Proceedings of Machine Learning Research vol 75:1–24, 2018 31st Annual Conference on Learning Theory

Underdamped Langevin MCMC: A non-asymptotic analysis

Xiang Cheng * X.CHENG@BERKELEY.EDU
University of California Berkeley & Adobe Research

Niladri S. Chatterji * CHATTERJI@BERKELEY.EDU
University of California Berkeley

Peter L. Bartlett PETER@BERKELEY.EDU
University of California Berkeley

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California Berkeley

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the underdamped Langevin diffusion when the log of the target distribution is smooth and strongly concave. We present a MCMC algorithm based on its discretization and show that it achieves ε error (in 2-Wasserstein distance) in $\mathcal{O}(\sqrt{d}/\varepsilon)$ steps. This is a significant improvement over the best known rate for overdamped Langevin MCMC, which is $\mathcal{O}(d/\varepsilon^2)$ steps under the same smoothness/concavity assumptions. The underdamped Langevin MCMC scheme can be viewed as a version of Hamiltonian Monte Carlo (HMC) which has been observed to outperform overdamped Langevin MCMC methods in a number of application areas. We provide quantitative rates that support this empirical wisdom.

1. Introduction

In this paper, we study the continuous time *underdamped* Langevin diffusion represented by the following stochastic differential equation (SDE):

$$\begin{aligned} dv_t &= -\gamma v_t dt - u\nabla f(x_t)dt + (\sqrt{2\gamma u})dB_t \\ dx_t &= v_t dt, \end{aligned} \tag{1}$$

where $(x_t, v_t) \in \mathbb{R}^{2d}$, f is a twice continuously-differentiable function and B_t represents standard Brownian motion in $\mathbb{R}^d$. Under fairly mild conditions, it can be shown that the invariant distribution of the continuous-time process (1) is proportional to $\exp(-(f(x) + \|v\|_2^2/2u))$. Thus the marginal distribution of x is proportional to $\exp(-f(x))$. There is a discretized version of (1) which can be implemented algorithmically, and provides a useful way to sample from $p^*(x) \propto e^{-f(x)}$ when the normalization constant is not known.

Our main result establishes the convergence of SDE (1) as well as its discretization, to the invariant distribution. This provides explicit rates for sampling from log-smooth and strongly log-concave

* X. Cheng and N. S. Chatterji contributed equally to this work.

distributions using the underdamped Langevin Markov chain Monte Carlo (MCMC) algorithm (Algorithm 2.1).

Underdamped Langevin diffusion is particularly interesting because it contains a Hamiltonian component, and its discretization can be viewed as a form of Hamiltonian MCMC. Hamiltonian MCMC (see review of HMC in Neal, 2011; Betancourt et al., 2017) has been empirically observed to converge faster to the invariant distribution compared to standard Langevin MCMC which is a discretization of *overdamped* Langevin diffusion,

$$dx_t = -\nabla f(x_t)dt + \sqrt{2}dB_t,$$

the first order SDE corresponding to the high friction limit of (1). This paper provides a non-asymptotic quantitative explanation for this statement.

1.1. Related Work

The first explicit proof of non-asymptotic convergence of overdamped Langevin MCMC for log-smooth and strongly log-concave distributions was given by Dalalyan (2017), where it was shown that discrete, overdamped Langevin diffusion achieves ε error, in total variation distance, in $\mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$ steps. Following this, Durmus and Moulines (2016) proved that the same algorithm achieves ε error, in 2-Wasserstein distance, in $\mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$ steps. Cheng and Bartlett (2017) obtained results similar to those by Dalalyan (2017) when the error is measured by KL-divergence. Recently Raginsky et al. (2017) and Dalalyan and Karagulyan (2017) also analyzed convergence of overdamped Langevin MCMC with stochastic gradient updates. Asymptotic guarantees for overdamped Langevin MCMC was established much earlier by Gelfand and Mitter (1991); Roberts and Tweedie (1996).

Hamiltonian Monte Carlo (HMC) is a broad class of algorithms which involve Hamiltonian dynamics in some form. We refer to Ma et al. (2015) for a survey of the results in this area. Among these, the variant studied in this paper (Algorithm 2.1), based on the discretization of (1), has a natural physical interpretation as the evolution of a particle's dynamics under a force field and drag. This equation was first proposed by Kramers (1940) in the context of chemical reactions. The continuous-time process has been studied extensively (Brockett, 1997; Hérau, 2002; Villani, 2009; Bolley et al., 2010; Calogero, 2012; Mischler and Mouhot, 2014; Dolbeault et al., 2015; Gorham et al., 2016; Baudoin, 2016; Eberle et al., 2017).

However, to the best of our knowledge, prior to this work, there was no polynomial-in-dimension convergence result for any version of HMC under a log-smooth or strongly log-concave assumption for the target distribution. Most closely related to our work is the recent paper Eberle et al. (2017) who demonstrated a contraction property of the continuous-time process defined (1). That result deals, however, with a much larger class of functions, and because of this the distance to the invariant distribution scales exponentially with dimension d. Subsequent to the appearance of the arXiv version of this work, two recent papers also analyzed and provided non-asymptotic guarantees for different versions of HMC. Lee and Vempala (2017) analyzed Riemannian HMC for sampling from polytopes using a logarithmic barrier function. Mangoubi and Smith (2017) studied a different variant of HMC under similar assumptions to this paper to get a mixing time bound of $\mathcal{O}(\frac{\sqrt{d}\kappa^{6.5}}{\varepsilon})$ in 1-Wasserstein distance (same as our result in d and ε but worse in the condition number κ). They also establish mixing time bounds for higher order integrators (both with and without a Metropolis correction) which have improved dependence in both d and ε but under a much stronger separability assumption[1].

Also related is the recent work on understanding acceleration of first-order optimization methods as discretizations of second-order differential equations (Su et al., 2014; Krichene et al., 2015; Wibisono et al., 2016).

1.2. Contributions

Our main contribution in this paper is to prove that Algorithm 2.1, a variant of HMC algorithm, converges to ε error in 2-Wasserstein distance after $\mathcal{O}\left(\frac{\sqrt{d}\kappa^2}{\varepsilon}\right)$ iterations, under the assumption that the target distribution is of the form $p^* \propto \exp(-(f(x))$, where f is L smooth and m strongly convex (see section 1.4.1), with $\kappa = L/m$ denoting the condition number. Compared to the results of Durmus and Moulines (2016) on the convergence of Langevin MCMC in W_2 in $\mathcal{O}\left(\frac{d\kappa^2}{\varepsilon^2}\right)$ iterations, this is an improvement in both d and ϵ. We also analyze the convergence of chain when we have noisy gradients with bounded variance and establish non-asymptotic convergence guarantees in this setting.

1.3. Organization of the Paper

In the next subsection we establish the notation and assumptions that we use throughout the paper. In Section 2 we present the discretized version of (1) and state our main results for convergence to the invariant distribution. Section 3 then establishes exponential convergence for the continuous-time process and in Section 4 we show how to control the discretization error. Finally in Section 5 we prove the convergence of the discretization (1). We defer technical lemmas to the appendix.

1.4. Notation and Definitions

In this section, we present basic definitions and notational conventions. Throughout, we let $\|v\|_2$ denotes the Euclidean norm, for a vector $v \in \mathbb{R}^d$.

1.4.1. ASSUMPTIONS ON f

We make the following assumptions regarding the function f.

(A1) The function f is twice continuously-differentiable on $\mathbb{R}^d$ and has Lipschitz continuous gradients; that is, there exists a positive constant $L > 0$ such that for all $x, y \in \mathbb{R}^d$ we have

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

(A2) f is m-strongly convex, that is, there exists a positive constant $m > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|x - y\|_2^2.$$

1. They assume that the potential function f is a sum of d/c functions $\{f_i\}_{i=1}^{\lceil \frac{d}{c} \rceil}$, where each f_i only depends on a distinct set of c coordinates, for some constant $c \in \mathbb{N}$.

It is fairly easy to show that under these two assumptions the Hessian of f is positive definite throughout its domain, with $mI_{d\times d} \preceq \nabla^2 f(x) \preceq LI_{d\times d}$. We define $\kappa = L/m$ as the condition number. Throughout the paper we denote the minimum of $f(x)$ by x^*. Finally, we assume that we have a gradient oracle $\nabla f(\cdot)$; that is, we have access to $\nabla f(x)$ for all $x \in \mathbb{R}^d$.

1.4.2. COUPLING AND WASSERSTEIN DISTANCE

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. Given probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we define a *transference plan* ζ between μ and ν as a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for all sets $A \in \mathcal{B}(\mathbb{R}^d)$, $\zeta(A \times \mathbb{R}^d) = \mu(A)$ and $\zeta(\mathbb{R}^d \times A) = \nu(A)$. We denote $\Gamma(\mu, \nu)$ as the set of all transference plans. A pair of random variables (X, Y) is called a coupling if there exists a $\zeta \in \Gamma(\mu, \nu)$ such that (X, Y) are distributed according to ζ. (With some abuse of notation, we will also refer to ζ as the coupling.)

We define the Wasserstein distance of order two between a pair of probability measures as follows:

$$W_2(\mu, \nu) \triangleq \left(\inf_{\zeta \in \Gamma(\mu,\nu)} \int \|x - y\|_2^2 d\zeta(x, y) \right)^{1/2}.$$

Finally we denote by $\Gamma_{opt}(\mu, \nu)$ the set of transference plans that achieve the infimum in the definition of the Wasserstein distance between μ and ν (for more properties of $W_2(\cdot, \cdot)$ see Villani, 2008).

1.4.3. UNDERDAMPED LANGEVIN DIFFUSION

Throughout the paper we use B_t to denote standard Brownian motion (Mörters and Peres, 2010). Next we set up the notation specific to the continuous and discrete processes that we study in this paper.

1. Consider the exact underdamped Langevin diffusion defined by the SDE (1), with an initial condition $(x_0, v_0) \sim p_0$ for some distribution p_0 on $\mathbb{R}^{2d}$. Let p_t denote the distribution of (x_t, v_t) and let Φ_t denote the operator that maps from p_0 to p_t:

$$\Phi_t p_0 = p_t. \tag{2}$$

2. One step of the discrete underdamped Langevin diffusion is defined by the SDE

$$\begin{aligned} d\tilde{v}_t &= -\gamma \tilde{v}_t dt - u \nabla f(\tilde{x}_0) dt + (\sqrt{2\gamma u}) dB_t \\ d\tilde{x}_t &= \tilde{v}_t dt, \end{aligned} \tag{3}$$

   with an initial condition $(\tilde{x}_0, \tilde{v}_0) \sim \tilde{p}_0$. Let $\tilde{p}_t$ and $\tilde{\Phi}_t$ be defined analogously to p_t and Φ_t for (x_t, v_t).

   Note 1: The discrete update differs from (1) by using $\tilde{x}_0$ instead of $\tilde{x}_t$ in the drift of $\tilde{v}_s$.

   Note 2: We will only be analyzing the solutions to (3) for small t. Think of an integral solution of (3) as a single step of the discrete Langevin MCMC.

Algorithm 1: Underdamped Langevin MCMC

Input : Step size $\delta < 1$, number of iterations n, initial point $(x^{(0)}, 0)$, smoothness parameter L and gradient oracle $\nabla f(\cdot)$

for $i = 0, 1, \ldots, n-1$ **do**

  Sample $(x^{i+1}, v^{i+1}) \sim Z^{i+1}(x^i, v^i)$

end

1.4.4. STATIONARY DISTRIBUTIONS

Throughout the paper, we denote by p^* the unique distribution which satisfies $p^*(x, v) \propto \exp -(f(x) + \frac{1}{2u}\|v\|_2^2)$. It can be shown that p^* is the unique invariant distribution of (1) (see Proposition 6.1 in Pavliotis, 2016). Let $g(x, v) = (x, x + v)$. We let q^* be the distribution of $g(x, v)$ when $(x, v) \sim p^*$.

2. Results

2.1. Algorithm

The underdamped Langevin MCMC algorithm that we analyze in this paper is shown in Algorithm 2.1.

The random vector $Z^{i+1}(x_i, v_i) \in \mathbb{R}^{2d}$, conditioned on (x^i, v^i), has a Gaussian distribution with conditional mean and covariance obtained from the following computations:

$$
\begin{aligned}
&\mathbb{E}\left[v^{i+1}\right] = v^i e^{-2\delta} - \frac{1}{2L}(1 - e^{-2\delta})\nabla f(x^i) \\
&\mathbb{E}\left[x^{i+1}\right] = x^i + \frac{1}{2}(1 - e^{-2\delta})v^i - \frac{1}{2L}\left(\delta - \frac{1}{2}\left(1 - e^{-2\delta}\right)\right)\nabla f(x^i) \\
&\mathbb{E}\left[\left(x^{i+1} - \mathbb{E}\left[x^{i+1}\right]\right)\left(x^{i+1} - \mathbb{E}\left[x^{i+1}\right]\right)^\top\right] = \frac{1}{L}\left[\delta - \frac{1}{4}e^{-4\delta} - \frac{3}{4} + e^{-2\delta}\right] \cdot I_{d\times d} \\
&\mathbb{E}\left[\left(v^{i+1} - \mathbb{E}\left[v^{i+1}\right]\right)\left(v^{i+1} - \mathbb{E}\left[v^{i+1}\right]\right)^\top\right] = \frac{1}{L}(1 - e^{-4\delta}) \cdot I_{d\times d} \\
&\mathbb{E}\left[\left(x^{i+1} - \mathbb{E}\left[x^{i+1}\right]\right)\left(v^{i+1} - \mathbb{E}\left[v^{i+1}\right]\right)^\top\right] = \frac{1}{2L}\left[1 + e^{-4\delta} - 2e^{-2\delta}\right] \cdot I_{d\times d}.
\end{aligned}
$$

The distribution is obtained by integrating the discrete underdamped Langevin diffusion (3) up to time δ, with the specific choice of $\gamma = 2$ and $u = 1/L$. In other words, if $p^{(i)}$ is the distribution of (x^i, v^i), then $Z^{i+1}(x^i, v^i) \sim p^{(i+1)} = \tilde{\Phi}_\delta p^{(i)}$. Refer to Lemma 11 in Appendix A for the derivation.

2.2. Main Result

Theorem 1 *Let $p^{(n)}$ be the distribution of the iterate of Algorithm 2.1 after n steps starting with the initial distribution $p^{(0)}(x, v) = 1_{x=x^{(0)}} \cdot 1_{v=0}$. Let the initial distance to optimum satisfy $\|x^{(0)} - x^*\|_2^2 \le \mathcal{D}^2$. If we set the step size to be*

$$\delta = \min\left\{\frac{\varepsilon}{104\kappa}\sqrt{\frac{1}{d/m + \mathcal{D}^2}}, 1\right\}$$

and run Algorithm 2.1 for n iterations with

$$n \geq \max\left\{\frac{208\kappa^2}{\varepsilon} \cdot \sqrt{\frac{d}{m} + \mathcal{D}^2}, 2\kappa\right\} \cdot \log\left(\frac{24\left(\frac{d}{m} + \mathcal{D}^2\right)}{\varepsilon}\right),$$

then we have the guarantee that

$$W_2(p^{(n)}, p^*) \leq \varepsilon.$$

Remark 2 *The dependence of the runtime on d, ε is thus $\tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon}\right)$, which is a significant improvement over the corresponding $\mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$ runtime of (overdamped) Langevin diffusion by Durmus and Moulines (2016). Also note that in almost all regimes of interest $\delta \ll 1$.*

We note that the $\log(24(d/m + \mathcal{D}^2)/\varepsilon)$ factor can be shaved off by using a time-varying step size. We present this result as Theorem 14 in Appendix C. In neither theorem have we attempted to optimize the constants.

2.2.1. RESULT WITH STOCHASTIC GRADIENTS

Now we state convergence guarantees when we have access to noisy gradients, $\hat{\nabla} f(x) = \nabla f(x) + \xi$, where ξ is a independent random variable that satisfies

1. The noise is unbiased – $\mathbb{E}\left[\xi\right] = 0$.
2. The noise has bounded variance – $\mathbb{E}[\|\xi\|_2^2] \leq d\sigma^2$.

Each step of the dynamics is now driven by the SDE,

$$\begin{aligned} d\hat{v}_t &= -\gamma \hat{v}_t dt - u\hat{\nabla} f(\hat{x}_0)dt + (\sqrt{2\gamma u})dB_t \\ d\hat{x}_t &= \hat{v}_s dt, \end{aligned} \tag{4}$$

with an initial condition $(\hat{x}_0, \hat{v}_0) \sim \hat{p}_0$. Let $\hat{p}_t$ and $\hat{\Phi}_t$ be defined analogously to p_t and Φ_t for (x_t, v_t) in Section 1.4.3.

Theorem 3 (Proved in Appendix D) *Let $p^{(n)}$ be the distribution of the iterate of Algorithm D (presented in Appendix D) after n steps starting with the initial distribution $p^{(0)}(x, v) = 1_{x=x^{(0)}} \cdot 1_{v=0}$. Let the initial distance to optimum satisfy $\|x^{(0)} - x^*\|_2^2 \leq \mathcal{D}^2$. If we set the step size to be*

$$\delta = \min\left\{\frac{\varepsilon}{310\kappa}\sqrt{\frac{1}{d/m + \mathcal{D}^2}}, \frac{\varepsilon^2 L^2}{1440\sigma^2 d\kappa}, 1\right\},$$

and run Algorithm 2.1 for n iterations with

$$n \geq \max\left\{\frac{2880\kappa^2\sigma^2 d}{\varepsilon^2 L^2}, \frac{620\kappa^2}{\varepsilon} \cdot \sqrt{\frac{d}{m} + \mathcal{D}^2}, 2\kappa\right\} \cdot \log\left(\frac{36\left(\frac{d}{m} + \mathcal{D}^2\right)}{\varepsilon}\right),$$

then we have the guarantee that

$$W_2(p^{(n)}, p^*) \leq \varepsilon.$$

Remark 4 *Note that when the variance in the gradients – $\sigma^2 d$ is large we recover back the rate of overdamped Langevin diffusion and we need $\tilde{\mathcal{O}}(\sigma^2\kappa^2 d/\epsilon^2)$ steps to achieve accuracy of ε in W_2.*

3. Convergence of the Continuous-Time Process

In this section we prove Theorem 5, which demonstrates a contraction for solutions of the SDE (1). We will use Theorem 5 along with a bound on the discretization error between (1) and (3) to establish guarantees for Algorithm 2.1.

Theorem 5 *Let (x_0, v_0) and (y_0, w_0) be two arbitrary points in $\mathbb{R}^{2d}$. Let p_0 be the Dirac delta distribution at (x_0, v_0) and let p_0' be the Dirac delta distribution at (y_0, w_0). We let $u = 1/L$ and $\gamma = 2$. Then for every $t > 0$, there exists a coupling $\zeta_t(x_0, v_0, y_0, w_0) \in \Gamma(\Phi_t p_0, \Phi_t p_0')$ such that*

$$\mathbb{E}_{(x_t,v_t,y_t,w_t)\sim\zeta_t((x_0,v_0,y_0,w_0))}\left[\|x_t - y_t\|_2^2 + \|(x_t + v_t) - (y_t + w_t)\|_2^2\right] \\ \leq e^{-t/\kappa}\left\{\|x_0 - y_0\|_2^2 + \|(x_0 + v_0) - (y_0 + w_0)\|_2^2\right\}. \tag{5}$$

Remark 6 *A similar objective function was used in Eberle et al. (2017) to prove contraction.*

Given this theorem it is fairly easy to establish the exponential convergence of the continuous-time process to the stationary distribution in W_2.

Corollary 7 *Let p_0 be arbitrary distribution with $(x_0, v_0) \sim p_0$. Let q_0 and $\Phi_t q_0$ be the distributions of $(x_0, x_0 + v_0)$ and $(x_t, x_t + v_t)$, respectively (i.e., the images of p_0 and $\Phi_t p_0$ under the map $g(x, v) = (x, x + v)$). Then*

$$W_2(\Phi_t q_0, q^*) \leq e^{-t/2\kappa} W_2(q_0, q^*).$$

Proof We let $\zeta_0 \in \Gamma(p_0, p^*)$ such that $\mathbb{E}_{\zeta_0}\left[\|x_0 - y_0\|_2^2 + \|x_0 - y_0 + v_0 - w_0\|_2^2\right] = W_2^2(q_0, q^*)$. For every x_0, v_0, y_0, w_0 we let $\zeta_t(x_0, v_0, y_0, w_0)$ be the coupling as prescribed by Theorem 5. Then we have,

$$\begin{aligned}
&W_2^2(q_t, q^*) \\
&\overset{(i)}{\leq} \mathbb{E}_{(x_0,v_0,y_0,w_0)\sim\zeta_0}\left[\mathbb{E}_{(x_t,v_t,y_t,w_t)\sim\zeta_t(x_0,v_0,y_0,w_0)}\left[\|x_t - y_t\|_2^2 + \|x_t - y_t + v_t - w_t\|_2^2 \Big| x_0, y_0, v_0, w_0\right]\right] \\
&\overset{(ii)}{\leq} \mathbb{E}_{(x_0,v_0,y_0,w_0)\sim\zeta_0}\left[e^{-t/\kappa}\left(\|x_0 - y_0\|_2^2 + \|x_0 - y_0 + v_0 - w_0\|_2^2\right)\right] \\
&\overset{(iii)}{=} e^{-t/\kappa} W_2^2(q_0, q^*),
\end{aligned}$$

where (i) follows as the Wasserstein distance is defined by the optimal coupling and by the tower property of expectation, (ii) follows by applying Theorem 5 and finally (iii) follows by choice of ζ_0 to be the optimal coupling. One can verify that the random variables $(x_t, x_t + v_t, y_t, y_t + w_t)$ defines a valid coupling between q_t and q^*. Taking square roots completes the proof. ∎

Lemma 8 (Sandwich Inequality) *The triangle inequality for the Euclidean norm implies that*

$$\frac{1}{2} W_2(p_t, p^*) \leq W_2(q_t, q^*) \leq 2W_2(p_t, p^*). \tag{6}$$

Thus we also get convergence of $\Phi_t p_0$ to p^*:

$$W_2(\Phi_t p_0, p^*) \leq 4e^{-t/2\kappa} W_2(p_0, p^*).$$

Proof Using Young's inequality, we have

$$\|x + v - (x' + v')\|_2^2 \leq 2\|x - x'\|_2^2 + 2\|v - v'\|_2^2.$$

Let $\gamma_t \in \Gamma_{opt}(p_t, p^*)$. Then

$$\begin{aligned} W_2(q_t, q^*) &\leq \sqrt{\mathbb{E}_{(x,v,x',v')\sim\gamma_t}\left[\|x - x'\|_2^2 + \|x + v - (x' + v')\|_2^2\right]} \\ &\leq \sqrt{\mathbb{E}_{(x,v,x',v')\sim\gamma_t}\left[3\|x - x'\|_2^2 + 2\|v - v'\|_2^2\right]} \\ &\leq 2\sqrt{\mathbb{E}_{(x,v,x',v')\sim\gamma_t}\left[\|x - x'\|_2^2 + \|v - v'\|_2^2\right]} = 2W_2(p_t, p^*). \end{aligned}$$

The other direction follows identical arguments, using instead the inequality

$$\|v - v'\|_2^2 \leq 2\|x + v - (x' + v')\|_2^2 + 2\|x - x'\|_2^2.$$

■

We now turn to the proof of Theorem 5.

Proof [Proof of Theorem 5] We will prove Theorem 5 in four steps. Our proof relies on a synchronous coupling argument, where p_t and p'_t are coupled (trivially) through independent p_0 and p'_0, and through shared Brownian motion B_t.

Step 1: By the definition of the continuous time process (1), we get

$$\frac{d}{dt}\left[(x_t + v_t) - (y_t + w_t)\right] = -(\gamma - 1)v_t - u\nabla f(x_t) - \left\{-(\gamma - 1)w_t - u\nabla f(y_t)\right\}.$$

The two processes are coupled synchronously which ensures that the Brownian motion terms cancel out. For ease of notation, we define $z_t \triangleq x_t - y_t$ and $\psi_t \triangleq v_t - w_t$. As f is twice differentiable, by Taylor's theorem we have

$$\nabla f(x_t) - \nabla f(y_t) = \underbrace{\left[\int_0^1 \nabla^2 f(x_t + h(y_t - x_t))dh\right]}_{\triangleq \mathcal{H}_t} z_t.$$

Using the definition of $\mathcal{H}_t$ we obtain

$$\frac{d}{dt}\left[z_t + \psi_t\right] = -\left((\gamma - 1)\psi_t + u\mathcal{H}_t z_t\right).$$

Similarly we also have the following derivative for the position update:

$$\frac{d}{dt}\left[x_t - y_t\right] = \frac{d}{dt}\left[z_t\right] = \psi_t.$$

Step 2: Using the result from Step 1, we get

$$\begin{aligned} &\frac{d}{dt}\left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2\right] \\ &= -2\langle (z_t + \psi_t, z_t), ((\gamma - 1)\psi_t + u\mathcal{H}_t z_t, -\psi_t)\rangle \\ &= -2\begin{bmatrix} z_t + \psi_t & z_t \end{bmatrix} \underbrace{\begin{bmatrix} (\gamma - 1)I_{d\times d} & u\mathcal{H}_t - (\gamma - 1)I_{d\times d} \\ -I_{d\times d} & I_{d\times d} \end{bmatrix}}_{\triangleq S_t} \begin{bmatrix} z_t + \psi_t \\ z_t \end{bmatrix} \end{aligned} \quad (7)$$

Here $(z_t + \psi_t, z_t)$ denotes the concatenation of $z_t + \psi_t$ and z_t.

Step 3: Note that for any vector $x \in \mathbb{R}^{2d}$ the quadratic form $x^\top S_t x$ is equal to

$$x^\top S_t x = x^\top \left(\frac{S_t + S_t^\top}{2} \right) x.$$

Let us define the symmetric matrix $Q_t = (S_t + S_t^\top)/2$. We now compute and lower bound the eigenvalues of the matrix Q_t by making use of an appropriate choice of the parameters γ and u. The eigenvalues of Q_t are given by the characteristic equation

$$\det \begin{bmatrix} (\gamma - 1 - \lambda) I_{d\times d} & \frac{u\mathcal{H}_t - (\gamma) I_{d\times d}}{2} \\ \frac{u\mathcal{H}_t - (\gamma) I_{d\times d}}{2} & (1-\lambda) I_{d\times d} \end{bmatrix} = 0.$$

By invoking a standard result of linear algebra (stated in the Appendix as Lemma 18), this is equivalent to solving the equation

$$\det \left((\gamma - 1 - \lambda)(1-\lambda) I_{d\times d} - \frac{1}{4} \left(u\mathcal{H}_t - \gamma I_{d\times d} \right)^2 \right) = 0.$$

Next we diagonalize $\mathcal{H}_t$ and get d equations of the form

$$(\gamma - 1 - \lambda)(1-\lambda) - \frac{1}{4} \left(u\Lambda_j - \gamma \right)^2 = 0,$$

where Λ_j with $j \in \{1, \dots d\}$ are the eigenvalues of $\mathcal{H}_t$. By the strong convexity and smoothness assumptions we have $0 < m \le \Lambda_j \le L$. We plug in our choice of parameters, $\gamma = 2$ and $u = 1/L$, to get the following solutions to the characteristic equation:

$$\lambda_j^* = 1 \pm \left(1 - \frac{\Lambda_j}{2L} \right).$$

This ensures that the minimum eigenvalue of Q_t satisfies $\lambda_{min}(Q_t) \ge 1/2\kappa$.

Step 4: Putting this together with our results in Step 2 we have the lower bound

$$[z_t + \psi_t, z_t]^\top S_t [z_t + \psi_t, z_t] = [z_t + \psi_t, z_t]^\top Q_t [z_t + \psi_t, z_t] \ge \frac{1}{2\kappa} \left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2 \right].$$

Combining this with (7) yields

$$\frac{d}{dt} \left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2 \right] \le -\frac{1}{\kappa} \left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2 \right].$$

The convergence rate of Theorem 5 follows immediately from this result by applying Grönwall's inequality (Corollary 3 in Dragomir, 2003). ■

4. Discretization Analysis

In this section, we study the solutions of the discrete process (3) up to $t = \delta$ for some small δ. Here, δ represents a *single step* of the Langevin MCMC algorithm. In Theorem 9, we will bound the discretization error between the continuous-time process (1) and the discrete process (3) starting from the same initial distribution. In particular, we bound $W_2(\Phi_\delta p_0, \tilde{\Phi}_\delta p_0)$. This will be sufficient to get the convergence rate stated in Theorem 1. Recall the definition of Φ_t and $\tilde{\Phi}_t$ from (2).

Furthermore, we will assume for now that the kinetic energy (second moment of velocity) is bounded for the continuous-time process,

$$\forall t \in [0, \delta] \quad \mathbb{E}_{p_t}\left[\|v\|_2^2\right] \leq \mathcal{E}_K. \tag{8}$$

We derive an explicit bound on $\mathcal{E}_K$ (in terms of problem parameters d, L, m etc.) in Lemma 12 in Appendix B.

In this section, we will repeatedly use the following inequality:

$$\left\|\int_0^t v_s ds\right\|_2^2 = \left\|\frac{1}{t}\int_0^t t \cdot v_s ds\right\|_2^2 \leq t\int_0^t \|v_s\|_2^2 ds,$$

which follows from Jensen's inequality using the convexity of $\|\cdot\|_2^2$.

We now present our main discretization theorem:

Theorem 9 *Let Φ_t and $\tilde{\Phi}_t$ be as defined in (2) corresponding to the continuous-time and discrete-time processes respectively. Let p_0 be any initial distribution and assume that the step size $\delta \leq 1$. As before we choose $u = 1/L$ and $\gamma = 2$. Then the distance between the continuous-time process and the discrete-time process is upper bounded by*

$$W_2(\Phi_\delta p_0, \tilde{\Phi}_\delta p_0) \leq \delta^2\sqrt{\frac{2\mathcal{E}_K}{5}}.$$

Proof We will once again use a standard synchronous coupling argument, in which $\Phi_\delta p_0$ and $\tilde{\Phi}_\delta p_0$ are coupled through the same initial distribution p_0 and common Brownian motion B_t.

First, we bound the error in velocity. By using the expression for v_t and $\tilde{v}_t$ from Lemma 10, we have

$$\begin{aligned}
\mathbb{E}\left[\|v_s - \tilde{v}_s\|_2^2\right] &\overset{(i)}{=} \mathbb{E}\left[\left\|u\int_0^s e^{-2(s-r)}\left(\nabla f(x_r) - \nabla f(x_0)\right)dr\right\|_2^2\right] \\
&= u^2\mathbb{E}\left[\left\|\int_0^s e^{-2(s-r)}\left(\nabla f(x_r) - \nabla f(x_0)dr\right)\right\|_2^2\right] \\
&\overset{(ii)}{\leq} su^2\int_0^s \mathbb{E}\left[\left\|e^{-2(s-r)}\left(\nabla f(x_r) - \nabla f(x_0)\right)\right\|_2^2\right]dr \\
&\overset{(iii)}{\leq} su^2\int_0^s \mathbb{E}\left[\|(\nabla f(x_r) - \nabla f(x_0))\|_2^2\right]dr \overset{(iv)}{\leq} su^2L^2\int_0^s \mathbb{E}\left[\|x_r - x_0\|_2^2\right]dr \\
&\overset{(v)}{=} su^2L^2\int_0^s \mathbb{E}\left[\left\|\int_0^r v_w dw\right\|_2^2\right]dr \overset{(vi)}{\leq} su^2L^2\int_0^s r\left(\int_0^r \mathbb{E}\left[\|v_w\|_2^2\right]dw\right)dr \\
&\overset{(vii)}{\leq} su^2L^2\mathcal{E}_K\int_0^s r\left(\int_0^r dw\right)dr = \frac{s^4u^2L^2\mathcal{E}_K}{3},
\end{aligned}$$

where (i) follows from the Lemma 10 and $v_0 = \tilde{v}_0$, (ii) follows from application of Jensen's inequality, (iii) follows as $|e^{-4(s-r)}| \leq 1$, (iv) is by application of the L-smoothness property of $f(x)$, (v) follows from the definition of x_r, (vi) follows from Jensen's inequality and (vii) follows by the uniform upper bound on the kinetic energy assumed in (8), and proven in Lemma 12. This completes the bound for the velocity variable. Next we bound the discretization error in the position variable:

$$\begin{aligned}
\mathbb{E}\left[\|x_s - \tilde{x}_s\|_2^2\right] &= \mathbb{E}\left[\left\|\int_0^s (v_r - \tilde{v}_r) dr\right\|_2^2\right] \leq s\int_0^s \mathbb{E}\left[\|v_r - \tilde{v}_r\|_2^2\right] dr \\
&\leq s\int_0^s \frac{r^4 u^2 L^2 \mathcal{E}_K}{3} dr = \frac{s^6 u^2 L^2 \mathcal{E}_K}{15},
\end{aligned}$$

where the first line is by coupling through the initial distribution p_0, the second line is by Jensen's inequality and the third inequality uses the preceding bound. Setting $s = \delta$ and by our choice of $u = 1/L$ we have that the squared Wasserstein distance is bounded as

$$W_2^2(\Phi_\delta p_0, \tilde{\Phi}_\delta p_0) \leq \mathcal{E}_K \left(\frac{\delta^4}{3} + \frac{\delta^6}{15}\right).$$

Given our assumption that δ is chosen to be smaller than 1, this gives the upper bound:

$$W_2^2(\Phi_\delta p_0, \tilde{\Phi}_\delta p_0) \leq \frac{2\mathcal{E}_K \delta^4}{5}.$$

Taking square roots establishes the desired result. ■

5. Proof of Theorem 1

Having established the convergence rate for the continuous-time SDE (1) and having proved a discretization error bound in Section 4 we now put these together and establish our main result for underdamped Langevin MCMC.

Proof [Proof of Theorem 1] From Corollary 7, we have that for any $i \in \{1, \ldots, n\}$

$$W_2(\Phi_\delta q^{(i)}, q^*) \leq e^{-\delta/2\kappa} W_2(q^{(i)}, q^*).$$

By the discretization error bound in Theorem 9 and the Sandwich Inequality (6), we get

$$W_2(\Phi_\delta q^{(i)}, \tilde{\Phi}_\delta q^{(i)}) \leq 2W_2(\Phi_\delta p^{(i)}, \tilde{\Phi}_\delta p^{(i)}) \leq \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}}.$$

By the triangle inequality for W_2,

$$\begin{aligned}
W_2(q^{(i+1)}, q^*) = W_2(\tilde{\Phi}_\delta q^{(i)}, q^*) &\leq W_2(\Phi_\delta q^{(i)}, \tilde{\Phi}_\delta q^{(i)}) + W_2(\Phi_\delta q^{(i)}, q^*) && (9) \\
&\leq \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + e^{-\delta/2\kappa} W_2(q^{(i)}, q^*). && (10)
\end{aligned}$$

Let us define $\eta = e^{-\delta/2\kappa}$. Then by applying (10) n times we have:

$$W_2(q^{(n)}, q^*) \leq \eta^n W_2(q^{(0)}, q^*) + \left(1 + \eta + \ldots + \eta^{n-1}\right) \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}}$$
$$\leq 2\eta^n W_2(p^{(0)}, p^*) + \left(\frac{1}{1-\eta}\right) \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}},$$

where the second step follows by summing the geometric series and by applying the upper bound (6). By another application of (6) we get:

$$W_2(p^{(n)}, p^*) \leq \underbrace{4\eta^n W_2(p^{(0)}, p^*)}_{\triangleq T_1} + \underbrace{\left(\frac{1}{1-\eta}\right) \delta^2 \sqrt{\frac{32\mathcal{E}_K}{5}}}_{\triangleq T_2}. \tag{11}$$

Observe that, $1 - \eta = 1 - e^{-\delta/2\kappa} \geq \delta/(4\kappa)$. This inequality follows as $\delta/\kappa < 1$. We now bound both terms T_1 and T_2 at a level $\varepsilon/2$ to bound the total error $W_2(p^{(n)}, p^*)$ at a level ε. Note that choice of $\delta = \varepsilon\kappa^{-1}\sqrt{1/10816\,(d/m + \mathcal{D}^2)} \leq \varepsilon\kappa^{-1}\sqrt{5/2048\mathcal{E}_K}$ (by upper bound on $\mathcal{E}_K$ in Lemma 12) ensures that,

$$T_2 = \left(\frac{1}{1-\eta}\right) \delta^2 \sqrt{\frac{32\mathcal{E}_K}{5}} \leq \frac{4\kappa}{\delta} \left(\delta^2 \sqrt{\frac{32\mathcal{E}_K}{5}}\right) \leq \frac{\varepsilon}{2}.$$

To control $T_1 < \varepsilon/2$ it is enough to ensure that

$$n > \frac{1}{\log(\eta)} \log\left(\frac{8W_2(p^{(0)}, p^*)}{\varepsilon}\right).$$

In Lemma 13 we establish a bound on $W_2^2(p^{(0)}, p^*) \leq 3(d/m + \mathcal{D}^2)$. This motivates our choice of $n > \frac{2\kappa}{\delta} \log\left(\frac{24\left(\frac{d}{m} + \mathcal{D}^2\right)}{\varepsilon}\right)$, which establishes our claim. ∎

6. Conclusion

We present an MCMC algorithm based on the underdamped Langevin diffusion and provide guarantees for its convergence to the invariant distribution in 2-Wasserstein distance. Our result is a quadratic improvement in both dimension ($\sqrt{d}$ instead of d) as well as error ($1/\varepsilon$ instead of $1/\varepsilon^2$) for sampling from strongly log-concave distributions compared to the best known results for overdamped Langevin MCMC. In its use of underdamped, second-order dynamics, our work also has connections to Nesterov acceleration (Nesterov, 1983) and to Polyak's heavy ball method (Polyak, 1964), and adds to the growing body of work that aims to understand acceleration of first-order methods as a discretization of continuous-time processes.

An interesting open question is whether we can improve the dependence on the condition number from κ^2 to κ. Another interesting direction would to explore if our approach can be used to sample efficiently from *non-log-concave* distributions. Also, lower bounds in the MCMC field are largely unknown and it would extremely useful to understand the gap between existing algorithms and optimal achievable rates. Another question could be to explore the wider class of second-order Langevin equations and study if their discretizations provide better rates for sampling from particular distributions.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the NSF through grant IIS-1619362 and the Army Research Office under grant number W911NF-17-1-0304.

References

Fabrice Baudoin. Wasserstein contraction properties for hypoelliptic diffusions. *arXiv preprint arXiv:1602.04177*, 2016.

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

François Bolley, Arnaud Guillin, and Florent Malrieu. Trend to equilibrium and particle approximation for a weakly self-consistent Vlasov-Fokker-Planck equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(5):867–884, 2010.

Roger Brockett. Oscillatory descent for function minimization. In *Current and future directions in applied mathematics*, pages 65–82. Springer, 1997.

Simone Calogero. Exponential convergence to equilibrium for kinetic Fokker-Planck equations. *Communications in Partial Differential Equations*, 37(8):1357–1390, 2012.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B*, 79(3):651–676, 2017.

Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Transactions of the American Mathematical Society*, 367(6):3807–3828, 2015.

Sever S. Dragomir. *Some Gronwall Type Inequalities and Applications*. Nova Science Publishers, 2003.

Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

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

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Jack Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *arXiv preprint arXiv:1611.06972*, 2016.

Thomas H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20:292–296, 1919.

Frédéric Hérau. Isotropic hypoellipticity and trend to the equilibrium for the Fokker-Planck equation with high degree potential. pages 1–13, 2002.

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

Walid Krichene, Alexandre Bayen, and Peter Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, pages 2845–2853, 2015.

Yin Tat Lee and Santosh S Vempala. Convergence Rate of Riemannian Hamiltonian Monte Carlo and Faster Polytope Volume Computation. *arXiv preprint arXiv:1710.06261*, 2017.

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

Oren Mangoubi and Aaron Smith. Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Stéphane Mischler and Clément Mouhot. Exponential stability of slowly decaying solutions to the kinetic Fokker-Planck equation. *arXiv preprint arXiv:1412.7487*, 2014.

Peter Mörters and Yuval Peres. *Brownian Motion*. Cambridge University Press, 2010.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Grigorios A. Pavliotis. *Stochastic Processes and Applications*. Springer, 2016.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

John R Silvester. Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467, 2000.

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

Cédric Villani. *Optimal Transport: Old and New*. Springer Science and Business Media, 2008.

Cedric Villani. *Hypocoercivity*. American Mathematical Society, 2009.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113. URL http://www.pnas.org/content/113/47/E7351.abstract.

Appendix A. Explicit Discrete Time Updates

In this section we calculate integral representations of the solutions to the continuous-time process (1) and the discrete-time process (3).

Lemma 10 *The solution (x_t, v_t) to the underdamped Langevin diffusion (1) is*

$$
\begin{aligned}
v_t &= v_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)} \nabla f(x_s) ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)} dB_s \\
x_t &= x_0 + \int_0^t v_s ds.
\end{aligned}
\tag{12}
$$

The solution $(\tilde{x}_t, \tilde{v}_t)$ of the discrete underdamped Langevin diffusion (3) is

$$
\begin{aligned}
\tilde{v}_t &= \tilde{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)} \nabla f(\tilde{x}_0) ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)} dB_s \\
\tilde{x}_t &= \tilde{x}_0 + \int_0^t \tilde{v}_s ds.
\end{aligned}
\tag{13}
$$

Proof It can be easily verified that the above expressions have the correct initial values (x_0, v_0) and $(\tilde{x}_0, \tilde{v}_0)$. By taking derivatives, one also verifies that they satisfy the differential equations in (1) and (3). ∎

Next we calculate the moments of the Gaussian used in the updates of Algorithm 2.1. These are obtained by integrating the expression for the discrete-time process presented in Lemma 10.

Lemma 11 *Conditioned on $(\tilde{x}_0, \tilde{v}_0)$, the solution $(\tilde{x}_t, \tilde{v}_t)$ of (3) with $\gamma = 2$ and $u = 1/L$ is a Gaussian with conditional mean,*

$$
\begin{aligned}
\mathbb{E}\left[\tilde{v}_t\right] &= \tilde{v}_0 e^{-2t} - \frac{1}{2L}(1-e^{-2t})\nabla f(\tilde{x}_0) \\
\mathbb{E}\left[\tilde{x}_t\right] &= \tilde{x}_0 + \frac{1}{2}(1-e^{-2t})\tilde{v}_0 - \frac{1}{2L}\left(t - \frac{1}{2}\left(1-e^{-2t}\right)\right)\nabla f(\tilde{x}_0),
\end{aligned}
$$

and with conditional covariance,

$$
\begin{aligned}
\mathbb{E}\left[\left(\tilde{x}_t - \mathbb{E}\left[\tilde{x}_t\right]\right)\left(\tilde{x}_t - \mathbb{E}\left[\tilde{x}_t\right]\right)^\top\right] &= \frac{1}{L}\left[t - \frac{1}{4}e^{-4t} - \frac{3}{4} + e^{-2t}\right] \cdot I_{d\times d} \\
\mathbb{E}\left[\left(\tilde{v}_t - \mathbb{E}\left[\tilde{v}_t\right]\right)\left(\tilde{v}_t - \mathbb{E}\left[\tilde{v}_t\right]\right)^\top\right] &= \frac{1}{L}(1-e^{-4t}) \cdot I_{d\times d} \\
\mathbb{E}\left[\left(\tilde{x}_t - \mathbb{E}\left[\tilde{x}_t\right]\right)\left(\tilde{v}_t - \mathbb{E}\left[\tilde{v}_t\right]\right)^\top\right] &= \frac{1}{2L}\left[1 + e^{-4t} - 2e^{-2t}\right] \cdot I_{d\times d}.
\end{aligned}
$$

Proof It follows from the definition of Brownian motion that the distribution of $(\tilde{x}_t, \tilde{v}_t)$ is a $2d$-dimensional Gaussian distribution. We will compute its moments below, using the expression in Lemma 10 with $\gamma = 2$ and $u = 1/L$.

Computation of the conditional means is straightforward, as we can simply ignore the zero-mean Brownian motion terms:

$$
\mathbb{E}\left[\tilde{v}_t\right] = \tilde{v}_0 e^{-2t} - \frac{1}{2L}(1-e^{-2t})\nabla f(\tilde{x}_0) \tag{14}
$$

$$
\mathbb{E}\left[\tilde{x}_t\right] = \tilde{x}_0 + \frac{1}{2}(1-e^{-2t})\tilde{v}_0 - \frac{1}{2L}\left(t - \frac{1}{2}\left(1-e^{-2t}\right)\right)\nabla f(\tilde{x}_0). \tag{15}
$$

The conditional variance for $\tilde{v}_t$ only involves the Brownian motion term:

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{v}_t - \mathbb{E}\left[\tilde{v}_t\right]\right)\left(\tilde{v}_t - \mathbb{E}\left[\tilde{v}_t\right]\right)^\top\right] =& \frac{4}{L}\mathbb{E}\left[\left(\int_0^t e^{-2(t-s)} dB_s\right)\left(\int_0^t e^{-2(s-t)} dB_s\right)^\top\right]\\ =& \frac{4}{L}\left(\int_0^t e^{-4(t-s)} ds\right) \cdot I_{d\times d}\\ =& \frac{1}{L}(1-e^{-4t}) \cdot I_{d\times d}.\end{aligned}$$

The Brownian motion term for $\tilde{x}_t$ is given by

$$\sqrt{\frac{4}{L}}\int_0^t \left(\int_0^r e^{-2(r-s)} dB_s\right) dr = \sqrt{\frac{4}{L}}\int_0^t e^{2s}\left(\int_s^t e^{-2r} dr\right) dB_s = \sqrt{\frac{1}{L}}\int_0^t \left(1 - e^{-2(t-s)}\right) dB_s.$$

Here the second equality follows by Fubini's theorem. The conditional covariance for $\tilde{x}_t$ now follows as

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{x}_t - \mathbb{E}\left[\tilde{x}_t\right]\right)\left(\tilde{x}_t - \mathbb{E}\left[\tilde{x}_t\right]\right)^\top\right] =& \frac{1}{L}\mathbb{E}\left[\left(\int_0^t \left(1-e^{-2(t-s)}\right) dB_s\right)\left(\int_0^t \left(1-e^{-2(t-s)}\right) dB_s\right)^\top\right]\\ =& \frac{1}{L}\left[\int_0^t \left(1-e^{-2(t-s)}\right)^2 ds\right] \cdot I_{d\times d}\\ =& \frac{1}{L}\left[t - \frac{1}{4}e^{-4t} - \frac{3}{4} + e^{-2t}\right] \cdot I_{d\times d}.\end{aligned}$$

Finally we compute the cross-covariance between $\tilde{x}_t$ and $\tilde{v}_t$,

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{x}_t - \mathbb{E}\left[\tilde{x}_t\right]\right)\left(\tilde{v}_t - \mathbb{E}\left[\tilde{v}_t\right]\right)^\top\right] =& \frac{2}{L}\mathbb{E}\left[\left(\int_0^t \left(1-e^{-2(t-s)}\right) dB_s\right)\left(\int_0^t e^{-2(t-s)} dB_s\right)^\top\right]\\ =& \frac{2}{L}\left[\int_0^t (1-e^{-2(t-s)})(e^{-2(t-s)}) ds\right] \cdot I_{d\times d}\\ =& \frac{1}{2L}\left[1 + e^{-4t} - 2e^{-2t}\right] \cdot I_{d\times d}.\end{aligned}$$

We thus have an explicitly defined Gaussian. Notice that we can sample from this distribution in time linear in d, since all d coordinates are independent. ∎

Appendix B. Controlling the Kinetic Energy

In this section, we establish an explicit bound on the kinetic energy $\mathcal{E}_K$ in (8) which is used to control the discretization error at each step.

Lemma 12 (Kinetic Energy Bound) *Let $p^{(0)}(x, v) = 1_{x=x^{(0)}} \cdot 1_{v=0}$— the Dirac delta distribution at $(x^{(0)}, 0)$. Let the initial distance from the optimum satisfy $\|x^{(0)} - x^*\|_2^2 \leq \mathcal{D}^2$ and $u = 1/L$ as before. Further let $p^{(i)}$ be defined as in Theorem 1 for $i = 1, \ldots n$, with step size δ and number* *of iterations n as stated in Theorem 1. Then for all $i = 1, \ldots n$ and for all $t \in [0, \delta]$, we have the bound*

$$\mathbb{E}_{(x,v)\sim \Phi_t p^{(i)}} \left[\|v\|_2^2 \right] \leq \mathcal{E}_K,$$

with $\mathcal{E}_K = 26(d/m + \mathcal{D}^2)$.

Proof We first establish an inequality that provides an upper bound on the kinetic energy for any distribution p.

Step 1: Let p be any distribution over (x, v), and let q be the corresponding distribution over $(x, x + v)$. Let (x', v') be random variables with distribution p^*. Further let $\zeta \in \Gamma_{opt}(p, p^*)$ such that,

$$\mathbb{E}_\zeta \left[\|x - x'\|_2^2 + \|(x - x') + (v - v')\|_2^2 \right] = W_2^2(q, q^*).$$

Then we have,

$$\begin{aligned}
\mathbb{E}_p \left[\|v\|_2^2 \right] &= \mathbb{E}_\zeta \left[\|v - v' + v'\|_2^2 \right] \\
&\leq 2\mathbb{E}_{p^*} \left[\|v\|_2^2 \right] + 2\mathbb{E}_\zeta \left[\|v - v'\|_2^2 \right] \\
&\leq 2\mathbb{E}_{p^*} \left[\|v\|_2^2 \right] + 4\mathbb{E}_\zeta \left[\|x + v - (x' + v')\|_2^2 + \|x - x'\|_2^2 \right] \\
&= 2\mathbb{E}_{p^*} \left[\|v\|_2^2 \right] + 4W_2^2(q, q^*),
\end{aligned} \tag{16}$$

where for the second and the third inequality we have used Young's inequality, while the final line follows by optimality of ζ.

Step 2: We know that $p^* \propto \exp(-(f(x) + \frac{L}{2}\|v\|_2^2))$, so we have $\mathbb{E}_{p^*} \left[\|v\|_2^2 \right] = d/L$.

Step 3: For our initial distribution $p^{(0)}(q^{(0)})$ we have the bound

$$W_2^2(q^{(0)}, q^*) \leq 2\mathbb{E}_{p^*} \left[\|v\|_2^2 \right] + 2\mathbb{E}_{x\sim p^{(0)}, x'\sim p^*} \left[\|x - x'\|_2^2 \right] = \frac{2d}{L} + 2\mathbb{E}_{p^*} \left[\|x - x^{(0)}\|_2^2 \right],$$

where the first inequality is an application of Young's inequality. The second term is bounded below,

$$\mathbb{E}_{p^*} \left[\|x - x^{(0)}\|_2^2 \right] \leq 2\mathbb{E}_{p^*} \left[\|x - x^*\|_2^2 \right] + 2\|x^{(0)} - x^*\|_2^2 \leq \frac{2d}{m} + 2\mathcal{D}^2,$$

where the first inequality is again by Young's inequality. The second line follows by applying Theorem 17 to control $\mathbb{E}_{p^*} \left[\|x - x^*\|_2^2 \right]$. Combining these we have the bound,

$$W_2^2(q^{(0)}, q^*) \leq 2d \left(\frac{1}{L} + \frac{2}{m} \right) + 4\mathcal{D}^2.$$

Putting all this together along with (16) we have

$$\mathbb{E}_{p^{(0)}} \left[\|v\|_2^2 \right] \leq \frac{10d}{L} + \frac{16d}{m} + 16\mathcal{D}^2 \leq 26 \left(\frac{d}{m} + \mathcal{D}^2 \right).$$

Step 4: By Theorem 5, we know that $\forall t > 0$,

$$W_2^2(\Phi_t q^{(i)}, q^*) \leq W_2^2(q^{(i)}, q^*).$$

This proves the theorem statement for $i = 0$. We will now prove it for $i > 0$ via induction. We have proved it for the base case $i = 0$, let us assume that the result holds for some $i > 0$. Then by (11) (along with our choice of step-size δ) applied up to the $(i+1)^{th}$ iteration, we know that

$$W_2^2(q^{(i+1)}, q^*) = W_2^2(\tilde{\Phi}_\delta q^{(i)}, q^*) \leq W_2^2(q^{(i)}, q^*).$$

Thus by (16) we have,

$$\mathbb{E}_{\Phi_t p^{(i)}} \left[\|v\|_2^2\right] \leq \mathcal{E}_K,$$

for all $t > 0$ and $i \in \{0, 1, \ldots, n\}$. ■

Next we prove that the distance of the initial distribution $p^{(0)}$ to the optimum distribution p^* is bounded.

Lemma 13 *Let $p^{(0)}(x, v) = 1_{x=x^{(0)}} \cdot 1_{v=0}$— the Dirac delta distribution at $(x^{(0)}, 0)$. Let the initial distance from the optimum satisfy $\|x^{(0)} - x^*\|_2^2 \leq \mathcal{D}^2$ and $u = 1/L$ as before. Then*

$$W_2^2(p^{(0)}, p^*) \leq 3\left(\mathcal{D}^2 + \frac{d}{m}\right).$$

Proof As $p^{(0)}(x, v)$ is a delta distribution, there is only one valid coupling between $p^{(0)}$ and p^*. Thus we have

$$\begin{aligned} W_2^2(p^{(0)}, p^*) = \mathbb{E}_{(x,v)\sim p^*}\left[\|x - x^{(0)}\|_2^2 + \|v\|_2^2\right] &= \mathbb{E}_{(x,v)\sim p^*}\left[\|x - x^* + x^* - x^{(0)}\|_2^2 + \|v\|_2^2\right] \\ &\leq 2\mathbb{E}_{x\sim p^*(x)}\left[\|x - x^*\|_2^2\right] + 2\mathcal{D}^2 + \mathbb{E}_{v\sim p^*(v)}\left[\|v\|_2^2\right], \end{aligned}$$

where the final inequality follows by Young's inequality and by the definition of $\mathcal{D}^2$. Note that $p^*(v) \propto \exp(-L\|v\|_2^2/2)$, therefore $\mathbb{E}_{v\sim p^*(v)}\left[\|v\|_2^2\right] = d/L$. By invoking Theorem 17 the first term $\mathbb{E}_{x\sim p^*(x)}\left[\|x - x^*\|_2^2\right]$ is bounded by d/m. Putting this together we have,

$$W_2^2(p^{(0)}, p^*) \leq 2\frac{d}{m} + \frac{d}{L} + 2\mathcal{D}^2 \leq 3\left(\frac{d}{m} + \mathcal{D}^2\right).$$

■

Appendix C. Varying Step Size

Here we provide a sharper analysis of underdamped Langevin MCMC by using a varying step size. By choosing an adaptive step size we are able to shave off the log factor appearing in Theorem 1.

Theorem 14 *Let the initial distribution $p^{(0)}(x, v) = 1_{x=x^{(0)}} \cdot 1_{v=0}$ and let the initial distance to optimum satisfy $\|x^{(0)} - x^*\|_2^2 \leq \mathcal{D}^2$. Also let $W_2(p^{(0)}, p^*) \leq 3\left(\frac{d}{m} + \mathcal{D}^2\right) < \epsilon_0$. We set the initial step size to be*

$$\delta_1 = \frac{\epsilon_0}{2 \cdot 104\kappa}\sqrt{\frac{1}{d/m + \mathcal{D}^2}},$$

and initial number of iterations,

$$n_1 = \frac{208\kappa^2}{\epsilon_0} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right) \cdot \log(16).$$

We define a sequence of ℓ epochs with step sizes $(\delta_1, \ldots, \delta_\ell)$ and number of iterations $(n_1, \ldots, n_\ell)$ where δ_1 and n_1 are defined as above. Choose $\ell = \lceil \log(\epsilon^0/\varepsilon)/\log(2) \rceil$ and, for $i \geq 1$ set $\delta_{i+1} = \delta_i/2$ and $n_{i+1} = 2n_i$.

We run ℓ epochs of underdamped Langevin MCMC (Algorithm 2.1) with step size sequence $(\delta_1, \delta_2, \ldots, \delta_\ell)$ with number of iterations $(n_1, n_2, \ldots, n_\ell)$ corresponding to each step size. Then we have the guarantee

$$W_2(p^{(n)}, p^*) \leq \varepsilon,$$

with total number of steps $n = n_1 + n_2 + \ldots + n_\ell$ being

$$n = \frac{416 \log(16) \kappa^2}{\varepsilon} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right).$$

Proof Let the initial error in the probability distribution be $W_2(p^{(0)}, p^*) = \epsilon_0$. Then by the results of Theorem 1 if we choose the step size to be

$$\delta_1 = \frac{\epsilon_0}{2 \cdot 104\kappa} \sqrt{\frac{1}{d/m + \mathcal{D}^2}},$$

then we have the guarantee that in

$$n_1 = \frac{208\kappa^2}{\epsilon_0} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right) \cdot \log(16)$$

steps the error will be less than $\epsilon_1 = \epsilon_0/2$. At this point we half the step size $\delta_2 = \delta_1/2$ and run for $n_2 = 2n_1$ steps. After that we set $\delta_3 = \delta_2/2$ and run for double the steps $n_3 = 2n_2$ and so on. We repeat this for ℓ steps. Then at the end if the probability distribution is $p^{(n)}$ by Theorem 1 we have the guarantee that $W_2(p^{(n)}, p^*) \leq \epsilon_0/2^\ell < \varepsilon$. The total number of steps taken is

$$\begin{aligned}
n_1 + n_2 \ldots + n_\ell &= \sum_{i=1}^{\ell} n_i \\
&= \frac{208\kappa^2}{\epsilon_0} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right) \cdot \log(16) \left\{ \sum_{i=0}^{\ell-1} 2^i \right\} \\
&= 104 \log(16) \kappa^2 \cdot \frac{2^\ell}{\epsilon_0} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right) \left\{ \sum_{i=0}^{\ell-1} 2^{-i} \right\} \\
&\leq 104 \log(16) \kappa^2 \cdot \frac{2}{\varepsilon} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right) \{2\} \\
&= \frac{416 \log(16) \kappa^2}{\varepsilon} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2} \right),
\end{aligned}$$

Algorithm 2: Stochastic Gradient Underdamped Langevin MCMC

Input : Step size $\delta < 1$, number of iterations n, initial point $(x^{(0)}, 0)$, smoothness parameter L and stochastic gradient oracle $\hat{\nabla} f(\cdot)$

for $i = 0, 1, \ldots, n-1$ **do**

  Sample $(x^{i+1}, v^{i+1}) \sim Z^{i+1}(x^i, v^i)$

end

where the inequality follows by the choice of ℓ and an upper bound on the sum of the geometric series. ■

Appendix D. Analysis with Stochastic Gradients

Here we state the underdamped Langevin MCMC algorithm with stochastic gradients. We will borrow notation and work under the assumptions stated in Section 2.2.1.

DESCRIPTION OF ALGORITHM D

The random vector $Z^{i+1}(x_i, v_i) \in \mathbb{R}^{2d}$, conditioned on (x^i, v^i), has a Gaussian distribution with conditional mean and covariance obtained from the following computations:

$$
\begin{aligned}
&\mathbb{E}\left[v^{i+1}\right] = v^i e^{-2\delta} - \frac{1}{2L}(1 - e^{-2\delta})\hat{\nabla} f(x^i) \\
&\mathbb{E}\left[x^{i+1}\right] = x^i + \frac{1}{2}(1 - e^{-2\delta})v^i - \frac{1}{2L}\left(\delta - \frac{1}{2}\left(1 - e^{-2\delta}\right)\right)\hat{\nabla} f(x^i) \\
&\mathbb{E}\left[\left(x^{i+1} - \mathbb{E}\left[x^{i+1}\right]\right)\left(x^{i+1} - \mathbb{E}\left[x^{i+1}\right]\right)^\top\right] = \frac{1}{L}\left[\delta - \frac{1}{4}e^{-4\delta} - \frac{3}{4} + e^{-2\delta}\right] \cdot I_{d\times d} \\
&\mathbb{E}\left[\left(v^{i+1} - \mathbb{E}\left[v^{i+1}\right]\right)\left(v^{i+1} - \mathbb{E}\left[v^{i+1}\right]\right)^\top\right] = \frac{1}{L}(1 - e^{-4\delta}) \cdot I_{d\times d} \\
&\mathbb{E}\left[\left(x^{i+1} - \mathbb{E}\left[x^{i+1}\right]\right)\left(v^{i+1} - \mathbb{E}\left[v^{i+1}\right]\right)^\top\right] = \frac{1}{2L}\left[1 + e^{-4\delta} - 2e^{-2\delta}\right] \cdot I_{d\times d}.
\end{aligned}
$$

The distribution is obtained by integrating the discrete underdamped Langevin diffusion (4) up to time δ, with the specific choice of $\gamma = 2$ and $u = 1/L$. In other words, if $p^{(i)}$ is the distribution of (x^i, v^i), then $Z^{i+1}(x^i, v^i) \sim p^{(i+1)} = \hat{\Phi}_\delta p^{(i)}$. Derivation is identical to the calculation in Appendix A by replacing exact gradients $\nabla f(\cdot)$ with stochastic gradients $\hat{\nabla} f(\cdot)$. A key ingredient as before in understanding these updates is the next lemma which calculates the exactly the update at each step when we are given stochastic gradients.

Lemma 15 *The solution $(\hat{x}_t, \hat{v}_t)$ of the stochastic gradient underdamped Langevin diffusion (4) is*

$$
\begin{aligned}
\hat{v}_t &= \hat{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\hat{\nabla} f(\hat{x}_0)ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s \qquad (17)\\
\hat{x}_t &= \hat{x}_0 + \int_0^t \hat{v}_s ds.
\end{aligned}
$$

Proof Note that they have the right initial values, by setting $t = 0$. By taking derivatives, one can also verify that they satisfy the differential equation (4). ■

D.1. Discretization Analysis

In Theorem 16, we will bound the discretization error between the discrete process without noise in the gradients (3) and the discrete process (4) starting from the same initial distribution.

Lemma 16 *Let q_0 be some initial distribution. Let $\tilde{\Phi}_\delta$ and $\hat{\Phi}_\delta$ be as defined in (2) corresponding to the discrete time process without noisy gradients and discrete-time process with noisy gradients respectively. For any $1 > \delta > 0$,*

$$W_2^2(\hat{\Phi}_\delta q_0, q^*) = W_2^2(\tilde{\Phi}_\delta q_0, q^*) + \frac{5\delta^2 d\sigma^2}{L^2}.$$

Proof Taking the difference of the dynamics in (13) and (17), and using the definition of $\hat{\nabla} f(x)$. We get that

$$\begin{aligned}\hat{v}_\delta &= \tilde{v}_\delta + u\left(\int_0^\delta e^{-\gamma(s-\delta)}ds\right)\xi \\ \hat{x}_\delta &= \tilde{x}_\delta + u\left(\int_0^\delta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi,\end{aligned} \tag{18}$$

where ξ is a zero-mean random variance with variance bounded by $\sigma^2 d$ and is independent of the Brownian motion. Let Γ_1 be the set of all couplings between $\tilde{\Phi}_\delta q_0$ and q^* and let Γ_2 be the set of all couplings between $\hat{\Phi}_\delta q_0$ and q^*. Let $\gamma_1(\theta, \psi) \in \Gamma_1$ be the optimal coupling between $\tilde{\Phi}_\delta q_0$ and q^*, i.e.

$$\mathbb{E}_{(\theta,\psi)\sim\gamma_1}\left[\|\theta - \psi\|_2^2\right] = W_2^2(\tilde{\Phi}_\delta q_0, q^*).$$

Let $\left(\begin{bmatrix}\tilde{x}\\ \tilde{w}\end{bmatrix}, \begin{bmatrix}x\\ w\end{bmatrix}\right) \sim \gamma_1$. By the definition of γ_1 we have the marginal distribution of $\begin{bmatrix}\tilde{x}\\ \tilde{w}\end{bmatrix} \sim \tilde{\Phi}_\delta q_0$. Finally let us define the random variables

$$\begin{bmatrix}\hat{x}\\ \hat{w}\end{bmatrix} \triangleq \begin{bmatrix}\tilde{x}\\ \tilde{w}\end{bmatrix} + u\begin{bmatrix}\left(\int_0^\delta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi \\ \left(\int_0^\delta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\delta e^{-\gamma(s-\delta)}ds\right)\xi\end{bmatrix}.$$

By (18), it follows that $\begin{bmatrix} \hat{x} \\ \hat{w} \end{bmatrix} \sim \hat{\Phi}_\delta p_0$. Thus $\left(\begin{bmatrix} \hat{x} \\ \hat{w} \end{bmatrix}, \begin{bmatrix} x \\ w \end{bmatrix} \right)$ defines a valid coupling between $\hat{\Phi}_t q_0$ and q^*. Let us now analyze the distance between q^* and $\hat{\nabla}_\delta q_0$,

$$
\begin{aligned}
&W_2^2(\hat{\Phi}_\delta q_0, q^*) \\
&\overset{(i)}{\leq} \mathbb{E}_{\gamma_1} \left[\left\| \begin{bmatrix} \tilde{x} \\ \tilde{v} \end{bmatrix} + u \begin{bmatrix} \left(\int_0^\delta \left(\int_0^r e^{-\gamma(s-r)} ds\right) dr\right) \xi \\ \left(\int_0^\delta \left(\int_0^r e^{-\gamma(s-r)} ds\right) dr + \int_0^\delta e^{-\gamma(s-\delta)} ds\right) \xi \end{bmatrix} - \begin{bmatrix} x \\ v \end{bmatrix} \right\|_2^2 \right] \\
&\overset{(ii)}{=} \mathbb{E}_{\gamma_1} \left[\left\| \begin{bmatrix} \tilde{x} \\ \tilde{v} \end{bmatrix} - \begin{bmatrix} x \\ v \end{bmatrix} \right\|_2^2 \right] + u \cdot \mathbb{E}_{\gamma_1} \left[\left\| \begin{bmatrix} \left(\int_0^\delta \left(\int_0^r e^{-\gamma(s-r)} ds\right) dr\right) \xi \\ \left(\int_0^\delta \left(\int_0^r e^{-\gamma(s-r)} ds\right) dr + \int_0^\delta e^{-\gamma(s-\delta)} ds\right) \xi \end{bmatrix} \right\|_2^2 \right] \\
&\overset{(iii)}{\leq} \mathbb{E}_{\gamma_1} \left[\left\| \begin{bmatrix} \tilde{x} \\ \tilde{v} \end{bmatrix} - \begin{bmatrix} x \\ v \end{bmatrix} \right\|_2^2 \right] + 4u^2 \left(\left(\int_0^\delta \left(\int_0^r e^{-\gamma(s-r)} ds \right) dr \right)^2 + \left(\int_0^\delta e^{-\gamma(s-\delta)} ds \right)^2 \right) d\sigma^2 \\
&\overset{(iv)}{\leq} \mathbb{E}_{\gamma_1} \left[\left\| \begin{bmatrix} \tilde{x} \\ \tilde{v} \end{bmatrix} - \begin{bmatrix} x \\ v \end{bmatrix} \right\|_2^2 \right] + 4u^2 \left(\frac{\delta^4}{4} + \delta^2 \right) d\sigma^2 \\
&\overset{(v)}{\leq} W_2^2(\tilde{\Phi}_t q_0, q^*) + 5u^2\delta^2 d\sigma^2,
\end{aligned}
$$

where (i) is by definition of W_2, (ii) is by independence and unbiasedness of ξ, (iii) is by Young's inequality and because $\mathbb{E}\left[\|\xi\|_2^2\right] \leq d\sigma^2$, (iv) uses the upper bound $e^{-\gamma(s-r)} \leq 1$ and $e^{-\gamma(s-t)} \leq 1$, and finally (v) is by definition of γ_1 being the optimal coupling and the fact that $\delta \leq 1$. The choice of $u = 1/L$ yields the claim.

■

Given the bound on the discretization error between the discrete processes with and without the stochastic gradient we are now ready to prove Theorem 3.

Proof [Proof of Theorem 3] From Corollary 7, we have that for any $i \in \{1, \ldots, n\}$

$$W_2(\Phi_\delta q^{(i)}, q^*) \leq e^{-\delta/2\kappa} W_2(q^{(i)}, q^*).$$

By the discretization error bound in Theorem 9 and the sandwich inequality (6), we get

$$W_2(\Phi_\delta q^{(i)}, \tilde{\Phi}_\delta q^{(i)}) \leq 2W_2(\Phi_\delta p^{(i)}, \tilde{\Phi}_\delta p^{(i)}) \leq \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}}.$$

By the triangle inequality for W_2,

$$W_2(\tilde{\Phi}_\delta q^{(i)}, q^*) \leq W_2(\Phi_\delta q^{(i)}, \tilde{\Phi}_\delta q^{(i)}) + W_2(\Phi_\delta q^{(i)}, q^*) \overset{(i)}{\leq} \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + e^{-\delta/2\kappa} W_2(q^{(i)}, q^*)$$

Combining this with the discretization error bound established in Lemma 16 we have,

$$W_2^2(\hat{\Phi}_t q^{(i)}, q^*) \leq \left(e^{-\delta/2\kappa} W_2(q^{(i)}, q^*) + \delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} \right)^2 + \frac{5\delta^2 d\sigma^2}{L^2}.$$

By invoking Lemma 19 we can bound the value of this recursive sequence by,

$$W_2(q^{(n)}, q^*) \leq e^{-n\delta/2\kappa} W_2(q^{(0)}, q^*) + \frac{\delta^2}{1 - e^{-\delta/2\kappa}} \sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{5\delta^2 d\sigma^2}{L^2 \left(\delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{1 - e^{-\delta/\kappa}} \sqrt{\frac{5\delta^2 d\sigma^2}{L^2}}\right)}.$$

By using the sandwich inequality (Lemma 8) we get,

$$W_2(p^{(n)}, p^*) \leq \underbrace{4e^{-n\delta/2\kappa} W_2(p^{(0)}, p^*)}_{T_1} + \underbrace{\frac{4\delta^2}{1 - e^{-\delta/2\kappa}} \sqrt{\frac{8\mathcal{E}_K}{5}}}_{T_2} + \underbrace{\frac{20\delta^2 d\sigma^2}{L^2 \left(\delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{1 - e^{-\delta/\kappa}} \sqrt{\frac{5\delta^2 d\sigma^2}{L^2}}\right)}}_{T_3}.$$

We will now control each of these terms at a level $\varepsilon/3$. By Lemma 13 we know $W_2^2(p^{(0)}, p^*) \leq 3\left(\frac{d}{m} + \mathcal{D}^2\right)$. So the choice,

$$n \leq \frac{2\kappa}{\delta} \log\left(\frac{36\left(\frac{d}{m} + \mathcal{D}^2\right)}{\varepsilon}\right)$$

ensures that T_1 is controlled below the level $\varepsilon/3$. Note that $1 - e^{-\delta/2\kappa} \geq \delta/4\kappa$ as $\delta/\kappa < 1$. So the choice $\delta < \varepsilon\kappa^{-1}\sqrt{5/479232(d/m + \mathcal{D}^2)} \leq \varepsilon\kappa^{-1}\sqrt{5/18432\mathcal{E}_K}$ (by upper bound on $\mathcal{E}_K$ in Lemma 12) ensures,

$$T_2 \leq \frac{16\delta^2\kappa}{\delta} \sqrt{\frac{8\mathcal{E}_K}{5}} \leq \frac{\varepsilon}{3}.$$

Finally $\delta \leq \varepsilon^2\kappa^{-1}L^2/1440d\sigma^2$ ensures T_3 is bounded,

$$T_3 = \frac{20\delta^2 d\sigma^2}{L^2 \left(\delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{1 - e^{-\delta/\kappa}} \sqrt{\frac{5\delta^2 d\sigma^2}{L^2}}\right)} \leq \frac{20\delta^2 d\sigma^2}{L^2 \left(\delta^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{\frac{5\delta^3 d\sigma^2}{2L^2\kappa}}\right)} \leq \frac{20\delta^2 d\sigma^2}{L^2 \sqrt{\frac{5\delta^3 d\sigma^2}{2L^2\kappa}}} \leq \frac{\varepsilon}{3}.$$

This establishes our claim. ■

Appendix E. Technical Results

We state this Theorem by Durmus and Moulines (2016) used in the proof of Lemma 12.

Theorem 17 *(Theorem 1 in Durmus and Moulines, 2016) For all $t \geq 0$ and $x \in \mathbb{R}^d$,*

$$\mathbb{E}_{p^*}\left[\|x - x^*\|_2^2\right] \leq \frac{d}{m}.$$

The following lemma is a standard result in linear algebra regarding the determinant of a block matrix. We apply this result in the proof of Theorem 5.

Lemma 18 *(Theorem 3 in Silvester, 2000) If A, B, C and D are square matrices of dimension d, and C and D commute, then we have*

$$\det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \det(AD - BC).$$

We finally present a useful lemma from (Dalalyan and Karagulyan, 2017) that we will use in the proof of Theorem 3.

Lemma 19 *(Lemma 7 in Dalalyan and Karagulyan, 2017) Let A, B and C be given non-negative numbers such that $A \in \{0, 1\}$. Assume that the sequence of non-negative numbers $\{x_k\}_{k\in\mathbb{N}}$ satisfies the recursive inequality*

$$x_{k+1}^2 \leq [(A)x_k + C]^2 + B^2$$

for every integer $k \geq 0$. Then

$$x_k \leq A^k x_0 + \frac{C}{1-A} + \frac{B^2}{C + \sqrt{(1-A^2)}B} \tag{19}$$

for all integers $k \geq 0$.