gamma^2}{\xi^3 L}\left(\xi\eta - \left(1 - e^{-\xi\eta}\right)\right)^2,$$

$$\sigma_{22} = \frac{2\gamma^2}{\xi}\eta - \frac{4\gamma^2}{\xi^2}\left(1 - e^{-\xi\eta}\right) + \frac{\gamma^2}{\xi^2}\left(1 - e^{-2\xi\eta}\right),$$

$$\sigma_{13} = -\frac{\gamma^3}{\xi^2}\eta^2\left(2e^{-\xi\eta} + 1\right) + \left(\frac{2\gamma^3}{\xi^3} - \frac{\gamma^3}{\xi^3}e^{-2\xi\eta} - \frac{4\gamma^3}{\xi^3}e^{-\eta\xi} - \frac{2\gamma}{\xi}e^{-\eta\xi}\right)\eta + \left(\frac{3\gamma^3}{2\xi^4} + \frac{\gamma}{\xi^2}\right)\left(1 - e^{-2\xi\eta}\right),$$

$$\sigma_{23} = \frac{\gamma^3}{\xi^2}\left(e^{-2\xi\eta} - 2e^{-\xi\eta} - 2\right)\eta + \frac{3\gamma^3}{2\xi^3}\left(e^{-2\xi\eta} - 4e^{-\xi\eta} + 3\right) + \frac{\gamma}{\xi}\left(1 - e^{-\xi\eta}\right)^2, \quad \text{and}$$

$$\sigma_{33} = -\frac{\gamma^4}{\xi^2}\eta^2 e^{-2\xi\eta} + \left(-\frac{2\gamma^2}{\xi}e^{-2\xi\eta} + \frac{\gamma^4}{\xi^3}\left(-3e^{-2\xi\eta} + 4e^{-\xi\eta} + 2\right)\right)\eta$$
$$+ \frac{\gamma^4}{2\xi^4}\left(-5e^{-2\eta\xi} + 16e^{-\eta\xi} - 11\right) + \frac{\gamma^2}{\xi^2}\left(-3e^{-2\eta\xi} + 4e^{-\eta\xi} - 1\right) + \left(1 - e^{-2\eta\xi}\right).$$