The Interplay between Sampling and Optimization

Cheng Xiang

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-145
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-145.html

August 12, 2020

Acknowledgement

To Professor Peter Bartlett and Professor Michael Jordan, I could not have asked for better advisors. Thank you for your mentorship throughout my PhD, for sharing your wisdom on research and on life, and for your unconditional support in all my pursuits over the years. I hope that I could one day inspire others as you have inspired me.

To Professor Satish Rao, thank you for your guidance at the very start of my journey; I would have been very lost if you were not there to show me the way.

To my collaborators, Dong, Niladri, Yian, and Yasin, thank you for the great fun we had, and for all the things you taught me.

To all my friends at Berkeley, thank you for making this a wonderful experience.

To my father, my mother, and my fiancee, I could not have done this without you.

The Interplay between Sampling and Optimization

by

Xiang Cheng

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter Bartlett, Chair
Professor Michael Jordan
Professor Steven Evans

Summer 2020

The dissertation of Xiang Cheng, titled The Interplay between Sampling and Optimization, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

The Interplay between Sampling and Optimization

Abstract

The Interplay between Sampling and Optimization

by

Xiang Cheng

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Peter Bartlett, Chair

We study the connections between optimization and sampling. In one direction, we study sampling algorithms from an optimization perspective. We will see how the Langevin MCMC algorithm can be viewed as a deterministic gradient descent in probability space, which enables us to do convergence analysis in KL divergence. We will also see how adding a momentum term improves the convergence rate of Langevin MCMC, much like acceleration in gradient descent. Finally, we will study the problem of sampling from non-logconcave distributions, which is roughly analogous to non-convex optimization.

Conversely, we will also study optimization algorithms from a sampling perspective. We will approximate stochastic gradient descent by a Langevin-like stochastic differential equation, and use this to explain some of its remarkable generalization properties.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

To Professor Peter Bartlett and Professor Michael Jordan, I could not have asked for better advisors. Thank you for your mentorship throughout my PhD, for sharing your wisdom on research and on life, and for your unconditional support in all my pursuits over the years. I hope that I could one day inspire others as you have inspired me.

To Professor Satish Rao, thank you for your guidance at the very start of my journey; I would have been very lost if you were not there to show me the way.

To my collaborators, Dong, Niladri, Yian, and Yasin, thank you for the great fun we had, and for all the things you taught me.

To all my friends at Berkeley, thank you for making this a wonderful experience.

To my father, my mother, and my fiancee, I could not have done this without you.

Part I

Introduction

Chapter 1

Preliminaries

1.1 Convexity and Smoothness

We say that a twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}^d$ is m strongly convex if

$$\forall x \qquad \nabla^2 f(x) \succ mI \tag{1.1}$$

A very useful consequence of (1.1) is that for all x, y, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m\|x - y\|_2^2$, which ensures that the gradient flow is contractive.

We say that $f : \mathbb{R}^d \to \mathbb{R}$ has L-Lipschitz gradients, or equivalently L-smooth, if

$$\forall x, y \qquad \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \tag{1.2}$$

If $f(x)$ is twice differentiable, the above is equivalent to $\nabla^2 f(x) \prec LI$ for all x.

1.2 Sampling

We will be concerned with the problem of sampling from a distribution

$$p^*(x) \propto e^{-U(x)}, \tag{1.3}$$

where $x \in \mathbb{R}^d$. We will refer to $U(x)$ as the *potential function*, and we will assume that we can compute $\nabla U(x)$ for all x.

We will mostly focus on the problem of sampling from $p^*(x)$ when $U(x)$ is m strongly convex and has L Lipschitz Gradients, given access to the gradient oracle of U, though in certain parts of this thesis, we will relax the convexity/regularity assumptions on $U(x)$.

1.3 The Langevin Dynamics

The Langevin Diffusion, with respect to the potential $U(x)$, is given by the following stochastic differential equation:

$$dx_t = -\nabla U(x_t)dt + \sqrt{2}dB_t, \tag{1.4}$$

where B_t is the standard Brownian motion. The invariant distribution of (1.4) is $p^*(x) \propto e^{-U(x)}$. For the purposes of this thesis, we can assume that $U(x)$ is smooth, and that $e^{-U(x)}$ is integrable.

To convert (1.4) to a computationally tractable algorithm, a common approach is to use the Euler-Murayama discretization scheme:

$$x_{k+1} = x_k - \delta \nabla U(x_k) + \sqrt{2\delta}\xi_k, \tag{1.5}$$

where δ is the step-size and $\xi_k \sim \mathcal{N}(0, I)$.

1.4 The Wasserstein Metric

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 by $\Gamma(\mu, \nu)$ 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.)

The k-Wasserstein distance between μ and ν is given by

$$W_k(\mu, \nu) = \inf_{\zeta \in \Gamma(\mu,\nu)} \left(\mathbb{E}_{(x,y)\sim\zeta} \left[\|x - y\|_2^k \right] \right)^{1/k}.$$

In this thesis, we will mainly be concerned with W_1 and W_2.

1.5 Optimization

Given an optimization objective $f(x) : \Xi \to \mathbb{R}$, the goal is to find

$$x^* \in \arg\min_{\Xi} f(x). \tag{1.6}$$

Very commonly, Ξ is a subset of $\mathbb{R}^d$. In this thesis, we will also consider the specific case of $\Xi = \mathscr{P}(\mathbb{R}^d)$, the space of densities over $\mathbb{R}^d$, metrized by the 2-Wasserstein distance.

1.6 Overview

The goal of this thesis is to explore various connections between sampling and optimization. We will see how optimization ideas can help us analyze and design better sampling algorithms. Conversely, stochastic optimization algorithms such as SGD can be better understood by analogy with sampling procedures. This thesis is organized as follows:

- In Chapter 2, we show how the Langevin MCMC algorithm (1.5) is equivalent to a deterministic gradient descent algorithm over $\mathscr{P}(\mathbb{R}^d)$. This allows us to bound the distance between (1.5) and (1.4) in terms of KL divergence. Our analysis relies largely on the view of Langevin diffusion as the gradient flow of KL divergence with respect to the Wasserstein distance, shown in [45].

- In Chapter 3, we study a different SDE derived from adding momentum to (1.4). The SDE, given in (3.1), is often known as the *underdamped Langevin diffusion*. We give a MCMC algorithm, based on discretizing (3.1), and show that the convergence rate is quadratically faster in dimension d and target accuracy ε. This curiously mirrors the acceleration phenomenon in optimization.

- In Chapter 4, we study the problem of sampling from a distribution $p^*(x) \propto e^{-U(x)}$ when the potential $U(x)$ is non-convex. This is roughly analogous to optimizing a non-convex objective. We show that the standard Langevin MCMC algorithm converges to $p^*(x)$, but the number of steps required is exponential in a quantity that measures how non-convex $U(x)$ is. We also show that underdamped Langevin MCMC achieves a quadratic speed-up even in this non-convex setting.

- In Chapter 5, we instead try to analyze stochastic gradient descent, an optimization algorithm, from a sampling point of view. We are motivated by the empirical observation that SGD solutions sometimes generalize better than gradient descent solutions. We show that SGD can be viewed as the discrete-time approximation of a SDE with a state-dependent diffusion coefficient. This view allows us to characterize the distribution of SGD solutions, in spite of the irregularity of SGD noise.

Part II

Sampling as Optimization

Chapter 2

Langevin MCMC as Gradient Descent over $\mathscr{P}(\mathbb{R}^d)$

2.1 Introduction

In this chapter, we study the problem of sampling from

$$\mathbf{p}^*(x) \propto e^{-U(x)},$$

where $U(x)$ is strongly convex (see 1.1).

We will study the MCMC algorithm given in (1.5), reproduced below for ease of reference:

$$u^{i+1} = u^i - h \cdot \nabla U(u^i) + \sqrt{2h}\xi^i, \tag{2.1}$$

where h is a step-size, and $\xi^i \overset{iid}{\sim} N(0,1)$.

Recall that (2.1) is the Euler-Murayama discretization of the Langevin SDE:

$$d\bar{x}_t = -\nabla U(\bar{x}_t)dt + \sqrt{2}dB_t, \tag{2.2}$$

where B_t is the standard Brownian motion.

We verify that (2.1) is equivalent to the following **Discretized Langevin SDE**:

$$dx_t = -\nabla U(x_{\tau(t)})dt + \sqrt{2}dB_t, \tag{2.3}$$

where $\tau(t) \triangleq \lfloor \frac{t}{h} \rfloor \cdot h$ (note that $\tau(t)$ is parameterized by h). Note that the difference between (2.2) and (2.3) is in the drift term: one is $\nabla U(\bar{x}_t)$, the other is $\nabla U(x_{\tau(t)})$

Let $\mathbf{p}_t$ denote the distribution of x_t. Our main goal is to establish the convergence of $\mathbf{p}_t$ in (2.3) in $KL(\mathbf{p}_t \| \mathbf{p}^*)$. KL-divergence is perhaps the most natural notion of distance between probability distributions in this context, because of its close relationship to maximum likelihood estimation, its interpretation as information gain in Bayesian statistics, and its central role in information theory. Convergence in KL-divergence implies convergence in total variation and 2-Wasserstein distance, thus we are able to obtain convergence rates in total variation and 2-Wasserstein that are comparable to the results shown in [20, 27, 29].

2.2 Assumptions and Definitions

We denote by $\mathscr{P}(\mathbb{R}^d)$ the space of all probability distributions over $\mathbb{R}^d$. In this chapter, only distributions with densities wrt the Lebesgue measure will appear (see Lemma 20), both in the algorithm and in the analysis. With abuse of notation, we use the same symbol (e.g. $\mathbf{p}$) to denote both the probability distribution and its density wrt the Lebesgue measure.

For the rest of this chapter, we will use $\mathbf{p}_t$ to exclusively denote the distribution of x_t in (2.3)**.**

We assume without loss of generality that

$$\arg\min_x U(x) = 0,$$

and that

$$U(0) = 0.$$

(We can always shift the origin to achieve this, and the minimizer of U is easy to find using, say, gradient descent.)

For the rest of this chapter, we will let

$$F(\boldsymbol{\mu}) = \begin{cases} \int \boldsymbol{\mu}(x) \log\left(\frac{\boldsymbol{\mu}(x)}{\mathbf{p}^*(x)}\right) dx, & \text{if } \boldsymbol{\mu} \text{ has a density wrt} \\ & \text{Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$

be the KL-divergence between $\boldsymbol{\mu}$ and $\mathbf{p}^*$. It is known that F is minimized by $\mathbf{p}^*$, and $F(\mathbf{p}^*) = 0$.

Finally, given a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, we define the $L^2(\boldsymbol{\mu})$-norm of v as

$$\|v\|_{L^2(\boldsymbol{\mu})} \triangleq \sqrt{\mathbb{E}_{\boldsymbol{\mu}}[\|v(x)\|_2^2]}.$$

Recall that the Wasserstein distance defined in 1.4. In this chapter, we will use Wasserstein distance to exclusively refer to the 2-Wasserstein distance, given by

$$W_2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sqrt{\inf_{\gamma \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})} \int (\|x - y\|_2^2), d\gamma(x, y)}$$

where $\Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$ is the set of all couplings between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Let $(X_1, \mathscr{B}(X_1))$ and $(X_2, \mathscr{B}(X_2))$ be two measurable spaces, $\boldsymbol{\mu}$ be a measure on X_1, and $r : X_1 \to X_2$ be a measurable map. The **push-forward measure** of $\boldsymbol{\mu}$ through r is defined as

$$r_{\#}\boldsymbol{\mu}(B) = \boldsymbol{\mu}(r^{-1}(B)) \quad \forall B \in \mathscr{B}(X_2) l.$$

Intuitively, for any f, $\mathbb{E}_{r_{\#}\boldsymbol{\mu}}[f(x)] = \mathbb{E}_{\boldsymbol{\mu}}[f(r(x))]$.

It is known that for any two distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ which have density wrt the Lebesgue measure, the optimal coupling is induced by a map $T_{opt} : \mathbb{R}^d \to \mathbb{R}^d$; i.e., $W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int (\|x - y\|_2^2) d\gamma^*(x, y)$ for

$$\gamma^* = (Id, T_{opt})_{\#}\boldsymbol{\mu},$$

where Id is the identity map, and T_{opt} satisfies $T_{opt\#}\boldsymbol{\mu} = \boldsymbol{\nu}$, so by definition, $\gamma^* \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$. We call T_{opt} **the optimal transport map**, and $T_{opt} - Id$ the **optimal displacement map**.

Given two points $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ in $\mathscr{P}(\mathbb{R}^d)$, a curve $\boldsymbol{\mu}_t : [0, 1] \to \mathscr{P}(\mathbb{R}^d)$ is a **constant-speed-geodesic** between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ if $\boldsymbol{\mu}_0 = \boldsymbol{\nu}$, $\boldsymbol{\mu}_1 = \boldsymbol{\pi}$ and $W_2(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t) = (t - s)W_2(\boldsymbol{\nu}, \boldsymbol{\pi})$ for all $0 \le s \le t \le 1$. If $v_{\boldsymbol{\nu}}^{\boldsymbol{\pi}}$ is the optimal displacement map between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$, then the constant-speed-geodesic $\boldsymbol{\mu}_t$ is characterized by

$$\boldsymbol{\mu}_t = (Id + tv_{\boldsymbol{\nu}}^{\boldsymbol{\pi}})_{\#}\boldsymbol{\nu}. \tag{2.4}$$

Given a curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$, we define its **metric derivative** as

$$|\boldsymbol{\mu}_t'| \triangleq \limsup_{s \to t} \frac{W_2(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t)}{|s - t|}. \tag{2.5}$$

Intuitively, this is the speed of the curve in 2-Wasserstein distance. We say that a curve $\boldsymbol{\mu}_t$ is **absolutely continuous** if $\int_a^b |\boldsymbol{\mu}_t'|^2 dt < \infty$ for all $a, b \in \mathbb{R}$.

Given a curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$ and a sequence of velocity fields $v_t : \mathbb{R}^+ \to (\mathbb{R}^d \to \mathbb{R}^d)$, we say that $\boldsymbol{\mu}_t$ and v_t satisfy the **continuity equation** at t if

$$\frac{d}{dt}\boldsymbol{\mu}_t(x) + \mathbf{div}(\boldsymbol{\mu}_t(x) \cdot v_t(x)) = 0. \tag{2.6}$$

(We assume that $\boldsymbol{\mu}_t$ has density wrt Lebesgue measure for all t).

Remark 1 *If $\boldsymbol{\mu}_t$ is a constant-speed-geodesic between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$, then $\boldsymbol{\mu}_t$ and $v_{\boldsymbol{\nu}}^{\boldsymbol{\pi}}$ satisfy* (2.6) *at $t = 0$, by the characterization in* (2.4)*.*

We say that v_t is tangent to $\boldsymbol{\mu}_t$ at t if the continuity equation holds and $\|v_t + w\|_{L^2(\boldsymbol{\mu}_t)} \ge \|v_t\|_{L^2(\boldsymbol{\mu}_t)}$ for all w such that $\mathbf{div}(\boldsymbol{\mu}_t \cdot w) = 0$. Intuitively, v_t is tangent to $\boldsymbol{\mu}_t$ if it minimizes $\|v_t\|_{L^2(\boldsymbol{\mu}_t)}$ among all velocity fields v that satisfy the continuity equation.

2.3 Main Results

In Section 2.3.1, we state Theorem 1, which establishes a nonasymptotic convergence in Kullback-Leibler divergence for (2.3) when $U(x)$ is m strongly convex and L smooth (see (1.1) and (1.2)). As a consequence, we also unify the proof of convergence in total variation and W_2 as simple corollaries to the convergence in KL.

The following table compares the number of iterations of (2.1) required to achieve ε error in each of the three quantities according to the analysis of various papers.

In Section 2.3.2, we state Theorem 2, which establishes a convergence rate for when U is not strongly convex. The corollary for convergence in total variation has a better dependence on the dimension than the corresponding result in [20], but a worse dependence on ε.

	TV	W_2	KL
[20, 27]	$\tilde{O}(\frac{d}{\varepsilon^2})$	-	-
[29]	$\tilde{O}(\frac{d}{\varepsilon^2})$	$\tilde{O}(\frac{d}{\varepsilon^2})$	-
our result	$\tilde{O}(\frac{d}{\varepsilon^2})$	$\tilde{O}(\frac{d}{\varepsilon^2})$	$\tilde{O}(\frac{d}{\varepsilon})$

Table 2.1: Comparison of iteration complexity

2.3.1 Strong Convexity Result

In this section, we assume that $U(x)$ is m strongly convex and L smooth (see (1.1) and (1.2)).

Theorem 1 *Let $U(x)$ be m strongly convex and have L Lipschitz gradient. Let x_t and $\mathbf{p}_t$ be as defined in* (2.3) *with $\mathbf{p}_0 = N(0, \frac{1}{m})$.*

If

$$h = \frac{m\varepsilon}{16dL^2}$$

and

$$k = 16\frac{L^2}{m^2}\frac{d\log\frac{dL}{m\varepsilon}}{\varepsilon},$$

then $KL\left(\mathbf{p}_{kh}\|\mathbf{p}^*\right) \leq \varepsilon$.

This theorem immediately allows us to obtain the convergence rate of $\mathbf{p}_{kh}$ in both total variation and 2-Wasserstein distance.

Corollary 1 *Using the choice of k and h in Theorem 1, we get*

1. $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \leq \sqrt{\varepsilon}$
2. $W_2(\mathbf{p}_{kh}, \mathbf{p}^*) \leq \sqrt{\frac{2\varepsilon}{m}}$.

The first assertion follows from Pinsker's inequality [69]. The second assertion follows from (2.15), where we take $\boldsymbol{\mu}_0$ to be $\mathbf{p}^*$ and $\boldsymbol{\mu}_1$ to be $\mathbf{p}_{kh}$. To achieve δ accuracy in total variation or W_2, we apply Theorem 1 with $\varepsilon = \delta^2$ and $\varepsilon = m\delta^2$ respectively.

Remark 2 *The* log *term in Theorem 1 is not crucial. To avoid the log term, one can run* (2.1) *a few times, each time aiming to only halve the objective $KL\left(\mathbf{p}_t\|\mathbf{p}^*\right)$ (thus the stepsize starts out large and is also halved each subsequent run). The proof is straightforward and will be omitted.*

2.3.2 Weak Convexity Result

In this section, we study the case when $\log \mathbf{p}^*$ is not m strongly convex (but still convex and L smooth). Let $\boldsymbol{\pi}_h$ be the stationary distribution of (2.3) with stepsize h.

We will assume that we can choose an initial distribution $\mathbf{p}_0$ which satisfies

$$W_2(\mathbf{p}_0, \mathbf{p}^*) = C_1 \tag{2.7}$$

and

$$\sqrt{E_{\mathbf{p}^*}\|x\|_2^2} = C_2. \tag{2.8}$$

Let h' be the largest stepsize such that

$$W_2(\boldsymbol{\pi}_h, p^*) \leq C_1 \quad , \forall h \leq h'. \tag{2.9}$$

Theorem 2 *Let x_t and $\mathbf{p}_t$ be as defined in (2.3) with $\mathbf{p}_0$ satisfying (2.7). If*

$$\begin{aligned} h &= \frac{1}{48}\min\left\{\frac{\varepsilon}{C_1(C_1+C_2)L^2}, \frac{\varepsilon^2}{C_1^2 dL^2}, h'\right\} \\ &= \frac{1}{48}\min\left\{\frac{\varepsilon}{C_1 C_2 L^2}, \frac{\varepsilon^2}{C_1^2 dL^2}, h'\right\} \end{aligned}$$

and

$$k = \frac{2C_1^2}{\varepsilon h} + \frac{2C_1^2 \log(F(\mathbf{r}^0) - F(\mathbf{p}^*))}{h},$$

then $KL\left(\mathbf{p}_{kh} \| \mathbf{p}^*\right) \leq \varepsilon$

Once again, applying Pinsker's inequality, we get that the above choice of k and t yields $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \leq \sqrt{\varepsilon}$. Without strong convexity, we cannot get a bound on W_2 from bounding $KL(\mathbf{p}_{kh} || \mathbf{p}^*)$ like we did in Corollary 1.

In [20], a proof in the non-strongly-convex case was obtained by running Langevin MCMC on

$$\tilde{\mathbf{p}}^* \propto \mathbf{p}^* \cdot \exp(-\frac{\delta}{d}\|x\|_2^2).$$

We see that $\log \tilde{\mathbf{p}}^*$ is thus strongly convex with $m = \frac{\delta}{d}$, and $d_{TV}(\mathbf{p}^*, \tilde{\mathbf{p}}^*) \leq \delta$. By the results of [20], or [27], or Theorem 1, we need

$$k = \tilde{O}(\frac{d^3}{\delta^4}) \tag{2.10}$$

iterations to get $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \leq \delta$.

On the other hand, if we assume $\log(F(\mathbf{p}_0) - F(\mathbf{p}^*)) \leq \frac{1}{\varepsilon}$ and $h' \geq \frac{1}{10}\min\left\{\frac{\varepsilon}{C_1 C_2 L^2}, \frac{\varepsilon^2}{C_1^2 dL^2}\right\}$, the results of Theorem 2 implies that with

$$h = O\left(\frac{\varepsilon}{L^2 C_1}\min\left\{\frac{1}{C_2}, \frac{\varepsilon}{dC_1}\right\}\right),$$

to get $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \leq \delta$, we need

$$k = \Omega\left(\frac{L^2 C_1^3}{\delta^4} \max\left\{C_2, \frac{dC_1}{\delta^2}\right\}\right).$$

Even if we ignore C_1 and C_2, our result is not strictly better than (2.10) as we have a worse dependence on δ. However, we do have a better dependence on d.

2.4 Proof Outline

In Section 2.4.1, we establish preliminary results which characterize curves over $\mathscr{P}(\mathbb{R}^d)$. In Section 2.4.2, we study the curve of steepest descent for $F(\cdot)$, as well as its discretization.

In section 2.4.3, we outline the proofs of Theorem 1 and 2.

2.4.1 Calculus over $\mathscr{P}(\mathbb{R}^d)$

In this section, we present some crucial lemmas which allow us to study the evolution of $F(\boldsymbol{\mu}_t)$ along a curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$. These results are all immediate consequences of results proven in [2].

Lemma 2 *For any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, let $\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu}) : \mathbb{R}^d \to \mathbb{R}$ be the first variation of F at $\boldsymbol{\mu}$ defined as $\left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu})\right)(x) \triangleq \log\left(\frac{\boldsymbol{\mu}(x)}{\mathbf{p}^*(x)}\right) + 1$. Let the subdifferential of F at $\boldsymbol{\mu}$ be given by*

$$w_{\boldsymbol{\mu}} \triangleq \nabla\left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu})\right) : \mathbb{R}^d \to \mathbb{R}^d.$$

For any curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$, and for any v_t that satisfies the continuity equation for $\boldsymbol{\mu}_t$ (see Equation (2.6)), the following holds:

$$\frac{d}{dt} F(\boldsymbol{\mu}_t) = \mathbb{E}_{\boldsymbol{\mu}_t}\left[\langle w_{\boldsymbol{\mu}_t}(x), v_t(x)\rangle\right].$$

Based on Lemma 2, we define (for any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$) the operator

$$\mathcal{D}_{\boldsymbol{\mu}}(v) \triangleq \mathbb{E}_{\boldsymbol{\mu}}\left[\langle w_{\boldsymbol{\mu}}(x), v(x)\rangle\right] : (\mathbb{R}^d \to \mathbb{R}^d) \to \mathbb{R}. \tag{2.11}$$

$\mathcal{D}_{\boldsymbol{\mu}}(v)$ is linear in v.

Lemma 3 *Let $\boldsymbol{\mu}_t$ be an absolutely continuous curve in $\mathscr{P}(\mathbb{R}^d)$ with tangent velocity field v_t. Let $|\boldsymbol{\mu}_t'|$ be the metric derivative of $\boldsymbol{\mu}_t$, then*

$$\|v_t\|_{L^2(\boldsymbol{\mu}_t)} = |\boldsymbol{\mu}_t'|.$$

Lemma 4 *For any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, let $\|\mathcal{D}_{\boldsymbol{\mu}}\|_* \triangleq \sup_{\|v\|_{L^2(\boldsymbol{\mu})} \leq 1} \mathcal{D}_{\boldsymbol{\mu}}(v)$, then*

$$\|\mathcal{D}_{\boldsymbol{\mu}}\|_* = \sqrt{\int \left\| \nabla \left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu}) \right)(x) \right\|_2^2 \boldsymbol{\mu}(x) dx}.$$

Furthermore, for any absolutely continuous curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$ with tangent velocity v_t, we have

$$\left| \frac{d}{dt} F(\boldsymbol{\mu}_t) \right| \leq \|\mathcal{D}_{\boldsymbol{\mu}_t}\|_* \|v_t\|_{L^2(\boldsymbol{\mu}_t)}.$$

As a corollary of Lemma 3 and Lemma 4, we have the following result:

Corollary 5 *Let $\boldsymbol{\mu}_t$ be an absolutely continuous curve with tangent velocity field v_t. Then*

$$\frac{d}{dt} F(\boldsymbol{\mu}_t) \leq \|\mathcal{D}_{\boldsymbol{\mu}_t}\|_* \cdot |\boldsymbol{\mu}_t'|.$$

2.4.2 Exact and Discrete Gradient Flow for $F(\mathbf{p})$

In this section, we will study the curve $\mathbf{p}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$ defined in (2.3). Unless otherwise specified, we will assume that $\mathbf{p}_0$ is an arbitrary distribution.

Let x_t be as defined in (2.3). For any given t and for all s, we define a stochastic process y_s^t as

$$\begin{aligned} y_s^t &= x_s && \text{for } s \leq t \\ dy_s^t &= -\nabla U(y_s^t) ds + \sqrt{2} dB_s && \text{for } s \geq t \\ &\text{let } \mathbf{q}_s^t \text{ denote the distribution for } y_s^t. \end{aligned} \tag{2.12}$$

From $s = t$ onwards, this is the exact Langevin diffusion with $\mathbf{p}_t$ as the initial distribution (compare with expression (2.2)).

Finally, for each t, we define a sequence z_s^t by

$$\begin{aligned} z_s^t &= x_s && \text{for } s \leq t \\ dz_s^t &= (-\nabla U(z_{\tau(t)}^t) + \nabla U(z_s^t)) ds, && \text{for } s \geq t \\ &\text{let } \mathbf{g}_s^t \text{ denote the distribution for } z_s^t. \end{aligned} \tag{2.13}$$

z_s^t represents the discretization error of $\mathbf{p}_s$ through the divergence between $\mathbf{q}_s^t$ and $\mathbf{p}_s$ (formally stated in Lemma 6). Note that $z_{\tau(t)}^t = x_{\tau(t)}^t$ because $\tau(t) \leq t$.

Remark 3 *The B_s in* (2.3), (2.12) *and* (2.13) *are the same. Thus, x_s (from* (2.3)*), y_s^t (from* (2.12)*) and z_s^t (from* (2.13)*) define a coupling between the curves $\mathbf{p}_s$, $\mathbf{q}_s^t$ and $\mathbf{g}_s^t$.*

Our proof strategy is as follows:

1. In Lemma 6, we demonstrate that the divergence between $\mathbf{p}_s$ (discretized Langevin) and $\mathbf{q}_s^t$ (exact Langevin) can be represented as a curve $\mathbf{g}_s^t$.
2. In Lemma 7, we demonstrate that the "decrease in $F(\mathbf{p}_t)$ due to exact Langevin" given by $\frac{d}{ds}F(\mathbf{q}_s^t)\big|_{s=t}$ is sufficiently negative.
3. In Lemma 8, we show that the "discretization error" given by $\frac{d}{ds}(F(\mathbf{p}_s) - F(\mathbf{q}_s^t))\big|_{s=t}$ is small.
4. Added together, they imply that $\frac{d}{ds}F(\mathbf{p}_s)\big|_{s=t}$ is sufficiently negative.

Lemma 6 *For all $x \in \mathbb{R}^d$ and $t \in \mathbb{R}^+$,*

$$\frac{d}{ds}\mathbf{g}_s^t(x)\bigg|_{s=t} = \left(\frac{d}{ds}\mathbf{p}_s(x) - \frac{d}{ds}\mathbf{q}_s^t(x)\right)\bigg|_{s=t}.$$

Lemma 7 *For all $s, t \in \mathbb{R}^+$,*

$$\frac{d}{ds}F(\mathbf{q}_s^t) = -\|\mathcal{D}_{\mathbf{q}_s^t}\|_*^2.$$

Lemma 8 *For all $t \in \mathbb{R}^+$,*

$$\frac{d}{ds}\left(F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\right)\bigg|_{s=t} \le \left(2L^2h\sqrt{\mathbb{E}_{\mathbf{p}_{\tau(t)}}[\|x\|_2^2]} + 2L\sqrt{hd}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_*.$$

2.4.3 Proof of Main Theorems

We now state the lemmas needed to prove Theorem 1. We first establish a notion of strong convexity of $F(\boldsymbol{\mu})$ with respect to W_2 metric.

Lemma 9 *For all $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathscr{P}(\mathbb{R}^d)$ and $t \in [0,1]$, let $\boldsymbol{\mu}_t : [0,1] \to \mathscr{P}(\mathbb{R}^d)$ be the constant-speed geodesic between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. If $\log \mathbf{p}^*(x)$ is m strongly convex, then*

$$F(\boldsymbol{\mu}_t) \le (1-t)F(\boldsymbol{\mu}_0) + tF(\boldsymbol{\mu}_1) - \frac{m}{2}t(1-t)W_2^2(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1). \tag{2.14}$$

(Recall from (2.4) that if $v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1}$ is the optimal displacement map from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_1$, then $\boldsymbol{\mu}_t = (Id + t \cdot v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1})_{\#}\boldsymbol{\mu}_0$.)

Equivalently,

$$F(\boldsymbol{\mu}_1) \ge F(\boldsymbol{\mu}_0) + \mathcal{D}_{\boldsymbol{\mu}_0}(v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1}) + \frac{m}{2}W_2^2(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1). \tag{2.15}$$

We call this the "m strong geodesic convexity" of F wrt the W_2 distance.

Next, we use the m strong geodesic convexity of F to upper bound $F(\boldsymbol{\mu}) - F(\mathbf{p}^*)$ by $\frac{1}{2m}\|\mathcal{D}_{\boldsymbol{\mu}}\|_*^2$ (for any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$). This is analogous to how $f(x) - f(x^*) \le \frac{1}{2m}\|\nabla f(x)\|_2^2$ for standard m strongly convex functions in R^d.

Lemma 10 *Under our assumption that $-\log \mathbf{p}^*(x)$ is m strongly convex, we have that for all $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$,*

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq \frac{1}{2m}\|\mathcal{D}_{\boldsymbol{\mu}}\|_*^2.$$

Now, recall $\mathbf{p}_t$ from (2.3). We use strong convexity to obtain a bound on $\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right]$ for all t. This will be important for bounding the discretization error in conjunction with Lemma 8.

Lemma 11 *Let $\mathbf{p}_t$ be as defined in (2.3). If $\mathbf{p}_0$ is such that $\mathbb{E}_{\mathbf{p}_0}\left[\|x\|_2^2\right] \leq \frac{4d}{m}$, and $h \leq \frac{1}{L}$ in the definition of (2.3), then for all $t \in \mathbb{R}^+$,*

$$E_{\mathbf{p}_t}\|x\|^2 \leq \frac{4d}{m}.$$

Finally, we put everything together to prove Theorem 1.

Proof of Theorem 1

We first note that $h = \frac{m\varepsilon}{16L^2} \leq \frac{1}{L}$.

By Lemma 11, for all t, $\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right] \leq \frac{4d}{m}$. Combined with Lemma 8, we get that for all $t \in \mathbb{R}^+$

$$\frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\bigg|_{s=t} \leq \left(4L^2h\sqrt{\frac{d}{m}} + 2L\sqrt{hd}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_*.$$

Suppose that $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq \varepsilon$, and let

$$h = \frac{m\varepsilon}{16dL^2} \leq \frac{1}{16}\min\left\{\frac{m}{L^2}\sqrt{\frac{\varepsilon}{d}}, \frac{m\varepsilon}{L^2 d}\right\}.$$

Then $\forall t$

$$\begin{aligned}\frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\bigg|_{s=t} &\leq \left(4L^2h\sqrt{\frac{d}{m}} + 2L\sqrt{hd}\right) \\ &\leq \frac{1}{2}\sqrt{m\varepsilon}\|\mathcal{D}_{\mathbf{p}_t}\|_* \leq \frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2,\end{aligned}$$

where the last inequality holds because Lemma 10 and the assumption that $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq \varepsilon$ together imply that $\|\mathcal{D}_{\mathbf{p}_t}\|_* \geq \sqrt{2m\varepsilon}$.

So combining Lemma 7 and Lemma 6, we have

$$\begin{aligned}\frac{d}{dt}F(\mathbf{p}_t) &= \frac{d}{ds}F(\mathbf{q}_s^t)\bigg|_{s=t} + \frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\bigg|_{s=t} \\ &\leq -\|\mathcal{D}_{\mathbf{p}_t}\|_*^2 + \frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2 \\ &= -\frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2 \\ &\leq -m(F(\mathbf{p}_t) - F(\mathbf{p}^*)),\end{aligned} \tag{2.16}$$

where the last line once again follows from Lemma 10.

To handle the case when $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq \varepsilon$, we use the following argument:

1. We can conclude that $F(\mathbf{p}_t) - F(\mathbf{p}^*) > \varepsilon$ implies $\frac{d}{dt}F(\mathbf{p}_t) \leq 0$.
2. By the results of Lemma 20 and Lemma 21, for all t, $|\mathbf{p}_t'|$ is finite and $\|\mathcal{D}_{\mathbf{p}_t}\|$ is finite, so $\frac{d}{dt}F(\mathbf{p}_t)$ is finite and $F(\mathbf{p}_t)$ is continuous in t.
3. Thus, if $F(\mathbf{p}_t) \leq \varepsilon$ for some $t \leq kh$, then $F(\mathbf{p}_s) \leq \varepsilon$ for all $s \geq t$ as $F(\mathbf{p}_t) > \varepsilon$ implies $\frac{d}{dt}F(\mathbf{p}_t) \leq 0$ and $F(\mathbf{p}_t)$ is continuous in t. Thus $F(\mathbf{p}_{kh}) - F(\mathbf{p}^*) \leq \varepsilon$.

Thus, we need only consider the case that $F(\mathbf{p}_t) > \varepsilon$ for all $t \leq kh$. This means that (2.16) holds for all $t \leq kh$.

By Gronwall's inequality, we get

$$F(\mathbf{p}_{kh}) - F(\mathbf{p}^*) \leq (F(\mathbf{p}_0) - F(\mathbf{p}^*)) \exp(-mkh).$$

We thus need to pick

$$k = \frac{\frac{1}{m}\log\frac{F(\mathbf{p}_0)-F(\mathbf{p}^*)}{\varepsilon}}{h} = 16\frac{L^2}{m^2}\frac{d\log\frac{F(\mathbf{p}_0)-F(\mathbf{p}^*)}{\varepsilon}}{\varepsilon}.$$

Using the fact that $\mathbf{p}_0 = N(0, \frac{1}{m})$. Using L-smoothness and m strong convexity, we can show that

$$-\log \mathbf{p}^*(x) \leq \frac{L}{2}\|x\|_2^2 + \frac{d}{2}\log(\frac{2\pi}{m}),$$

and

$$\log \mathbf{p}_0(x) = -\frac{m}{2}\|x\|_2^2 - \frac{d}{2}\log(\frac{2\pi}{m}).$$

We thus get that $F(\mathbf{p}_0) - F(\mathbf{p}^*) = KL(\mathbf{p}_0\|\mathbf{p}^*) \leq \frac{dL}{m}$, so

$$k = 16\frac{L^2}{m^2}\frac{d\log\frac{dL}{m\varepsilon}}{\varepsilon}.$$

□

The proof of Theorem 2 is quite similar to that of Theorem 1, so we defer it to Appendix A.

2.5 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 [20], 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 et al. [29] proved that the same algorithm

achieves ε error, in 2-Wasserstein distance, in $\mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$ steps. We remark that the proofs of Lemma 8, 11 and 17 are essentially taken from [29]. Recently Raginsky et al. [72] and Dalalyan and Karagulyan [21] also analyzed convergence of Langevin MCMC with stochastic gradient updates. Asymptotic guarantees for Langevin MCMC was established much earlier by Gelfand and Mitter [38], Roberts and Tweedie [75].

Our work also relies heavily on the theory established in the book of Ambrosio, Gigli and Savare [2], which studies the underlying probability distribution $\bar{\mathbf{p}}_t$ induced by (2.2) as a gradient flow in probability space. This allows us to view (2.3) as a *deterministic* convex optimization procedure over the probability space, with KL-divergence as the objective. This beautiful line of work relating SDEs with gradient flows in probability space was begun by Jordan, Kinderlehrer and Otto [45]. We refer any interested reader to an excellent survey by Santambrogio in [76].

Finally, we remark that the theory in [2] has some very interesting connections with the study of normalization flows in [73] and [55]. For example, the tangent velocity of (2.2), given by $v_t = \nabla \log \mathbf{p}^* - \nabla \log \mathbf{p}_t$, can be thought of as a deterministic transformation that induces a normalizing flow.

Chapter 3

Underdamped Langevin MCMC and Acceleration

3.1 Introduction

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

$$\begin{aligned} dv_t &= -\gamma v_t dt - u\nabla U(x_t)dt + (\sqrt{2\gamma u})dB_t \\ dx_t &= v_t dt, \end{aligned} \tag{3.1}$$

where $(x_t, v_t) \in \mathbb{R}^{2d}$, U 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 (3.1) is proportional to $\exp(-(U(x) + \|v\|_2^2/2u))$. Thus the marginal distribution of x is proportional to $\exp(-U(x))$. There is a discretized version of (3.1) which can be implemented algorithmically, and provides a useful way to sample from $p^*(x) \propto e^{-U(x)}$ when the normalization constant is not known.

We establish the convergence of SDE (3.1) as well as its discretization to the invariant distribution p^*. This provides explicit rates for sampling from log-smooth and strongly log-concave distributions using the underdamped Langevin Markov chain Monte Carlo (MCMC) algorithm (Algorithm 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 5, 63] has been empirically observed to converge faster to the invariant distribution compared to standard Langevin MCMC which is a discretization of *overdamped* Langevin diffusion that we studied in Chapter 2:

$$dx_t = -\nabla U(x_t)dt + \sqrt{2}dB_t, \tag{3.2}$$

the first order SDE corresponding to the high friction limit of (3.1). This chapter provides a non-asymptotic quantitative explanation for these observations.

3.2 Assumptions and Definitions

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

3.2.1 Assumptions on U

We make the following assumptions regarding the function U.

(A1) The function U 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

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

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

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

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

3.2.2 Underdamped Langevin Diffusion

Throughout the paper we use B_t to denote standard Brownian motion [62]. Next we set up the notation specific to the continuous and discrete processes that we study in this chapter.

1. Consider the exact underdamped Langevin diffusion defined by the SDE (3.1), with an initial condition $(x_0, v_0) \sim p_0$ for some distribution p_0 on $\mathcal{R}^{2d}$. 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:

$$\Phi_t p_0 = p_t. \tag{3.3}$$

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

$$\begin{aligned} d\tilde{v}_t &= -\gamma\tilde{v}_t dt - u\nabla U(\tilde{x}_0)dt + (\sqrt{2\gamma u})dB_t \\ d\tilde{x}_t &= \tilde{v}_s dt, \end{aligned} \tag{3.4}$$

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 (3.1) 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 (3.4) for small t. Think of an integral solution of (3.4) 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 $\nabla U(\cdot)$

for $i = 0, 1, \ldots, n-1$ **do**
  Sample $(x^{i+1}, v^{i+1}) \sim Z^{i+1}(x^i, v^i)$
end

3.2.3 Stationary Distributions

Throughout the chapter, we denote by p^* the unique distribution which satisfies $p^*(x, v) \propto \exp -(U(x) + \frac{1}{2u}\|v\|_2^2)$. It can be shown that p^* is the unique invariant distribution of (3.1) [see Proposition 6.1 in 68]. Let $g(x, v) = (x, x + v)$. We let q^* be the distribution of $g(x, v)$ when $(x, v) \sim p^*$.

3.3 Main Results

Our main result is a proof that Algorithm 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 $p^* \propto \exp(-(U(x))$, where U is L smooth and m strongly convex (see section 3.2.1), with $\kappa = L/m$ denoting the condition number. Compared to the results of [26] 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 when we have noisy gradients with bounded variance and establish non-asymptotic convergence guarantees in this setting.

3.3.1 Algorithm

The underdamped Langevin MCMC algorithm that we analyze in this chapter in shown in Algorithm 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:

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

The distribution is obtained by integrating the discrete underdamped Langevin diffusion (3.4) 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 $Z^{i+1}(x^i, v^i) \sim p^{(i+1)} = \tilde{\Phi}_\nu p^{(i)}$. Refer to Lemma 23 in Appendix B.1 for the derivation.

3.3.2 Convergence under Exact Gradient

Theorem 3 *Let $p^{(n)}$ be the distribution of the iterate of Algorithm 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*

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

and run Algorithm 1 for n iterations with

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

then we have the guarantee that

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

Remark 4 *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 [26].*

We note that the $\log(24(d/m+\mathcal{D}^2)/\varepsilon)$ factor can be shaved off by using a time-varying step size.

3.3.3 Convergence under Stochastic Gradient

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

1. The noise is unbiased : $\mathbb{E}[\xi] = 0$.
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,

$$\begin{aligned} d\hat{v}_t &= -\gamma \hat{v}_t dt - u\hat{\nabla} U(\hat{x}_0)dt + (\sqrt{2\gamma u})dB_t \\ d\hat{x}_t &= \hat{v}_s dt, \end{aligned} \tag{3.5}$$

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 3.2.2.

Theorem 4 *Let $p^{(n)}$ be the distribution of the iterate of Algorithm 3 (presented in Appendix B.3) 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*

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

and run Algorithm 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 5 *Note that when the variance in the gradients $\sigma^2 d$ is large we recover the rate of overdamped Langevin diffusion and we need $\tilde{\mathcal{O}}(\sigma^2\kappa^2 d/\varepsilon^2)$ steps to achieve accuracy of ε in W_2.*

3.4 Proof Outline

In this section, we outline the proof of Theorem 3. The proof of Theorem 4 is similar, and will be relegated to Appendix B.3.

In Section 3.4.1, we establish the convergence rate for the continuous-time SDE (3.1).

In Section 3.4.2, we bound the discretization error between (3.1) and (4.7).

The proof of Theorem 3, given in Section 3.4.3, follows from combining these two results.

3.4.1 Convergence of the Continuous-Time Process

In this section we prove Theorem 5, which demonstrates a contraction for solutions of the SDE (3.1). We will use Theorem 5 along with a bound on the discretization error between (3.1) and (3.4) to establish guarantees for Algorithm 1.

Theorem 5 *Let (x_0, v_0) and (y_0, w_0) be two arbitrary points in $\mathcal{R}^{2d}$. 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*

$$\begin{aligned}\mathbb{E}_{(x_t,v_t,y_t,w_t)\sim\zeta_t((x_0,v_0,y_0,w_0))}\left[\|x_t - y_t\|_2^2 + \|(x_t + v_t) - (y_t + w_t)\|_2^2\right] \\ \leq e^{-t/\kappa}\left\{\|x_0 - y_0\|_2^2 + \|(x_0 + v_0) - (y_0 + w_0)\|_2^2\right\}.\end{aligned} \tag{3.6}$$

Remark 6 *A similar objective function was used in [33] 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 12 *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 $\zeta_0 \in \Gamma(p_0, p^*)$ be 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,

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

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

$$\frac{1}{2}W_2(p_t, p^*) \leq W_2(q_t, q^*) \leq 2W_2(p_t, p^*). \tag{3.7}$$

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 of Lemma 13
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 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 (3.1), we get

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

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 U is twice differentiable, by Taylor's theorem we have

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

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

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

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

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

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

$$\begin{aligned}
&\frac{d}{dt}\left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2\right] \\
&= -2\langle (z_t + \psi_t, z_t), ((\gamma - 1)\psi_t + u\mathcal{H}_t z_t, -\psi_t)\rangle \\
&= -2\begin{bmatrix} z_t + \psi_t & z_t \end{bmatrix} \underbrace{\begin{bmatrix} (\gamma - 1)I_{d\times d} & u\mathcal{H}_t - (\gamma - 1)I_{d\times d} \\ -I_{d\times d} & I_{d\times d} \end{bmatrix}}_{\triangleq S_t} \begin{bmatrix} z_t + \psi_t \\ z_t \end{bmatrix}
\end{aligned} \tag{3.8}$$

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

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

By invoking a standard result of linear algebra (stated in the appendix as Lemma 28), 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 $0 < m \leq \Lambda_j \leq L$. 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 $\lambda_{min}(Q_t) \geq 1/2\kappa$.

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

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

Combining this with (3.8) yields

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

The convergence rate of Theorem 5 follows immediately from this result by applying Grönwall's inequality [Corollary 3 in 25]. □

3.4.2 Discretization Analysis

In this section, we study the solutions of the discrete process (3.4) up to $t = \delta$ for some small δ. Here, δ represents a *single step* of the Langevin MCMC algorithm. In Theorem 6, we will bound the discretization error between the continuous-time process (3.1) and the discrete process (3.4) 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 3. Recall the definition of Φ_t and $\tilde{\Phi}_t$ from (3.3).

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

$$\forall t \in [0, \nu] \quad \mathbb{E}_{p_t} \left[\|v\|_2^2\right] \leq \mathcal{E}_K. \tag{3.9}$$

We derive an explicit bound on $\mathcal{E}_K$ (in terms of problem parameters d, L, m etc.) in Lemma 24 in Appendix B.2.

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 6 *Let Φ_t and $\tilde{\Phi}_t$ be as defined in (3.3) 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*

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

Proof
We will once again use a standard synchronous coupling argument, in which $\Phi_\nu p_0$ and $\tilde{\Phi}_\nu p_0$ 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 63, we have

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

where (i) follows from Lemma 63 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 $U(x)$, (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 (3.9), and proven in Lemma 24. This completes the bound for the velocity variable. Next we bound the discretization error in the position variable:

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

where the equality is by coupling through the initial distribution p_0, the first inequality is by Jensen's inequality and the second 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

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

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

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

Taking square roots establishes the desired result. □

3.4.3 Proof of Theorem 3

From Corollary 12, we have that for any $i \in \{1, \ldots, n\}$

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

By the discretization error bound in Theorem 6 and the Sandwich Inequality (3.7), we get

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

By the triangle inequality for W_2,

$$W_2(q^{(i+1)}, q^*) = W_2(\tilde{\Phi}_\nu q^{(i)}, q^*) \leq W_2(\Phi_\nu q^{(i)}, \tilde{\Phi}_\nu q^{(i)}) + W_2(\Phi_\nu q^{(i)}, q^*) \tag{3.10}$$

$$\leq \nu^2 \sqrt{\frac{8\mathcal{E}_K}{5}} + e^{-\delta/2\kappa} W_2(q^{(i)}, q^*). \tag{3.11}$$

Let us define $\eta = e^{-\delta/2\kappa}$. Then by applying (3.11) n times we have:

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

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

where the second step follows by summing the geometric series and by applying the upper bound (3.7). By another application of (3.7) we get:

$$W_2(p^{(n)}, p^*) \leq \underbrace{4\eta^n W_2(p^{(0)}, p^*)}_{\triangleq T_1} + \underbrace{\left(\frac{1}{1-\eta}\right) \nu^2 \sqrt{\frac{32\mathcal{E}_K}{5}}}_{\triangleq T_2}. \tag{3.12}$$

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 24) ensures that,

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

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

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

In Lemma 25 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.

3.5 Related Work

Hamiltonian Monte Carlo (HMC) is a broad class of algorithms which involve Hamiltonian dynamics in some form. We refer to Ma et al. [56] for a survey of the results in this area. Among these, the variant studied in this paper (Algorithm 2), based on the discretization of (3.1), 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 [49] in the context of chemical reactions. The continuous-time process has been studied extensively [4, 6, 11, 24, 33, 39, 42, 61, 83].

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. [33] who demonstrated a contraction property of the continuous-time process defined in (3.2). 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 [52] analyzed Riemannian HMC for sampling from polytopes using a logarithmic barrier function. Mangoubi and Smith [59] 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 [50, 81, 84].

[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 $c \in \mathbb{N}$.

Chapter 4

Non-convex Sampling

4.1 Introduction

We study the problem of sampling from a target distribution of the following form:

$$p^*(x) \propto \exp\left(-U(x)\right),$$

where $x \in \mathbb{R}^d$, and the *potential function* $U : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth everywhere and m strongly convex outside a ball of radius R (see detailed assumptions in Section 4.2.1).

In both optimization and sampling, while the classical theory focused on convex problems, recent attention has turned to the more broadly useful setting of non-convex problems. While general non-convex problems are infeasible, it is possible to make reasonable assumptions that allow theory to proceed while still making contact with practice.

We again study the overdamped Langevin MCMC algorithm, which is a discretization of the following SDE:

$$dx_t = -\nabla U(x_t)dt + \sqrt{2}dB_t, \tag{4.1}$$

whose invariant distribution is $p^*(x)$. We will also study the *underdamped Langevin diffusion*, which can be represented by the following SDE:

$$\begin{aligned} dx_t &= u_t dt\,, \\ du_t &= -\lambda_1 u_t - \lambda_2 \nabla U(x_t)dt + \sqrt{2\lambda_1\lambda_2}dB_t, \end{aligned} \tag{4.2}$$

where $\lambda_1, \lambda_2 > 0$ are free parameters. This SDE can also be discretized appropriately to yield a corresponding MCMC algorithm (Algorithm 1). Second-order methods such as underdamped Langevin MCMC are particularly interesting as it has been previously observed both empirically [63] and theoretically [17, 59] that these methods can be faster to converge than the classical first-order methods.

In this chapter, we show that it is possible to sample from p^* in time polynomial in the dimension d and the target accuracy ε (as measured in 1-Wasserstein distance). We also

show that the convergence depends exponentially on the product LR^2. Intuitively, LR^2 is a measure of the non-convexity of U. Our results establish rigorously that as long as the problem is not "too badly non-convex," sampling is provably tractable.

Our main results are presented in Theorem 8 and Theorem 9, and can be summarized informally as follows:

Theorem 7 (informal) *Given a potential U that is L-smooth everywhere and strongly-convex outside a ball of radius R, we can output a sample from a distribution which is ε-close to $p^*(x) \propto \exp(-U(x))$ in W_1 distance by running $\widetilde{\mathcal{O}}\left(e^{cLR^2} d/\varepsilon^2\right)$ steps of overdamped Langevin MCMC (4.4), or $\widetilde{\mathcal{O}}\left(e^{cLR^2}\sqrt{d}/\varepsilon\right)$ steps of underdamped Langevin MCMC (Algorithm 2). Here, c is an explicit positive constant.*

For the case of strongly convex U, it has been shown by [17] that the iteration complexity of Algorithm 2 is $\widetilde{\mathcal{O}}(\sqrt{d}/\varepsilon)$, improving quadratically upon the best known iteration complexity of $\widetilde{\mathcal{O}}(d/\varepsilon^2)$ for overdamped Langevin MCMC [27]. We will find this quadratic speed-up in d and ε in our setting as well (see Theorem 8 versus Theorem 9).

4.2 Assumptions and Definitions

In this section, we present the basic definitions, notational conventions and assumptions used throughout the paper. For $q \in \{1, 2\}$ and $v \in \mathbb{R}^d$ we let $\|v\|_q$ denote the q-norm. For $M \in \mathbb{R}^{d\times d}$, we let $\|M\|_2 := \sup_{\|v\|_2\leq 1, v\in\mathbb{R}^d} \|Mv\|_2$. We use B_t to denote standard Brownian motion [see, e.g., 62].

4.2.1 Assumptions on the potential U

We make the following assumptions on the *potential function* U:

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

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

(A2) The function has a stationary point at zero:

$$\nabla U(0) = 0.$$

(A3) The function is strongly convex outside of a ball; that is, there exist constants $m, R > 0$ such that for all $x, y \in \mathbb{R}^d$ with $\|x - y\|_2 > R$, we have:

$$\langle \nabla U(x) - \nabla U(y), x - y\rangle \geq m\|x - y\|_2^2.$$

Finally we define the condition number as $\kappa := L/m$. Observe that Assumption (A2) is imposed without loss of generality, because we can always find a stationary point in polynomial time and shift the coordinate system so that this stationary point of U is at zero. These conditions are similar to the assumptions made by Eberle [32]. Note that crucially Assumption (A3) is *strictly stronger* than the assumption made in recent papers by Durmus and Moulines [27], Raginsky et al. [72] and Zhang et al. [87]. To see this observe that these papers only require Assumption (A3) to hold for a fixed $y = 0$, while we require this condition to hold for all $y \in \mathbb{R}^d$. One can also think of the difference between these two conditions as being analogous to the difference between strong convexity (outside a ball) and one-point strong convexity (outside a ball).

4.3 Main Results

In this section, we state our main theorems. Our first result, Theorem 8, shows that overdamped Langevin MCMC needs $O(d/\varepsilon^2)$ steps to achieve ε error in Wasserstein distance. The proof is quite simple, and serves as a warm-up to Theorem 9, which shows that under Langevin MCMC needs $O(\sqrt{d}/\varepsilon)$ steps to achieve ε error in Wasserstein distance. This is a quadratic improvement over the overdamped Langevin MCMC algorithm in terms of d and ε, which mirrors the results when U is convex.

4.3.1 Overdamped Langevin diffusion

Our first result studies the convergence of overdamped Langevin diffusion, given in (1.4), and reproduced below for ease of reference:

$$dy_t = -\nabla U(y_t)dt + \sqrt{2}dB_t. \tag{4.3}$$

The discretized overdamped Langevin diffusion as

$$dx_t = -\nabla U\Big(x_{\lfloor \frac{t}{\delta} \rfloor}\Big)dt + \sqrt{2}dB_t, \tag{4.4}$$

where δ is the step-size of the discretization and $\lfloor \cdot \rfloor$ denotes the floor function.

Our first result, stated as Theorem 8, establishes the rate at which the distribution of the solution of Eq. (4.4) converges to p^*.

Theorem 8 *Assume that* $m \geq \frac{\exp\left(-LR^2/2\right)}{R^2}$, *and let* $0 < \varepsilon \leq \frac{dR^2}{\sqrt{d/m+R^2}}$ *be the desired accuracy. Also let the initial point* $x^{(0)}$ *be such that* $\|x^{(0)}\|_2 \leq R$. *Then if the step size scales as:*

$$\delta = \frac{\varepsilon^2 \exp\left(-LR^2\right)}{2^{10}R^2 d},$$

and number of iterations scales as:

$$n = \widetilde{\Omega}\left(\exp\left(\frac{3LR^2}{2}\right) \cdot \frac{d}{\varepsilon^2}\right),$$

we have the following guarantee:

$$W_1(p_{n\delta}, p^*) \leq \varepsilon,$$

where $p_{n\delta}$ is the distribution of $x_{n\delta}$ in (4.4) and the distribution $p^*(y) \propto e^{-U(y)}$.

For potentials where LR^2 is a constant, the number of iterations taken by overdamped MCMC scales as $\widetilde{\Omega}(d/\varepsilon^2)$. This matches the rate obtained in the strongly log-concave setting by Durmus and Moulines [27].

Intuitively, LR^2 measures the extent of non-convexity. When this quantity is large, it is possible for U to contain numerous local minima that are deep. It is therefore reasonable that the runtime of the algorithm should be exponential in this quantity.

The assumption on the strong convexity parameter, m, is made to simplify the presentation of the theorem. Note that this assumption is without loss of generality, since we can always take the radius R to be sufficiently large in Assumption (A3). Similarly, our assumption on the target accuracy can also be easily removed, but we make this assumption in the interest of clarity.

The proof of Theorem 8 is relegated to Appendix C.3. The proof follows by carefully combining the continuous-time argument of [32] together with the discretization bound of [27].

4.3.2 Underdamped Langevin diffusion

In this section, we present our results for *underdamped Langevin diffusion*. The underdamped Langevin diffusion is a second-order stochastic process described by the following SDE:

$$\begin{aligned} dy_t &= v_t dt, \\ dv_t &= -2v_t - \frac{c_\kappa}{L}\nabla U(y_t)dt + \sqrt{\frac{4c_\kappa}{L}}dB_t, \end{aligned} \tag{4.5}$$

where we define the constant:

$$c_\kappa := 1/(1000\kappa), \tag{4.6}$$

where $\kappa = L/m$ is the condition number. Similar to the case of overdamped Langevin diffusion, it can be verified that the invariant distribution of the SDE is $p^*(y, v) \propto e^{-U(y) - \frac{L}{2c_\kappa}\|v\|_2^2}$. This ensures that the marginal along y is the distribution that we are interested in. Based on the SDE in Eq. (4.5), we define the discretized underdamped Langevin diffusion as:

$$\begin{aligned} dx_t &= u_t dt, \\ du_t &= -2u_t - \frac{c_\kappa}{L}\nabla U\Big(x_{\lfloor \frac{t}{\delta} \rfloor \delta}\Big)dt + \sqrt{\frac{4c_\kappa}{L}}dB_t, \end{aligned} \tag{4.7}$$

where δ is the step size of discretization. The SDE in Eq. (4.7) is implementable as the following algorithm:

Algorithm 2: Underdamped Langevin MCMC

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

for $i = 0, 1, \ldots, n-1$ **do**
  Sample $(x_{(i+1)\delta}, u_{(i+1)\delta}) \sim Z^{(i)}(x_{i\delta}, u_{i\delta})$
end

In this algorithm $Z^{(i)}(x_{i\delta}, u_{i\delta}) \in \mathbb{R}^{2d}$ is a Gaussian random vector with the following mean and covariance (which are functions of the previous iterates $(x_{i\delta}, u_{i\delta})$):

$$\mathbb{E}\left[u_{(i+1)\delta}\right] = u_{i\delta}e^{-2\delta} - \frac{c_\kappa}{2L}(1 - e^{-2\delta})\nabla U(x_{i\delta})\,,$$

$$\mathbb{E}\left[x_{(i+1)\delta}\right] = x_{i\delta} + \frac{1}{2}(1 - e^{-2\delta})u_{i\delta} - \frac{c_\kappa}{2L}\left(\delta - \frac{1}{2}\left(1 - e^{-2\delta}\right)\right)\nabla U(x_{i\delta})\,,$$

$$\mathbb{E}\left[\left(x_{(i+1)\delta} - \mathbb{E}\left[x_{(i+1)\delta}\right]\right)\left(x_{(i+1)\delta} - \mathbb{E}\left[x_{(i+1)\delta}\right]\right)^\top\right] = \frac{c_\kappa}{L}\left[\delta - \frac{1}{4}e^{-4\delta} - \frac{3}{4} + e^{-2\delta}\right]\cdot I_{d\times d}\,,$$

$$\mathbb{E}\left[\left(u_{(i+1)\delta} - \mathbb{E}\left[u_{(i+1)\delta}\right]\right)\left(u_{(i+1)\delta} - \mathbb{E}\left[u_{(i+1)\delta}\right]\right)^\top\right] = \frac{c_\kappa}{L}(1 - e^{-4\delta})\cdot I_{d\times d}\,,$$

$$\mathbb{E}\left[\left(x_{(i+1)\delta} - \mathbb{E}\left[x_{(i+1)\delta}\right]\right)\left(u_{(i+1)\delta} - \mathbb{E}\left[u_{(i+1)\delta}\right]\right)^\top\right] = \frac{c_\kappa}{2L}\left[1 + e^{-4\delta} - 2e^{-2\delta}\right]\cdot I_{d\times d}\,.$$

We show that the iterates at round i of Algorithm 2 and the solution to the SDE in Eq. (4.7) at time $t = i\delta$ have the same distribution (see Lemma 64 in Appendix C.8).

In Theorem 9, we establish a bound on the rate at which the distribution of the iterates produced by this algorithm converge to the target distribution p^*.

Theorem 9 *Assume that* $m \geq \frac{\exp\left(-6LR^2\right)}{64R^2}$ *and let* $0 < \varepsilon \leq \frac{dR^2}{\sqrt{d/m+R^2}}$ *be the desired accuracy. Also let the initial point* $x^{(0)}$ *be such that* $\|x^{(0)}\|_2 \leq R$*. Assume also that* $e^{72LR^2} \geq 2$*.*

Then if the step size scales as:

$$\delta = \frac{\varepsilon}{R + \sqrt{d/m}} \cdot e^{-12LR^2} \cdot 2^{-35} \min\left(\frac{1}{LR^2}, \frac{1}{\kappa}\right),$$

and the number of iterations as:

$$n = \widetilde{\Omega}\left(\frac{\sqrt{d}}{\varepsilon}\exp\left(18LR^2\right)\right),$$

we have the guarantee that

$$W_1(p_{n\delta}, p^*) \leq \varepsilon,$$

where $p_{n\delta}$ *is the distribution of* $x_{n\delta}$ *and we have* $p^*(y) \propto e^{-U(y)}$*.*

If we consider potentials for which LR^2 is a constant, the iteration complexity of underdamped Langevin MCMC grows as $\widetilde{\mathcal{O}}(\sqrt{d}/\varepsilon)$, which is a quadratic improvement over the first-order overdamped Langevin MCMC algorithm. Again, the iteration complexity grows exponentially in LR^2 which is to be expected. As before, the condition on the strong convexity parameter and the target accuracy is made in the interest of clarity and can be removed..

4.4 Proof Outline

In this section, we sketch the proof of Theorem 9. The heart of the proof of this theorem is a somewhat intricate coupling argument. We begin by defining two processes, (x_t, u_t) and (y_t, v_t), and then couple them appropriately. The first set of variables, (x_t, u_t), represent a solution to the discretized SDE in Eq. (4.7). On the other hand, the variables (y_t, v_t) represent a solution of the continuous-time SDE in Eq. (4.5) with the initial conditions being $(y_0, v_0) \sim p^*(y, v)$. Thus the variables (y_t, v_t) evolve according to the invariant distribution for all $t > 0$. The noise that underlies both processes is *coupled*, and with an appropriate choice of a Lyapunov function we are able to demonstrate that the distributions of these variables converge in 1-Wasserstein distance.

In section D.1.2 we explicitly construct a coupling. In section 4.4.3, we describe the Lyapunov function that we will use. Finally, in the section 4.4.4, we describe how the Lyapunov function contracts to 0 under the stated coupling, which in turns implies a convergence in Wasserstein distance.

In Section D.1.2,

4.4.1 A coupling construction

Let $\beta = 1/poly(L, 1/m, d, R, 1/C_m)$ be a small constant (see proof of Theorem 9 for the exact value), and let $\ell(x) = q(\|x\|_2)$ be a smoothed approximation of $\|x\|_2$ at a scale of β, as defined in (C.2).

Additionally, let $\nu = 1/poly(L, 1/m, d, R, 1/C_m)$ be another small constant (see proof of Theorem 9 for the exact value). In designing our coupling, we ensure that certain values are only updated at intervals of size ν. These are needed to ensure that the stochastic process that we work with is sufficiently regular.

While reading the proofs it might be convenient for the reader to think of both β and ν to be arbitrarily close to zero, and to think of $\ell(x)$ as equal to $\|x\|_2$; β and ν do not impact the bound on the iteration complexity in Theorem 9. For a detailed discussion see Appendix C.2.

We define a time T_{sync} as

$$T_{sync} := \frac{3 \log 100}{c_\kappa^2}. \tag{4.8}$$

We then choose ν to be such that $\frac{T_{sync}}{\nu}$ is a positive integer, and define the constant

$$\begin{aligned} C_m :=& \min\left\{\frac{e^{-6LR^2}}{6000\kappa(1+LR^2)}, \frac{e^{-6LR^2}}{200T_{sync}}, \frac{c_\kappa^2}{3}\right\} \\ =& \min\left\{\frac{e^{-6LR^2}}{2^{13}\kappa(1+LR^2)}, \frac{e^{-6LR^2}}{2^{29}\cdot\log(100)\cdot\kappa^2}, \frac{1}{2^{22}\kappa^2}\right\}. \end{aligned} \tag{4.9}$$

This constant C_m will be the rate at which our Lyapunov function contracts.

With these definitions in place we are ready to define a coupling between variables (x_t, u_t) that evolve according to the discretized process described in Eq. (4.11), and variables (y_t, v_t) that evolve according to the SDE in Eq. (4.13).

Let the initial conditions for these processes be given by,

$$\begin{aligned} (x_0, u_0) &= (x^{(0)}, 0), \\ (y_0, v_0) &\sim p^*(y, v). \end{aligned} \tag{4.10}$$

Define a variable τ_t that will be useful in determining how the noise underlying the processes is coupled. We initialize this variable as follows: $\tau_0 = 0$, if $\sqrt{\|x_0 - y_0\|_2^2 + \|x_0 - y_0 + u_0 - w_0\|_2^2} \geq \sqrt{5}R$, and $\tau_0 = -T_{sync}$ otherwise.

Let A_t and B_t denote independent d-dimensional Brownian motions. We then let the complete set of variables $\left(x_t, u_t, y_t, v_t, \tau_{\lfloor \frac{t}{\nu} \rfloor}\right)$ evolve according to the following stochastic dynamics:

$$dx_t = u_t dt \tag{4.11}$$

$$du_t = -2u_t dt - \frac{c_\kappa}{L}\nabla U\left(x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right)dt + 2\sqrt{\frac{c_\kappa}{L}}dB_t \tag{4.12}$$

$$dy_t = v_t dt \tag{4.13}$$

$$\begin{aligned} dv_t =& -2v_t - \frac{c_\kappa}{L}\nabla U(y_t)dt + 2\sqrt{\frac{c_\kappa}{L}}dB_t \\ &- \mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\} \cdot \left(4\sqrt{\frac{c_\kappa}{L}}\gamma_t\gamma_t^T dB_t + 2\sqrt{\frac{c_\kappa}{L}}\bar{\gamma}_t\bar{\gamma}_t^T dA_t\right), \end{aligned} \tag{4.14}$$

where the functions $\mathcal{M}$, γ_t and $\bar{\gamma}_t$ are defined as follows:

$$\begin{aligned} \mathcal{M}(r) :=& \begin{cases} 1, & \text{for } r \in [\beta, \infty) \\ \frac{1}{2} + \frac{1}{2}\cos\left(r \cdot \frac{2\pi}{\beta}\right), & \text{for } r \in [\beta/2, \beta] \\ 0, & \text{for } r \in [0, \beta/2] \end{cases} \\ \gamma_t :=& (\mathcal{M}(\|z_t + w_t\|_2))^{1/2} \frac{z_t + w_t}{\|z_t + w_t\|_2} \\ \bar{\gamma}_t :=& \left(1 - (1 - 2\mathcal{M}(\|z_t + w_t\|_2))^2\right)^{1/4} \frac{z_t + w_t}{\|z_t + w_t\|_2}, \end{aligned} \tag{4.15}$$

and where for convenience we have defined

$$\begin{aligned} z_t &:= x_t - y_t \\ w_t &:= u_t - v_t. \end{aligned} \tag{4.16}$$

Note that the function $\mathcal{M}$ essentially is a Lipschitz approximation to the indicator function $\mathbb{1}\{r > 0\}$.

Let us unpack the definition of the SDE. First, note that when the indicator $\mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\}$ is equal to zero, then both (x_t, u_t) and (y_t, v_t) are evolved by the same Brownian motion B_t. This is called a *synchronous coupling* between the processes.

Second, when this indicator is equal to one, the processes are evolved by the same Brownian motion in the directions perpendicular to $z_t + w_t$, and (roughly) by the reflected Brownian motion along the direction $z_t + w_t$. This is called a *reflection coupling* between the two processes.

In the following lemma, we show that the variables (y_t, v_t) have the same marginal distributions as the solution to the SDE defined in Eq. (4.13).

Lemma 14 *The dynamics in defined by Eq.* (4.13) *and Eq.* (4.14) *is distributionally equivalent to the dynamics defined by Eq.* (4.5)*.*

We give the proof in Appendix C.8. It is easy to verify that (x_t, u_t) have the same marginal distribution as the solution to the SDE defined in Eq. (4.11) so we omit the proof.

Finally, we define an update rule for τ which dictates how the noise is coupled. For any $k \in \mathbb{Z}^+$, τ_k is defined as follows:

$$\tau_k := \begin{cases} k\nu \\ \text{if } \left(k\nu - \tau_{k-1} \geq T_{sync} \text{ AND } \sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} \geq \sqrt{5}R\right) & \tau_{k-1} otherwise. \end{cases} \tag{4.17}$$

From the dynamics in Eq. (4.14), we see that τ_k is used for determining whether (x_t, y_t, u_t, v_t) evolves by synchronous or reflection coupling over the interval $t \in [k\nu, (k+1)\nu)$. From its definition in Eq. (4.17), we see that, roughly speaking, τ_k is "the last time (up to $k\nu$) that (z_t, w_t) ends up outside the ball $\sqrt{\|z_t\|_2^2 + \|z_t + w_t\|_2^2} = \sqrt{5}R$," but with a caveat: we do not update the value of τ_k more than once in a T_{sync} interval of time.

Let $(\Omega, \mathcal{F}_t, P)$ be the probability space, where $\mathcal{F}_t$ is the σ-algebra generated by (y_0, v_0), B_s and A_s for all $s \in [0, t)$. In the following Lemma, we prove that $\left(x_t, u_t, v_t, y_t, \tau_{\lfloor \frac{t}{\nu} \rfloor}\right)$ has a unique strong solution $(x_t, u_t, y_t, v_t, \tau_{\lfloor \frac{t}{\nu} \rfloor})(\omega)$ $(\omega \in \Omega)$, which is adapted to the filtration $\mathcal{F}_t$. Furthermore, with probability one, $(x_t, u_t, y_t, v_t)(\omega)$ is t-continuous:

Lemma 15 *Let B_t and A_t be two independent Brownian motions, and let $\mathcal{F}_t$ be the σ-algebra generated by B_s, A_s; $s \leq t$, and (x_0, u_0, y_0, v_0).*

For all $t \geq 0$, the stochastic process $(x_t, u_t, y_t, v_t, \tau_{\lfloor \frac{t}{\nu} \rfloor})(\omega)$ defined in Eqs. (4.11)–(4.17) has a unique solution such that (x_s, u_s, y_s, v_s) is t-continuous with probability one, and satisfies the following, for all $s \geq 0$,

1. *$(x_s, u_s, y_s, v_s, \tau_{\lfloor \frac{s}{\nu} \rfloor})$ is adapted to the filtration $\mathcal{F}_s$.*
2. *$\mathbb{E}\left[\|x_s\|_2^2 + \|y_s\|_2^2 + \|u_s\|_2^2 + \|v_s\|_2^2\right] \leq \infty$.*

We defer the proof of this lemma to Appendix C.7.

Finally, for notational convenience, we define the following quantities, for any $k \in \mathbb{Z}^+$:

$$\begin{aligned} \mu_k :=& \mathbb{1}\left\{k\nu \geq \tau_k + T_{sync}\right\} && (4.18)\\ r_t :=& (1 + 2c_\kappa)\ell(z_t) + \ell(z_t + w_t) && (4.19)\\ \nabla_t :=& \nabla U(x_t) - \nabla U(y_t) \\ \Delta_t :=& \nabla U(x_{\lfloor \frac{t}{\delta} \rfloor \delta}) - \nabla U(x_t). && (4.20) \end{aligned}$$

As described above, when $\mu_k = 0$ the processes are synchronously coupled, and when $\mu_k = 1$ they are coupled via reflection coupling. Roughly, r_t corresponds to the sum of $\|z_t\|_2$ and $\|z_t + w_t\|_2$. ∇_t is the difference of the gradients of U at x_t and y_t, while Δ_t is the difference of the gradients at $x_{\lfloor \frac{t}{\delta} \rfloor \delta}$ and x_t.

4.4.2 A concave upper bound on $\|\cdot\|_2$

We follow Eberle [32] in our specification of the *distance function* $f : \mathbb{R}^+ \to \mathbb{R}$ that is used in the definition of our Lyapunov function.

Let α_f and $\mathcal{R}_f$ be two positive constants, to be specified later.

We define auxiliary functions ψ, Ψ and g, all mapping from $\mathbb{R}^+$ to $\mathbb{R}^+$:

$$\begin{aligned} h(r) :=& \begin{cases} 1, & \text{for } r \in [0, \mathcal{R}_f] \\ 1 - \frac{1}{\mathcal{R}_f}(r - \mathcal{R}_f), & \text{for } r \in [\mathcal{R}_f, 2\mathcal{R}_f] \\ 0, & \text{for } r \in [2\mathcal{R}_f, \infty) \end{cases} \\ \psi(r) :=& e^{-2\alpha_f \int_0^r h(s)ds}, \qquad \Psi(r) := \int_0^r \psi(s)ds, && (4.21) \\ g(r) :=& 1 - \frac{1}{2}\frac{\int_0^r h(s)\frac{\Psi(s)}{\psi(s)}ds}{\int_0^\infty h(s)\frac{\Psi(s)}{\psi(s)}ds}. \end{aligned}$$

Let us summarize some important properties of the functions ψ and g:

- ψ is decreasing, $\psi(0) = 1$, and $\psi(r) = \psi(2\mathcal{R}_f)$ for any $r > 2\mathcal{R}_f$.
- g is decreasing, $g(0) = 1$, and $g(r) = \frac{1}{2}$ for any $r > 2\mathcal{R}_f$.

Finally we define f as

$$f(r) := \int_0^r \psi(s)g(s)ds. \tag{4.22}$$

In Lemma 55 in Appendix C.5, we state and prove various several useful properties of the distance function f. Most importantly, f is designed to be concave, and $f(r) \geq C \cdot r$ for some constant C.

4.4.3 Lyapunov Function

In this section, we define a Lyapunov function that will be useful in demonstrating that the distributions of (x_t, u_t) and (y_t, v_t) converge in 1-Wasserstein distance. Let f be as defined in (4.22), with α_f and $\mathcal{R}_f$ defined as:

$$\alpha_f := \frac{L}{4}, \qquad \text{and,} \qquad \mathcal{R}_f := 12R, \tag{4.23}$$

Recall the constant C_m defined in (4.9).

Additionally define the stochastic processes:

$$\xi_t = \int_0^t e^{-C_m(t-s)} c_\kappa \left\| x_s - x_{\lfloor \frac{s}{\delta} \rfloor \delta} \right\|_2 ds, \tag{4.24}$$

$$\sigma_t = \int_0^t \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(t-s)} \cdot \mathbb{1}\left\{ r_s \geq \sqrt{12}R \right\} 4 r_s ds, \tag{4.25}$$

$$\phi_t = \int_0^t \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(t-s)} \left\langle \nabla_{w_s}(f(r_s)), 4\sqrt{\frac{c_\kappa}{L}} \left(\gamma_s \gamma_s^T dB_s + \frac{1}{2} \bar{\gamma}_s \bar{\gamma}_s^T dA_s \right) \right\rangle. \tag{4.26}$$

The processes ξ_t and σ_t will be used to track the discretization error arising due to the time increments δ and ν. We refer to Lemma 62 in Appendix C.7 for a proof of the existence of ϕ_t.

The following stochastic process $\mathcal{L}_t$ acts as our Lyapunov function:

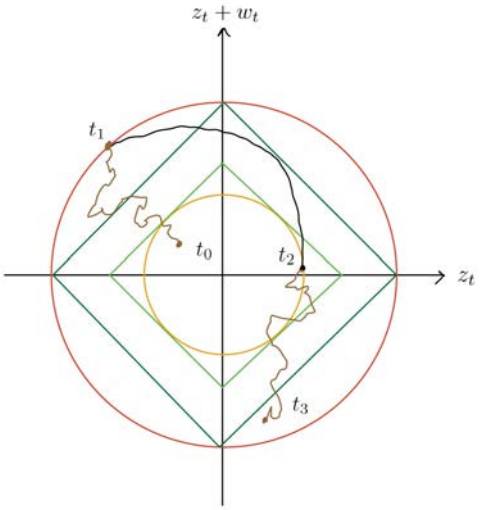
$$\mathcal{L}_t := \mu_k \cdot (f(r_t) - \xi_t) + (1 - \mu_k) \cdot \exp\left(-C_m(t - \tau_k)\right) \cdot (f(r_{\tau_k}) - \xi_{\tau_k}) - (\sigma_t + \phi_t), \tag{4.27}$$

where $k := \lfloor \frac{t}{\nu} \rfloor$. Note that $\mathcal{L}_t$ (the Lyapunov function at time t) depends on r_{τ_k} (at time τ_k). In Lemma 50, we demonstrate that this function contracts at a rate of $e^{-C_m t}$. The convergence bound then follows by showing that the convergence of this Lyapunov function implies convergence of the distributions in 1-Wasserstein distance.

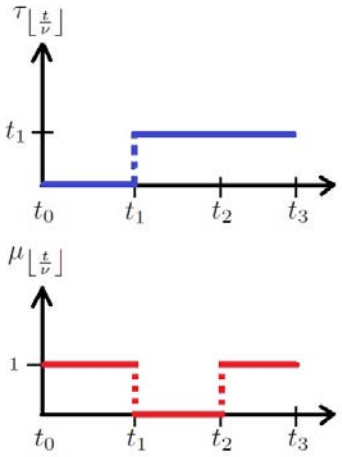
4.4.4 Proof Sketch

We present a full proof of Theorem 9 in Appendix C.4. In this section we provide a high-level sketch of our proof.

The proof proceeds by a path-wise analysis of the evolution of the Lyapunov function. In Figure 4.1a, we illustrate a sample path of the process when $d = 1$ (the $d > 1$ case is identical, but harder to draw on paper).



(a) Illustration of Coupling



(b) Update for τ and ν

First, let us highlight the features of the figure.

1. The red circle represents the set $\sqrt{\|z_t\|_2^2 + \|z_t + w_t\|_2^2} = \sqrt{5}R$. It affects the updates of $\tau_{\lfloor \frac{t}{\nu} \rfloor}$, which, in turn, dictates how the processes are coupled.

2. The orange circle represents $\sqrt{\|z_t\|_2^2 + \|z_t + w_t\|_2^2} = \frac{23}{50} \cdot \sqrt{5}R$. In relation to the red circle, it represents the contraction of $\sqrt{\|z_t\|_2^2 + \|z_t + w_t\|_2^2}$ when evolved according to synchronous coupling.

3. The space enclosed by the dark green diamond represents $(1 + 2c_\kappa)\|z_t\|_2 + \|z_t + w_t\|_2 \leq \sqrt{5}R$. It is is contained in the set $\sqrt{\|z_t\|_2^2 + \|z_t + w_t\|_2^2} \leq \sqrt{5}R$.

4. The space enclosed by the light green diamond represents $2((1 + 2c_\kappa)\|z_t\|_2 + \|z_t + w_t\|_2) \leq 2 \cdot \frac{23}{50}\sqrt{5}R$. It contains the set $(1 + 2c_\kappa)\|z_t\|_2 + \|z_t + w_t\|_2 \leq \frac{23}{50} \cdot \sqrt{5}R$.

5. It is not shown, but note that the red quarter circle is contained in $(1 + 2c_\kappa)\|z_t\|_2 + \|z_t + w_t\|_2 \leq \sqrt{12}R$, which is the radius $\mathcal{R}_f$ used for defining f in Eq. (4.23).

6. The brown squiggly lines $(t_0 \to t_1)$ and $(t_2 \to t_3)$ represent the evolution of the process under *reflection coupling*.

7. The black line $t_1 \to t_2$ represents the evolution of the process under *synchronous coupling*.

Below, we describe how (z_t, w_t) evolves over $t \in [t_0, t_3]$, and illustrate the main ideas behind the proof. To simplify matters, assume that

1. $k_i := t_i/\nu$ are integers, for $i = 0, 1, 2, 3$.
2. $t_3 - t_2 = T_{sync}$.
3. $\xi_t = \sigma_t = 0$ as these terms correspond to discretization errors.
4. $r_t \approx \|z_t\|_2 + (1 + 2c_\kappa)\|z_t + w_t\|_2$.

Then

- From $t_0 \to t_1$:
  Suppose that the process starts somewhere inside the red circle and stays inside for until time t_1, then $\tau_{\lfloor \frac{t}{\nu} \rfloor} = t_0$ and $\mu_{\lfloor \frac{t}{\nu} \rfloor} = 1$ for $t \in [t_0, t_1)$, and the process (z_t, w_t) undergoes reflection coupling.

  In this case, we can show that when $r_t \leq \sqrt{12}R$ then $f(r_t) - \phi_t$ contracts at a rate of $\exp(-C_m t)$ with probability one (see Lemma 33). This in turn implies that our Lyapunov function $\mathcal{L}_t$ also contracts at the same rate with probability one (see Lemma 53 and Lemma 54).

- From $t_1 \to t_2$:
  At $t = t_1$, we update τ_{k_1} so that $\tau_{k_1} = t_1$. Thus $\mu_s = 0$ for all $s \in [t_1, t_2)$. During this period, (z_t, w_t) evolves under synchronous coupling. In Lemma 37, we show that $\sqrt{\|z_{t_2}\|_2^2 + \|z_{t_2} + w_{t_2}\|_2^2} \leq \frac{23}{50}\sqrt{\|z_{t_1}\|_2^2 + \|z_{t_1} + w_{t_1}\|_2^2}$. This implies that $f(r_{t_2}) \leq e^{-C_m(t_2 - t_1)} f(r_{t_1})$ (Lemma 34). Again, this contraction is with probability one. Intuitively, we use synchronous coupling because when the value of $\|z_t\|_2 + \|z_t + w_t\|_2$ is large, Assumption (A3) guarantees contraction even in the absence of noise.

  This contraction in f consequently results in a contraction of the Lyapunov function (see Lemma 52).

- After a duration T_{sync} of synchronous coupling, we have $\mu_{k_2} = 1$ and we resume reflection coupling over $[t_2, t_3]$. Note that at $t = t_2$, the Lyapunov function $\mathcal{L}_t$, undergoes a jump in value, from $\exp(-C_m(t_2 - t_1))f(r_{t_1})$ to $f(r_{t_2})$ (see (4.27)). We show in Lemma 51 that this jump is negative with probability one.

4.5 Related Work

A convergence rate for overdamped Langevin diffusion, under assumptions (A1) – (A3) (see Section 4.2.1) has been established by [32], but the continuous-time diffusion studied in that paper is not implementable algorithmically. In a more algorithmic line of work, [20] bounded the discretization error of overdamped Langevin MCMC, and provided the first nonasymptotic convergence rate of overdamped Langevin MCMC under log-concavity assumptions. This was followed by a sequence of papers in the strongly log-concave setting [see, e.g., 16, 21, 27, 30].

Our result for overdamped Langevin MCMC is in line with this existing work; indeed, we combine the continuous-time convergence rate of Eberle [32] with a variant of the discretization error analysis by Durmus and Moulines [27]. The final number of timesteps needed is $\widetilde{\mathcal{O}}(e^{cLR^2}d/\varepsilon^2)$, which is expected, as the rate of [32] is $\mathcal{O}(e^{-cLR^2})$ (for the continuous-time process) and the iteration complexity established by [27] is $\widetilde{\mathcal{O}}(d/\varepsilon^2)$.

On the other hand, convergence of underdamped Langevin MCMC under (strongly) log-concave assumptions was first established by Cheng et al. [17]. Also very relevant to our results is the work of Eberle et al. [33], who demonstrated a contraction property of the continuous-time process stated in Eq. (4.2). That result deals, however, with a much larger class of potential functions, and accordingly the distance to the invariant distribution scales exponentially with dimension d. Our analysis yields a more favorable result by combining ideas from both Eberle et al. [33] and Cheng et al. [17], under new assumptions; see Section 4.3.2 for a full discussion.

Also noteworthy is the fact that the problem of sampling from non-log-concave distributions has been studied by [72], but under weaker assumptions, with a worst-case convergence rate that is exponential in d. In [85], this technique is used to study the application of Stochastic Gradient Langevin Diffusion (and its variance-reduced version) to non-convex optimization. Similarly, Durmus and Moulines [27] analyze the overdamped Langevin MCMC algorithm under the assumption that U is superlinear outside a ball. This is more general than our assumption of "strong convexity outside a ball"; in this setting, the authors prove a rate that is exponential in dimension. On the other hand, [37] established a $poly(d, 1/\varepsilon)$ convergence rate for sampling from a distribution that is close to a mixture of Gaussians, where the mixture components have the same variance (which is subsumed by our assumptions).

Finally, there is a large class of sampling algorithms known as Hamiltonian Monte Carlo (HMC), which involve Hamiltonian dynamics in some form. We refer to Ma et al. [56] for a survey of the results in this area. Among these, the variant studied in this paper (Algorithm 2), based on the discretization of the SDE in Eq. (4.2), has a natural physical interpretation as the evolution of a particle's dynamics under a viscous force field. This model was first studied by Kramers [49] in the context of chemical reactions. The continuous-time process has been studied extensively [4, 6, 11, 24, 33, 39, 42, 61, 83]. Four recent papers—Mangoubi and Smith [59], Lee and Vempala [52], Mangoubi and Vishnoi [60] and [23]—study the convergence rate of (variants of) HMC under log-concavity assumptions. In [34], the authors study the convergence of HMC on general metric state spaces. Bou-Rabee et al. [8] study the convergence of HMC under assumptions similar to ours, and prove a convergence rate that depends on e^{cLR^2} for some constant c. We remark that the algorithm studied in this case is different from the underdamped Langevin MCMC algorithm, because of the incorporation of an accept-reject step.

Part III

Optimization as Sampling

Chapter 5

Stochastic Gradient and Langevin Processes

5.1 Introduction

Stochastic Gradient Descent (SGD) is one of the workhorses of modern machine learning. In many non-convex optimization problems, such as training deep neural networks, SGD is able to produce solutions with good generalization error; indeed, there is evidence that the generalization error of an SGD solution can be significantly better than that of Gradient Descent (GD) [41, 44, 47]. This suggests that, to understand the behavior of SGD, it is not enough to consider the limiting cases such as small step size or large batch size where it degenerates to GD. In this paper, we take an alternate view of SGD as a sampling algorithm, and aim to understand its convergence to an appropriate stationary distribution.

There has been rapid recent progress in understanding the finite-time behavior of MCMC methods, by comparing them to stochastic differential equations (SDEs), such as the Langevin diffusion. It is natural in this context to think of SGD as a discrete-time approximation of an SDE. There are, however, two significant barriers to extending previous analyses to the case of SGD. First, these analysis are often restricted to isotropic Gaussian noise, whereas the noise in SGD can be far from Gaussian. Second, the noise depends significantly on the current state (the optimization variable). For instance, if the objective is an average over training data with a nonnegative loss, as the objective approaches zero the variance of minibatch SGD goes to zero. Any attempt to cast SGD as an SDE must be able to handle this kind of noise.

This motivates the study of Langevin MCMC-like methods that have a state-dependent noise term:

$$w_{(k+1)\delta} = w_{k\delta} - \delta \nabla U(w_{k\delta}) + \sqrt{\delta}\xi(w_{k\delta}, \eta_k), \tag{5.1}$$

where $w_t \in \mathbb{R}^d$ is the state variable at time t, δ is the step size, $U : \mathbb{R}^d \to \mathbb{R}$ is a (possibly non-convex) potential, $\xi : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ is the *noise function*, and η_k are sampled i.i.d. according to some distribution over Ω (for example, in minibatch SGD, Ω is the set of subsets of indices in the training sample).

Throughout this paper, we assume that $\mathbb{E}_\eta[\xi(x,\eta)] = 0$ for all x. We define a matrix-valued function $M(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ to be the square root of the covariance matrix of ξ; i.e., for all x, $M(x) := \sqrt{\mathbb{E}_\eta[\xi(x,\eta)\xi(x,\eta)^T]}$, where for a positive semidefinite matrix G, $A = \sqrt{G}$ is the unique positive semidefinite matrix such that $A^2 = G$.

In studying the generalization behavior of SGD, earlier work [41, 44] propose that (5.1) be approximated by the stochastic process $y_{(k+1)\delta} = y_{k\delta} - \delta\nabla U(y_{k\delta}) + \sqrt{\delta}M(y_{k\delta})\theta_k$ where $\theta_k \sim \mathcal{N}(0, I)$, or, equivalently:

$$\begin{aligned} dy_t &= -\nabla U(y_{k\delta})dt + M(y_{k\delta})dB_t \\ &\quad \text{for } t \in [k\delta, (k+1)\delta], \end{aligned} \tag{5.2}$$

with B_t denoting standard Brownian motion [46]. Specifically, the non-Gaussian noise $\xi(\cdot,\eta)$ is approximated by a Gaussian variable $M(\cdot)\theta$ with the same covariance, via an assumption that the minibatch size is large and an appeal to the central limit theorem.

The process in (5.2) can be seen as the Euler-Murayama discretization of the following SDE:

$$dx_t = -\nabla U(x_t)dt + M(x_t)dB_t. \tag{5.3}$$

We let p^* denote the invariant distribution of (5.3). Note the similarity between (5.3) and (1.4); the key difference is in $M(x_t)$, the diffusion coefficient.

We prove quantitative bounds on the discretization error between (5.2), (5.1) and (5.3), as well as convergence rates of (5.2) and (5.1) to p^*. Our bounds are in 1-Wasserstein distance (denoted by $W_1(\cdot,\cdot)$ in the following). We present the full theorem statements in Section 5.4, and summarize our contributions below:

1. In Theorem 10, we bound the discretization error between (5.2) and (5.3). Informally, Theorem 10 states:

   1. *If $x_0 = y_0$, then for all k, $W_1(x_{k\delta}, y_{k\delta}) = O(\sqrt{\delta})$;*
   2. *For $n \geq \tilde{O}\left(\frac{1}{\delta}\right)$, $W_1(p^*, \mathsf{Law}(y_{n\delta})) = O(\sqrt{\delta})$,*

   where $\mathsf{Law}(\cdot)$ denotes the distribution of a random vector. This is a crucial intermediate result that allows us to prove the convergence of (5.1) to (5.3). We highlight that the variable diffusion matrix: 1) leads to a very large discretization error, due to the scaling factor of $\sqrt{\delta}$ in the $M(y_{k\delta})\theta_k$ noise term, and 2) makes the stochastic process non-contractive (this is further compounded by the non-convex drift). Our convergence proof relies on a carefully constructed Lyapunov function together with a specific coupling. Remarkably, the ε dependence in our iteration complexity is the same as that in Langevin MCMC with constant isotropic diffusion [29].

2. In Theorem 11, we bound the discretization error between (5.1) and (5.3). Informally, Theorem 11 states:

   1. *If* $x_0 = w_0$*, then for all* k*,* $W_1(x_{k\delta}, w_{k\delta}) = O(\delta^{1/8})$;
   2. *For* $n \geq \tilde{O}\left(\frac{1}{\delta}\right)$, $W_1(p^*, \mathsf{Law}(w_{n\delta})) = O(\delta^{1/8})$.

   Notably, the noise in each step of (5.1) may be far from Gaussian, but for sufficiently small step size, (5.1) is nonetheless able to approximate (5.3). This is a weaker condition than earlier work, which must assume that the batch size is sufficiently large so that CLT ensures that the per-step noise is approximately Gaussian.

3. Based on Theorem 10, we predict that for sufficiently small δ, two different processes of the form (5.1) will have similar distributions if their noise terms ξ have the same covariance matrix, as that leads to the same limiting SDE (5.3). In Section 5.5, we evaluate this claim empirically: we design a family of SGD-like algorithms and evaluate their test error at convergence. We observe that the noise covariance alone is a very strong predictor for the test error, regardless of higher moments of the noise. This corroborates our theoretical prediction that the noise covariance approximately determines the distribution of the solution. This is also in line with, and extends upon, observations in earlier work that the ratio of batch size to learning rate correlates with test error [41, 44].

5.2 Motivating Example

It is generally difficult to write down the invariant distribution of (5.3). In this section, we consider a very simple one-dimensional setting which does admit an explicit expression for p^*, and serves to illustrate some remarkable properties of anisotropic diffusion matrices.

Let us define $D(x) := M^2(x)$. Our analysis will be based on the Fokker-Planck equation, which states that p^* is the invariant distribution of (5.3) if

$$0 = \mathbf{div}(p^*(x)\nabla U(x)) + \mathbf{div}(p^*(x)\Gamma(x) + D(x)\nabla p^*(x)), \tag{5.4}$$

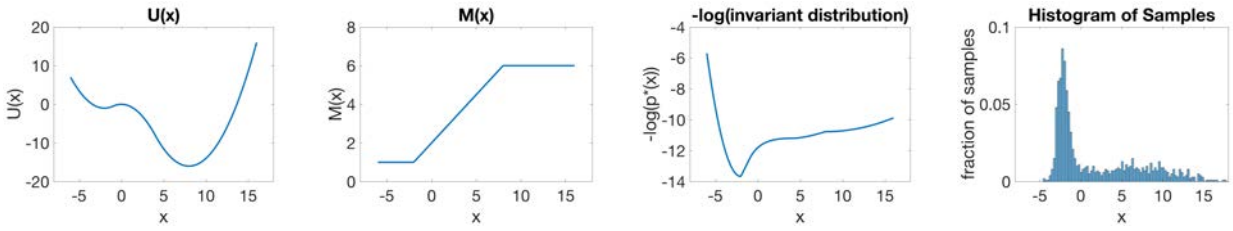
where $\Gamma(x)$ is a vector whose i^{th} coordinate equals $\sum_{j=1}^{d} \frac{\partial}{\partial x_j}[D(x)]_{i,j}$. In the one-dimensional setting, we can explicitly write down the density of $p^*(x)$. Note that in this case, $\Gamma(x) = \nabla D(x)$. Let $V(x) := \int_0^x \left(\frac{\nabla U(x)}{D(x)} + \frac{\nabla D(X)}{D(x)}\right)dx = \int_0^x \left(\frac{\nabla U(x)}{D(x)}\right)dx + \log D(x) - \log D(0)$. We can verify that $p^*(x) \propto e^{-V(x)}$ satisfies (5.4).

For a concrete example, let the potential $U(x)$ and the diffusion function $M(x)$ be defined

as

$$U(x) := \begin{cases} \frac{1}{2}x^2, & \text{for } x \in [-1, 4] \\ \frac{1}{2}(x+2)^2 - 1, & \text{for } x \leq -1 \\ \frac{1}{2}(x-8)^2 - 16, & \text{for } x \geq 4 \end{cases}$$

$$M(x) = \begin{cases} \frac{1}{2}(x+2), & \text{for } x \in [-2, 8] \\ 1, & \text{for } x \leq -2 \\ 6, & \text{for } x \geq 8 \end{cases}.$$



(a) $U(x)$ (b) $M(x)$ (c) $V(x)$ (d) Samples

Figure 5.1: One-dimensional example exhibiting the importance of state-dependent noise: A simple construction showing how $M(x)$ can affect the shape of the invariant distribution. While $U(x)$ has two local minima, $V(x)$ only has the smaller minimum at $x = -2$. Figure 5.1d represents samples obtained from simulating using the process (5.2). We can see that most of the samples concentrate around $x = -2$.

We plot $U(x)$ in Figure 5.1a. Note that $U(x)$ has two local minima: a shallow minimum at $x = -2$ and a deeper minimum at $x = 8$. A plot of $M(x)$ can be found in Figure 5.1b. $M(x)$ is constructed to have increasing magnitude at larger values of x. This has the effect of biasing the invariant distribution towards smaller values of x.

We plot $V(x)$ in Figure 5.1c. Remarkably, $V(x)$ has only one local minimum at $x = -2$. The larger minimum of $U(x)$ at $x = 8$ has been smoothed over by the effect of the large diffusion $M(x)$. This is very different from when the noise is homogeneous (e.g., $M(x) = I$), in which case $p^*(x) \propto e^{-U(x)}$. We also simulate (5.3) (using (5.2)) for the given $U(x)$ and $M(x)$ for 1000 samples (each simulated for 1000 steps), and plot the histogram in Figure 5.1d.

5.3 Assumptions and Definitions

In this section, we state the assumptions and definitions that we need for our main results in Theorem 10 and Theorem 11.

Assumption A *We assume that $U(x)$ satisfies*

1. *The function $U(x)$ is continuously-differentiable on $\mathbb{R}^d$ and has Lipschitz continuous gradients; that is, there exists a positive constant $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla U(x) - \nabla U(y)\|_2 \leq L\|x - y\|_2$.*

2. *U has a stationary point at zero: $\nabla U(0) = 0$.*

3. *There exist positive constants m, L_R, R such that for all $\|x - y\|_2 \geq R$,*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m\|x - y\|_2^2. \tag{5.5}$$

and for all $\|x - y\|_2 \leq R$, $\|\nabla U(x) - \nabla U(y)\|_2 \leq L_R\|x - y\|_2$.

Remark 7 *The assumption in (5.5) roughly states that "$U(x)$ is convex outside a ball of radius R". This assumption, and minor variants, is common in the non-convex sampling literature [18, 32, 32, 36, 39, 57].*

Assumption B *We make the following assumptions on ξ and M:*

1. *For all x, $\mathbb{E}[\xi(x, \eta)] = 0$.*

2. *For all x, $\|\xi(x, \eta)\|_2 \leq \beta$ almost surely.*

3. *For all x, y, $\|\xi(x, \eta) - \xi(y, \eta)\|_2 \leq L_\xi\|x - y\|_2$ almost surely.*

4. *There is a positive constant c_m such that for all x, $2c_m I \prec M(x)$.*

Remark 8 *We discuss these assumptions in a specific setting in Section 5.5.2.*

For convenience we define a matrix-valued function $N(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$:

$$N(x) := \sqrt{M(x)^2 - c_m^2 I}. \tag{5.6}$$

Under Assumption A, we can prove that $N(x)$ and $M(x)$ are bounded and Lipschitz (see Lemma 79 and 80 in Appendix D.4). These properties will be crucial in ensuring convergence.

Recall that, given an arbitrary sample space Ω and any two distribution $p \in \mathscr{P}(\Omega)$ and $q \in \mathscr{P}(\Omega)$, a joint distribution $\zeta \in \mathscr{P}(\Omega \times \Omega)$ is a *coupling* between p and q if its marginals are equal to p and q respectively.

For a matrix, we use $\|G\|_2$ to denote the operator norm: $\|G\|_2 = \sup_{v\in\mathbb{R}^d, \|v\|_2=1} \|Gv\|_2$.

Finally, we define a few useful constants which will be used throughout the rest of the chapter:

$$\begin{aligned}
L_N &:= \frac{4\beta L_\xi}{c_m}, \quad \alpha_q := \frac{L_R + L_N^2}{2c_m^2}, \\
\mathcal{R}_q &:= \max\left\{R, \frac{16\beta^2 L_N}{m \cdot c_m}\right\} \\
\lambda &:= \min\left\{\frac{m}{2}, \frac{2c_m^2}{32\mathcal{R}_q{}^2}\right\} \exp\left(-\frac{7}{3}\alpha_q \mathcal{R}_q{}^2\right).
\end{aligned} \tag{5.7}$$

L_N is the smoothness parameter of the matrix $N(x)$, and we show in Lemma 80 that $\mathrm{tr}\big((N(x)-N(y))^2\big) \le L_N^2\|x-y\|_2^2$. The constants α_q and $\mathcal{R}_q$ are used to define a Lyapunov function q in Appendix D.5.1. A key step in our proof uses the fact that, under the dynamics (5.2), q contracts at a rate of $e^{-\lambda}$, plus discretization error.

5.4 Main Results

In this section, we present our main convergence results beginning with convergence under Gaussian noise and proceeding to the non-Gaussian case.

Theorem 10 *Let x_t and y_t have dynamics as defined in (5.3) and (5.2) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|x_0\|_2^2\right] \le R^2+\beta^2/m$ and $\mathbb{E}\left[\|y_0\|_2^2\right] \le R^2+\beta^2/m$. Let $\hat{\varepsilon}$ be a target accuracy satisfying $\hat{\varepsilon} \le \left(\frac{16\left(L+L_N^2\right)}{\lambda}\right)\cdot \exp\left(7\alpha_q\mathcal{R}_q/3\right)\cdot \frac{\mathcal{R}_q}{\alpha_q{\mathcal{R}_q}^2+1}$. Let δ be a step size satisfying*

$$\delta \le \min\left\{ \begin{array}{l} \frac{\lambda^2\hat{\varepsilon}^2}{512\beta^2\left(L^2+L_N^4\right)\exp\left(\frac{14\alpha_q{\mathcal{R}_q}^2}{3}\right)} \\ \frac{2\lambda\hat{\varepsilon}}{\left(L^2+L_N^4\right)\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\sqrt{R^2+\beta^2/m}} \end{array}\right. .$$

If we assume that $x_0 = y_0$, then there exists a coupling between x_t and y_t such that for any k,

$$\mathbb{E}\left[\|x_{k\delta}-y_{k\delta}\|_2\right] \le \hat{\varepsilon}$$

Alternatively, if we assume $n \ge \frac{3\alpha_q{\mathcal{R}_q}^2}{\delta}\log\frac{R^2+\beta^2/m}{\hat{\varepsilon}}$, then

$$W_1(p^*, p^y_{n\delta}) \le 2\hat{\varepsilon}$$

where $p_t^y := \mathsf{Law}(y_t)$.

Remark 9 *Note that m, L, R are from Assumption A, L_N is from (5.7), c_m, β, L_ξ are from Assumption B)*

Remark 10 *Finding a suitable y_0 can be done very quickly using gradient descent wrt $U(\cdot)$. The convergence rate to the ball of radius R is very fast, due to Assumption A.3.*

After some algebraic simplifications, we see that for a sufficiently small $\hat{\varepsilon}$, achieving $W_1(p^y_{n\delta}, p^*) \le \hat{\varepsilon}$ requires number of steps

$$n = \tilde{O}\left(\frac{\beta^2}{\hat{\varepsilon}^2}\cdot\exp\left(\frac{14}{3}\cdot\left(\frac{L_R}{c_m^2}+\frac{16\beta^2L_\xi^2}{c_m^4}\right)\right.\right.$$
$$\left.\left.\cdot\max\left\{R^2, \frac{2^{12}\beta^6L_\xi^2}{m^2c_m^4}\right\}\right)\right).$$

Remark 11 *The convergence rate contains a term e^{R^2}; this term is also present in all of the work cited in the previous section under Remark 1. Given our assumptions, including the "convexity outside a ball of radius R" assumption, this dependence is unavoidable as it describes the time to transit between two modes of the invariant distribution for a simple double-well potential.*

Remark 12 *As illustrated in Section 5.5.2, the m from Assumption B.3 should be thought of as a regularization term which can be set arbitrarily large. In the following discussion, we will assume that* $\max\left\{R^2, \frac{\beta^6 L_\xi^2}{m^2 c_m^4}\right\}$ *is dominated by the* R^2 *term.*

To gain intuition about this term, let's consider what it looks like under a sequence of increasingly weak assumptions:

a. Strongly convex, constant noise: $U(x)$ m strongly convex, L smooth, $\xi(x,\eta) \sim \mathcal{N}(0,I)$ for all x. (In reality we need to consider a truncated Gaussian so as not to violate Assumption B.2, but this is a minor issue). In this case, $L_\xi = 0$, $c_m = 1$, $R = 0$, $\beta = \tilde{O}(\sqrt{d})$, so $k = O(\frac{d}{\hat{\varepsilon}^2})$. This is the same rate as obtained by [29]. We remark that [29] obtains a W_2 bound which is stronger than our W_1 bound.

b. Non-convex, constant noise: $U(x)$ not strongly convex but satisfies Assumption A, and $\xi(x,\eta) \sim \mathcal{N}(0,I)$. In this case, $L_\xi = 0$, $c_m = 1$, $\beta = \tilde{O}(\sqrt{d})$ This is the setting studied by [18] and [57]. The rate we recover is $k = \tilde{O}\Big(\frac{d}{\hat{\varepsilon}^2} \cdot \exp\left(\frac{14}{3} L R^2\right)\Big)$, which is in line with [18], and is the best W_1 rate obtainable from [57].

c. Non-convex, state-dependent noise: $U(x)$ satisfies Assumption A, and ξ satisfies Assumption B. To simplify matters, suppose the problem is rescaled so that $c_m = 1$. Then the main additional term compared to setting b. above is $\exp\left(\frac{64\beta^2 L_\xi^2 R^2}{c_m^4}\right)$. This suggests that the effect of a L_ξ-Lipschitz noise can play a similar role in hindering mixing as a L_R-Lipschitz non-convex drift.

When the dimension is high, computing $M(y_k)$ can be difficult, but if for each x, one has access to samples whose covariance is $M(x)$, then one can approximate $M(y_k)\theta_k$ via the central limit theorem by drawing a sufficiently large number of samples. The proof of Theorem 10 can be readily modified to accommodate this (see Appendix D.1.5).

We now turn to the non-Gaussian case.

Theorem 11 *Let x_t and w_t have dynamics as defined in* (5.3) *and* (5.1) *respectively, and suppose that the initial conditions satisfy* $\mathbb{E}\left[\|x_0\|_2^2\right] \le R^2 + \beta^2/m$ *and* $\mathbb{E}\left[\|w_0\|_2^2\right] \le R^2 + \beta^2/m$. *Let $\hat{\varepsilon}$ be a target accuracy satisfying* $\hat{\varepsilon} \le \left(\frac{16\left(L+L_N^2\right)}{\lambda}\right) \cdot \exp\left(7\alpha_q \mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q {\mathcal{R}_q}^2 + 1}$. *Let* $\varepsilon := \frac{\lambda}{16(L+L_N^2)} \exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\hat{\varepsilon}$.

Let $T := \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}, \frac{\varepsilon}{32\sqrt{L}\beta}, \frac{\varepsilon^2}{128\beta^2}, \frac{\varepsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$ *and let* δ *be a step size satisfying*

$$\delta \leq \min\left\{\frac{T\varepsilon^2 L}{36d\beta^2 \log\left(\frac{36d\beta^2}{\varepsilon^2 L}\right)}, \frac{T\varepsilon^4 L^2}{2^{14}d\beta^4 \log\left(\frac{2^{14}d\beta^4}{\varepsilon^4 L^2}\right)}\right\}.$$

If we assume that $x_0 = w_0$*, then there exists a coupling between* x_t *and* w_t *such that for any* k*,*

$$\mathbb{E}\left[\|x_{k\delta} - w_{k\delta}\|_2\right] \leq \hat{\varepsilon}.$$

Alternatively, if we assume that $n \geq \frac{3\alpha_q {\mathcal{R}_q}^2}{\delta} \cdot \log\frac{R^2+\beta^2/m}{\hat{\varepsilon}}$*, then*

$$W_1(p^*, p^w_{n\delta}) \leq 2\hat{\varepsilon},$$

where $p^w_t := \mathsf{Law}(w_t)$*.*

Remark 13 *To achieve* $W_1(p^*, p^w_{n\delta}) \leq \hat{\varepsilon}$*, the number of steps needed is of order* $n = \tilde{O}\Big(\frac{1}{\hat{\varepsilon}^8} \cdot e^{29\alpha_q {\mathcal{R}_q}^2}\Big)$*. The* $\hat{\varepsilon}$ *dependency is considerably worse than in Theorem 10. This is because we need to take many steps of* (5.1) *in order to approximate a single step of* (5.2)*. For details, see the coupling construction in Equations* (D.8)–(D.12) *of Appendix D.2.*

5.5 Application to Stochastic Gradient Descent

In this section, we will cast SGD in the form of (5.1). We consider an objective of the form

$$U(w) = \frac{1}{n}\sum_{i=1}^{n} U_i(w). \tag{5.8}$$

We reserve the letter η to denote a random minibatch from $\{1, \ldots, n\}$, sampled with replacement, and define $\zeta(w, \eta)$ as follows:

$$\zeta(w, \eta) := \nabla U(w) - \frac{1}{|\eta|}\sum_{i\in\eta} \nabla U_i(w) \tag{5.9}$$

For a sample of size one, i.e. $|\eta| = 1$, we define

$$H(w) := \mathbb{E}\left[\zeta(w, \eta)\zeta(w, \eta)^T\right] \tag{5.10}$$

as the covariance matrix of the difference between the true gradient and a single sampled gradient at w. A standard run of SGD, with minibatch size $b := |\eta_k|$, then has the following form:

$$\begin{aligned} w_{k+1} &= w_k - \delta\frac{1}{b}\sum_{i\in\eta_k} \nabla U_i(w_k) \\ &= w_k - \delta\nabla U(w_k) + \sqrt{\delta}\Big(\sqrt{\delta}\zeta(w_k, \eta_k)\Big). \end{aligned} \tag{5.11}$$

We refer to an SGD algorithm with step size δ and minibatch size b a (δ, b)-SGD. Notice that (5.11) is in the form of (5.1), with $\xi(w, \eta) = \sqrt{\delta}\zeta(w, \eta)$. The covariance matrix of the noise term is

$$\mathbb{E}\left[\xi(w,\eta)\xi(w,\eta)^T\right] = \frac{\delta}{b}H(w). \tag{5.12}$$

Because the magnitude of the noise covariance scales with $\sqrt{\delta}$, it follows that as $\delta \to 0$, (5.11) converges to deterministic gradient flow. However, the loss of randomness as $\delta \to 0$ is not desirable as it has been observed that as SGD approaches GD, through either small step size or large batch size, the generalization error goes up [41, 43, 44, 47]; this is also consistent with our experimental observations in Section 5.5.3.1.

Therefore, a more meaningful way to take the limit of SGD is to hold the noise term constant in (5.11). More specifically, we define the ***constant-noise limit*** of (5.11) as

$$dx_t = -\nabla U(x_t)dt + M(x_t)dB_t, \tag{5.13}$$

where $M(x) := \sqrt{\frac{\delta}{b}H(x)}$. Note that this is in the form of (5.3), with noise covariance $M(x_t)^2$ matching that of SGD in (5.11). Using Theorem 11, we can bound the W_1 distance between the SGD iterates w_k from (5.11), and the continuous-time SDE x_t from (5.13).

5.5.1 Importance of Noise Covariance

We highlight the fact that the limiting SDE of a discrete process,

$$w_{k+1} = w_k - s\nabla U(w_k) + \sqrt{s}\xi(w_k, \eta_k), \tag{5.14}$$

depends only on the covariance matrix of ξ. More specifically, as long as ξ satisfies $\sqrt{\mathbb{E}\left[\xi(w,\eta)\xi(w,\eta)^T\right]} = M(w)$, (5.14) will have (5.13) as its limiting SDE, *regardless of higher moments of* ξ. This fact, combined with Theorem 11, means that in the limit of $\delta \to 0$ and $k \to \infty$, the distribution of w_k will be determined by the covariance of ξ alone. An immediate consequence is the following: *at convergence, the test performance of any Langevin MCMC-like algorithm is almost entirely determined by the covariance of its noise term.*

Returning to the case of SGD algorithms, since the noise covariance is $M(x)^2 = \frac{\delta}{b}H(x)$ (see (5.12)), we know that the ratio of step size δ to batch size b is an important quantity which can dictate the test error of the algorithm; this observation has been made many times in prior work [41, 44], and our results in this paper are in line with these observations. Here, we move one step further, and provide experimental evidence to show that more fundamentally, it is the noise covariance in the constant-noise limit that controls the test error.

To verify this empirically, we propose the following algorithm called *large-noise SGD*.

Definition 1 *An (s, σ, b_1, b_2)-large-noise SGD is an algorithm that aims to minimize* (5.8) *using the following updates:*

$$w_{k+1} = w_k - \frac{s}{b_1} \sum_{i \in \eta_k} \nabla U_i(w_k) + \frac{\sigma\sqrt{s}}{b_2} \left(\sum_{i \in \eta_k'} \nabla U_i(w_k) - \sum_{i \in \eta_k''} \nabla U_i(w_k) \right), \tag{5.15}$$

where η_k, η_k', and η_k'' are minibatches of sizes b_1, b_2, and b_2, sampled uniformly at random from $\{1, \ldots, n\}$ with replacement. The three minibatches are sampled independently and are also independent of other iterations.

Intuitively, an (s, σ, b_1, b_2)-large-noise SGD should be considered as an SGD algorithm with step size s and minibatch size b_1 and an additional noise term. The noise term computes the difference of two independent and unbiased estimates of the full gradient $\nabla U(w_k)$, each using a batch of b_2 data points. Using the definition of ζ in (5.9), we can verify that the update (5.15) is equivalent to

$$w_{k+1} = w_k - s\nabla U(w_k) + s\zeta(w_k, \eta_k) + \sigma\sqrt{s}(\zeta(w_k, \eta_k'') - \zeta(w_k, \eta_k')), \tag{5.16}$$

which is in the form of (5.1), with

$$\xi(w, \tilde{\eta}) = \sqrt{s}\zeta(w, \eta) + \sigma(\zeta(w, \eta'') - \zeta(w, \eta')), \tag{5.17}$$

where $\tilde{\eta} = (\eta, \eta', \eta'')$, and $|\eta| = b_1$, $|\eta'| = |\eta''| = b_2$. Further, the noise covariance matrix is

$$\mathbb{E}\left[\xi(w, \tilde{\eta})\xi(w, \tilde{\eta})^T\right] = (\frac{s}{b_1} + \frac{2\sigma^2}{b_2})H(w). \tag{5.18}$$

Therefore, if we have

$$\frac{s}{b_1} + \frac{2\sigma^2}{b_2} = \frac{\delta}{b}, \tag{5.19}$$

then an (s, σ, b_1, b_2)-large-noise SGD should have the same noise covariance as a (δ, b)-SGD *(but very different higher noise moments due to the injected noise)*, and based on our theory, the large-noise SGD should have similar test error to that of the SGD algorithm, even if the step size and batch size are different. In Section 5.5.3, we verify this experimentally. We stress that we are not proposing the large-noise SGD as a practical algorithm. The reason that this algorithm is interesting is that it gives us a family of $(w_k)_{k=1,2,\ldots}$ which converges to (5.13), and is implementable in practice. Thus this algorithm helps us uncover the importance of noise covariance (and the unimportance of higher noise moments) in Langevin MCMC-like algorithms. We also remark that [43] proposed a different way of injecting noise, multiplying the sampled gradient with a suitably scaled Gaussian noise. This has a similar effect on maintain the noise covariance independent of changes to step-size/batch-size; they observed that under this injection, generalization error remained almost constant across different batch sizes.

5.5.2 Satisfying the Assumptions

Before presenting the experimental results, we remark on a particular way that a function $U(w)$ defined in (5.8), along with the stochastic sequence w_k defined in (5.15), can satisfy the assumptions in Section 5.3.

Suppose first that we shift the coordinate system so that $\nabla U(0) = 0$. Let us additionally assume that for each i, $U_i(w)$ has the form

$$U_i(w) = U_i'(w) + V(w),$$

where $V(w) := m(\|x\|_2 - R/2)^2$ is a m strongly convex regularizer outside a ball of radius R, and each $U_i'(w)$ has L_R-Lipschitz gradients. Suppose further that $m \geq 4 \cdot L_R$. These additional assumptions make sense when we are only interested in $U(w)$ over $B_R(0)$, so $V(w)$ plays the role of a barrier function that keeps us within $B_R(0)$. Then, it can immediately be verified that $U(w)$ satisfies Assumption A with $L = m + L_R$.

The noise term ξ in (5.17) satisfies Assumption B.1 by definition, and satisfies Assumption B.3 with $L_\xi = (\sqrt{s} + 2\sigma)L$. Assumption B.2 is satisfied if $\zeta(w, \eta)$ is bounded for all w, i.e. the sampled gradient does not deviate from the true gradient by more than a constant. We will need to assume directly Assumption B.4, as it is a property of the distribution of $\nabla U_i(w)$ for $i = 1, \ldots, n$.

5.5.3 Experiments

In this section, we present experimental results that validate the importance of noise covariance in predicting the test error of Langevin MCMC-like algorithms. In all experiments, we use two different neural network architectures on the CIFAR-10 dataset [51] with the standard test-train split. The first architecture is a simple convolutional neural network, which we call CNN in the following, and the other is the VGG19 network [79]. To make our experiments consistent with the setting of SGD, we do not use batch normalization or dropout, and use constant step size. In all of our experiments, we run SGD algorithm 2000 epochs such that the algorithm converges sufficiently. Since in most of our experiments, the accuracies on the training dataset are almost 100%, we use the test accuracy to measure the generalization performance.

Recall that according to (5.12) and (5.18), for both SGD and large-noise SGD, the noise covariance is a scalar multiple of $H(w)$. For simplicity, in the following, we will slightly abuse our terminology and call this scalar the *noise covariance*; more specifically, for (δ, b)-SGD, the noise covariance is δ/b, and for an (s, σ, b_1, b_2)-large-noise SGD, the noise covariance is $\frac{s}{b_1} + \frac{2\sigma^2}{b_2}$.

5.5.3.1 Accuracy vs Noise Covariance

In our first experiment, we focus on the SGD algorithm, and show that there is a positive correlation between the noise covariance and the final test accuracy of the trained model. One

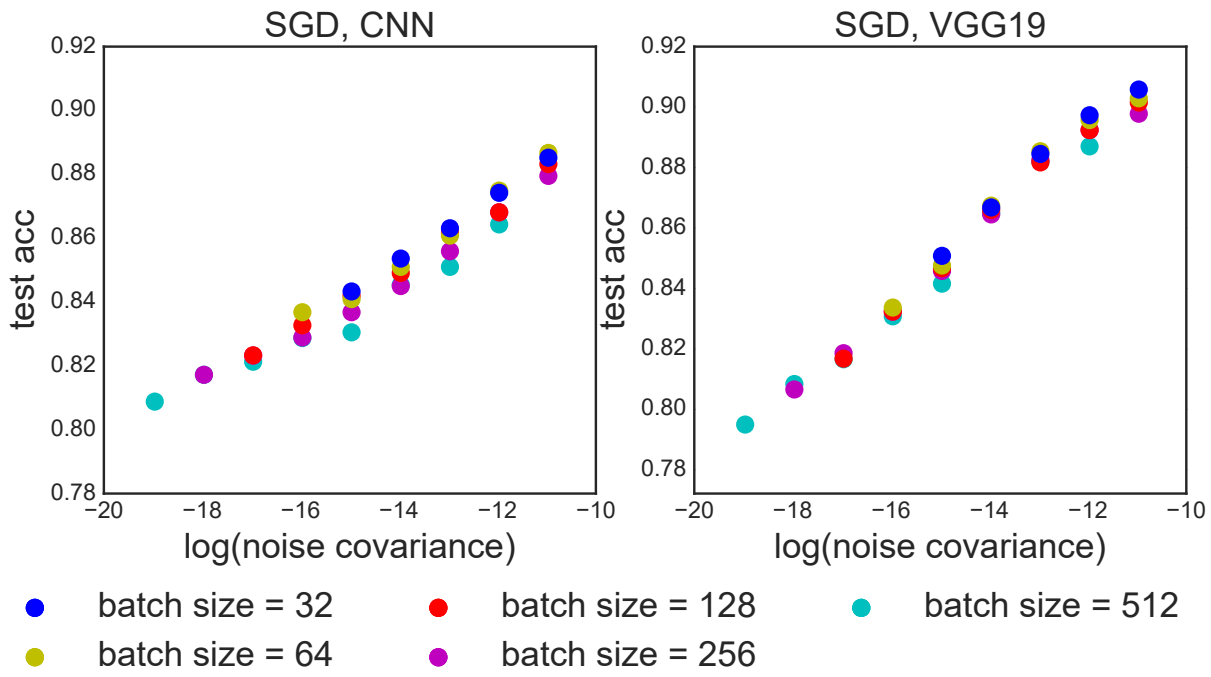


Figure 5.2: Relationship between test accuracy and the noise covariance of SGD algorithm. In each plot, the dots with the same color correspond to SGD runs with the same batch size but different step sizes.

major purpose of this experiment is to establish baselines for our experiments on large-noise SGD.

We choose constant step size δ from

$$\{0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128\}$$

and minibatch size b from $\{32, 64, 128, 256, 512\}$. For each (step size, batch size) pair, we plot its final test accuracy against its noise covariance in Figure 5.2. From the plot, we can see that higher noise covariance leads to better final test accuracy, and there is a linear trend between the test accuracy and the logarithm. We also highlight the fact that conditioned on the noise covariance, the test accuracy is not significantly correlated with either the step size or the minibatch size. In other words, similar to the observations in prior work [41, 44], there is a strong correlation between relative variance of an SGD sequence and its test accuracy, regardless of the combination of minibatch size and step size.

5.5.3.2 Large-Noise SGD

In this section, we implement and examine the performance of the large-noise SGD algorithm proposed in (5.15). We select a subset of SGD runs with relatively small noise covariance in the experiment in the previous section (we call them *baseline SGD runs*), and implement

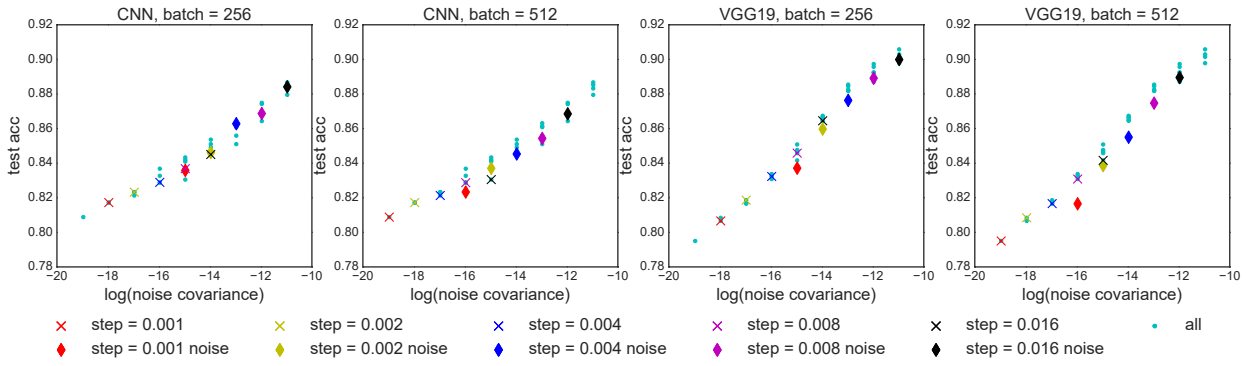


Figure 5.3: Large-noise SGD. Small dots correspond to all the baseline SGD runs in Figure 5.2. Each $\times$ corresponds to a baseline SGD run whose step size is specified in the legend and batch size is specified in the title. Each $\diamond$ corresponds to a large-noise SGD run whose noise covariance is 8 times that of the $\times$ with the same color. As we can see, injecting noise improves test accuracy, and the large-noise SGD runs fall close to the linear trend.

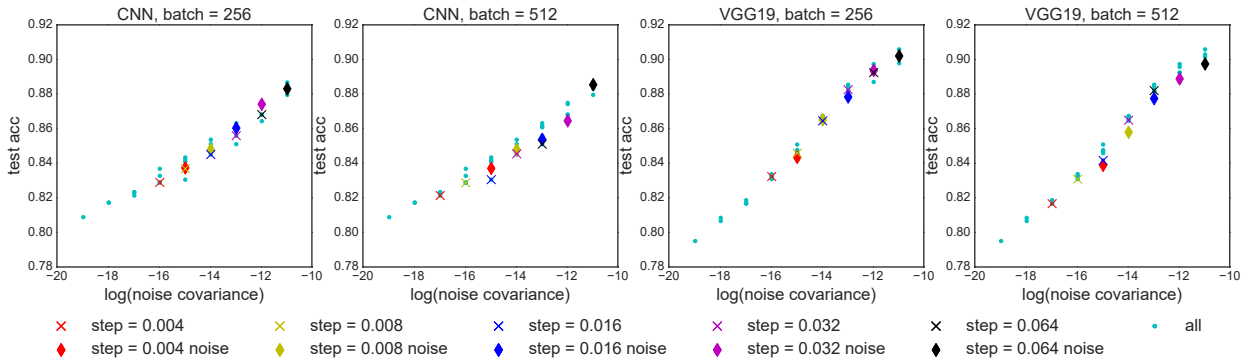


Figure 5.4: Large-noise SGD. Batch size in the titles represents the batch size of $\times$ runs. Each $\diamond$ corresponds to a large-noise SGD run whose noise covariance matches that of a baseline SGD run whose step size is the same as the $\times$ run with the same color and batch size is 128. Again, large-noise SGD falls close to the linear trend.

large-noise SGD by injecting noise. Our goal is to see, for a particular noise covariance, whether large-noise SGD has test accuracy that is similar to SGD, *in spite of significant differences in third-and-higher moments of the noise in large-noise SGD compared to standard SGD*.

Our first experiment is to add noise with the same minibatch size to the (δ, b) baseline SGD run such that the new noise covariance matches that of an $(8\delta, b)$-SGD (an SGD run with larger step size). In other words, we implement $(\delta, \sqrt{7\delta/2}, b, b)$-large-noise SGD, whose noise covariance is 8 times that of the baseline. Our results are shown in Figure 5.3. Our second experiment is similar: we add noise with minibatch size 128 to the (δ, b) baseline SGD run with $b \in \{256, 512\}$ such that the new noise covariance matches that of a $(\delta, 128)$-SGD (an

SGD run with smaller batch size). More specifically, we implement $(\delta, \sqrt{\frac{1}{2}(1-\frac{128}{b})\delta}, b, 128)$-large-noise SGD runs. The results are shown in Figure 5.4. In these figures, each $\times$ denotes a baseline SGD run, with step size specified in the legend and minibatch size specified by plot title. For each baseline SGD run, we have a corresponding large-noise SGD run, denoted by $\diamond$ with the same color. As mentioned, these $\diamond$ runs are designed to match the noise covariance of SGD with larger step size or smaller batch size. In addition to $\times$ and $\diamond$, we also plot using a small teal marker all the other runs from Section 5.5.3.1. This helps highlight the linear trend between the logarithm of noise covariance and test accuracy that we observed in Section 5.5.3.1.

As can be seen, the (noise variance, test accuracy) values for the $\diamond$ runs fall close to the linear trend. More specifically, a run of large-noise SGD produces similar test accuracy to vanilla SGD runs with the same noise variance. We highlight two potential implications: First, just like in Section 5.5.3.1, we observe that the test accuracy strongly correlates with relative variance, even for noise of the form (5.17), which can have rather different higher moments than ζ (standard SGD noise); Second, since the $\diamond$ points fall close to the linear trend, we hypothesize that the constant-noise limit SDE (5.13) should also have similar test error. If true, then this implies that we only need to study the potential $U(x)$ and noise covariance $M(x)$ to explain the generalization properties of SGD.

5.6 Related Work

Previous work has drawn connections between SGD noise and generalization [41, 43, 44, 47, 58]. Notably, He et al. [41], Jastrzebski et al. [44], Mandt et al. [58] analyze favorable properties of SGD noise by arguing that in the neighborhood of a local minimum, (5.2) is roughly the discretization of an Ornstein-Uhlenbeck (OU) process, and so the distribution of $y_{k\delta}$ approximates is approximately Gaussian. However, empirical results [43, 47] suggest that SGD generalizes better by finding better local minima, which may require us to look beyond the "OU near local minimum" assumption to understand the global distributional properties of SGD. Indeed, Hoffer et al. [43] suggest that SGD performs a random walk on a random loss landscape, Kleinberg et al. [48] propose that SGD noise helps smooth out "sharp minima." Jastrzebski et al. [44] further note the similarity between (5.1) and an Euler-Murayama approximation of (5.3). Chaudhari and Soatto [14] also made connections between SGD and SDE. Our work tries to make these connections rigorous, by quantifying the error between (5.3), (5.2) and (5.1), without any assumptions about (5.3) being close to an OU process or being close to a local minimum.

Our work builds on a long line of work establishing the convergence rate of Langevin MCMC in different settings [18, 20, 29, 36, 39, 54, 57]. We will discuss our rates in relation to some of this work in detail following our presentation of Theorem 10. We note here that some of the techniques used in this paper were first used by Eberle [32], Gorham et al. [39], who analyzed the convergence of (5.3) to p^* without log-concavity assumptions. Erdogdu et al. [36] studied processes of the form (5.2) as an approximation to (5.3) under a distant-dissipativity assumption, which is similar to the assumptions made in this paper. For the sequence (5.2),

they prove an $O(1/\varepsilon^2)$ iteration complexity to achieve ε integration error for any loss f which has bounded derivatives up to fourth order. In comparison, we prove W_1 convergence between $\mathsf{Law}(y_{k\delta})$ and p^*, which is equivalent to $\sup_{\|\nabla f\|_\infty \leq 1} |\mathbb{E}[f(y_{k\delta})] - \mathbb{E}_{y\sim p^*}[f(y)]|$, also with rate $\tilde{O}(1/\varepsilon^2)$. By smoothing the W_1 test function to have bounded derivatives up to fourth order, we believe that the results by Erdogdu et al. [36] can imply a qualitatively similar result to Theorem 10, but with a worse dimension and ε dependence.

In concurrent work by Li et al. [54], the authors study a process based on a stochastic Runge-Kutta discretization scheme of (5.3). They prove an $\tilde{O}\big(\frac{d}{\varepsilon^{-2/3}}\big)$ iteration complexity to achieve ε error in W_2 for an algorithm based on Runge-Kutta discretization of (5.3). They make a strong assumption of *uniform dissipativity* (essentially assuming that the process (5.3) is uniformly contractive), which is much stronger than the assumptions in this paper, and may be violated in the settings of interest considered in this paper.

There has been a number of works [3, 15, 53] which establish CLT results for SGD with very small step size (rescaled to have constant variance). This work generally focuses on the setting of "OU process near a local minimum", in which the diffusion matrix is constant.

Finally, a number of authors have studied the setting of heavy-tailed gradient noise in neural network training. [86] showed that in some cases, the heavy-tailed noise can be detrimental to training, and a clipped version of SGD performs much better. [80] argue that when the SGD noise is heavy-tailed, it should not be modeled as a Gaussian random variable, but instead as an α-stable random variable, and propose a Generalized Central Limit Theorem to analyze the convergence in distribution. Our paper does not handle the setting of heavy-tailed noise; our theorems require that the norm of the noise term is uniformly bounded, which will be satisfied, for example, if gradients are explicitly clipped at a threshold, or if the optimization objective has Lipschitz gradients and the SGD iterates stay within a bounded region.

Bibliography

[1] Yasin Abbasi, Peter L. Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems*, pages 2508–2516, 2013.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[3] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. *arXiv preprint arXiv:1904.02130*, 2019.

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

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

[6] Francois 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.

[7] Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018.

[8] Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *ArXiv e-prints*, May 2018.

[9] Sébastien Bubeck. Lecture notes: Introduction to online optimization, 2011.

[10] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

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

[12] Nicolo Cèsa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[13] Ioannis Chatzigeorgiou. Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

[14] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[15] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

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

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

[18] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[19] Xiang Cheng, Peter L Bartlett, and Michael I Jordan. Quantitative central limit theorems for discrete stochastic processes. *arXiv preprint arXiv:1902.00832*, 2019.

[20] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *arXiv preprint arXiv:1412.7392*, 2014.

[21] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129 (12):5278–5311, 2019.

[22] Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

[23] George Deligiannidis, Daniel Paulin, Alexandre Bouchard-Côté, and Arnaud Doucet. Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *arXiv preprint arXiv:1808.04299*, 2018.

[24] 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.

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

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

[27] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[28] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte marlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[29] Alain Durmus, Eric Moulines, et al. High-dimensional bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[30] Raaz Dwivedi, Yuansi Chen, Martin Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018.

[31] Andreas Eberle. Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathematique*, 349(19-20):1101–1104, 2011.

[32] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3-4):851–886, 2016.

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

[34] Andreas Eberle, Mateusz B Majka, et al. Quantitative contraction rates for markov chains on general state spaces. *Electronic Journal of Probability*, 24, 2019.

[35] Ronen Eldan, Dan Mikulincer, and Alex Zhai. The CLT in high dimensions: quantitative bounds via martingale embedding. *arXiv preprint arXiv:1806.09087*, 2018.

[36] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018.

[37] Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. *arXiv preprint arXiv:1710.02736*, 2017.

[38] 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.

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

[40] 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.

[41] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, pages 1141–1150, 2019.

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

[43] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

[44] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[45] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[46] Ioannis Karatzas and Steven E Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.

[47] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[48] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

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

[50] 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.

[51] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[52] Yin Tat Lee and Santosh Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. *arXiv preprint arXiv:1710.06261*, 2017.

[53] Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[54] Xuechen Li, Yi Wu, and Lester Mackey. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, pages 7746–7758, 2019.

[55] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

[56] 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.

[57] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

[58] Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 354–363, 2016.

[59] Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

[60] Oren Mangoubi and Nisheeth K Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. *arXiv preprint arXiv:1802.08898*, 2018.

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

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

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

[64] Edward Nelson. *Dynamical theories of Brownian motion*, volume 3. Princeton University Press, 1967.

[65] 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.

[66] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2013.

[67] Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180 (3):378–384, 1981.

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

[69] D Pollard. Asymptopia: an exposition of statistical asymptotic theory, 2013.

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

[71] Philip E. Protter. Stochastic differential equations. In *Stochastic Integration and Differential Equations*, pages 249–361. Springer, 2005.

[72] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

[73] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[74] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

[75] Gareth Roberts and Richard Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[76] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[77] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

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

[79] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[80] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*, 2019.

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

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

[83] Cédric Villani. *Hypocoercivity*. American Mathematical Society, 2009.

[84] 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. URL `http://www.pnas.org/content/113/47/E7351.abstract`.

[85] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

[86] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why ADAM beats SGD for attention models. *arXiv preprint arXiv:1912.03194*, 2019.

[87] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.

Appendix A

Proofs for Chapter 2

Proof of Lemma 2
The proof is directly from results in [2]. See Theorem 10.4.9, with $\mathcal{F}(\mu|\gamma) = KL(\mu\|\gamma)$, with $\mu = \boldsymbol{\mu}$, $\gamma = \mathbf{p}^*$, $\sigma = \frac{\boldsymbol{\mu}}{\mathbf{p}^*}$, $F(\rho) = \rho\log\rho$, $L_F(\sigma) = \sigma$, and $w_{\boldsymbol{\mu}} = \frac{\nabla L_F(\sigma)}{\sigma} = \nabla\log\frac{\boldsymbol{\mu}}{\mathbf{p}^*}$. The expression for $\frac{d}{dt}F(\mathbf{p}_t)$ comes from expression 10.1.16 (section E of chapter 10.1.2, page 233). See also expressions 10.4.67 and 10.4.68.

(One can also refer to Theorem 10.4.13 and Theorem 10.4.17 for proofs of $w_{\boldsymbol{\mu}}$ for the KL-divergence functional in more general settings.) By Lemma 20, $w_{\mathbf{p}_t}$ is well defined for all t. □

Proof of Lemma 3
Theorem 8.3.1 of [2].

□

Proof of Lemma 4
By definition of $\mathcal{D}_{\boldsymbol{\mu}_t}(v)$ in (2.11) and Lemma 2 and Cauchy-Schwarz. □

Proof
Proof of Lemma 6 In this proof, we treat t as a fixed but arbitrary number, and prove the Lemma for all $t \in \mathbb{R}^+$. We will use x_s, y_s^t, z_s^t, $\mathbf{p}_s$, $\mathbf{q}_s^t$ and $\mathbf{g}_s^t$ as defined in (2.3), (2.12) and (2.13).

First, consider the case when $t = \tau(t)$. By definition, $x_t = y_t^t = z_t^t$, and $\mathbf{p}_t = \mathbf{q}_t^t = \mathbf{g}_t^t$. By Fokker Planck,

$$\begin{aligned}\frac{d}{ds}\mathbf{p}_s(x)\Big|_{s=t} &= -\nabla U(x_t) + \mathrm{tr}(\nabla^2\mathbf{p}_t)\\ &= -\nabla U(y_t^t) + \mathrm{tr}(\nabla^2\mathbf{q}_t^t)\\ &= \frac{d}{ds}\mathbf{q}_s^t(x)\Big|_{s=t}.\end{aligned}$$

On the other hand

$$dz_s^t\big|_{s=t} = -\nabla U(z_{\tau(t)}^t) + \nabla U(z_t^t) = -\nabla U(x_t) + \nabla U(x_t) = 0.$$

Thus $\frac{d}{ds}\mathbf{g}_s^t\big|_{s=t} = 0$ So Lemma (6) holds.

In the remainder of this proof, we assume that $t \neq \tau(t)$.

For a given $\Theta \in \mathbb{R}^{2d}$, we let $\Pi_1(\Theta)$ denote the projection of Θ onto its first d coordinates, and $\Pi_2(\Theta)$ denote the projection of Θ onto its last d coordinates. With abuse of notation, for $\mathbf{P} \in \mathscr{P}(\mathbb{R}^{2d})$, we let $\Pi_1(\mathbf{P})$ and $\Pi_2(\mathbf{P})$ denote the corresponding marginal densities.

We will consider three stochastic processes: $\Theta_s, \Lambda_s^t, \Psi_s^t$ over $\mathbb{R}^{2d}$ for $s \in [\tau(t), \tau(t)+h)$.

First, we introduce the stochastic process Θ_s for $s \in [\tau(t), \tau(t)+h)$

$$\begin{aligned}
\Theta_{\tau(t)} &= \begin{bmatrix} x_{\tau(t)} \\ -\nabla U(x_{\tau(t)}) \end{bmatrix} \\
d\Theta_s &= \begin{bmatrix} \Pi_2(\Theta_s) \\ 0 \end{bmatrix} dt + \begin{bmatrix} \sqrt{2} dB_t \\ 0 \end{bmatrix} \quad \text{for } s \in [\tau(t), \tau(t)+h).
\end{aligned}$$

We let $\mathbf{P}_s$ denote the density for Θ_s. Intuitively, $\mathbf{P}_s$ is the joint density between x_s and $-\nabla U(x_{\tau(t)})$. One can verify that $\Pi_1(\Theta_s) = x_s$ and $\Pi_1(\mathbf{P}_s) = \mathbf{p}_s$. By Fokker-Planck, we have $\forall \Theta \in \mathbb{R}^{2d}$

$$\begin{aligned}
\frac{d}{ds}\mathbf{P}_s(\Theta)\Big|_{s=t} = &- \nabla \cdot \left(\mathbf{P}_t(\Theta) \cdot \begin{bmatrix} \Pi_2(\Theta) \\ 0 \end{bmatrix} \right) \\
&+ \sum_{i=1}^{d} \frac{\partial^2}{\partial \Theta_i^2} \mathbf{P}_t(\Theta). \qquad \text{(A.1)}
\end{aligned}$$

Next, for any given t, we introduce the stochastic process Λ_s^t for $s \in [\tau(t), \tau(t)+h)$.

$$\begin{aligned}
\Lambda_s^t &= \Theta_s && \text{for } s \leq t \\
d\Lambda_s^t &= \begin{bmatrix} -\nabla U(\Pi_1(\Lambda_s^t)) \\ 0 \end{bmatrix} ds + \begin{bmatrix} \sqrt{2} dB_s \\ 0 \end{bmatrix} && \text{for } s \geq t.
\end{aligned}$$

Let $\mathbf{Q}_s^t$ denote the density for Λ_s^t. One can verify that $\Pi_1(\Lambda_s^t) = y_s^t$ and $\Pi_1(\mathbf{Q}_s^t) = \mathbf{q}_s^t$. By Fokker-Planck, we have $\forall \Theta \in \mathbb{R}^{2d}$

$$\begin{aligned}
\frac{d}{ds}\mathbf{Q}_s^t(\Theta)\Big|_{s=t} = &- \nabla \cdot \left(\mathbf{Q}_t^t(\Theta) \cdot \begin{bmatrix} -\nabla U(\Pi_1(\Theta)) \\ 0 \end{bmatrix} \right) \\
&+ \sum_{i=1}^{d} \frac{\partial^2}{\partial \Theta_i^2} \mathbf{Q}_t^t(\Theta). \qquad \text{(A.2)}
\end{aligned}$$

Finally, define

$$\begin{aligned}\Psi_s^t &= \Theta_s && \text{for } s \le t\\ d\Psi_s^t &= \begin{bmatrix}\Pi_2(\Psi_s^t) + \nabla U(\Pi_1(\Psi_s^t))\\ 0\end{bmatrix} ds\\ &\quad + \begin{bmatrix}\sqrt{2}dB_s\\ 0\end{bmatrix} && \text{for } s \ge t.\end{aligned}$$

Let $\mathcal{G}_s^t$ denote the density for Ψ_s^t. One can verify that $\Pi_1(\Psi_s^t) = z_s^t$ and $\Pi_1(\mathcal{G}_s^t) = \mathbf{g}_s^t$. By Fokker-Planck, we have $\forall \Theta \in \mathbb{R}^{2d}$.

$$\frac{d}{ds}\mathcal{G}_s^t(\Theta)\Big|_{s=t} = -\nabla \cdot \left(\mathcal{G}_t^t(\Theta) \cdot \begin{bmatrix}\Pi_2(\Theta) + \nabla U(\Pi_1(\Theta))\\ 0\end{bmatrix}\right). \tag{A.3}$$

By definition, $\Theta_t = \Lambda_t^t = \Psi_t^t$ almost surely, and $\mathbf{P}_t = \mathbf{Q}_t^t = \mathcal{G}_t^t$. Taking the difference between (A.1), (A.2) thus gives

$$\begin{aligned}\frac{d}{ds}\mathbf{P}_s(\Theta) - \mathbf{Q}_s^t(\Theta)\Big|_{s=t} &= -\nabla \cdot \left(\mathbf{P}_t(\Theta) \cdot \begin{bmatrix}\Pi_2(\Theta) + \nabla U(\Pi_1(\Theta))\\ 0\end{bmatrix}\right)\\ &= \frac{d}{ds}\mathcal{G}_s^t(\Theta)\Big|_{s=t}\end{aligned}$$

Finally, marginalizing out the last d coordinates on both sides, and recalling that $\Pi_1(\mathbf{P}_s) = \mathbf{p}_s$, $\Pi_1(\mathbf{Q}_s^t) = \mathbf{q}_s^t$ and $\Pi_1(\mathcal{G}_s^t) = \mathbf{g}_s^t$, we prove the Lemma. □

Proof of Lemma 7

The fact that $\mathbf{q}_s^t$ is the steepest descent follows from the fact that Fokker-Planck equation for Langevin diffusion yields, for all $x \in \mathbb{R}^d$

$$\begin{aligned}\frac{d}{ds}\mathbf{q}_s^t(x) =& \mathbf{div}(\mathbf{q}_s^t(x)\nabla \log \mathbf{p}^*(x)) + \operatorname{tr}(\nabla^2 \mathbf{q}_s^t(x))\\ =& \mathbf{div}\left(\mathbf{q}_s^t(x)\left(\nabla \log \frac{\mathbf{p}^*(x)}{\mathbf{q}_s^t(x)}\right)\right).\end{aligned}$$

By definition of (2.6), we get that

$$v_s = \nabla \log \frac{\mathbf{p}^*(x)}{\mathbf{q}_s^t(x)} \tag{A.4}$$

satisfies the continuity equation for $\mathbf{q}_s^t$. By Lemma 2,

$$w_{\mathbf{q}_s^t} = \nabla \log \left(\frac{\mathbf{q}_s^t}{\mathbf{p}^*}\right).$$

Thus

$$\frac{d}{ds}F(\mathbf{q}_s^t) = \mathcal{D}_{\mathbf{q}_s^t}(v_s) = -\mathbb{E}_{\mathbf{q}_s^t}\left[\|w_{\mathbf{q}_s^t}\|_2^2\right] = -\|\mathcal{D}_{\mathbf{q}_s^t}\|_*^2,$$

where the last equality is by Cauchy-Schwarz.

□

Proof of Lemma 8
Consider z_s^t and $\mathbf{g}_s^t$ as defined in (2.13). By Lemma 6, $\frac{d}{ds}\mathbf{g}_s^t\big|_{s=t} = (\frac{d}{ds}\mathbf{p}_s - \frac{d}{ds}\mathbf{q}_s^t)\big|_{s=t}$. The first variation of F, defined by

$$\lim_{\varepsilon\to 0}\frac{F(\boldsymbol{\mu}+\varepsilon\Delta)-F(\boldsymbol{\mu})}{\varepsilon} = \int\left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu})\right)(x)\cdot\Delta(x)dx,$$

is linear (see Chapter 7.2 of [76]). (In the above, $\Delta : \mathbb{R}^d \to \mathbb{R}$ is an arbitrary 0-mean perturbation). In addition, because $\mathbf{p}_t = \mathbf{q}_t^t = \mathbf{g}_t^t$, we have $\frac{\delta F}{\delta \boldsymbol{\mu}}(\mathbf{p}_t) = \frac{\delta F}{\delta \boldsymbol{\mu}}(\mathbf{q}_t^t) = \frac{\delta F}{\delta \boldsymbol{\mu}}(\mathbf{g}_t^t)$, we get that

$$\frac{d}{ds}F(\mathbf{g}_s^t)\bigg|_{s=t} = \left(\frac{d}{ds}F(\mathbf{p}_s) - \frac{d}{ds}F(\mathbf{q}_s^t)\right)\bigg|_{s=t}.$$

We will upper bound $|\mathbf{g}_s^{t\prime}|\big|_{s=t}$, then apply Corollary 5.

$$\begin{aligned}
&|\mathbf{g}_s^{t\prime}|\big|_{s=t}\\
=&\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}W_2(\mathbf{g}_{t+\varepsilon}^t, \mathbf{g}_t^t)\\
\le&\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\sqrt{\mathbb{E}\left[\left\|\varepsilon(\nabla U(x_t) - \nabla U(x_{\tau(t)}))\right\|_2^2\right]}\\
=&\sqrt{\mathbb{E}\left[\left\|\nabla U(x_t) - \nabla U(x_{\tau(t)})\right\|_2^2\right]}\\
\le&\sqrt{\mathbb{E}\left[L^2\|x_t - x_{\tau(t)}\|_2^2\right]}\\
=&L\sqrt{\mathbb{E}\left[\|(t-\tau(t))\nabla U(x_{\tau(t)}) + \sqrt{2}(B_t - B_{\tau(t)})\|_2^2\right]}\\
\le&2L(t-\tau(t))\sqrt{\mathbb{E}\left[\|\nabla U(x_{\tau(t)})\|_2^2\right]} + 2L\sqrt{(t-\tau(t))d}\\
\le&2L(t-\tau(t))\sqrt{L^2\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]} + 2L\sqrt{(t-\tau(t))d},
\end{aligned}$$

where the first line is by definition of metric derivative, second line is by the coupling between $\mathbf{g}_t^t$ and $\mathbf{g}_{t+\varepsilon}^t$) induced by the joint distribution $(z_t^t, z_{t+\varepsilon}^t)$ and the fact that $z_{\tau(t)}^t = x_t$. The fourth line is by Lipschitz-gradient of $U(x)$, fifth line is by definition of x_t, sixth line is by variance of $B_t - B_0$, seventh line is once again by Lipschitz-gradient of $U(x)$.

Thus, we upper bound $|\mathbf{g}_s^{t\prime}|\big|_{s=t}$ by $2L^2(t-\tau(t))\sqrt{\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]} + 2L\sqrt{(t-\tau(t))d}$. Applying Corollary 5, and using the fact that for all t, $t-\tau(t) \le h$, we get

$$\begin{aligned}
&\frac{d}{ds}F(\mathbf{g}_s^t)\bigg|_{s=t}\\
\le&\left(2L^2h\sqrt{\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]} + 2L\sqrt{hd}\right)\|\mathcal{D}_{\mathbf{g}_t^t}\|_*\\
\le&\left(2L^2h\sqrt{\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]} + 2L\sqrt{hd}\right)\|\mathcal{D}_{\mathbf{p}_t}\|_*.
\end{aligned}$$

The last line is because $\mathbf{g}_t^t = \mathbf{p}_t$ by definition.

□

Proof of Lemma 9
By Theorem 9.4.11 of [2], m-strong-convexity of $\log \mathbf{p}^*$ implies geodesic convexity. Expression (2.14) then follows from the definition of geodesic convexity in definition 9.1.1 of [2].

Rearrranging terms, dividing by t and taking limit as $t \to 0$, we get

$$\begin{aligned} F(\boldsymbol{\mu}_1) &\geq F(\boldsymbol{\mu}_0) + \lim_{t\to 0} \frac{F(\boldsymbol{\mu}_t) - F(\boldsymbol{\mu}_0)}{t} + \frac{m}{2} W_2^2(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) \\ &= F(\boldsymbol{\mu}_0) + \mathcal{D}_{\boldsymbol{\mu}_0}(v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1}) + \frac{m}{2} W_2^2(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1). \end{aligned}$$

The last equality follows by Lemma 2 and by the remark immediately following (2.6).

We remark that the proof of (2.15) is completely analogous to the proof of first-order characterization of strongly convex functions over $\mathbb{R}^d$.

□

Proof of Lemma 10
We consider (2.15), and use two facts

1. For any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, $\mathcal{D}_{\boldsymbol{\mu}}(v)$ is linear in v. (see (2.11))
2. For any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathbb{R}^d)$, $W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbb{E}_{\boldsymbol{\mu}} \|v_{\boldsymbol{\mu}}^{\boldsymbol{\nu}}(x)\|_2^2$, by definition of W_2 and $v_{\boldsymbol{\mu}}^{\boldsymbol{\nu}}$ as the optimal displacement map.

We apply Lemma (10) with $\boldsymbol{\mu}_0 = \mathbf{p}^*$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$. Let $v_{\boldsymbol{\mu}}^{\mathbf{p}^*}$ be the optimal displacement map from $\boldsymbol{\mu}$ to $\mathbf{p}^*$, so (2.15) gives

$$\begin{aligned} F(\boldsymbol{\mu}) - F(\mathbf{p}^*) &\leq -\mathcal{D}_{\boldsymbol{\mu}}(v_{\boldsymbol{\mu}}^{\mathbf{p}^*}) - \frac{m}{2} W_2^2(\boldsymbol{\mu}, \mathbf{p}^*) \\ &= -\mathcal{D}_{\boldsymbol{\mu}}(v_{\boldsymbol{\mu}}^{\mathbf{p}^*}) - \frac{m}{2} \mathbb{E}_{\boldsymbol{\mu}} \|v_{\boldsymbol{\mu}}^{\mathbf{p}^*}(x)\|_2^2. \end{aligned}$$

Let $v^* \triangleq \arg\max_{\|v\|_{L^2(\boldsymbol{\mu})} \leq 1} -\mathcal{D}_{\boldsymbol{\mu}}(v)$, so $\mathcal{D}_{\boldsymbol{\mu}}(v^*) = -\|\mathcal{D}_{\boldsymbol{\mu}}\|_*$ by linearity. We know that the maximizer of

$$\arg\max_v -\mathcal{D}_{\boldsymbol{\mu}}(v) - \frac{m}{2} \mathbb{E}_{\boldsymbol{\mu}} \|v_{\boldsymbol{\mu}}^{\mathbf{p}^*}(x)\|_2^2 = c \cdot v^*,$$

for some real number c. Taking derivatives wrt c gives $c = \frac{1}{m}\|\mathcal{D}_{\boldsymbol{\mu}}\|_*$. Thus we get

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq \frac{m}{2} \|\mathcal{D}_{\boldsymbol{\mu}}\|_*^2.$$

□

Proof of Lemma 11
We prove this by induction on k. First, by definition of $\mathbf{p}_0 = N(0, \frac{1}{m})$, we get that

$$\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right] = \frac{d}{m} \leq \frac{4d}{m} \quad , \forall t \leq 0h.$$

Next, we assume that for some k, and for all $t \leq kh$, $\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right] \leq \frac{4d}{m}$.
For the inductive step, we consider $t \in (kh, (k+1)h]$
From (2.3),

$$x_t = x_{kh} - (t - kh)\nabla U(x_{kh}) + \sqrt{2}(B_t - B_{kh}).$$

By smoothness and strong convexity and the assumption that $\arg\min_x U(x) = 0$, we get that for all x and for all t:

$$\|(x - (t - kh)\nabla U(x)) - 0\|_2 \leq (1 - mt)\|x - 0\|_2.$$

(Note that $h \leq \frac{1}{L}$ implies that $t - kh \leq \frac{1}{L}$.) So for all t

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathbf{p}_t}\|x\|_2^2 \\
=&\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x - (t - kh)\nabla U(x) + \sqrt{2}(B_t - B_{kh})\|_2^2 \\
=&\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x - (t - kh)\nabla U(x)\|_2^2 + \mathbb{E}\|\sqrt{2}(B_t - B_{kh})\|_2^2 \\
\leq&(1 - mt)\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x\|_2^2 + 2dt \\
=&\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x\|_2^2 + (2dt - mt\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x\|_2^2).
\end{aligned}$$

By our inductive hypothesis, we have $\mathbb{E}_{x\sim\mathbf{p}_t}\|x\|_2^2 \leq \frac{4d}{m}$ for all $t \leq kh$
If $\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x\|_2^2 \geq \frac{2d}{m}$, then $\mathbb{E}_{x\sim\mathbf{p}_t}\|x\|_2^2 \leq \mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x\|_2^2 \leq \frac{4d}{m}$.
If $\mathbb{E}_{x\sim\mathbf{p}_{kh}}\|x\|_2^2 \leq \frac{2d}{m}$, then $\mathbb{E}_{x\sim\mathbf{p}_t}\|x\|_2^2 \leq \frac{2d}{m} + \frac{2d}{L} \leq \frac{4d}{m}$ (by $t - kh \leq \frac{1}{L}$ and by $L \geq m$).
Thus if $\mathbf{p}_{kh}$ is such that $E_{\mathbf{p}_{kh}}\|x\|_2^2 \leq \frac{4d}{m}$, then it must be that $E_{\mathbf{p}_t}\|x\|^2 \leq \frac{4d}{m}$ for all $t \in (kh, (k+1)h]$, thus proving the inductive step. □

A.1 Proof of Theorem 2

First, we present a lemma for upper bounding $F(\boldsymbol{\mu}) - F(\mathbf{p}^*)$ for $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$ in the absence of strong convexity. The following lemma plays an analogous role to Lemma 10.

Lemma 16 *Let F be convex in W_2, then for all $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$,*

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq \|\mathcal{D}_{\boldsymbol{\mu}}\|_* W_2(\boldsymbol{\mu}, \mathbf{p}^*).$$

Proof of Lemma 16
Similar to the proof of Lemma 10, we consider (2.15), but with $m = 0$, (and once again $v_{\boldsymbol{\mu}}^{\mathbf{p}^*}$ denotes the optimal displacement map from $\boldsymbol{\mu}$ to $\mathbf{p}^*$):

$$\begin{aligned} F(\boldsymbol{\mu}) - F(\mathbf{p}^*) &\leq -\mathcal{D}_{\mathbf{p}}(v_{\boldsymbol{\mu}}^{\mathbf{p}^*}) \\ &\leq \|\mathcal{D}_{\boldsymbol{\mu}}\|_* \cdot \|v_{\boldsymbol{\mu}}^{\mathbf{p}^*}\|_{L^2(\boldsymbol{\mu})} \\ &\leq \|\mathcal{D}_{\boldsymbol{\mu}}\|_* \cdot W_2(\boldsymbol{\mu}, \mathbf{p}^*), \end{aligned}$$

where first inequality is from (2.15), second line is by definition of $\|\mathcal{D}_{\boldsymbol{\mu}}\|_*$, third line is by defintion of Wasserstein distance and the fact that $v_{\boldsymbol{\mu}}^{\mathbf{p}^*}$ is the optimal transport map.
□

Next, we establish that for a fixed stepsize h, $W_2(\mathbf{p}_t, \boldsymbol{\pi}_h)$ is nonincreasing, using a synchronous coupling technique taken from [29].

Lemma 17 *Let $\mathbf{p}_t$ be defined as in the statement of Theorem 2. Let h be a fixed stepsize satisfying $h \leq \min\{\frac{1}{L}, h'\}$. Then for all k,*

$$W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) \leq W_2(\mathbf{p}_0, \boldsymbol{\pi}_h).$$

Proof of Lemma 17

First, we demonstrate that (2.3) is contractive in W_2.
We will prove this by induction.
Base case: trivially true.
Inductive Hypothesis: $W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) \leq W_2(\mathbf{p}_0, \boldsymbol{\pi}_h)$ for some k.
Inductive Step: Let T be the optimal transport map from $\mathbf{p}_{kh}$ to $\boldsymbol{\pi}_h$. We will demonstrate a coupling between $\mathbf{p}_{(k+1)h}$ and $\boldsymbol{\pi}_h$ with cost less than $W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h)$. The lemma then follows from induction.

Since $x_{kh} \sim \mathbf{p}_{kh}$ (see (2.3)), the optimal coupling between $\mathbf{p}_{kh}$ and $\boldsymbol{\pi}_h$ is given by the pair of random variables $(x_{kh}, T(x_{kh}))$. For $t \in [kh, (k+1)h]$,

$$x_{(k+1)h} = x_{kh} - h\nabla U(x_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})).$$

Consider the coupling γ between $\mathbf{p}_{kh}$ and $\boldsymbol{\pi}_h$ defined by the following pair of random variables

$$\Big(x_{kh} - h\nabla U(x_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}),$$
$$T(x_{kh}) - h\nabla U(T(x_{kh})) + \sqrt{2}(B_{(k+1)h} - B_{kh})\Big).$$

(Note that $\boldsymbol{\pi}_h$ is stationary under the discrete Langevin diffusion with stepsize h, so γ does have the right marginals).

To demonstrate contraction in W_2:

$$\begin{aligned}
&W_2^2(\mathbf{p}_{(k+1)h}, \boldsymbol{\pi}_h) \\
\leq &\mathbb{E}\left[\left\|\Big(x_{kh} - h\nabla U(x_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})\Big)\right.\right. \\
&\left.\left. - \Big(T(x_{kh}) - h\nabla U(T(x_{kh})) + \sqrt{2}(B_{(k+1)h} - B_{kh})\Big)\right\|_2^2\right] \\
= &\mathbb{E}\left[\|(x_{kh} - h\nabla U(x_{kh})) - (T(x_{kh}) - h\nabla U(T(x_{kh})))\|_2^2\right] \\
\leq &\mathbb{E}[\|x_{kh} - T(x_{kh})\|_2^2 - 2h\left\langle \nabla U(x_{kh}) - \nabla U(T(x_{kh})), x_{kh} - T(x_{kh})\right\rangle \\
&+ h^2\|\nabla U(x_{kh}) - \nabla U(T(x_{kh}))\|_2^2] \\
\leq &\mathbb{E}\left[\|x_{kh} - T(x_{kh})\|_2^2\right] \\
= &W_2^2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h),
\end{aligned}$$

where the last equality follows by optimality of T, and the last inequality follows because L-smoothness of $U(x)$ implies

$$\begin{aligned}
&- 2h\left\langle \nabla U(x_{kh}) - \nabla U(T(x_{kh})), x_{kh} - T(x_{kh})\right\rangle \\
\leq &-\frac{h}{L}\|\nabla U(x_{kh}) - \nabla U(T(x_{kh}))\|_2^2 \\
\leq &-h^2\|\nabla U(x_{kh}) - \nabla U(T(x_{kh}))\|_2^2.
\end{aligned}$$

This completes the inductive step. □

Corollary 18 *Let $\mathbf{p}_t$ be as defined in (2). Then for all t,*

$$W_2(\mathbf{p}_t, \mathbf{p}^*) \leq 4C_1.$$

Proof of Corollary 18
First, if $t = \tau(t)$, then by Lemma 17 and (2.9) and triangle inequality, we get our conclusion.

So assume that $t \neq \tau(t)$. Using identical arguments as in Lemma 17, and noting the assumption on h' in (2.9) and the fact that $h \leq h'$, we can show that

$$W_2(\mathbf{p}_t, \boldsymbol{\pi}_{t-\tau(t)}) \leq W_2(\mathbf{p}_{\tau(t)}, \boldsymbol{\pi}_{t-\tau(t)}) \tag{A.5}$$

By triangle inequality and the assumption in (2.9), we have

$$\begin{aligned}
&W_2(\mathbf{p}_t, \mathbf{p}^*) \\
\leq &W_2(\mathbf{p}_t, \boldsymbol{\pi}_{t-\tau(t)}) + W_2(\boldsymbol{\pi}_{t-\tau(t)}, \mathbf{p}^*) \\
\leq &W_2(\mathbf{p}_{\tau(t)}, \boldsymbol{\pi}_{t-\tau(t)}) + W_2(\boldsymbol{\pi}_{t-\tau(t)}, \mathbf{p}^*) \\
\leq &W_2(\mathbf{p}_{\tau(t)}, \boldsymbol{\pi}_h) + W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) \\
&+ W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) + W_2(\boldsymbol{\pi}_{t-\tau(t)}, \mathbf{p}^*) \\
\leq &4C_1
\end{aligned}$$

Where the first inequality is by triangle inequality, the second inequality is by (A.5), third inequality is by triangle inequality, fourth inequality is by assumption (2.9) and the fact that $t - \tau(t) \leq h \leq h'$. □

Next, we use Lemma 17, to bound $\mathbb{E}\left[\|x_{kh}\|_2^2\right]$ for all k:

Lemma 19 *Let h, x_t and $\mathbf{p}_t$ be as defined in the statement of Theorem 2. Then for all k*

$$\mathbb{E}\left[\|x_{kh}\|_2^2\right] \leq 4(C_1^2 + C_2^2).$$

Proof of Lemma 19
Let $\gamma(x, y)$ be the optimal coupling between $\mathbf{p}_{kh}$ and $\boldsymbol{\pi}_h$. Let $\gamma'(x, y)$ be the optimal coupling between $\boldsymbol{\pi}_h$ and $\mathbf{p}^*$. Then

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}_{kh}}\left[\|x\|_2^2\right] &= \mathbb{E}_{\gamma}\left[\|x\|_2^2\right] \\
&= \mathbb{E}_{\gamma}\left[\|x - y + y\|_2^2\right] \\
&\leq 2\mathbb{E}_{\gamma}\left[\|x - y\|_2^2\right] + 2\mathbb{E}_{\gamma}\left[\|y\|_2^2\right] \\
&= 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 2\mathbb{E}_{\boldsymbol{\pi}_h}\left[\|y\|_2^2\right] \\
&= 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 2\mathbb{E}_{\gamma'}\left[\|x\|_2^2\right] \\
&= 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 2\mathbb{E}_{\gamma'}\left[\|x - y + y\|_2^2\right] \\
&\leq 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 4\mathbb{E}_{\gamma'}\left[\|x - y\|_2^2\right] + 4\mathbb{E}_{\gamma'}\left[\|y\|_2^2\right] \\
&\leq 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 4W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) + 4\mathbb{E}_{\mathbf{p}^*}\left[\|x\|_2^2\right].
\end{aligned}$$

By definition of C_2 at the start of Section 2.3.2, we have

$$\mathbb{E}_{\mathbf{p}^*}\left[\|x\|_2^2\right] \leq C_2^2.$$

By Lemma 17, we have

$$W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) \leq W_2(\mathbf{p}_0, \boldsymbol{\pi}_h) \leq C_1.$$

By definition of h' at the start of Section 2.3.2, and h in Theorem 2 (which ensures $h \leq h'$), we have

$$W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) \leq C_1$$

□

Proof of Theorem 2
First, we bound the discretization error (for an arbitrary t). By Lemma 8:

$$\begin{aligned}
\frac{d}{ds}\left(F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\right)\Big|_{s=t} &\leq \left(2L^2 t\sqrt{\mathbb{E}_{\mathbf{p}_{\tau(t)}}\|x_{\tau(t)}\|_2^2} + 2L\sqrt{td}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_* \\
&\leq \left(2L^2 t\sqrt{\mathbb{E}_{\mathbf{p}_{\tau(t)}}\|x_{\tau(t)}\|_2^2} + 2L\sqrt{td}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_*.
\end{aligned}$$

Given the choice of

$$h = \frac{1}{48}\min\left\{\frac{\varepsilon}{C_1(C_1+C_2)L^2}, \frac{\varepsilon^2}{L^2C_1^2 d}, h'\right\},$$

we can ensure that

$$\begin{aligned}\left(L^2 h\sqrt{E\|x_{\tau(t)}\|_2^2} + 2L\sqrt{hd}\right) \leq& \frac{1}{4}\left(L^2 h\sqrt{18(C_1^2+C_2^2)} + 2L\sqrt{hd}\right)\\ \leq& \frac{\varepsilon}{8C_1},\end{aligned}$$

where the first inequality comes from Lemma 19.

Assume that $F(\mathbf{p}_s) - F(\mathbf{p}^*) \geq \varepsilon$. By Lemma 16 and Corollary 18, we have

$$\begin{aligned}\|\mathcal{D}_{\mathbf{p}_s}\|_* \geq& \frac{F(\mathbf{p}_s) - F(\mathbf{p}^*)}{W_2(\mathbf{p}_s, \mathbf{p}^*)}\\ \geq& \frac{\varepsilon}{W_2(\mathbf{p}_s, \mathbf{p}^*)}\\ \geq& \frac{\varepsilon}{4C_1}.\end{aligned} \tag{A.6}$$

This implies that

$$\frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\Big|_{s=t} \leq \frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2$$

The rate of decrease of $F(\mathbf{p}_t)$ thus satisfies

$$\begin{aligned}\frac{d}{dt}F(\mathbf{p}_t) - F(\mathbf{p}^*) =& \frac{d}{dt}F(\mathbf{q}_s^t) - F(\mathbf{p}^*)\Big|_{s=t} + \frac{d}{dt}(F(\mathbf{p}_s) - F(\mathbf{q}_s^t))\Big|_{s=t}\\ =& -\|\mathcal{D}_{\mathbf{p}_t}\|_*^2 + \frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2\\ \leq& -\frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2\\ \leq& -\frac{1}{2C_1^2}(F(\mathbf{p}_t) - F(\mathbf{p}^*))^2.\end{aligned}$$

We now study two regimes. The first regime is when $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq 1$, $\frac{d}{dt}F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq -\frac{1}{2C_1^2}(F(\mathbf{p}_t) - F(\mathbf{p}^*))$, which implies

$$F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq (F(\mathbf{p}_0) - F(\mathbf{p}^*))\exp(-\frac{t}{2C_1^2}).$$

We thus achieve $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq 1$ after time

$$t \geq 2C_1^2 \log(F(\mathbf{p}_0) - F(\mathbf{p}^*)).$$

In the second regime, $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq 1$. By noting that $f_t = \frac{1}{t}$ is the solution to $\frac{d}{dt} f_t = -f_t^2$, and letting $f_t = \frac{1}{2C_1^2}(F(\mathbf{p}_t) - F(\mathbf{p}^*))$, we get $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq \frac{2C_1^2}{t}$. To achieve $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq \varepsilon$, we set $t = \frac{2C_1^2}{\varepsilon}$. Overall, we just need to set

$$t \geq \frac{2C_1^2}{\varepsilon} + 2C_1^2 \log(F(\mathbf{p}_0) - F(\mathbf{p}^*)).$$

This, combined with the choice of h earlier, proves the theorem.

□

A.1.1 Some regularity results

In this subsection, we provide some regularity results needed in various parts of the paper.

Lemma 20 *Let $w_{\boldsymbol{\mu}}$ be as defined in Lemma 2. Let $\mathbf{p}_t$ be as defined in 2.3. For all t, $w_{\mathbf{p}_t}$ is well defined, and $\mathbb{E}_{\mathbf{p}_t}\left[\|w_{\mathbf{p}_t}\|_2^2\right]$ is finite.*

Proof of Lemma 20
First, we establish the following statement: For any t, there exists a $\delta \in \mathbb{R}$ with $\boldsymbol{\mu}_{\delta,y}(x)$ being the distribution of $N(y, \delta)$ and $\mathbf{p} \in \mathscr{P}(\mathbb{R}^d)$ such that

1. For all $x \in \mathbb{R}^d$, $\mathbf{p}_t(x) = \mathbb{E}_{y\sim\mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right]$
2. $\mathbb{E}_{\mathbf{p}}\left[\|x\|_2^2\right]$ is finite.

If $t = \tau(t)$, then let $\mathbf{p} = (Id(\cdot) - h\nabla U(\cdot))_{\#}\mathbf{p}_{\tau(t)-1}$ and let $\delta = 2h$. Otherwise, if $t \neq \tau(t)$, then let $\mathbf{p} = (Id(\cdot) - (t - \tau(t))\nabla U(\cdot))_{\#}\mathbf{p}_{\tau(t)}$ and $\delta = 2(t - \tau(t))$. Where we used the definition of push-forward distribution from (2.2). 1. now can be easily verified.

To see 2, let $t' = \tau(t) - 1$ in case 1 and let $t' = \tau(t)$ in case 2.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{p}}\left[\|x\|_2^2\right] \\
=&\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x - h\nabla U(x)\|_2^2\right] \\
\leq&2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x\|_2^2\right] + 2h^2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|\nabla U(x)\|_2^2\right] \\
\leq&2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x\|_2^2\right] + 2h^2L^2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x\|_2^2\right] \\
\leq&(2 + 2h^2L^2)\frac{4d}{m},
\end{aligned}$$

where the last inequality follows by Lemma 11.

Since $\boldsymbol{\mu}_{\delta,y}(x)$ for all x, y, $\mathbb{E}_{\mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right]$ is differentiable for all x. This proves the first part of the Lemma.

Next, a nice property of Gaussians is that

$$\nabla_x \boldsymbol{\mu}_{\delta,y}(x) = -\frac{\boldsymbol{\mu}_{\delta,y}}{\delta}(x - y).$$

Thus,

$$\begin{aligned}
&\nabla \log \mathbf{p}_t(x) \\
=&\frac{1}{\delta}\nabla \log \mathbb{E}_{y\sim\mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right] \\
=&\frac{1}{\delta}\frac{1}{\mathbb{E}_{y\sim\mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right]}\mathbb{E}_{y\sim\mathbf{p}}\left[((y-x)\boldsymbol{\mu}_{\delta,y}(x))\right] \\
=&\frac{1}{\delta}\mathbb{E}_{y\sim\boldsymbol{\mu}_\delta^x}\left[y\right]-x,
\end{aligned}$$

where $\boldsymbol{\mu}_\delta^x$ denotes the conditional distribution of y given x, when $y\sim\mathbf{p}$ and $x\sim\boldsymbol{\mu}_{\delta,y}$.

Thus

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathbf{p}_t(x)}\left[\|\nabla \log \mathbf{p}_t(x)\|_2^2\right] \\
\leq&\frac{1}{\delta}\mathbb{E}_{x\sim\mathbf{p}_t(x)}\left[2\mathbb{E}_{y\sim\boldsymbol{\mu}_\delta^x}\left[\|y\|_2^2\right]+2\|x\|_2^2\right] \\
=&\frac{2}{\delta}\mathbb{E}_{y\sim\mathbf{p}}\left[\|y\|_2^2\right]+\frac{2}{\delta}\mathbb{E}_{x\sim\mathbf{p}_t}\left[\|x\|_2^2\right] \\
<&\infty,
\end{aligned}$$

where the first inequality is by Jensen's inequality and Young's inequality and the preceding result, the second inequality is by definition of conditional distribution, the third inequality is by the fact that $\delta>0$ (by definition at the start of the proof), the fact that $\mathbb{E}_{x\sim\mathbf{p}_t}\left[\|x\|_2^2\right]\leq\frac{4d}{m}$ (by Lemma 11), and by the fact that $\mathbb{E}_{y\sim\mathbf{p}}\left[\|y\|_2^2\right]<\infty$ (see item 2. at the start of the proof)

Finally, we have that

$$\begin{aligned}
&\|w_{\mathbf{p}_t}\|_{L^2(\mathbf{p}_t)}^2 \\
=&\mathbb{E}_{\mathbf{p}_t}\left[\|w_{\mathbf{p}_t}(x)\|_2^2\right] \\
=&\mathbb{E}_{\mathbf{p}_t}\left[\|\nabla \log \mathbf{p}_t(x)-\nabla \log \mathbf{p}^*(x)\|_2^2\right] \\
\leq&2\mathbb{E}_{\mathbf{p}_t}\left[\|\nabla \log \mathbf{p}_t(x)\|_2^2\right]+2\mathbb{E}_{\mathbf{p}_t}\left[\|\nabla \log \mathbf{p}^*(x)\|_2^2\right] \\
<&\infty,
\end{aligned}$$

where the last inequality uses the fact that $\|\nabla \log \mathbf{p}^(x)\|_2=\|\nabla U(x)\|_2\leq L\|x\|_2$ and $\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right]\leq\frac{4d}{m}$. □

Lemma 21 *Let $\mathbf{p}_t$ be as defined in* (2.3)*. Then $|\mathbf{p}_t'|$ is finite for all t, where $|\mathbf{p}_t'|$ is the metric derivative of $\mathbf{p}_t$, as defined in* (2.5)*.*

Proof of Lemma 21

We define the random variable ξ to be distributed as $N(0,1)$. For all t, let $\mathbf{x}_t$ be as defined in (2.3). One can verify that the random variable $y_t\triangleq x_{\tau(t)}-\nabla(t-\tau(t))U(x_{\tau(t)})+\sqrt{2(t-\tau(t))}\xi$ has the same distribution as x_t. Thus y_t and $y_{t+\varepsilon}$ define a coupling between $\mathbf{p}_t$ and $\mathbf{p}_{t+\varepsilon}$.

Let $h \triangleq t - \tau(t)$, then

$$
\begin{aligned}
&|\mathbf{p}_t'| \\
=&\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} W_2(\mathbf{p}_t, \mathbf{p}_{t+\varepsilon}) \\
\leq&\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \sqrt{\mathbb{E}\left[\|y_t - y_{t+\varepsilon}\|_2^2\right]} \\
=&\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \sqrt{\mathbb{E}_{x \sim \mathbf{p}_{\tau(t)}} \left[\|\varepsilon \nabla U(x) + (\sqrt{2(h+\varepsilon)} - \sqrt{2h})\xi\|_2^2\right]} \\
=&\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \sqrt{\mathbb{E}_{x \sim \mathbf{p}_{\tau(t)}} \left[\|\varepsilon \nabla U(x)\|_2^2\right] + \mathbb{E}\left[\|(\sqrt{2(h+\varepsilon)} - \sqrt{2h})\xi\|_2^2\right]} \\
\leq&\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \sqrt{\mathbb{E}_{x \sim \mathbf{p}_{\tau(t)}} \left[\|\varepsilon \nabla U(x)\|_2^2\right]} \\
&\quad + \frac{1}{\varepsilon} \sqrt{\mathbb{E}\left[\|(\sqrt{2(h+\varepsilon)} - \sqrt{2h})\xi\|_2^2\right]} \\
=&\sqrt{\mathbb{E}_{x \sim \mathbf{p}_{\tau(t)}} \left[\|\nabla U(x)\|_2^2\right]} + \frac{1}{\sqrt{8h}} \sqrt{\mathbb{E}\left[\|\xi\|_2^2\right]},
\end{aligned}
$$

where the last inequality follows by Taylor expansion of $\sqrt{2h + 2\varepsilon}$. We can bound the first term by a finite number using $\|\nabla U(x)\|_2^2 \leq L^2 \|x\|_2^2$, then applying Lemma 11. The second term is finite for $h \neq 0$.

For the case $h = 0$, we know that $w_{\mathbf{p}_t}$ satisfies the continuity equation for $\mathbf{p}_t$ at t, and so $|\mathbf{p}_t'| = \|w_{\mathbf{p}_t}\|_{L^2(\mathbf{p}_t)} < \infty$, by Lemma 3 and Lemma 20. $\square$

Appendix B

Proofs for Chapter 3

B.1 Explicit Discrete Time Updates

In this section we calculate integral representations of the solutions to the continuous-time process (3.1) and the discrete-time process (3.4).

Lemma 22 *The solution* (x_t, v_t) *to the underdamped Langevin diffusion* (3.1) *is*

$$v_t = v_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)} \nabla U(x_s) ds\right) + \sqrt{2\gamma u} \int_0^t e^{-\gamma(t-s)} dB_s \tag{B.1}$$
$$x_t = x_0 + \int_0^t v_s ds.$$

The solution $(\tilde{x}_t, \tilde{v}_t)$ *of the discrete underdamped Langevin diffusion* (3.4) *is*

$$\tilde{v}_t = \tilde{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)} \nabla U(\tilde{x}_0) ds\right) + \sqrt{2\gamma u} \int_0^t e^{-\gamma(t-s)} dB_s \tag{B.2}$$
$$\tilde{x}_t = \tilde{x}_0 + \int_0^t \tilde{v}_s ds.$$

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 (3.1) and (3.4). □

Next we calculate the moments of the Gaussian used in the updates of Algorithm 1. These are obtained by integrating the expression for the discrete-time process presented in Lemma 22.

Lemma 23 *Conditioned on $(\tilde{x}_0, \tilde{v}_0)$, the solution $(\tilde{x}_t, \tilde{v}_t)$ of (3.4) with $\gamma = 2$ and $u = 1/L$ is a Gaussian with conditional mean,*

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

and with conditional covariance,

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

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 22 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:

$$\mathbb{E}\left[\tilde{v}_t\right] = \tilde{v}_0 e^{-2t} - \frac{1}{2L}(1 - e^{-2t})\nabla U(\tilde{x}_0) \tag{B.3}$$
$$\mathbb{E}\left[\tilde{x}_t\right] = \tilde{x}_0 + \frac{1}{2}(1 - e^{-2t})\tilde{v}_0 - \frac{1}{2L}\left(t - \frac{1}{2}\left(1 - e^{-2t}\right)\right)\nabla U(\tilde{x}_0). \tag{B.4}$$

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

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

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

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

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. □

B.2 Controlling the Kinetic Energy

In this section, we establish an explicit bound on the kinetic energy $\mathcal{E}_K$ in (3.9) which is used to control the discretization error at each step.

Lemma 24 (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 3 for $i = 1, \ldots n$, with step size ν and number of iterations n as stated in Theorem 3. Then for all $i = 1, \ldots n$ and for all $t \in [0, \nu]$, 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,

$$\begin{aligned}\mathbb{E}_p\left[\|v\|_2^2\right] &= \mathbb{E}_\zeta\left[\|v - v' + v'\|_2^2\right] \\ &\leq 2\mathbb{E}_{p^*}\left[\|v\|_2^2\right] + 2\mathbb{E}_\zeta\left[\|v - v'\|_2^2\right] \\ &\leq 2\mathbb{E}_{p^*}\left[\|v\|_2^2\right] + 4\mathbb{E}_\zeta\left[\|x + v - (x' + v')\|_2^2 + \|x - x'\|_2^2\right] \\ &= 2\mathbb{E}_{p^*}\left[\|v\|_2^2\right] + 4W_2^2(q, q^*), \end{aligned} \tag{B.5}$$

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 $p^* \propto \exp(-(U(x) + \frac{L}{2}\|v\|_2^2))$, 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 12 to control $\mathbb{E}_{p^*}[\|x - x^*\|_2^2]$. 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 (B.5) 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 (3.12) applied up to the $(i+1)^{th}$ iteration, we know that

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

Thus by (B.5) we have,

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

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

Algorithm 3: 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 $\hat{\nabla} U(\cdot)$

for $i = 0, 1, \ldots, n-1$ **do**
  Sample $(x^{i+1}, v^{i+1}) \sim Z^{i+1}(x^i, v^i)$
end

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

Lemma 25 *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 12 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,

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

□

B.3 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 3.3.3.

Description of Algorithm 3

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

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

The distribution is obtained by integrating the discrete underdamped Langevin diffusion (3.5) 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 $Z^{i+1}(x^i, v^i) \sim p^{(i+1)} = \hat{\Phi}_\nu p^{(i)}$. Derivation is identical to the calculation in Appendix B.1 by replacing exact gradients $\nabla U(\cdot)$ with stochastic gradients $\hat{\nabla}U(\cdot)$. 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 26 *The solution $(\hat{x}_t, \hat{v}_t)$ of the stochastic gradient underdamped Langevin diffusion (3.5) is*

$$\hat{v}_t = \hat{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\hat{\nabla}U(\hat{x}_0)ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s \tag{B.6}$$
$$\hat{x}_t = \hat{x}_0 + \int_0^t \hat{v}_s ds.$$

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 (3.5). □

B.3.1 Discretization Analysis

In Theorem 27, we will bound the discretization error between the discrete process without noise in the gradients (3.4) and the discrete process (3.5) starting from the same initial distribution.

Lemma 27 *Let q_0 be some initial distribution. Let $\tilde{\Phi}_\delta$ and $\hat{\Phi}_\delta$ be as defined in (3.3) 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 (B.2) and (B.6), and using the definition of $\hat{\nabla}U(x)$. We get that

$$\begin{aligned}\hat{v}_\delta &= \tilde{v}_\delta + u\left(\int_0^\delta e^{-\gamma(s-\delta)}ds\right)\xi \\ \hat{x}_\delta &= \tilde{x}_\delta + u\left(\int_0^\delta \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi,\end{aligned} \tag{B.7}$$

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 (B.7), 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 4.

Proof of Theorem 4
From Corollary 12, we have that for any $i \in \{1, \ldots, n\}$

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

By the discretization error bound in Theorem 6 and the sandwich inequality (3.7), we get

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

By the triangle inequality for W_2,

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

Combining this with the discretization error bound established in Lemma 27 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 29 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 13) 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 25 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 24) 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. □

B.4 Technical Results

We state this theorem by Durmus and Moulines [27] used in the proof of Lemma 24.

Theorem 12 *[Theorem 1 in 27] 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 28 *[Theorem 3 in 78] 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 [21] that we will use in the proof of Theorem 4.

Lemma 29 *[Lemma 7 in 21] 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*

$$x_k \leq A^k x_0 + \frac{C}{1-A} + \frac{B^2}{C + \sqrt{(1-A^2)}B} \tag{B.8}$$

for all integers $k \geq 0$.

Appendix C

Proofs for Chapter 4

We outline here the organization of the Appendix.

In Appendix C.1, we list the variables used in this paper, with references to their definitions. In Appendix C.2, we give a description of two small constants, β and ν, which are used throughout our analysis to ensure regularity in time and space.

In Appendix C.3, we give a proof of Theorem 8. In Appendix C.4, we give a proof of Theorem 9.

In Appendix C.5, we specify the construction of the distance function f, which is used to demonstrate contraction. In Appendix C.6, we bound the moments of some of the relevant quantities; these are used in discretization bounds of Appendix C.3 and C.4. In Appendix C.7, we gives proofs of the existence of our coupling constructions. In Appendix C.8, we prove that our coupling constructions have the correct marginals. We also prove that Algorithm 2 exactly implements (4.7).

C.1 Index of Notation

α_f	Parameter of f (4.22). See (C.7)(overdamped) and (4.23) (underdamped).
β	Constant in defining ℓ. See also Section C.2.
c_κ	See (4.6)
C_m	Underdamped contraction rate, see (4.9).
C_o	Overdamped contraction rate, see (C.8).
d	Dimension of x
f	See (4.22)
κ	Condition number, defined after Assumption (A3)
ℓ	Twice continuously differentiable approximation to $\|\cdot\|_2$ with β error. See Lemma 31.
L	Lipschitz gradient parameter, see Assumption (A1).
$\mathcal{L}$	Lyapunov function. See (C.11) (overdamped) and (4.27) (underdamped)
m	Contraction parameter outside the R ball. See Assumption (A3)
$\mathcal{M}$	See (C.5) (overdamped) (4.15) (underdamped)
q	See Lemma 30
r	See (4.19).
R	See Assumption (A3)
$\mathcal{R}_f$	Parameter of f (4.22). See (C.7)(overdamped) and (4.23) (underdamped).
T_{sync}	See (4.8).
τ_k	See (4.17)
w_t	Short for $u_t - v_t$, defined in (4.16)
z_t	Short for $x_t - y_t$, defined in (4.16)

C.2 Two Small Constants

On ℓ and β:
In this paper, we will take $\beta = 1/poly(L, 1/m, d, R, e^{LR^2})$ to be a small constant. See the proofs of Theorem 8 and Theorem 9 for the exact values of β. Intuitively, β is a radius inside of which we perform the following smoothing:

We define a function $q(r)$ in (C.1), which is a smoothed approximation of $|r|$, such that it has continuous second derivatives everywhere. Specifically, for $r \leq \beta/2$, $q(r)$ is a cubic spline.

$$q(r) = \begin{cases} \frac{\beta}{3} + \frac{8}{3\beta^2} \cdot r^3, & \text{for } r \in [0, \beta/4] \\ \frac{5\beta}{12} - r + \frac{4}{\beta} \cdot r^2 - \frac{8}{3\beta^2} \cdot r^3, & \text{for } r \in [\beta/4, \beta/2] \\ r, & \text{for } r \in [\beta/2, \infty]. \end{cases} \tag{C.1}$$

This allows us to define a smoothed version of $\|x\|_2$, which has continuous second derivatives everywhere:

$$\ell(x) = q(\|x\|_2). \tag{C.2}$$

In various parts of our proof, we replace $\|\cdot\|_2$ by its smooth approximation $\ell(\cdot)$, defined in Lemma 31, parametrized by β; a small β means that $\ell(\cdot)$ and $\|\cdot\|_2$ are close. We need to be careful as $\ell(\cdot)$ is strongly convex, with parameter $1/\beta^2$, in a $\beta/2$ radius around zero. We thus need to design our dynamics to ensure that the coupling has no noise in this region (see Eq. (4.15)).

When reading the proofs, it helps to think of $\ell(\cdot) = \|\cdot\|_2$ and $\beta = 0$, as we can take β to be arbitrarily small without additional computation costs.

On ν:
In order to demonstrate the existence of a strong solution to the coupling presented in Section 4.4.1 (Lemma 15), we switch between synchronous and reflection coupling at deterministic, finite intervals of width ν.

This is not necessary strictly speaking, as there are results that ensure the existence of solutions of an SDE when the diffusion and drift coefficients are discontinuous but have finite variation. However, we choose to use a discretized coupling as the existence of its solution can be verified by using standard results.

This discretized coupling scheme adds an error term σ_t (see Eq. (4.25)). We show in Lemma 42 that this is $o(\nu^2)$.

When reading the proofs, it helps to think of $\nu = 0$ and $\sigma_t = 0$, as we can take ν to be arbitrarily small without additional computation costs. In the proof, it suffices to let $\nu = 1/poly(L, 1/m, d, R, e^{LR^2})$. See the proof and Theorem 9 for the exact value of ν.

Note that ν is distinct from (and unrelated to) δ, which is the step-size of the under-damped Langevin MCMC algorithm (Algorithm 2). The step size δ and the corresponding discretization error ξ_t, cannot be made arbitrarily small without additional computation costs.

Lemma 30 *Let β be any positive real. Let $q(r)$ be defined as in* (C.1)*, reproduced below for ease of reference:*

$$q(r) = \begin{cases} \frac{\beta}{3} + \frac{8}{3\beta^2} \cdot r^3, & \text{for } r \in [0, \beta/4] \\ \frac{5\beta}{12} - r + \frac{4}{\beta} \cdot r^2 - \frac{8}{3\beta^2} \cdot r^3, & \text{for } r \in [\beta/4, \beta/2] \\ r, & \text{for } r \in [\beta/2, \infty]. \end{cases}$$

Then,

1. *$q(r)$, $q'(r)/r$ and $q''(r)/r^2$ exist for all r, and are continuous.*
2. *For all r, $q(r)$ satisfies $\beta/3 \leq q(r)$ and $|r - q(r)| \leq \beta/3$. In addition, $q(r) = r$ for $r \geq \beta/2$.*
3. *$q'(r)$ is monotonically nondecreasing, $q'(r) = 1$ for $r \geq \beta/2$, and $q'(r) = 0$ for $r = 0$.*
4. *$q''(r) = 0$ for all $r \geq \beta/2$.*

Proof
Taking derivatives, we verify that

$$q'(r) = \begin{cases} \frac{8}{\beta^2} \cdot r^2, & \text{for } r \in [0, \beta/4] \\ -1 + \frac{8}{\beta} \cdot r - \frac{8}{\beta^2} \cdot r^2, & \text{for } r \in [\beta/4, \beta/2] \\ 1, & \text{for } r \in [\beta/2, \infty]; \end{cases}$$

$$q''(r) = \begin{cases} \frac{16}{\beta^2} \cdot r, & \text{for } r \in [0, \beta/4] \\ \frac{8}{\beta} - \frac{16}{\beta^2} \cdot r, & \text{for } r \in [\beta/4, \beta/2] \\ 0, & \text{for } r \in [\beta/2, \infty]. \end{cases}$$

All the claims can then be verified algebraically. □

Lemma 31 *For a given $\beta > 0$, let $q(r)$ be as defined above. Define $\ell(x) : \mathbb{R}^n \to \mathbb{R}^+$ as $\ell(x) = q(\|x\|_2)$. Then*

1. *For all x, $\ell(x)$ satisfies $\beta/3 \leq \ell(x)$ and $|\ell(x) - \|x\|_2| \leq \beta/3$. In addition, for $\|x\|_2 \geq \beta/2$, $\ell(x) = \|x\|_2$.*
2. *$\nabla \ell(x) = q'(\|x\|_2) \frac{x}{\|x\|_2}$, for all x, $\|\nabla \ell(x)\|_2 \leq 1$, for $\|x\|_2 \geq \beta/2$, $\nabla \ell(x) = \frac{x}{\|x\|_2}$.*
3. *for $\|x\|_2 \geq \beta/2$, $\nabla^2 \ell(x) = q''(\|x\|_2) \frac{xx^T}{\|x\|_2^2} + q'(\|x\|_2) \frac{1}{\|x\|_2} \left(I - \frac{xx^T}{\|x\|_2^2} \right)$.*

4. $\nabla \ell(x)$ *and* $\nabla^2 \ell(x)$ *are defined everywhere and continuous. In particular, for* $\|x\|_2 \leq \beta/2$,

$$\|\nabla \ell(x)\|_2 \leq 4, \text{ and } \left\|\nabla^2 \ell(x)\right\|_2 \leq \frac{8}{\beta}.$$

Proof

1. Immediate from Lemma 30.2.
2. By Chain rule, $\nabla \ell(x) = q'(\|x\|_2)\frac{x}{\|x\|_2}$. Furthermore, From the Lemma 30.1, we verify that $\nabla \ell(x)$ is defined everywhere, including at 0. The remaining claims follow from Lemma 30.3
3. This is just chain rule, together with Lemma 30.1, which guarantees the existence of $q''(\|x\|_2)/\|x\|_2^2$ for all x.
4. Existence and continuity follow from Lemma 30.

□

C.3 Proofs for overdamped Langevin Monte Carlo

C.3.1 Coupling construction for overdamped Langevin MCMC

Let β be a small constant (see proof of Theorem 8 for the exact value), and let $\ell(x) = q(\|x\|_2)$ be the smoothed approximation of $\|x\|_2$ as defined in Appendix C.2.

We begin by establishing the convergence of the continuous-time process in Eq. (4.3) to the invariant distribution. Similar to [32], we construct a coupling between the SDEs described by Eq. (4.3) and Eq. (4.4). We initialize the coupling at

$$\begin{aligned} x_0 &= 0 \\ y_0 &\sim p^*(y), \end{aligned}$$

and evolve the pair (x_t, y_t) according to the dynamics

$$dx_t = -\nabla U\left(x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right) dt + \sqrt{2} dB_t \tag{C.3}$$

$$dy_t = -\nabla U(y_t) dt + \sqrt{2} dB_t - 2\sqrt{2}\gamma_t \gamma_t^T dB_t + \sqrt{2}\bar{\gamma}_t \bar{\gamma}_t^T dA_t, \tag{C.4}$$

where the terms γ_t and $\bar{\gamma}_t$ are defined as:

$$\begin{aligned}\gamma_t :=& (\mathcal{M}(\|z_t\|_2))^{1/2} \frac{z_t}{\|z_t\|_2} \\ \bar{\gamma}_t :=& \left(1 - (1 - 2\mathcal{M}(\|z_t\|_2))^2\right)^{1/4} \frac{z_t}{\|z_t\|_2}, \\ \text{with} \qquad z_t :=& x_t - y_t,\end{aligned}$$

$$\mathcal{M}(r) := \begin{cases} 1, & \text{for } r \in [\beta, \infty) \\ \frac{1}{2} + \frac{1}{2}\cos\left(r \cdot \frac{2\pi}{\beta}\right), & \text{for } r \in [\beta/2, \beta] \\ 0, & \text{for } r \in [0, \beta/2]. \end{cases} \tag{C.5}$$

We use the convention that $0/0 = 0$ when $\|z_t\|_2 = 0$. It can be verified that γ_t and $\bar{\gamma}_t$ are Lipschitz and gradient-Lipschitz for all $z_t \in \mathbb{R}^d$.

The following lemma confirms that the dynamics (C.4) give the correct distribution.

Lemma 32 *The dynamics in Eq.* (C.4) *is distributionally equivalent to the dynamics defined in Eq.* (4.3).

We defer the proof to Appendix C.8.

For notational convenience, we define

$$\begin{aligned}\nabla_t :=& \nabla U(x_t) - \nabla U(y_t) \\ \Delta_t :=& \nabla U(x_{\lfloor \frac{t}{\delta} \rfloor \delta}) - \nabla U(x_t).\end{aligned} \tag{C.6}$$

Finally, we construct the Lyapunov function that we will use to show convergence. Let f be as defined in Eq. (4.22), with

$$\alpha_f := \frac{L}{4}, \quad \text{and,} \quad \mathcal{R}_f := R. \tag{C.7}$$

Define a constant,

$$C_o := \min\left\{\frac{1}{8R^2} e^{-LR^2/2}, m\right\}, \tag{C.8}$$

and finally, define two stochastic processes

$$\xi_t := L \int_0^t e^{-C_o(t-s)} \left\| x_s - x_{\lfloor \frac{s}{\delta} \rfloor \delta} \right\|_2 ds \tag{C.9}$$

$$\phi_t := \int_0^t e^{-C_o(t-s)} f'(\|z_s\|_2) \left\langle \frac{z_s}{\|z_s\|_2}, \left(2\sqrt{2}\gamma_s \gamma_s^T dB_s + \sqrt{2}\bar{\gamma}_s \bar{\gamma}_s^T dA_s\right) \right\rangle. \tag{C.10}$$

With these definitions, the following stochastic process $\mathcal{L}_t$ acts as our Lyapunov function:

$$\mathcal{L}_t := f(\ell(z_t)) - \xi_t - \phi_t. \tag{C.11}$$

C.3.2 Proof of Theorem 8

The proof follows in three steps. In Step 1 we analyze the evolution of $f(\ell(z_t))$ using Itô's Lemma. In Step 2 we use this to show that the Lyapunov function $\mathcal{L}_t$ which is defined in Eq. (C.11) contracts at a sufficiently fast rate. Finally in Step 3 we relate this contraction in the Lyapunov function to a bound on the iteration complexity of (4.4).

We note that the technique in establishing Step 1 is essentially taken from [32].

Step 1: By Itô's Lemma applied to $f(\ell(z_t))$,

$$df(\ell(z_t)) = \underbrace{\langle \nabla_z f(\ell(z_t)), -\nabla_t - \Delta_t \rangle}_{=:\spadesuit} dt + \underbrace{\frac{1}{2}\mathrm{tr}\big(\nabla_z^2 f(\ell(z_t))\big(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\big)\big)}_{=:\heartsuit} dt + \left\langle \nabla_z f(\ell(z_t)), 2\sqrt{2}\gamma_t\gamma_t^T dB_t - \sqrt{2}\bar{\gamma}_t\bar{\gamma}_t^T dA_t \right\rangle.$$

We first bound the term $\spadesuit$. Note that $\nabla_z f(\ell(z_t)) = f'(\ell(z_t))\nabla\ell(z_t)$. When $z_t = 0$, $\ell(z_t) = \beta/3 + 8/(3\beta^2)$ by (C.1) and (C.2), so $f(\ell(z_t))$ is well defined by Lemma 55.F2. By Lemma 31.2, $\nabla\ell(z_t) = \frac{q'(\|z_t\|_2}{\|z_t\|_2}$. Since $q'(r)/r$ is always well defined for all r (by Lemma 30.1), we conclude that when $z_t = 0$, $\nabla\ell(z_t) = 0$, and thus $\nabla f(\ell(z_t)) = 0$ as well. For the case when $\|z_t\|_2 \neq 0$ we have,

$$\nabla f(\ell(z_t)) = f'(\ell(z_t))q'(\|z_t\|_2)\frac{z_t}{\|z_t\|_2},$$

where $q(\cdot)$ is the function used to define ℓ (see Lemma 30). Thus

$$\begin{aligned}
\spadesuit &= \langle \nabla f(\ell(z_t)), -\nabla_t - \Delta_t \rangle \\
&= f'(\ell(z_t)) \cdot q'(\|z_t\|_2) \cdot \left\langle \frac{z_t}{\|z_t\|_2}, -\nabla_t - \Delta_t \right\rangle \\
&\overset{(i)}{\leq} f'(\ell(z_t)) \cdot q'(\|z_t\|_2) \left\langle \frac{z_t}{\|z_t\|_2}, -\nabla_t \right\rangle + \|\Delta_t\|_2 \\
&\overset{(ii)}{\leq} \mathbb{1}\{\|z_t\|_2 \in [0,\beta]\} \cdot L\beta + \mathbb{1}\{\|z_t\|_2 \in [R,\infty]\} \cdot (-m\|z_t\|_2) \\
&\quad + \mathbb{1}\{\|z_t\|_2 \in (\beta, R)\} \cdot f'(\ell(z_t)) \cdot (L\|z_t\|_2) + \|\Delta_t\|_2,
\end{aligned}$$

where (i) is by the Cauchy-Schwarz inequality, along with the fact that $|f'(r)| \leq 1$ (see (F2) of Lemma 55), and Lemma 30.3. The inequality in (ii) can be verified by considering three disjoint events. When $\|z_t\|_2 \in [0,\beta]$, the bound follows by Cauchy-Schwarz, (F2) of Lemma 55, combined with Lemma 30.3. When $\|z_t\|_2 \in [R,\infty]$ the bound follows from strong convexity (Assumption (A3)). When $\|z_t\|_2 \in (\beta, R]$, we bound the term using Cauchy-Schwarz, Assumption (A1), and Lemma 30.3.

Next, we consider the other term $\heartsuit = \frac{1}{2}\mathrm{tr}\big(\nabla_z^2 f(\ell(z_t))\big(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\big)\big)$.

First, consider the case when $\|z_t\|_2 = 0$. By chain rule, and by definition of ℓ in (C.2) and q in (C.1),

$$\nabla^2 f(\ell(z_t)) = f'(\ell(z_t))\nabla^2 q(\|z_t\|_2) + f''(\ell(z_t))\nabla q(\|z_t\|_2)(\nabla q(\|z_t\|_2))^T \tag{C.12}$$

We further verify that at $\|z_t\|_2 = 0$,

$$\nabla q(\|z_t\|_2) = \frac{8}{3\beta^2}\|z_t\|_2 z_t = 0 \tag{C.13}$$

$$\nabla^2 q(\|z_t\|_2) = \frac{8}{3\beta^2}\|z_t\|_2 I + \frac{z_t z_t^T}{\|z_t\|_2} = 0 \tag{C.14}$$

Thus for $\|z_t\|_2 = 0$, $\nabla^2 f(\ell(z_t)) = 0$, and the following holds:

$$\heartsuit = 0 = \mathbb{1}\{\|z_t\|_2 \in [\beta, R]\} 4 f''(\ell(z_t)).$$

Next, consider the case when $\|z_t\|_2 \neq 0$,

$$\begin{aligned}\nabla^2 f(\ell(z_t)) =& f''(\ell(z_t))q'(\|z_t\|_2)^2 \frac{z_t z_t^T}{\|z_t\|_2^2} + f'(\ell(z_t))q'(\|z_t\|_2)\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2}\right) \\ &+ f'(\ell(z_t))q''(\|z_t\|_2)\frac{z_t z_t^T}{\|z_t\|_2^2}.\end{aligned}$$

Exapanding using the definition of $\heartsuit$,

$$\begin{aligned}\heartsuit =& \frac{1}{2}\mathrm{tr}\left(\nabla_z^2 f(\ell(z_t))\left(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\right)\right) \\ \overset{(i)}{=}& \underbrace{\frac{1}{2}\mathrm{tr}\left(f''(\ell(z_t))\cdot q'(\|z_t\|_2)^2 \cdot \frac{z_t z_t^T}{\|z_t\|_2^2}\left(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\right)\right)}_{=:\heartsuit_1} \\ &+ \underbrace{\frac{1}{2}\mathrm{tr}\left(f'(\ell(z_t))\cdot q'(\|z_t\|_2)\cdot\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2}\right)\left(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\right)\right)}_{=:\heartsuit_2} \\ &+ \underbrace{\frac{1}{2}\mathrm{tr}\left(f'(\ell(z_t))q''(z_t)\frac{z_t z_t^T}{\|z_t\|_2^2}\left(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\right)\right)}_{=:\heartsuit_3},\end{aligned}$$

where (i) is by the expression for $\nabla^2 f(\ell(z_t))$ above.

Before proceeding, we verify by definition of γ_t and $\bar{\gamma}_t$ in Eq. (C.5) that

$$\mathrm{tr}\left(\frac{z_t z_t^T}{\|z_t\|_2^2}\left(8\gamma_t\gamma_t^T + 2\bar{\gamma}_t\bar{\gamma}_t^T\right)\right) = 8\|\gamma_t\|_2^2 + 2\|\bar{\gamma}_t\|_2^2. \tag{C.15}$$

First we simplify $\heartsuit_1$:

$$\begin{aligned}\heartsuit_1 =& \frac{1}{2} f''(\ell(z_t)) \cdot q'(\|z_t\|_2)^2 \cdot \left(8\|\gamma_t\|_2^2 + 2\|\bar{\gamma}_t\|_2^2\right) \\ \overset{(i)}{\leq}& \mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\}\left(f''(\ell(z_t)) \cdot \left(4\|\gamma_t\|_2^2 + \|\bar{\gamma}_t\|_2^2\right)\right) \\ \overset{(ii)}{=}& \mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\} 4 f''(\ell(z_t)),\end{aligned}$$

where the inequality (i) is because $f''(r) \leq 0$ for all $r > 0$ (by Lemma 55.(F5)), $q'(r) \geq 0$ for all r (by Lemma 30.3) and $q'(r) = 1$ for all $r \geq \beta/2$ (Lemma 30.3). The equality in (ii) is because $\gamma_t = 1$ and $\bar{\gamma}_t = 0$ for $\|z_t\|_2 \geq \beta$ (by their definition in Eq. (C.5)).

Next, using Eq. (C.15), we can immediately verify that $\heartsuit_2 = 0$.

Finally, we focus on $\heartsuit_3$,

$$\heartsuit_3 = \frac{1}{2} f'(\ell(z_t)) q''(z_t)\left(8\|\gamma_t\|_2^2 + 2\|\bar{\gamma}_t\|_2^2\right) = 0,$$

where we use the fact that $q''(\|z_t\|_2) = 0$ if $\|z_t\|_2 \geq \beta/2$ (by Lemma 30.4) and $\gamma_t = \bar{\gamma}_t = 0$ if $\|z_t\|_2 \leq \beta/2$ (by its definition in Eq. (C.5)).

Putting together the bounds on $\heartsuit_1$, $\heartsuit_2$ and $\heartsuit_3$, we can upper bound $\heartsuit$ as

$$\heartsuit \leq \mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\} 4 f''(\ell(z_t)).$$

Combining the upper bounds on $\spadesuit$ and $\heartsuit$,

$$\begin{aligned}\spadesuit + \heartsuit \leq & \underbrace{\mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\}\left(L\|z_t\|_2 f'(\ell(z_t)) + 4 f''(\ell(z_t))\right)}_{=:\clubsuit} \\ & + \mathbb{1}\left\{\|z_t\|_2 \in [R, \infty]\right\} \cdot (-m\|z_t\|_2) + \|\Delta_t\|_2 + L\beta.\end{aligned}$$

Let us now focus on $\clubsuit$. By Lemma 55,

$$\begin{aligned}\clubsuit =& \mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\} \cdot \left(L\|z_t\|_2 f'(\ell(z_t)) + 4 f''(\ell(z_t))\right) \\ \overset{(i)}{\leq}& \mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\} \cdot \left(L \cdot \ell(z_t) f'(\ell(z_t)) + 4 f''(\ell(z_t)) + L\beta/3\right) \\ \overset{(ii)}{\leq}& \mathbb{1}\left\{\|z_t\|_2 \in [\beta, R]\right\} \cdot \left(-C_o f(\ell(z_t))\right) + L\beta/3 \\ \overset{(iii)}{\leq}& \mathbb{1}\left\{\|z_t\|_2 \in [0, R]\right\}\left(-C_o f(\ell(z_t))\right) + (L + 8C_o)\beta \\ \overset{(iv)}{\leq}& \mathbb{1}\left\{\|z_t\|_2 \in [0, R]\right\}\left(-C_o f(\ell(z_t))\right) + 10L\beta,\end{aligned}$$

where (i) is because $\left|\|z_t\|_2 - \ell(z_t)\right| \leq \beta/3$ (by Lemma 31.1) and because $|f'(r)| \leq 1$ for all $r > 0$ (by Lemma 55.(F2)). The inequality in (ii) is by Lemma 55 (F4), our definition of C_o in (C.8), and the fact that $\|z_t\|_2 \in [\beta, R]$ implies $\ell(z_t) \leq R$ (Lemma 31.1). Inequality (iii) is again by Lemma 31.1 and Lemma 55 (F3). Finally, (iv) is by (C.8) and $m \leq L$, a consequence of Assumptions (A1) and (A3).

Thus,

$$\begin{aligned}\spadesuit + \heartsuit \leq & \mathbb{1}\left\{\|z_t\|_2 \in [0,R]\right\}(-C_o f(\ell(z_t))) + \mathbb{1}\left\{\|z_t\|_2 \in [R,\infty]\right\}\cdot(-m\|z_t\|_2) + 11L\beta + \|\Delta_t\|_2 \\ \leq & -C_o f(\ell(z_t)) + 12L\beta + \|\Delta_t\|_2,\end{aligned}$$

where the second line is by Lemma 31.1 and 55.(F3), and by $m \leq L$.

Putting this together with the expression for $df(\ell(z_t))$,

$$\begin{aligned}df(\ell(z_t)) \leq & (-C_o f(\ell(z_t)) + 12L\beta + \|\Delta_t\|_2)dt + \left\langle \nabla_z f(\ell(z_t)), 2\sqrt{2}\gamma_t\gamma_t^T dB_t + \sqrt{2}\bar{\gamma}_t\bar{\gamma}_t^T dA_t \right\rangle \\ \leq & \Big(-C_o f(\ell(z_t)) + 12L\beta + L\Big\|x_t - x_{\lfloor \frac{t}{\delta}\rfloor \delta}\Big\|_2\Big)dt \\ & + \left\langle \nabla_z f(\ell(z_t)), 2\sqrt{2}\gamma_t\gamma_t^T dB_t + \sqrt{2}\bar{\gamma}_t\bar{\gamma}_t^T dA_t \right\rangle.\end{aligned}$$

The second inequality uses the definition of Δ_t in Eq. (C.6) and Assumption (A1).

Step 2: If we consider the evolution of the Lyapunov function $\mathcal{L}_t$ (defined in Eq. (C.11)), we can verify that

$$\begin{aligned}d\mathcal{L}_t = & d(f(\ell(z_t)) - \phi_t - \xi_t) \\ \overset{(i)}{\leq} & -C_o(f(\ell(z_t)) - \phi_t - \xi_t)dt + 12L\beta dt \\ = & -C_o\mathcal{L}_t dt + 12L\beta dt,\end{aligned}$$

where the simplification in inequality (i) can be verified by taking time derivatives of stochastic processes ϕ_t and ξ_t defined in Eq. (C.10) and Eq. (C.9).

Applying Grönwall's inequality,

$$\mathcal{L}_t \leq e^{-C_o t}\mathcal{L}_0 + \int_0^t e^{-C_o(t-s)}12L\beta ds \leq e^{-C_o t}\mathcal{L}_0 + \frac{12L\beta}{C_o}.$$

Using the definition of $\mathcal{L}_t$ in Eq. (C.11) we get,

$$f(\ell(z_t)) \leq e^{-C_o t}f(\ell(z_0)) + \xi_t + \phi_t.$$

Taking expectations with respect to the Brownian motion yields:

$$\mathbb{E}\left[f(\ell(z_t))\right] \leq e^{-C_o t}\mathbb{E}\left[f(\ell(z_0))\right] + \mathbb{E}\left[\xi_t\right] + \mathbb{E}\left[\phi_t\right]. \tag{C.16}$$

By the definition of ϕ_t in Eq. (C.10), we verify that $\mathbb{E}\left[\phi_t\right] = 0$, and by definition of ξ_t in

Eq. (C.9),

$$\begin{aligned}
\mathbb{E}\left[\xi_t\right] &= \int_0^t e^{-C_o(t-s)}\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2\right] ds \\
&\leq \int_0^t e^{-C_o(t-s)}\mathbb{E}\left[\left\|\left(s - \left\lfloor\frac{s}{\delta}\right\rfloor\delta\right)\nabla U(x_{\lfloor \frac{s}{\delta}\rfloor\delta}) + \int_{\lfloor \frac{s}{\delta}\rfloor\delta}^s dB_r\right\|_2\right] ds \\
&\leq \int_0^t e^{-C_o(t-s)}\left(\mathbb{E}\left[\delta L\left\|x_{\lfloor \frac{s}{\delta}\rfloor\delta}\right\|_2\right] + \sqrt{\delta d}\right) ds \\
&\leq \int_0^t e^{-C_o(t-s)}\left(2\delta L\sqrt{R^2 + d/m} + \sqrt{\delta d}\right) ds \\
&\leq \frac{2\delta L\sqrt{R^2 + d/m} + \sqrt{\delta d}}{C_o}
\end{aligned}$$

We can also bound the initial value of $\mathbb{E}\left[f(\ell(z_0))\right]$ as follows:

$$\mathbb{E}\left[f(\ell(z_0))\right] \overset{(i)}{=} \mathbb{E}\left[f(\ell(y_0))\right] \overset{(ii)}{\leq} \mathbb{E}\left[\ell(y_0)\right] \overset{(iii)}{\leq} \mathbb{E}\left[\|y_0\|_2\right] + \beta/3 \overset{(iv)}{\leq} \sqrt{R^2 + \frac{d}{m}} + \beta/3,$$

where (i) is because $x^(0) = 0$ in Eq. (C.3), (ii) is by Lemma 55.(F3), (iii) is by Lemma 31.1, and finally (iv) is by Lemma 61.

Let n be the number of time steps, so that $t = n\delta$. Substituting into the inequality in Eq. (C.16), we get

$$\mathbb{E}\left[f(\ell(z_{n\delta}))\right] \leq e^{-C_o(n\delta)}\left(32\sqrt{R^2 + \frac{d}{m}} + \beta/3\right) + \frac{2\delta L\sqrt{R^2 + d/m} + \delta d}{C_o} + \frac{12L\beta}{C_o}.$$

Step 3: We translate our bound on $\mathbb{E}\left[f(\ell(z_{n\delta}))\right]$ to a bound on $\mathbb{E}\left[\|z_{n\delta}\|_2\right]$, which implies a bound in 1-Wasserstein distance. By Lemma 55(F3),

$$\begin{aligned}
&\mathbb{E}\left[\|z_{n\delta}\|_2\right] \\
&\quad\leq 2e^{L(R+\beta)^2/2}\left(e^{-C_o(n\delta)}\left(32\sqrt{R^2 + \frac{d}{m}} + \beta/3\right) + \frac{2\delta L\sqrt{R^2 + d/m} + \delta d}{C_o} + \frac{12L\beta}{C_o}\right) \\
&\quad\leq 4e^{LR^2/2}\left(e^{-C_o(n\delta)}\left(32\sqrt{R^2 + \frac{d}{m}}\right) + \frac{2\delta L\sqrt{R^2 + d/m} + \delta d}{C_o}\right),
\end{aligned}$$

where for the second inequality, it suffices to let $\beta = \delta d/6$

For a given ε, the first term is less than $\varepsilon/2$ if

$$n\delta \geq 10\left(\log\left(\frac{R^2 + d/m}{\varepsilon}\right) + LR^2\right)\cdot\frac{1}{C_o}.$$

The second term is less than $\varepsilon/2$ if

$$\delta \leq \frac{1}{10} e^{-LR^2/2} \min\left\{ \frac{\varepsilon}{\sqrt{R^2 + d/m}}, \frac{\varepsilon^2 C_o}{d} \right\}.$$

By the definition of C_o in Eq. (C.8),

$$C_o \leq \frac{1}{8} \min\left\{ \frac{\exp\left(-LR^2/2\right)}{R^2}, m \right\} = \frac{\exp\left(LR^2/2\right)}{8R^2},$$

where the equality is by our assumption on the strong convexity parameter m in the theorem statement. Recall that we also assume that $\varepsilon \leq \frac{dR^2}{\sqrt{d/m+R^2}}$. Thus we can verify that

$$\min\left\{ \frac{\varepsilon}{\sqrt{R^2 + d/m}}, \frac{\varepsilon^2 C_o}{d} \right\} = \frac{\varepsilon^2 C_o}{d}.$$

Putting everything together, we obtain a guarantee that $\mathbb{E}\left[\|z_t\|_2\right] \leq \varepsilon$ if

$$\delta = \frac{\varepsilon^2 \exp\left(-LR^2\right)}{2^{10} R^2 d},$$

and

$$n \geq 2^{18} \log\left(\frac{R^2 + d/m}{\varepsilon}\right) \cdot R^4 \cdot \exp\left(\frac{3LR^2}{2}\right) \cdot \frac{d}{\varepsilon^2},$$

as prescribed by the theorem statement.

C.4 Proofs for Underdamped Langevin Monte Carlo

C.4.1 Overview

The main idea behind the proof is to show that $\mathcal{L}_t$ contracts with probability one by a factor of $e^{-C_m\nu}$, going from $t = (k-1)\nu$ to $t = k\nu$. The result can be found in Lemma 50 in Section C.4.5. The proof considers four cases:

1. $\mu_{k-1} = 1, \mu_k = 1$. In Lemma 53 in Section C.4.5, we show that $\mathcal{L}_{k\nu} \leq e^{-C_m\nu}\mathcal{L}_{(k-1)\nu}$. The proof of this result in turn uses Lemma 33 in Section C.4.2, which shows that $\mathcal{L}_t$ contracts at a rate of $-C_m$ over the interval $t \in [(k-1)\nu, k\nu]$.

2. $\mu_{k-1} = 1, \mu_k = 0$. In Lemma 54 in Section C.4.5, we show that $\mathcal{L}_{k\nu} \leq e^{-C_m\nu}\mathcal{L}_{(k-1)\nu}$. The proof of this result is almost identical to the preceding case $\mu_{k-1} = 1, \mu_k = 1$. (In particular, $\mathcal{L}_t$ undergoes no jump in value at $t = k\nu$, in spite in the change in value from $\mu_{k-1} = 1$ to $\mu_k = 0$. See proof for details.)

3. $\mu_{k-1} = 0, \mu_k = 0$. In Lemma 52 in Section C.4.5, we show that $\mathcal{L}_{k\nu} \leq e^{-C_m\nu}\mathcal{L}_{(k-1)\nu}$. The proof of this result is mainly based on the definition of $\mathcal{L}_t$.

4. $\mu_{k-1} = 0, \mu_k = 1$. In Lemma 51 in Section C.4.5, we show that $\mathcal{L}_{k\nu} \leq e^{-C_m\nu}\mathcal{L}_{(k-1)\nu}$. This case is somewhat tricky, as $\mathcal{L}_t$ undergoes a jump in value at $t = k\nu$. Specifically, $\mathcal{L}_t$ jumps from $e^{-C_m T_{sync}}\big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\big) - (\sigma_{k\nu} + \phi_{k\nu})$ to $f(r_{k\nu}) - \xi_{k\nu} - (\sigma_{k\nu} + \phi_{k\nu})$. We prove that this jump is always negative (Lemma 34, Section C.4.3). The proof of Lemma 36 in turn relies on a contraction result in Lemma 37.

Having proven Lemma 50, we prove Theorem 9 by applying Lemma 50 recursively, and showing that $\mathbb{E}[\mathcal{L}_t]$ sandwiches the Wasserstein distance $W_1(p_t, p^*)$.

C.4.2 Contraction under Reflection Coupling

Our main result is stated as Lemma 33. It shows that $\mu_k f(r_t)$ contracts at a rate of $\exp(-C_m t)$, plus some discretization error terms.

Lemma 33 *For any positive integer k, with probability one we have,*

$$\begin{aligned}\mu_k \cdot \big(f(r_{(k+1)\nu}) - \xi_{(k+1)\nu}\big) - \big(\sigma_{(k+1)\nu} + \phi_{(k+1)\nu}\big) \\ \leq e^{-C_m\nu}(\mu_k \cdot (f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})) + 5\beta\nu.\end{aligned}$$

Proof
If $\mu_k = 0$, by definition of σ_t and ϕ_t in (4.26), $\sigma_{(k+1)\nu} = e^{-C_m\nu}\sigma_{k\nu}$ and $\phi_{(k+1)\nu} = e^{-C_m\nu}\phi_{k\nu}$, so the inequality reduces to $e^{-C_m\nu}(-(\sigma_{k\nu} + \phi_{k\nu})) \leq e^{-C_m\nu}(-(\sigma_{k\nu} + \phi_{k\nu})) + 5\beta\nu$ which is clearly true. To simplify notation, we leave out the factor of μ_k in subsequent expressions and assume that $\mu_k = 1$ unless otherwise stated.

For the rest of this proof, we will consider time $s \in [k\nu, (k+1)\nu)$ for some k.

Let us first establish some useful derivatives of the function f:

$$\begin{aligned}
\nabla_z f(r_s) &= f'(r_s)\cdot(1+2c_\kappa)q'(\|z_s\|_2)\cdot\frac{z_s}{\|z_s\|_2} + f'(r_s)\cdot q'(z_s+w_s)\cdot\frac{z_s+w_s}{\|z_s+w_s\|_2},\\
\nabla_w f(r_s) &= f'(r_s)\cdot q'(\|z_s+w_s\|_2)\cdot\frac{z_s+w_s}{\|z_s+w_s\|_2},\\
\nabla_w^2 f(r_s) &= f'(r_s)\cdot q'(\|z_s+w_s\|_2)\cdot\frac{1}{\|z_s+w_s\|_2}\left(I-\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\\
&\quad + f'(r_s)q''(\|z_s+w_s\|_2)\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\\
&\quad + f''(r_s)q'(\|z_s+w_s\|_2)^2\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}. \qquad \text{(C.17)}
\end{aligned}$$

The derivatives follow from Lemma 31 and by the definition of r_t in Eq. (4.19). From Lemma 31.3, $\nabla_w^2 f((1+2c_\kappa)\ell(z)+\ell(z+w))$ exists everywhere and is continuous, with $\nabla_w^2 f((1+2c_\kappa)\ell(z)+\ell(z+w))|_{z+w=0}=0$. Note that, we use the convention $0/0=0$.

For any $s\in[k\nu,(k+1)\nu)$, we have:

$$\begin{aligned}
d\mu_k\cdot f(r_s) &\overset{(i)}{=} \mu_k\cdot\langle\nabla_z f(r_s), dz_s\rangle + \langle\nabla_w f(r_s), dw_s\rangle\\
&\quad + \mu_k\cdot\frac{8c_\kappa}{L}\gamma_s^T\nabla_w^2 f(r_s)\gamma_s ds + \frac{2c_\kappa}{L}\bar\gamma_s^T\nabla_w^2 f(r_s)\bar\gamma_s ds\\
&\overset{(ii)}{=} \mu_k\cdot\underbrace{\left(\langle\nabla_z f(r_s), w_s\rangle + \left\langle\nabla_w f(r_s), -2w_s-\frac{c_\kappa}{L}\nabla_s-\frac{c_\kappa}{L}\Delta_s\right\rangle\right)}_{=:\spadesuit} ds\\
&\quad + \mu_k\cdot\underbrace{\left(\left(\frac{8c_\kappa}{L}\gamma_s^T\nabla_w^2 f(r_s)\gamma_s + \frac{2c_\kappa}{L}\bar\gamma_s^T\nabla_w^2 f(r_s)\bar\gamma_s\right)\right)}_{=:\heartsuit} ds\\
&\quad + \mu_k\cdot\left\langle\nabla_w f(r_s), \left(4\sqrt{\frac{c_\kappa}{L}}\gamma_t\gamma_t^T dB_t + 2\sqrt{\frac{c_\kappa}{L}}\bar\gamma_t\bar\gamma_t^T dA_t\right)\right\rangle ds, \qquad \text{(C.18)}
\end{aligned}$$

where (i) follows from Itô's Lemma, and (ii) follows from Eqs. (4.11) - (4.14), and the definition of ∇_t and Δ_t in Eq. (4.20).

In the sequel, we upper bound the terms $\spadesuit, \heartsuit, \clubsuit$ separately. Before we proceed, we verify the following inequalities:

$$q'(\|z_s\|_2)\left\langle\frac{z_s}{\|z_s\|_2}, w_s\right\rangle = q'(\|z_s\|_2)\left\langle\frac{z_s}{\|z_s\|_2}, z_s+w_s-z_s\right\rangle \overset{(i)}{\le} q'(\|z_s\|_2)(\|z_s+w_s\|_2-\|z_s\|_2),$$

where (i) is by Cauchy-Schwarz, and:

$$\begin{aligned}
&q'(\|z_s+w_s\|_2)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -w_s-\frac{c_\kappa}{L}\nabla_s\right\rangle \\
&\quad= q'(\|z_s+w_s\|_2)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -z_s-w_s+z_s-\frac{c_\kappa}{L}\nabla_s\right\rangle \\
&\quad\overset{(i)}{\leq} q'(\|z_s+w_s\|_2)\left(-\|z_s+w_s\|_2+\|z_s\|_2+\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -\frac{c_\kappa}{L}\nabla_s\right\rangle\right) \\
&\quad\overset{(ii)}{\leq} q'(\|z_s+w_s\|_2)(-\|z_s+w_s\|_2+(1+c_\kappa)\|z_s\|_2),
\end{aligned}$$

where (i) is again by Cauchy-Schwarz and (ii) is by Cauchy-Schwarz combined with Assumption (A1). Finally:

$$q'(z_s+w_s)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -\frac{c_\kappa}{L}\Delta_s\right\rangle \leq q'(\|z_s+w_s\|_2)\frac{c_\kappa}{L}\|\Delta_s\|_2, \tag{C.19}$$

where the inequality above is by Cauchy-Schwarz along with the fact that $q'(r)\geq 0$ for all r from Lemma 30.

Bounding ♠: From Eqs. (C.18) and (C.17):

$$\begin{aligned}
\spadesuit &= (1+2c_\kappa)f'(r_s)q'(\|z_s\|_2)\left\langle \frac{z_s}{\|z_s\|_2}, w_s\right\rangle \\
&\quad + f'(r_s)q'(\|z_s+w_s\|_2)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, w_s\right\rangle \\
&\quad + f'(r_s)q'(\|z_s+w_s\|_2)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -2w_s-\frac{c_\kappa}{L}\nabla_s-\frac{c_\kappa}{L}\Delta_s\right\rangle \\
&= (1+2c_\kappa)f'(r_s)q'(\|z_s\|_2)\left\langle \frac{z_s}{\|z_s\|_2}, w_s\right\rangle \\
&\quad + f'(r_s)q'(\|z_s+w_s\|_2)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -w_s-\frac{c_\kappa}{L}\nabla_s-\frac{c_\kappa}{L}\Delta_s\right\rangle =: \spadesuit_1.
\end{aligned} \tag{C.20}$$

We again highlight the fact that $q'(\|z\|_2)\frac{z}{\|z\|_2}$ is defined for all z, particularly at $\|z\|_2=0$, as $q(r)=o(r^2)$ near zero (see Lemma 30).

Substituting the inequality in Eq. (C.19) into $\spadesuit_1$:

$$\begin{aligned}
\spadesuit_1 &= (1+2c_\kappa)f'(r_s)q'(\|z_s\|_2)\left\langle \frac{z_s}{\|z_s\|_2}, w_s\right\rangle \\
&\quad + f'(r_s)q'(\|z_s+w_s\|_2)\left\langle \frac{z_s+w_s}{\|z_s+w_s\|_2}, -w_s-\frac{c_\kappa}{L}\nabla_s-\frac{c_\kappa}{L}\Delta_s\right\rangle \\
&\leq (1+2c_\kappa)f'(r_s)q'(\|z_s\|_2)(\|z_s+w_s\|_2-\|z_s\|_2) \\
&\quad + f'(r_s)q'(\|z_s+w_s\|_2)(-\|z_s+w_s\|_2+(1+c_\kappa)\|z_s\|_2)+\frac{c_\kappa}{L}\|\Delta_s\|_2,
\end{aligned}$$

where the inequality uses Cauchy-Schwarz and (F2) of Lemma 55.

Now consider a few cases. We will use the expression for $q'(r)$ from Eq. (30) a number of times:

1. If $\|z_s\|_2 \in [\beta, \infty), \|z_s + w_s\|_2 \in [\beta, \infty)$, then $q'(\|z_s\|_2) = q'(\|z_s + w_s\|_2) = 1$, so that

$$\begin{aligned}
\spadesuit_1 &\leq f'(r_s)(\|z_s + w_s\|_2 - \|z_s\|_2 - \|z_s + w_s\|_2 + (1 + c_\kappa)\|z_s\|_2) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&= f'(r_s)(c_\kappa\|z_s\|_2) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\leq 2c_\kappa f'(r_s)r_s + \beta + \frac{c_\kappa}{L}\|\Delta_s\|_2,
\end{aligned}$$

   where we use the definition of r_t defined in Eq. (4.19) and Lemma 31.1.

2. If $\|z_s\|_2 \in [0, \beta), \|z_s + w_s\|_2 \in [\beta, \infty)$, then $q'(\|z_s\|_2) \in [0, 1]$ and $q'(\|z_s + w_s\|_2) = 1$, so that

$$\begin{aligned}
\spadesuit_1 &\overset{(i)}{\leq} f'(r_s)((1 + 2c_\kappa)q'(\|z_s\|_2)\|w_s\|_2 - \|z_s + w_s\|_2 + (1 + c_\kappa)\|z_s\|_2) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(ii)}{\leq} f'(r_s)(2c_\kappa\|w_s\|_2 + 3\|z_s\|_2) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(iii)}{\leq} f'(r_s)(2c_\kappa\|w_s\|_2 + 3\beta) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(iv)}{\leq} 2c_\kappa f'(r_s)r_s + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2,
\end{aligned}$$

   where (i) uses $\|z_s + w_s\|_2 - \|z_s\|_2 \leq \|w_s\|_2$, (ii) uses $\|w_s\|_2 - \|z_s + w_s\|_2 \leq \|z)s\|_2$, (iii) uses our upper bound in $\|z_s\|_2$ and (iv) uses the definition of r_t in Eq. (4.19) and Lemma 31.1.

3. If $\|z_s\|_2 \in [\beta, \infty), \|z_s + w_s\|_2 \in [0, \beta)$, then $q'(\|z_s\|_2) = 1$ and $q'(\|z_s + w_s\|_2) \in [0, 1]$, so that

$$\begin{aligned}
\spadesuit_1 &\overset{(i)}{\leq} f'(r_s)((1 + 2c_\kappa)(\|z_s + w_s\|_2 - \|z_s\|_2) - \|z_s + w_s\|_2 + (1 + c_\kappa)\|z_s\|_2) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&= f'(r_s)(2c_\kappa\|z_s + w_s\|_2 - c_\kappa\|z_s\|_2) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(ii)}{\leq} f'(r_s)\Big(3c_\kappa\|z_s + w_s\|_2 - \frac{c_\kappa}{2}r_s + 2\beta\Big) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\leq f'(r_s)\Big(-\frac{c_\kappa}{2}r_s\Big) + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2,
\end{aligned}$$

   where (i) uses our expression for $q'(\cdot)$, and (ii) uses the expression for r_t in Eq. (4.19), the fact that $c_\kappa \leq 1/1000$ and Lemma 31.1.

4. Finally, if $\|z_s\|_2 \in [0,\beta)$, $\|z_s+w_s\|_2 \in [0,\beta)$, then $q'(\|z_s\|_2) \in [0,1]$ and $q'(\|z_s+w_s\|_2) \in [0,1]$, so that

$$\spadesuit_1 \leq f'(r_s)(3\beta) + \frac{c_\kappa}{L}\|\Delta_s\|_2 \leq f'(r_s)\Big(-\frac{c_\kappa}{2}r_s\Big) + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2,$$

where we again use the expression for r_s in Eq. (4.19) and Lemma 31.1.

Combining the four cases above we find that,

$$\begin{aligned}\spadesuit \leq \spadesuit_1 \leq& \mathbb{1}\left\{\|z_s+w_s\|_2 \in [0,\beta)\right\}\cdot\Big(f'(r_s)\Big(-\frac{c_\kappa}{2}r_s\Big) + 4\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2\Big)\\ &+ \mathbb{1}\left\{\|z_s+w_s\|_2 \in [\beta,\infty)\right\}\cdot\Big(2c_\kappa f'(r_s)r_s + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2\Big),\end{aligned} \tag{C.21}$$

where we use Lemma 55.(F2), Lemma 31.1 and Eq. (4.19).

Bounding $\heartsuit$:

$$\begin{aligned}\heartsuit &\overset{(i)}{=} \left(\frac{8c_\kappa}{L}\gamma_s^T\nabla_w^2 f(r_s)\gamma_s + \frac{2c_\kappa}{L}\bar{\gamma}_s^T\nabla_w^2 f(r_s)\bar{\gamma}_s\right)\\
&\overset{(ii)}{=} \frac{8c_\kappa}{L}\cdot\gamma_s^T\left(f'(r_s)\cdot q'(\|z_s+w_s\|_2)\cdot\frac{1}{\|z_s+w_s\|_2}\left(I - \frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\right)\gamma_s\\
&\quad + \frac{8c_\kappa}{L}\cdot\gamma_s^T\left(f'(r_s)q''(\|z_s+w_s\|_2)\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\gamma_s\\
&\quad + \frac{8c_\kappa}{L}\cdot\gamma_s^T\left(f''(r_s)q'(\|z_s+w_s\|_2)^2\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\gamma_s\\
&\quad + \frac{2c_\kappa}{L}\cdot\bar{\gamma}_s^T\left(f'(r_s)\cdot q'(\|z_s+w_s\|_2)\cdot\frac{1}{\|z_s+w_s\|_2}\left(I - \frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\right)\bar{\gamma}_s\\
&\quad + \frac{2c_\kappa}{L}\cdot\bar{\gamma}_s^T\left(f'(r_s)q''(\|z_s+w_s\|_2)\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\bar{\gamma}_s\\
&\quad + \frac{2c_\kappa}{L}\cdot\bar{\gamma}_s^T\left(f''(r_s)q'(\|z_s+w_s\|_2)^2\frac{(z_s+w_s)(z_s+w_s)^T}{\|z_s+w_s\|_2^2}\right)\bar{\gamma}_s\\
&\overset{(iii)}{=} \frac{8c_\kappa}{L}\cdot\Big(\big(f''(r_s)q'(\|z_s+w_s\|_2)^2 + f'(r_s)q''(\|z_s+w_s\|_2)\big)\Big)\cdot\|\gamma_s\|_2^2\\
&\quad + \frac{2c_\kappa}{L}\cdot\big(f''(r_s)q'(\|z_s+w_s\|_2)^2 + f'(r_s)q''(\|z_s+w_s\|_2)\big)\cdot\|\bar{\gamma}_s\|_2^2,\end{aligned}$$

where (i) is by Eq. (C.17), (ii) is by Lemma C.17 and (iii) is because $\left\langle\gamma_s, \frac{z_s+w_s}{\|z_s+w_s\|_2}\right\rangle = \|\gamma_s\|_2$ and $\left\langle\bar{\gamma}_s, \frac{z_s+w_s}{\|z_s+w_s\|_2}\right\rangle = \|\bar{\gamma}_s\|_2$ (see Eq. (4.15)).

From Lemma 30.4, $q''(\|z_s + w_s\|_2) = 0$ for $\|z_s + w_s\|_2 \geq \beta/2$ and from Eq. (4.15), $\gamma_s = \bar{\gamma}_s = 0$ for $\|z_s + w_s\|_2 \leq \beta/2$. Thus the above simplifies to

$$\begin{aligned}
\heartsuit &\overset{(i)}{\leq} \frac{8c_\kappa}{L} \cdot \left(f''(r_s)q'(\|z_s + w_s\|_2)^2\right) \cdot \|\gamma_s\|_2^2 + \frac{2c_\kappa}{L} \cdot \left(f''(r_s)q'(\|z_s + w_s\|_2)^2\right) \cdot \|\bar{\gamma}_s\|_2^2 \\
&\overset{(ii)}{\leq} \frac{8c_\kappa}{L} \cdot \left(f''(r_s)q'(\|z_s + w_s\|_2)^2\right) \cdot \|\gamma_s\|_2^2 \\
&\leq \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta\right\} \cdot \frac{8c_\kappa}{L} \cdot f''(r_s), \qquad \text{(C.22)}
\end{aligned}$$

where (i) is by Lemma 55 (F5), which implies that $\frac{2c_\kappa}{L} \cdot (f''(r_s)q'(\|z_s + w_s\|_2)^2) \cdot \|\bar{\gamma}_s\|_2^2 \leq 0$. The inequality in (ii) is because $f''(r) \leq 0$ for all r (Lemma 55.(F5)), along with the facts that $\mathbb{1}\{\|z_s + w_s\|_2 \geq \beta\} \cdot q'(\|z_s + w_s\|_2) = \mathbb{1}\{\|z_s + w_s\|_2 \geq \beta\}$ (by Lemma 30.3), and $\mathbb{1}\{r \geq \beta\}q'(r)^2 = \mathbb{1}\{r \geq \beta\}$ (by Eq. (4.15)).

Combining our upper bounds on ♠ and ♡ from Eq. (C.21) and Eq. (C.22),

$$\begin{aligned}
\spadesuit + \heartsuit &\leq \mathbb{1}\left\{\|z_s + w_s\|_2 < \beta\right\} \cdot \left(f'(r_s)\left(-\frac{c_\kappa}{2}r_s\right) + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2\right) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta\right\} \cdot \left(2c_\kappa f'(r_s)r_s + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2\right) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta\right\} \cdot \frac{8c_\kappa}{L} \cdot f''(r_s) \\
&\overset{(i)}{=} \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s \leq \sqrt{12}R\right\} \cdot \left(\frac{8c_\kappa}{L}f''(r_s) + 2c_\kappa f'(r_s) \cdot r_s\right) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s > \sqrt{12}R\right\} \cdot (2c_\kappa f'(r_s)r_s) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 < \beta\right\} \cdot \left(f'(r_s)\left(-\frac{c_\kappa}{2}r_s\right)\right) \\
&\quad + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(ii)}{=} \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s \leq \sqrt{12}R\right\} \cdot \left(\frac{8c_\kappa}{L}\left(f''(r_s) + \frac{L}{4}f'(r_s) \cdot r_s\right)\right) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s > \sqrt{12}R\right\} \cdot (2c_\kappa f'(r_s)r_s) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 < \beta\right\} \cdot \left(f'(r_s)\left(-\frac{c_\kappa}{2}r_s\right)\right) \\
&\quad + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2,
\end{aligned}$$

where (i) and (ii) follow from algebraic manipulations. Continuing forward we find that,

$$\begin{aligned}
\spadesuit + \heartsuit &\overset{(i)}{\leq} \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s \leq \sqrt{12}R\right\} \cdot \left(-\frac{8c_\kappa}{L} \cdot \frac{e^{-6LR^2}}{48R^2} f(r_s)\right) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s > \sqrt{12}R\right\} \cdot (2c_\kappa f'(r_s) r_s) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 < \beta\right\} \cdot \left(-\frac{c_\kappa e^{-6LR^2}}{4} f(r_s)\right) \\
&\quad + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(ii)}{\leq} \mathbb{1}\left\{\|z_s + w_s\|_2 \geq \beta, r_s \leq \sqrt{12}R\right\} \cdot (-C_m f(r_s)) \\
&\quad + \mathbb{1}\left\{\|z_s + w_s\|_2 < \beta\right\} \cdot (-C_m f(r_s)) \\
&\quad + \mathbb{1}\left\{r_s > \sqrt{12}R\right\} \cdot 2r_s + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(iii)}{\leq} -C_m f(r_s) + \mathbb{1}\left\{r_s > \sqrt{12}R\right\} \cdot (C_m f(r_s) + 2r_s) + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2 \\
&\overset{(iv)}{\leq} -C_m f(r_s) + \mathbb{1}\left\{r_s > \sqrt{12}R\right\} \cdot (4r_s) + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2, \qquad \text{(C.23)}
\end{aligned}$$

where (i) is by Lemma 55 (F4) combined with the choice of α_f and $\mathcal{R}_f$, third line is by Lemma 55 (F2) and Lemma 55 (F3). (ii) follows immediately from the definition of C_m in (4.9). (iii) can be verified from algebra, and finally (iv) is from the fact that $C_m \leq 1$ and $f(r) \leq r$ for all r (Lemma 55 (F3)).

Thus, by combining the bounds on $\spadesuit$ and $\heartsuit$ in Eqs. (C.23) back into Eq. (C.18),

$$\begin{aligned}
d\mu_k f(r_s) \leq& -\mu_k C_m f(r_s) ds \\
&+ \mu_k\Big(\mathbb{1}\left\{r_s > \sqrt{12}R\right\} \cdot 4r_s + 5\beta + \frac{c_\kappa}{L}\|\Delta_s\|_2\Big) ds \\
&+ \mu_k\left\langle \nabla_w f(r_s), 4\sqrt{\frac{c_\kappa}{L}}\left(\gamma_s \gamma_s^T dB_s + \frac{1}{2}\bar{\gamma}_s \bar{\gamma}_s^T dA_s\right)\right\rangle \\
\leq& -\mu_k C_m f(r_s) ds \\
&+ \mu_k\Big(\mathbb{1}\left\{r_s > \sqrt{12}R\right\} \cdot 4r_s ds + 5\beta + c_\kappa\left\|x_s - x_{\lfloor \frac{s}{\delta} \rfloor \delta}\right\|_2\Big) ds \\
&+ \mu_k\left\langle \nabla_w f(r_s), 4\sqrt{\frac{c_\kappa}{L}}\left(\gamma_s \gamma_s^T dB_s + \frac{1}{2}\bar{\gamma}_s \bar{\gamma}_s^T dA_s\right)\right\rangle. \qquad \text{(C.24)}
\end{aligned}$$

By taking the time derivative of Eq. (4.24)-(4.26), we can verify that for $s \in [k\nu, (k+1)\nu)$,

$$\begin{aligned}
d\mu_k\xi_s =& -\mu_k \cdot C_m\xi_s ds + \mu_k \cdot c_\kappa \left\| x_s - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 ds, \\
d\sigma_s =& -\mu_k C_m \sigma_s ds + \mu_k \cdot \mathbb{1}\left\{ r_s \geq \sqrt{12}R \right\} \cdot 4 r_s ds, \\
d\phi_s =& -\mu_k C_m \phi_s ds + \mu_k \cdot \left\langle \nabla_w f(r_s), 4\sqrt{\frac{c_\kappa}{L}} \left(\gamma_s \gamma_s^T dB_s + \frac{1}{2}\bar{\gamma}_s \bar{\gamma}_s^T dA_s \right) \right\rangle.
\end{aligned}$$

By combining with Eq. (C.24) we get

$$d(\mu_k \cdot (f(r_s) - \xi_s) - \sigma_s - \phi_s) \leq -C_m(\mu_k \cdot (f(r_s) - \xi_s) - (\sigma_s + \phi_s)) + 5\beta ds.$$

An application of Grönwall's Lemma over the interval $s \in [k\nu, (k+1)\nu)$ gives us the claimed result:

$$\begin{aligned}
&\mu_k \cdot \left(f(r_{(k+1)\nu}) - \xi_{(k+1)\nu} \right) - \left(\sigma_{(k+1)\nu} + \phi_{(k+1)\nu} \right) \\
&\qquad \leq e^{-C_m \nu} (\mu_k \cdot (f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})) + 5\beta\nu.
\end{aligned}$$

□

C.4.3 Main results for synchronous coupling

Our main result in this section is Lemma 34, which shows that over a period of T_{sync}, $f(r_s)$ contracts by an amount $\exp(-C_m T_{sync})$ with probability one. Note that this is weaker than showing a contraction rate of $\exp(-C_m t)$ for all t, but is sufficient for our purposes.

Lemma 34 *Assume that $e^{72LR^2} \geq 2$. With probability one, for all k,*

$$\begin{aligned}
&\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot (f(r_{k\nu}) - \xi_{k\nu}) \\
&\qquad \leq \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp(-C_m T_{sync}) \cdot \left(f\left(r_{\tau_{k-1}}\right) - \xi_{\tau_{k-1}} \right) + 5\beta.
\end{aligned}$$

Proof
From our definition of c_κ in Eq. (4.6), r_t in Eq. (4.19), and from Lemma 31.1, it can be verified that

$$\begin{aligned}
r_{k\nu} \leq& 1.002(\|z_{k\nu}\|_2 + \|z_{k\nu} + w_{k\nu}\|_2) + 2\beta \\
\leq& \sqrt{2.002}\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} + 2\beta.
\end{aligned}$$

On the other hand, by $\|\cdot\|_1 \geq \|\cdot\|_2$ and by Lemma 31,

$$r_{\tau_{k-1}} \geq \sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}} + w_{\tau_{k-1}}\right\|_2^2} - 2\beta.$$

Combining the inequality in the display above with the statement of Lemma 37 gives:

$$\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot r_{k\nu} \leq \sqrt{\frac{47}{50}} \cdot \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot r_{\tau_{k-1}} + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt + 5\beta.$$

Combining the above with (F2), (F3) and (F6) of Lemma 55, and by using the definition of f in Eq. (4.23),

$$\begin{aligned}
&\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot f(r_{k\nu}) \\
&\quad \leq \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp\left(-\frac{1 - \sqrt{47/50}}{4} e^{-6LR^2}\right) f(r_{\tau_{k-1}}) \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt + 5\beta \\
&\quad \leq \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp\left(-C_m T_{sync}\right) f\left(r_{\tau_{k-1}}\right) \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt + 5\beta \\
&\quad \overset{(i)}{\leq} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp\left(-C_m T_{sync}\right) f\left(r_{\tau_{k-1}}\right) \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-C_m(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt + 5\beta, \qquad \text{(C.25)}
\end{aligned}$$

where the first line in (i) follows from the definition of T_{sync} and C_m in Eq. (4.8) and Eq. (4.9) along with the fact that $(1 - \sqrt{47/50})/4 \geq 1/200$. The second line in (i) is because $C_m \leq \frac{c_\kappa^2}{3}$ from Eq. (4.9).

By definition of ξ_t in Eq. (4.24),

$$\begin{aligned}
&\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\}\xi_{k\nu} \\
&\quad= \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \int_0^{k\nu} e^{-C_m(k\nu - t)} c_\kappa \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt \\
&\quad= \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot e^{-C_m(k\nu - \tau_{k-1})} \int_0^{\tau_{k-1}} e^{-C_m(\tau_{k-1} - t)} c_\kappa \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt \\
&\qquad+ \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \int_{\tau_{k-1}}^{k\nu} e^{-C_m(k\nu - t)} c_\kappa \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt \\
&\quad= \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp\left(-C_m(k\nu - \tau_{k-1})\right)\xi_{\tau_{k-1}} \\
&\qquad+ c_\kappa \int_{\tau_{k-1}}^{k\nu} \exp\left(-C_m(k\nu - t)\right) \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt \\
&\quad= \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp\left(-C_m T_{sync}\right)\xi_{\tau_{k-1}} \\
&\qquad+ c_\kappa \int_{\tau_{k-1}}^{k\nu} \exp\left(-C_m(k\nu - t)\right) \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt. \qquad \text{(C.26)}
\end{aligned}$$

By subtracting the left and the right hand sides of Eq. (C.26) and Eq. (C.25) thus gives us that,

$$\begin{aligned}
&\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \left(f(r_{k\nu}) - \xi_{k\nu}\right) \\
&\qquad\leq \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \exp\left(-C_m T_{sync}\right) \cdot \left(f\big(r_{\tau_{k-1}}\big) - \xi_{\tau_{k-1}}\right) + 5\beta.
\end{aligned}$$

□

We now state and prove several auxillary lemmas which are required for the proof of Lemma 34.

Lemma 35 *If* $\|z_s\|_2^2 + \|z_s + w_s\|_2^2 \geq 2.2R^2$, *then*

$$\langle z_s, w_s \rangle + \left\langle z_s + w_s, -w_s - \frac{c_\kappa}{L}\nabla_s \right\rangle \leq -\frac{c_\kappa^2}{3}\big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big).$$

Proof
We begin by expanding the differentials $d\|z_s\|_2^2 + d\|z_s + w_s\|_2^2$:

$$\begin{aligned}
d\|z_s\|_2^2 + d\|z_s + w_s\|_2^2 &= 2\langle z_s, w_s \rangle + 2\left\langle z_s + w_s, -w_s - \frac{c_\kappa}{L}\nabla_s \right\rangle \\
&= -2\|w_s\|_2^2 - 2\left\langle z_s, \frac{c_\kappa}{L}\nabla_s \right\rangle - 2\left\langle w_s, \frac{c_\kappa}{L}\nabla_s \right\rangle \\
&= -2\|w_s\|_2^2 - 2\left\langle z_s, \frac{c_\kappa}{L}\nabla_s \right\rangle + \|w_s\|_2^2 + \frac{c_\kappa^2}{L^2}\|\nabla_t\|_2^2 - \|w_t + \frac{c_\kappa}{L}\nabla_t\|_2^2 \\
&\leq -\|w_s\|_2^2 - 2\left\langle z_s, \frac{c_\kappa}{L}\nabla_s \right\rangle + \frac{c_\kappa^2}{L^2}\|\nabla_s\|_2^2 \\
&\leq -\|w_s\|_2^2 - 2\left\langle z_s, \frac{c_\kappa}{L}\nabla_s \right\rangle + c_\kappa^2\|z_s\|_2^2 =: \spadesuit. \qquad \text{(C.27)}
\end{aligned}$$

Now consider two cases.

Case 1: ($\|z_s\|_2 \leq R$) By Young's inequality,

$$\|z_s + w_s\|_2^2 \leq 11\|w_s\|_2^2 + 1.1\|z_s\|_2^2.$$

Furthermore, by our assumption that $\|z_s\|_2^2 + \|z_s + w_s\|_2^2 \geq 2.2R^2$,

$$\begin{aligned} 11\|w_s\|_2^2 &\geq \|z_s + w_s\|_2^2 - 1.1\|z_s\|_2^2 \\ &= \|z_s\|_2^2 + \|z_s + w_s\|_2^2 - 1.1\|z_s\|_2^2 - \|z_s\|_2^2 \\ &\geq 2.2R^2 - 2.1R^2 \\ &\geq 0.1R^2 \\ &\geq 0.1\|z_s\|_2^2, \\ \implies \|z_s\|_2^2 &\leq \frac{1000}{9}\|w_s\|_2^2. \end{aligned} \tag{C.28}$$

With this implication ♠ can now be upper bounded by

$$\begin{aligned} ♠ &= -\|w_s\|_2^2 - 2\left\langle z_s, \frac{c_\kappa}{L}\nabla_s \right\rangle + c_\kappa^2\|z_s\|_2^2 \\ &\overset{(i)}{\leq} -\|w_s\|_2^2 + 2c_\kappa\|z_s\|_2^2 + c_\kappa^2\|z_s\|_2^2 \\ &\overset{(ii)}{\leq} -\|w_s\|_2^2 + 3c_\kappa\|z_s\|_2^2 \\ &\overset{(iii)}{\leq} -\frac{2}{3}\|w_s\|_2^2 \\ &\overset{(iv)}{\leq} -\frac{c_\kappa^2}{3}\big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big), \end{aligned}$$

where (i) is by Assumption (A1) and Cauchy-Schwarz, and (ii) is because $c_\kappa := \frac{1}{1000\kappa} \leq \frac{1}{1000}$. The inequality (iii) is by the implication in Eq. (C.28), which gives $3c_\kappa\|z_s\|_2^2 \leq \frac{1000c_\kappa}{3}\|w_s\|_2^2 \leq$

$\frac{1}{3}\|w_s\|_2^2$. Finally, (iv) can be verified as follows:

$$\begin{aligned}
\|z_s\|_2^2 + \|z_s + w_s\|_2^2 &\overset{(i)}{\leq} 3\|z_s\|_2^2 + 2\|w_s\|_2^2 \\
&\overset{(ii)}{\leq} \frac{1000}{3}\|w_t\|_2^2 + 2\|w_t\|_2^2 \\
&\leq \frac{1006}{3}\|w_t\|_2^2. \\
\implies \qquad \big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big) &\leq \frac{1006}{3}\|w_s\|_2^2 \\
&\leq \frac{1}{2c_\kappa}\|w_s\|_2^2. \\
\implies \qquad \frac{2}{3}\|w_s\|_2^2 &\geq \frac{4c_\kappa}{3}\big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big) \\
&\overset{(iii)}{\geq} \frac{c_\kappa^2}{3}\big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big),
\end{aligned}$$

where (i) is by Young's inequality, (ii) is by Eq. (C.28), and (iii) is by $c_\kappa \leq \frac{1}{1000}$.

Case 2: $(\|z_s\|_2 \geq R)$ We have,

$$\begin{aligned}
\spadesuit &= -\|w_s\|_2^2 - 2\left\langle z_s, \frac{c_\kappa}{L}\nabla_s \right\rangle + c_\kappa^2\|z_s\|_2^2 \\
&\overset{(i)}{\leq} -\|w_s\|_2^2 - 2c_\kappa^2\|z_s\|_2^2 + c_\kappa^2\|z_s\|_2^2 \\
&\leq -\|w_s\|_2^2 - c_\kappa^2\|z_s\|_2^2 \\
&\leq -c_\kappa^2\big(\|w_s\|_2^2 + \|z_s\|_2^2\big) \\
&\overset{(ii)}{\leq} -\frac{c_\kappa^2}{3}\big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big),
\end{aligned}$$

where (i) is by Assumption (A3) and (ii) is because

$$\begin{aligned}
\|z_s\|_2^2 + \|z_s + w_s\|_2^2 &\leq 3\|z_s\|_2^2 + 2\|w_s\|_2^2 \\
&\leq 3\big(\|z_s\|_2^2 + \|w_s\|_2^2\big).
\end{aligned}$$

Hence, we have proved the result under both cases. □

Lemma 36 *With probability one,*

$$\begin{aligned}
&(1-\mu_k)\cdot\left(\sqrt{\left\|z_{(k+1)\nu}\right\|_2^2 + \left\|z_{(k+1)\nu} + w_{(k+1)\nu}\right\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad \leq (1-\mu_k)\cdot e^{-\frac{c_\kappa^2\nu}{3}}\left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad\quad + (1-\mu_k)\cdot c_\kappa \int_{k\nu}^{(k+1)\nu} e^{-\frac{c_\kappa^2}{3}((k+1)\nu - t)}\left\|x_t - x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2 dt.
\end{aligned}$$

Proof
When $\mu_k = 1$, the inequality holds trivially ($0 = 0$), so for the rest of this proof, we consider the case $\mu_k = 0$. To simplify notation, we leave out the multiplier $(1 - \mu_k)$ in all subsequent expressions.

We can verify from Eqs. (4.11)-(4.14) and Eq. (4.18) that when $\mu_k = 0$, for any $s \in [k\nu, (k+1)\nu)$,

$$\begin{aligned} dz_s =& w_s ds \\ d(z_s + w_s) =& \Big(-w_s + \frac{c_\kappa}{L}(\nabla_s + \Delta_s)\Big) ds. \end{aligned}$$

Thus, for any $s \in [k\nu, (k+1)\nu)$,

$$\begin{aligned} & d\left(\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+^2\right) \\ & \overset{(i)}{=} \frac{\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+}{\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2}} \left\langle \begin{bmatrix} z_s \\ z_s + w_s \end{bmatrix}, \begin{bmatrix} w_s \\ -w_s - \frac{c_\kappa}{L}(\nabla_s + \Delta_s) \end{bmatrix} \right\rangle ds \\ & \overset{(ii)}{\leq} -\frac{c_\kappa^2}{3} \frac{\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+}{\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2}} \cdot \big(\|z_s\|_2^2 + \|z_s + w_s\|_2^2\big) ds \\ & \quad + \frac{\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+}{\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2}} \cdot (\|z_s + w_s\|_2 \|\Delta_s\|_2) ds \\ & \leq -\frac{c_\kappa^2}{3}\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+ \cdot \sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} ds \\ & \quad + \frac{c_\kappa}{L}\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+ \cdot \|\Delta_s\|_2 ds \\ & \leq -\frac{c_\kappa^2}{3} \cdot \left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+^2 ds \\ & \quad + c_\kappa\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+ \cdot \left\|x_s - x_{\lfloor \frac{s}{\delta} \rfloor \delta}\right\|_2 ds, \end{aligned}$$

where (i) is by the expression for dz_s and dw_s established above, and (ii) is by Lemma 35 and Cauchy-Schwarz, the last two inequalities follow by algebraic manipulations.

Dividing throughout by $\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+$ gives us that

$$\begin{aligned} & d\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+ \\ & \qquad \leq \left(-\frac{c_\kappa^2}{3}\left(\sqrt{\|z_s\|_2^2 + \|z_s + w_s\|_2^2} - \sqrt{2.2}R\right)_+ + c_\kappa \left\|x_s - x_{\lfloor \frac{s}{\delta} \rfloor \delta}\right\|_2\right) dt. \end{aligned}$$

We can verify that the inequality implies that

$$d\left(\left(\sqrt{\|z_s\|_2^2+\|z_s+w_s\|_2^2}-\sqrt{2.2}R\right)_+ - c_\kappa\int_{k\nu}^{s} e^{-\frac{c_\kappa^2}{3}(s-t)}\left\|x_t-x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right)$$
$$\leq -\frac{c_\kappa^2}{3}\left(\left(\sqrt{\|z_s\|_2^2+\|z_s+w_s\|_2^2}-\sqrt{2.2}R\right)_+ - c_\kappa\int_{k\nu}^{s} e^{-\frac{c_\kappa^2}{3}(s-t)}\left\|x_t-x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right)dt.$$

Thus by Grönwall's Lemma,

$$\left(\sqrt{\|z_{(k+1)\nu}\|_2^2+\|z_{(k+1)\nu}+w_{(k+1)\nu}\|_2^2}-\sqrt{2.2}R\right)_+ - c_\kappa\int_{k\nu}^{(k+1)\nu} e^{-\frac{c_\kappa^2}{3}((k+1)\nu-t)}\left\|x_t-x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt$$
$$\leq e^{-\frac{c_\kappa^2}{3}((k+1)\nu-k\nu)}\left(\left(\sqrt{\|z_{k\nu}\|_2^2+\|z_{k\nu}+w_{k\nu}\|_2^2}-\sqrt{2.2}R\right)_+ - c_\kappa\int_{k\nu}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t-x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right)$$
$$= e^{-\frac{c_\kappa^2\nu}{3}}\left(\sqrt{\|z_{k\nu}\|_2^2+\|z_{k\nu}+w_{k\nu}\|_2^2}-\sqrt{2.2}R\right)_+.$$

This proves the statement of the Lemma. □

Lemma 37 *Assume that $e^{72LR^2}\geq 2$. With probability one, for all positive integers k,*

$$\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\cdot\sqrt{\|z_{k\nu}\|^2+\|z_{k\nu}+w_{k\nu}\|^2}$$
$$\leq\sqrt{\frac{23}{50}}\cdot\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\cdot\sqrt{\left\|z_{\tau_{k-1}}\right\|^2+\left\|z_{\tau_{k-1}}+w_{\tau_{k-1}}\right\|^2}$$
$$+\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\cdot c_\kappa\int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t-x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt+3\beta.$$

Proof
By our choice ν we know that T_{sync}/ν is an integer, thus we have,

$$k\nu=\tau_{k-1}+T_{sync}\Rightarrow(k-1)\nu<\tau_{k-1}+T_{sync}.$$

Thus,

$$\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\overset{(i)}{=}\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\cdot\mathbb{1}\{(k-1)\nu<\tau_{k-1}+T_{sync}\}$$
$$\overset{(ii)}{=}\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\cdot(1-\mu_{k-1})$$
$$\overset{(iii)}{=}\mathbb{1}\{k\nu=\tau_{k-1}+T_{sync}\}\cdot\prod_{i\in S_{k-1}}(1-\mu_i),\qquad\text{(C.29)}$$

where $S_{k-1}:=\left\{\frac{\tau_{k-1}}{\nu},\frac{\tau_{k-1}}{\nu}+1,...,k-1\right\}$ (as defined in Lemma 38). Above, (i) is because $k\nu=\tau_{k-1}+T_{sync}\Rightarrow(k-1)\nu<\tau_{k-1}+T_{sync}$, (ii) is because $(k-1)\nu<\tau_{k-1}+T_{sync}\Rightarrow\mu_{k-1}=0$ (see Eq. (4.18)) and (iii) is by Part 2 of Lemma 38.

We can now recursively apply Lemma 36 as follows: (to simplify notation, let $\alpha := \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\}$):

$$\begin{aligned}
&\alpha \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot \left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\quad\le \alpha \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot e^{-\frac{c_\kappa^2}{3}\nu}\left(\sqrt{\left\|z_{(k-1)\nu}\right\|_2^2 + \left\|z_{(k-1)\nu}+w_{(k-1)\nu}\right\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \alpha \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot c_\kappa \int_{(k-1)\nu}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2 dt \\
&\quad\le \alpha \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot e^{-\frac{c_\kappa^2}{3}T_{sync}}\left(\sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}}+w_{\tau_{k-1}}\right\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \alpha \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2 dt, \qquad \text{(C.30)}
\end{aligned}$$

where the last inequality uses the fact that $\nu \cdot (k - \tau_{k-1}) = T_{sync}$ in the definition of α.

Thus, we have,

$$\begin{aligned}
&\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\quad\overset{(i)}{=} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot \left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\quad\overset{(ii)}{\le} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot \left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \prod_{i\in S_{k-1}} (1-\mu_i) \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2 dt \\
&\quad\overset{(iii)}{=} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot e^{-\frac{c_\kappa^2}{3}T_{sync}}\left(\sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}}+w_{\tau_{k-1}}\right\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2 dt \\
&\quad\overset{(iv)}{\le} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \frac{1}{100}\left(\sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}}+w_{\tau_{k-1}}\right\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2 dt, \qquad \text{(C.31)}
\end{aligned}$$

where (i) is by Eq. (C.29), (ii) is by Eq. (C.30), (iii) is by Eq. (C.29) again, and (iv) is by the definition $T_{sync} = \frac{3}{c_\kappa^2}\log(100)$.

Let $j := \tau_{k-1}/\nu$. Then by the first part of Lemma 38, we know that $\tau_j = \tau_{k-1} = j\nu$. From the update rule for τ_k, Eq. (4.17), this must imply that

$$\sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}} + w_{\tau_{k-1}}\right\|_2^2} = \sqrt{\|z_{j\nu}\|_2^2 + \|z_{j\nu} + w_{j\nu}\|_2^2} \geq \sqrt{5}R. \tag{C.32}$$

Thus finally,

$$\begin{aligned}
&\mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2}\right) \\
&\overset{(i)}{\leq} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \sqrt{2.2}R \\
&\overset{(ii)}{\leq} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \frac{1}{100} \left(\sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}} + w_{\tau_{k-1}}\right\|_2^2} - \sqrt{2.2}R\right)_+ \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \sqrt{2.2}R \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\|x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2 dt \\
&\overset{(iii)}{\leq} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \frac{1}{100} \left(\sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}} + w_{\tau_{k-1}}\right\|_2^2} - \sqrt{2.2}R\right) \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot \sqrt{\frac{22}{50}} \sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}} + w_{\tau_{k-1}}\right\|_2^2} \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\|x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2 dt \\
&\overset{(iv)}{\leq} \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \sqrt{\frac{23}{50}} \sqrt{\left\|z_{\tau_{k-1}}\right\|_2^2 + \left\|z_{\tau_{k-1}} + w_{\tau_{k-1}}\right\|_2^2} \\
&\qquad + \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\} \cdot c_\kappa \int_{\tau_{k-1}}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\|x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2 dt,
\end{aligned}$$

where (i) is by an algebraic manipulation, (ii) is by Eq. (C.31), (iii) is by Eq. (C.32) and (iv) is because $1/100 + \sqrt{22/50} \leq \sqrt{23/50}$. □

Lemma 38 *Let k be a positive integer, then:*

1. *Let $j = \tau_k/\nu$. Then for all $i \in \{j, j+1, ..., k\}$, $\tau_i = \tau_k = j\nu$.*
2. *If $\mu_k = 0$, then $\mu_i = 0$ for all $i \in \{\tau_k/\nu ... k\}$, $\mu_i = 0$. Equivalently,*

$$\mathbb{1}\{\mu_k = 0\} = \prod_{i \in S_k} \mathbb{1}\{\mu_i = 0\},$$

where $S_k := \left\{\frac{\tau_k}{\nu}, ..., k\right\}$.

Proof
For the first claim: By definition of the update for τ_k, if $j = \tau_k/\nu$ for any k, then $j\nu = \tau_j = \tau_k$. Note that τ_i is nondecreasing with i, so that $j = \tau_k \leq k$, which implies that $\tau_j \leq \tau_{j+1} \leq ... \leq \tau_j$. Since $\tau_j = \tau_k$, the inequalities must hold with equality.

For the second claim: By the definition of μ_k; $\mu_k = 0$ implies that $k\nu < \tau_k + T_{sync}$. From the first claim, we know that for all $i \in \{\tau_k/\nu ... k\}$, $\tau_i = \tau_k$. Thus $i\nu \leq k\nu < \tau_k + T_{sync} = \tau_i + T_{sync}$. □

C.4.4 Discretization Error Bound

In this section, we bound the various *discretization errors*. First, in Section C.4.4.1, we establish a bound on $\mathbb{E}[\xi_t]$. Then in Lemma 42, we bound $\mathbb{E}[\sigma_t]$. Finally, in Lemma 47, we show that $\mathbb{E}[\phi_t] = 0$ as it is a martingale.

C.4.4.1 Bound on $\mathbb{E}[\xi_t]$

In this subsection, we establish a bound on $\mathbb{E}[\xi_t]$. This term represents the discretization error that arises because in the SDE in Eq. (4.12), the update to u_t uses the gradient $\nabla U\left(x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right)$ instead of $\nabla U(x_t)$. Our main result is Lemma 39, which in turn relies on the uniform bound for all $t \geq 0$ on $\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right]$ established in Corollary 40 (based on the moment bounds established in Appendix C.6).

Lemma 39 *For all $t \geq 0$,*

$$\mathbb{E}[\xi_t] \leq \delta \cdot \frac{2^9 c_\kappa \left(R + \sqrt{d/m}\right)}{C_m}.$$

Proof
By the bound in Corollary 40,

$$\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right] \leq \delta^8 2^{72} \left(R^2 + \frac{d}{m}\right)^4.$$

Further, by Jensen's inequality,

$$\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2\right] \leq \delta \cdot 2^9 \left(R + \sqrt{\frac{d}{m}}\right).$$

By integrating from up to time t,

$$\begin{aligned}\mathbb{E}\left[\xi_t\right] &= \int_0^t e^{-C_m(t-s)}\mathbb{E}\left[c_\kappa\left\|x_t - x_{\lfloor \frac{s}{\delta}\rfloor\delta}\right\|_2\right] ds \\ &\leq \int_0^t e^{-C_m(t-s)} c_\kappa \delta 2^9\left(R + \sqrt{\frac{d}{m}}\right) ds \\ &\leq \delta \cdot \frac{2^9 c_\kappa\left(R+\sqrt{d/m}\right)}{C_m}.\end{aligned}$$

□

Corollary 40 *For all $t \geq 0$,*

$$\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8\right] \leq \delta^8 2^{72}\left(R^2 + \frac{d}{m}\right)^4.$$

Proof
This follows directly by combining the results of Lemma 56 and Lemma 41. □

Lemma 41 *Suppose that the step size $\delta \leq \frac{1}{1000}$. Then for all $t \in [\lfloor \frac{t}{\delta}\rfloor\delta, (\lfloor \frac{t}{\delta}\rfloor + 1)\delta]$,*

$$\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8\right] \leq \delta^8\left(1.1\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8\right)\right] + 2^{12}\left(R^2 + \frac{d}{m}\right)^4\right).$$

Proof

$$\begin{aligned}\mathbb{E}\left[\left\|x_t - x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8\right] &= \mathbb{E}\left[\left\|\int_{\lfloor \frac{t}{\delta}\rfloor\delta}^t w_s ds\right\|_2^8\right] \\ &\leq \delta^7 \int_{\lfloor \frac{t}{\delta}\rfloor\delta}^t \mathbb{E}\left[\|w_s\|_2^8\right] ds \\ &= \delta^8\left(1.1\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8\right)\right] + 2^{12}\left(R^2 + \frac{d}{m}\right)^4\right),\end{aligned}$$

where for the last inequality, we use Lemma 58.

□

C.4.4.2 Bounds on $\mathbb{E}[\sigma_t]$ and $\mathbb{E}[\phi_t]$

In this subsection, we bound $\mathbb{E}[\sigma_t]$ (Lemma 42). This term represents the discretization error that arises because τ_k (and hence μ_k) is updated at discrete time intervals of ν. We highlight the fact that $\mathbb{E}[\sigma_t]$ is bounded by a term that depends on ν, which can be made arbitrarily small. The main ingredient of this proof is a bound on $\mathbb{E}\left[\mu_k \cdot \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\}\right]$ in Lemma 44.

Lemma 42 *For $\beta \leq 0.0001R$. There exists a $C_5 = poly(L, 1/m, d, R, \frac{1}{C_m})$ and $C_3 = 1/poly(L, 1/m, d, R)$, such that for all $\nu \leq C_3$, for all positive integers k, and for all $t \geq 0$,*

$$\mathbb{E}[\sigma_t] \leq C_5 \nu^2.$$

Proof
By the definition of σ_t in Eq. (4.25),

$$\begin{aligned}
\mathbb{E}[\sigma_t] &= \mathbb{E}\left[\int_0^t \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(t-s)} \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\} \cdot 4 r_s ds\right] \\
&= 4\int_0^t e^{-C_m(t-s)} \mathbb{E}\left[\mu_{\lfloor \frac{s}{\nu} \rfloor} \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\} r_s\right] ds \\
&\overset{(i)}{\leq} 4\int_0^t e^{-C_m(t-s)} \nu^2 \cdot C_4 \\
&\leq \frac{4\nu^2 C_4}{C_m} \\
&= \nu^2 \cdot C_5,
\end{aligned}$$

where (i) is by Corollary 45. □

Lemma 43 *For all $s \geq 0$,*

$$\mathbb{E}\left[r_s^2\right] \leq 2^{32}\left(R^2 + \frac{d}{m}\right).$$

Proof
Recall that,

$$\begin{aligned}
r_s^2 &= \left((1+2c_\kappa)\|z_s\|_2 + \|z_s + w_s\|_2\right)^2 \\
&\leq \left((2+2c_\kappa)\|z_s\|_2 + \|w_s\|_2\right)^2 \\
&\leq \left(2.1\|x_s\|_2 + 2.1\|y_s\|_2 + \|u_s\|_2 + \|v_s\|_2\right)^2 \\
&\overset{(i)}{\leq} 16\left(\|x_s\|_2^2 + \|u_s\|_2^2 + \|y_s\|_2^2 + \|v_s\|_2^2\right) \\
&\leq 2^{16}\left(2^{72}\left(R^2 + \frac{d}{m}\right)^4\right)^{1/4} \\
&= 2^{32}\left(R^2 + \frac{d}{m}\right),
\end{aligned}$$

where (i) is by Lemma 56 and Lemma 57. □

Lemma 44 *For every $\beta \le 0.0001R$, there exists a $C_2 = poly(L, 1/m, d, R)$, $C_3 = 1/poly(L, 1/m, d, R)$, such that for all $\nu \le C_3$, for all positive integers k, and for all $s \in [k\nu, (k+1)\nu]$,*

$$\mathbb{E}\left[\mu_k \cdot \mathbb{1}\left\{r_s \ge \sqrt{12}R\right\}\right] \le C_2\nu^4.$$

Proof
By definition of μ_k in Eq. (4.18), we know that $\mu_k = 1$ implies that $k\nu - \tau_k \ge T_{sync}$ which further implies that $\tau_k = \tau_{k-1}$ (otherwise τ_k must equal $k\nu$ by the definition of τ_t, in which case $k\nu - \tau_k = 0 < T_{sync}$). This then implies that $k\nu - \tau_{k-1} \ge T_{sync}$. It must thus be the case that $\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} < \sqrt{5}R$, because otherwise $\tau_k = k\nu$, which contradicts $\mu_k = 1$. Thus,

$$\mu_k \le \mathbb{1}\left\{\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} < \sqrt{5}R\right\}. \tag{C.33}$$

By a standard inequality between $\|\cdot\|_1$ and $\|\cdot\|_2$,

$$\begin{aligned}
\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu} + w_{k\nu}\|_2^2} &\ge \frac{1}{\sqrt{2}}(\|z_{k\nu}\|_2 + \|z_{k\nu} + w_{k\nu}\|_2) \\
&\overset{(i)}{\ge} \frac{1}{\sqrt{2}}(\ell(z_{k\nu}) + \ell(z_{k\nu} + w_{k\nu})) - \beta \\
&\overset{(ii)}{\ge} \frac{1}{1.002\sqrt{2}} r_{k\nu} - \beta,
\end{aligned}$$

where (i) is by Lemma 31.1, and (ii) is by definition of r_t in Eq. (4.19) and by definition of c_κ.

Combining with the inequality (C.33),

$$\begin{aligned}
\mu_k &\le \mathbb{1}\left\{\frac{1}{1.002\sqrt{2}} r_{k\nu} - \beta < \sqrt{5}R\right\} \\
&= \mathbb{1}\left\{r_{k\nu} < 1.002\sqrt{10}R + \beta\right\} \\
&\le \mathbb{1}\left\{r_{k\nu} < \sqrt{11}R\right\},
\end{aligned} \tag{C.34}$$

where the final inequality uses our assumption that $\beta \le 0.0001R$. Thus,

$$\begin{aligned}
\mu_k \cdot \mathbb{1}\left\{r_s \ge \sqrt{12}R\right\} &\le \mathbb{1}\left\{r_{k\nu} < \sqrt{11}R\right\} \cdot \mathbb{1}\left\{r_s \ge \sqrt{12}R\right\} \\
&\le \mathbb{1}\left\{|r_s - r_{k\nu}| \ge 0.14R\right\}.
\end{aligned}$$

Taking expectations,

$$\begin{aligned}\mathbb{E}\left[\mu_k \cdot \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\}\right] &\leq \mathbb{E}\left[\mathbb{1}\left\{|r_s - r_{k\nu}| \geq 0.14R\right\}\right] \\ &\overset{(i)}{\leq} \frac{\mathbb{E}\left[(r_s - r_{k\nu})^8\right]}{(0.14R)^8} \\ &\overset{(ii)}{\leq} \frac{2^{10}\mathbb{E}\left[\|z_s - z_{k\nu}\|_2^8 + \|w_s - w_{k\nu}\|_2^8\right] + 2^{10}\beta^4}{(0.14R)^8}, \end{aligned} \tag{C.35}$$

where (i) by Markov's inequality, (ii) can be verified by using Lemma 31.1 and some algebra.

Next, by the dynamics of z_t we have that

$$\begin{aligned}\|z_s - z_{k\nu}\|_2^8 &= \left\|\int_{k\nu}^s w_s dt\right\|_2^8 \\ &\leq (s - k\nu)^7 \int_{k\nu}^s \|w_s\|_2^8 dt \\ &\leq 2^3(s - k\nu)^7 \int_{k\nu}^s \|u_s\|_2^8 + \|v_s\|_2^8 dt. \end{aligned} \tag{C.36}$$

Further by the definition of the dynamics of w_t we get,

$$\begin{aligned}&\|w_s - w_{k\nu}\|_2^8 \\ &= \left\|\int_{k\nu}^s -2w_t - \frac{c_\kappa}{L}\nabla U(x_{\lfloor \frac{t}{\delta} \rfloor}) + \frac{c_\kappa}{L}\nabla U(y_t)dt + 4\sqrt{\frac{c_\kappa}{L}}\int_{k\nu}^s \gamma_t\gamma_t^T dB_t + 2\sqrt{\frac{c_\kappa}{L}}\int_{k\nu}^s \bar{\gamma}_t\bar{\gamma}_t^T dA_t\right\|_2^8 \\ &\overset{(i)}{\leq} 2^{20}(s - k\nu)^7\left(\int_{k\nu}^s \|w_t\|_2^8 + \frac{c_\kappa^8}{L^8}\|\nabla U(y_t)\|_2^8 + \frac{c_\kappa^8}{L^8}\left\|\nabla U(x_{\lfloor \frac{t}{\nu} \rfloor})\right\|_2^8 dt\right) \\ &\quad + 2^{12}\frac{c_\kappa^4}{L^4}\left\|\int_{k\nu}^s \gamma_t\gamma_t^T dB_t\right\|_2^8 + 2^{12}\frac{c_\kappa^4}{L^4}\left\|\int_{k\nu}^s \bar{\gamma}_t\bar{\gamma}_t^T dA_t\right\|_2^8 \\ &\overset{(ii)}{\leq} 2^{30}(s - k\nu)^7\left(\int_{k\nu}^s \|u_t\|_2^8 + \|v_t\|_2^8 + c_\kappa^8\|y_t\|_2^8 + c_\kappa^8\left\|x_{\lfloor \frac{t}{\nu} \rfloor}\right\|_2^8 dt\right) \\ &\quad + 2^{12}\frac{c_\kappa^4}{L^4}\left\|\int_{k\nu}^s \gamma_t\gamma_t^T dB_t\right\|_2^8 + 2^{12}\frac{c_\kappa^4}{L^4}\left\|\int_{k\nu}^s \bar{\gamma}_t\bar{\gamma}_t^T dA_t\right\|_2^8 \\ &\overset{(iii)}{\leq} 2^{30}(s - k\nu)^7\left(\int_{k\nu}^s \|u_t\|_2^8 + \|v_t\|_2^8 + \|y_t\|_2^8 + \left\|x_{\lfloor \frac{t}{\nu} \rfloor}\right\|_2^8 dt\right) \\ &\quad + 2^{12}\frac{1}{L^4}\left\|\int_{k\nu}^s \gamma_t\gamma_t^T dB_t\right\|_2^8 + 2^{12}\frac{1}{L^4}\left\|\int_{k\nu}^s \bar{\gamma}_t\bar{\gamma}_t^T dA_t\right\|_2^8, \end{aligned} \tag{C.37}$$

where (i) is by the triangle inequality and Young's inequality, (ii) uses Assumption (A1), and (iii) uses the fact that $c_\kappa \leq 1$.

Therefore, summing the two inequalities above and taking expectations,

$$\begin{aligned}
&\mathbb{E}\left[\|z_s - z_{k\nu}\|_2^8 + \|w_s - w_{k\nu}\|_2^8\right] \\
&\quad\leq \mathbb{E}\left[2^{30}(s-k\nu)^7\left(\int_{k\nu}^{s}\left\|x_{\lfloor \frac{t}{\delta}\rfloor\delta}\right\|_2^8 + \|u_t\|_2^8 + \|y_t\|_2^8 + \|v_t\|_2^8 dt\right)\right] \\
&\qquad + \mathbb{E}\left[2^{12}\frac{1}{L^4}\left\|\int_{k\nu}^{s}\gamma_t\gamma_t^T dB_t\right\|_2^8 + 2^{12}\frac{1}{L^4}\left\|\int_{k\nu}^{s}\bar{\gamma}_t\bar{\gamma}_t^T dA_t\right\|_2^8\right] \\
&\quad\leq 2^{32}(s-k\nu)^8\left(R^2 + \frac{d}{m}\right)^4 + 2^{52}\cdot(s-k\nu)^4\cdot\frac{1}{L^4},
\end{aligned}$$

where the last inequlaity is by combining Lemma 56, Lemma 57 and Lemma 46 and by noting that by their definition in Eq. (4.15), $\|\gamma_t\|_2 \leq 1$ and $\|\bar{\gamma}_t\|_2 \leq 1$ for all t, with probability one.

There exists $C_1 = poly(R, d, \frac{1}{m})$ and $C_3 = 1/poly(R, d, \frac{1}{m})$, such that for all $\nu < C_3$ and for all $s \in [k\nu, (k+1)\nu]$, the right-hand side of the inequality above is upper bounded by

$$\mathbb{E}\left[\|z_s - z_{k\nu}\|_2^8 + \|w_s - w_{k\nu}\|_2^8\right] \leq \nu^4 C_1.$$

Combining the above with inequality (C.35), we find that there exists $C_2 = poly(R, d, \frac{1}{m})$ and $C_3 = 1/poly(R, d, \frac{1}{m})$, such that for all $\nu < C_3$ and for all $s \in [k\nu, (k+1)\nu]$

$$\begin{aligned}
\mathbb{E}\left[\mu_k \cdot \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\}\right] \leq& \frac{\mathbb{E}\left[(r_s - r_{k\nu})^8\right]}{(0.14R)^8} \\
\leq& \frac{2^{10}\mathbb{E}\left[\|z_s - z_{k\nu}\|_2^8 + \|w_s - w_{k\nu}\|_2^8\right] + 2^{10}\beta^4}{(0.14R)^8} \\
\leq& \nu^4 C_2,
\end{aligned}$$

where β is absorbed into C_2 due to our assumption that $\beta \leq 0.0001R$.

□

Corollary 45 *For $\beta \leq 0.0001R$. There exists constants, $C_3 = 1/poly(L, 1/m, d, R)$ and $C_4 = poly(L, 1/m, d, R)$, such that for all $\nu \leq C_3$, for all positive integers k, and for all $s \in [k\nu, (k+1)\nu]$,*

$$\mathbb{E}\left[\mu_k \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\} r_s\right] \leq \sqrt{\mathbb{E}\left[\mu_k \mathbb{1}\left\{r_s \geq \sqrt{12}R\right\}\right]}\sqrt{\mathbb{E}\left[r_s^2\right]} \leq C_4\nu^2.$$

Proof
Proof follows by combining the results of Lemma 43 and Lemma 44. □

Lemma 46 *Let γ_t be a d-dimensional adapted process satisfying $\|\gamma_t\|_2 \leq 1$ for all $t > 0$ with probability one. Then*

$$\mathbb{E}\left[\left\|\int_0^t \gamma_s\gamma_s^T dB_s\right\|_2^8\right] \leq 2^{20}t^4.$$

Proof
Let us define $\beta_t := \int_0^t \gamma_s \gamma_s^T dB_s$. Define the function $l(\beta) := \|\beta\|_2^8$ for this proof. The derivates of this function are,

$$\begin{aligned}\nabla l(\beta) &= 8l(\beta)^{3/4}\beta \\ \nabla^2 l(\beta) &= 8l(\beta)^{3/4} I + 48 l(\beta)^{2/4}\beta\beta^T.\end{aligned}$$

By Itô's Lemma,

$$\begin{aligned}dl(\beta_t) &= \left\langle 8l(\beta_t)^{3/4}\beta_t, \beta_t\beta_t^T dB_t \right\rangle + 4l(\beta_t)^{3/4}\|\gamma_t\|_2^2 dt + 24 l(\beta_t)^{2/4}(\langle \beta_t, \gamma_t\rangle)^2 \|\gamma_t\|_2^2 dt \\ &\leq \left\langle 8l(\beta_t)^{3/4}\beta_t, \beta_t\beta_t^T dB_t \right\rangle + 4l(\beta_t)^{3/4} dt + 24 l(\beta_t)^{2/4}\|\beta_t\|_2^2 dt \\ &= \left\langle 4l(\beta_t)^{3/4}\beta_t, \beta_t\beta_t^T dB_t \right\rangle + 28 l(\beta_t)^{3/4} dt.\end{aligned}$$

Taking expectations,

$$\frac{d}{dt}\mathbb{E}\left[l(\beta_t)\right] \leq 28\mathbb{E}\left[l(\beta_t)^{3/4}\right] \leq 28\mathbb{E}\left[l(\beta_t)\right]^{3/4}.$$

Thus,

$$\begin{aligned}& \frac{d}{dt}\mathbb{E}\left[l(\beta_t)\right]^{1/4} \leq 28 \\ \Rightarrow \quad & \mathbb{E}\left[l(\beta_t)\right]^{1/4} \leq 28t \\ \Rightarrow \quad & \mathbb{E}\left[l(\beta_t)\right] \leq 2^{20}t^4,\end{aligned}$$

as claimed. □

Lemma 47 *For all $t \geq 0$, $\mathbb{E}\left[\phi_t\right] = 0$.*

Proof
By the definition of ϕ_t it is a martingale. Hence, $\mathbb{E}\left[\phi_t\right] = 0$. □

C.4.5 Putting it all together

In this section, we combine the results from Appendices C.4.2, C.4.3 and C.4.4 to prove Theorem 9. The heart of the proof is Lemma 50, which shows that $\mathcal{L}_t$ contracts with probability one at a rate of $-C_m$. This lemma essentially combines the results of Lemmas 51, 52 (proved in Appendix C.4.2) and Lemmas 53, 54 (proved in Appendix C.4.3).

Proof of Theorem 9
From Lemma 50 we have,

$$\mathcal{L}_{k\nu} \leq e^{-C_m k\nu}\mathcal{L}_0 + \frac{1}{1-\exp\left(-C_m\nu\right)} \cdot (3\nu + 5)\beta. \tag{C.38}$$

while from Lemma 49,

$$f(r_{k\nu}) \leq 200\mathcal{L}_{k\nu} + 400\xi_{k\nu} + \sigma_{k\nu} + \phi_{k\nu} + 400\beta.$$

Taking expectations,

$$\begin{aligned}
&\mathbb{E}\left[f(r_{k\nu})\right] \\
&\quad \overset{(i)}{\leq} 200\mathbb{E}\left[\mathcal{L}_{k\nu}\right] + 400\mathbb{E}\left[\xi_{k\nu}\right] + \mathbb{E}\left[\sigma_{k\nu}\right] + \mathbb{E}\left[\phi_{k\nu}\right] + 400\beta \\
&\quad \overset{(ii)}{\leq} 200e^{-C_m k\nu}\mathbb{E}\left[\mathcal{L}_0\right] + 400\mathbb{E}\left[\xi_{k\nu}\right] + \mathbb{E}\left[\sigma_{k\nu}\right] + \mathbb{E}\left[\phi_{k\nu}\right] + \frac{2000(\nu+1)}{1-\exp\left(-C_m\nu\right)}\beta \\
&\quad = 200e^{-C_m k\nu}\mathbb{E}\left[f(r_0)\right] + 400\mathbb{E}\left[\xi_{k\nu}\right] + \mathbb{E}\left[\sigma_{k\nu}\right] + \mathbb{E}\left[\phi_{k\nu}\right] + \frac{2000(\nu+1)}{1-\exp\left(-C_m\nu\right)}\beta \\
&\quad \leq 200e^{-C_m k\nu}\mathbb{E}\left[r_0\right] + 400\mathbb{E}\left[\xi_{k\nu}\right] + \mathbb{E}\left[\sigma_{k\nu}\right] + \mathbb{E}\left[\phi_{k\nu}\right] + \frac{3000(\nu+1)}{1-\exp\left(-C_m\nu\right)}\beta,
\end{aligned}$$

where (i) is by Eq. (C.38) and (ii) can be verified from the initialization in Eq. (4.10) and the definition of the Lyapunov function $\mathcal{L}_t$ in Eq. (4.27).

From Lemmas 39, 42 and 47,

$$400\mathbb{E}\left[\xi_{k\nu}\right] + \mathbb{E}\left[\sigma_{k\nu}\right] + \mathbb{E}\left[\phi_{k\nu}\right] \leq \delta \cdot \frac{2^{18}c_\kappa\left(R + \sqrt{d/m}\right)}{C_m} + C_5\nu^2,$$

where $C_5 = poly(L, 1/m, d, R, 1/C_m)$ as defined in Lemma 42.

From Lemma 57, our choice of $x_0 = u_0 = 0$ in Eq. (4.10) and our definition of r_t in Eq. (4.19),

$$\mathbb{E}\left[r_0\right] \leq 3\mathbb{E}\left[\left\|y_0\right\|_2 + \left\|v_0\right\|_2\right] \leq 2^{10}\left(R + \sqrt{\frac{d}{m}}\right) + 3\beta.$$

By plugging the bound on $\mathbb{E}\left[r_0\right]$ and $\mathbb{E}\left[\xi_{k\nu}\right]$ into the bound on $\mathbb{E}\left[f(r_{k\nu})\right]$ above gives us that

$$\mathbb{E}\left[f(r_{k\nu})\right] \leq e^{-C_m k\nu}2^{18}\left(R + \sqrt{\frac{d}{m}}\right) + \delta \cdot \frac{2^{18}c_\kappa\left(R + \sqrt{d/m}\right)}{C_m} + C_5\nu^2 + \frac{3000(\nu+1)}{1-\exp\left(-C_m\nu\right)}\beta.$$

This inequality along with (F3) of Lemma 55, and Lemma 31.1 also implies that,

$$\begin{aligned}
\mathbb{E}\left[\left\|z_{k\nu}\right\|_2\right] &\leq \mathbb{E}\left[r_{k\nu}\right] + \beta \\
&\leq 2e^{6LR^2} \cdot \mathbb{E}\left[f(r_{ku})\right] + \beta \\
&\leq e^{6LR^2} \cdot e^{-C_m k\nu}2^{19}\left(R + \sqrt{\frac{d}{m}}\right) + e^{6LR^2} \cdot \delta \cdot \frac{2^{19}c_\kappa\left(R + \sqrt{d/m}\right)}{C_m} \\
&\qquad + 2e^{6LR^2} \cdot \left(C_5\nu^2 + \frac{3000(\nu+1)}{1-\exp\left(-C_m\nu\right)}\beta + \beta\right).
\end{aligned}$$

We can take ν and β to be arbitrarily small without any additional computation cost, so let $\nu = e^{-10}\min\left\{\frac{1}{C_m}, \left(e^{-6LR^2}\cdot\frac{\delta c_\kappa(R+\sqrt{d/m})}{C_m}\right)^{1/2}\right\}$ and $\beta = 2^{-20}e^{-6LR^2}\cdot\delta c_\kappa(R+\sqrt{d/m})\cdot\frac{\nu}{1+\nu}$.

In particular, note that our assumption that $\nu \le e^{-10}/C_m$ allows us to simplify $1/(1-\exp(-C_m\nu)) \le C_m\nu/2$. We have thus shown that the sum of terms containing β and ν are less than the sum of terms which do not contain β and ν.

We can ensure that the second term $\left(e^{6LR^2}\cdot\delta\cdot\frac{2^{19}c_\kappa\left(R+\sqrt{d/m}\right)}{C_m}\right)$ is less than $\varepsilon/2$ by setting

$$\delta = \varepsilon 2^{-20}e^{-6LR^2}\frac{C_m}{R+\sqrt{d/m}}\frac{1}{c_\kappa}.$$

We can ensure that the first term $\left(e^{6LR^2}\cdot e^{-C_mk\nu}2^{19}\left(R+\sqrt{\frac{d}{m}}\right)\right)$ is less than $\varepsilon/2$ by setting

$$k\nu \ge \frac{\log\frac{1}{\varepsilon} + 6LR^2 + \log\left(2^{20}\left(R^2+\frac{d}{m}\right)\right)}{C_m}.$$

Recalling the definition of $C_m := \min\left\{\frac{e^{-6LR^2}}{6000\kappa LR^2}, \frac{e^{-6LR^2}}{21\cdot 10^7\cdot\log(100)\cdot\kappa^2}, \frac{1}{3\cdot 10^6\kappa^2}\right\}$ in Eq. (4.9), and $c_\kappa := 1/(1000\kappa)$, some algebra shows that it suffices to let

$$\delta = \frac{\varepsilon}{R+\sqrt{d/m}}\cdot e^{-12LR^2}\cdot 2^{-35}\min\left(\frac{1}{LR^2}, \frac{1}{\kappa}\right).$$

The number of steps of the algorithm is thus

$$\begin{aligned} n = \frac{k\nu}{\delta} &\ge 2^{60}\cdot\frac{R+\sqrt{d/m}}{\varepsilon}\cdot e^{18LR^2}\cdot\kappa\cdot\max\left\{LR^2,\kappa\right\}^2\cdot\left(\log\frac{1}{\varepsilon}+LR^2+\log\left(R^2+\frac{d}{m}\right)\right)\\ &= \widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon}e^{18LR^2}\right). \end{aligned}$$

This completes the proof. □

Lemma 48 *With probability one, for all positive integers k,*

$$(1-\mu_k)\cdot f(r_{k\nu}) \le (1-\mu_k)\cdot 2\left(f(r_{\tau_k}) + c_\kappa\int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t - x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right) + 6\beta.$$

Proof
First, by Eq. (4.19) and Lemma 31.1,

$$\begin{aligned} (1-\mu_k)\cdot r_{k\nu} =& (1-\mu_k)\cdot\left((1+2c_\kappa)\ell(z_{k\nu})_2 + \ell(z_{k\nu}+w_{k\nu})\right)\\ \le& (1-\mu_k)\cdot\left((1+2c_\kappa)\|z_{k\nu}\|_2 + \|z_{k\nu}+w_{k\nu}\|\right) + 3\beta. \end{aligned}$$

Note that by Lemma 38 we have,

$$1-\mu_k = \mathbb{1}\{\mu_k = 0\} = \prod_{i\in S_k} \mathbb{1}\{\mu_i = 0\} = \prod_{i\in S_k}(1-\mu_i), \tag{C.39}$$

where $S_k := \left\{\frac{\tau_k}{\nu}, ..., k\right\}$. Thus using this characterization of $1-\mu_k$ we get,

$$\begin{aligned}
&(1-\mu_k)\cdot((1+2c_\kappa)\|z_{k\nu}\|_2 + \|z_{k\nu}+w_{k\nu}\|_2) \\
&\quad\overset{(i)}{\leq} (1-\mu_k)\cdot 2\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} \\
&\quad\overset{(ii)}{\leq} (1-\mu_k)\cdot 2\left(\left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ + \sqrt{2.2}R\right),
\end{aligned}$$

where (i) is by defintion of c_κ in Eq. (4.6) and (ii) inequality is by algebra. Unpacking this further we get that:

$$\begin{aligned}
&(1-\mu_k)\cdot((1+2c_\kappa)\|z_{k\nu}\|_2 + \|z_{k\nu}+w_{k\nu}\|_2) \\
&\overset{(i)}{\leq} (1-\mu_k)\cdot 2\left(\left(\prod_{i\in S_k}(1-\mu_i)\right)\cdot\left(\sqrt{\|z_{k\nu}\|_2^2 + \|z_{k\nu}+w_{k\nu}\|_2^2} - \sqrt{2.2}R\right)_+ + \sqrt{2.2}R\right) \\
&\overset{(ii)}{\leq} (1-\mu_k)\cdot 2\left(\left(\prod_{i\in S_k}(1-\mu_i)\right)\cdot e^{-\frac{c_\kappa^2}{3}(k\nu-\tau_k)}\left(\sqrt{\|z_{\tau_k}\|_2^2 + \|z_{\tau_k}+w_{\tau_k}\|_2^2} - \sqrt{2.2}R\right)_+\right) \\
&\quad + (1-\mu_k)\cdot 2\left(\left(\prod_{i\in S_k}(1-\mu_i)\right)\cdot c_\kappa\int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t - x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right) \\
&\quad + (1-\mu_k)\cdot 2\left(\sqrt{2.2}R\right) \\
&\overset{(iii)}{\leq} (1-\mu_k)\cdot 2\left(\left(\sqrt{\|z_{\tau_k}\|_2^2 + \|z_{\tau_k}+w_{\tau_k}\|_2^2} - \sqrt{2.2}R\right)_+ + \sqrt{2.2}R\right) \\
&\quad + (1-\mu_k)\cdot 2\left(c_\kappa\int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t - x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right) \\
&\overset{(iv)}{=} (1-\mu_k)\cdot 2\left(\sqrt{\|z_{\tau_k}\|_2^2 + \|z_{\tau_k}+w_{\tau_k}\|_2^2}\right) \\
&\quad + (1-\mu_k)\cdot 2\left(c_\kappa\int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t - x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right) \\
&\overset{(v)}{\leq} (1-\mu_k)\cdot 2\left(r_{\tau_k} + \left(c_\kappa\int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu-t)}\left\|x_t - x_{\lfloor\frac{t}{\delta}\rfloor\delta}\right\|_2 dt\right)\right) + 3\beta,
\end{aligned}$$

where (i) is by Eq. (C.39), (ii) follows by Lemma 36, applied recursively for $i \in \left\{\frac{\tau_k}{\nu}...k\right\}$, while (iii) is again by Eq. (C.39). The equality in (iv) can be verified as follows: By Lemma 38

we know that $\tau_{\tau_k/\nu} = \tau_k$, which implies that $\sqrt{\|z_{\tau_k}\|_2^2 + \|z_{\tau_k} + w_{\tau_k}\|_2^2} \geq \sqrt{5}R$ based on the dynamics of τ_k in Eq. (4.17). Finally (v) is by definition of r_t in Eq. (4.19).

Our conclusion thus follows from the concavity of f and the fact that $f(0) = 0$, so that for all $a, b, c \in \mathbb{R}^+$, $f(4b) \leq 4f(b)$ and $a \leq b + c$ implies that $f(a) \leq f(b) + c$:

$$(1-\mu_k)\cdot f(r_{k\nu}) \leq (1-\mu_k)\cdot 2\left(f(r_{\tau_k}) + c_\kappa \int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt\right) + 6\beta.$$

□

Lemma 49 *For all positive integer k, with probability one,*

$$f(r_{k\nu}) \leq 200\mathcal{L}_{k\nu} + 400\xi_{k\nu} + \sigma_{k\nu} + \phi_{k\nu} + 400\beta.$$

Proof
From Lemma 48,

$$\begin{aligned}(1-\mu_k)\cdot f(r_{k\nu}) &\leq 2(1-\mu_k)\cdot f(r_{\tau_k}) + 2(1-\mu_k)\cdot c_\kappa \int_{\tau_k}^{k\nu} e^{-\frac{c_\kappa^2}{3}(k\nu - t)} \left\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \right\|_2 dt + 6\beta \\ &\leq 2(1-\mu_k)\cdot f(r_{\tau_k}) + 2(1-\mu_k)\xi_{k\nu} + 6\beta, \end{aligned} \tag{C.40}$$

where the last inequality is by Eq. (4.24).

We can also verify from the definition of μ_t in Eq. (4.18) that $\mu_k = 0 \Leftrightarrow k\nu \leq \tau_k + T_{sync}$. Thus,

$$\begin{aligned}(1-\mu_k)\cdot e^{-C_m(k\nu - \tau_k)} &\overset{(i)}{\geq} (1-\mu_k)\cdot e^{-C_m T_{sync}} \\ &\overset{(ii)}{\geq} (1-\mu_k)\cdot \exp\left(-\frac{c_\kappa^2}{3}\cdot T_{sync}\right) \\ &= (1-\mu_k)\cdot \frac{1}{100}, \end{aligned} \tag{C.41}$$

where (i) is by Eq. (4.9) and (ii) line is by Eq. (4.8).

Combining the above with the definition of $\xi_{k\nu}$ in Eq. (4.24) we get,

$$\begin{aligned}(1-\mu_k)\xi_{k\nu} &= (1-\mu_k)e^{-C_m(k\nu-\tau_k)}\xi_{\tau_k} + \int_{\tau_k}^{k\nu} e^{-C_m(k\nu - s)} c_\kappa \left\| x_s - x_{\lfloor \frac{s}{\delta} \rfloor \delta} \right\|_2 ds \\ &\geq (1-\mu_k)e^{-C_m(k\nu-\tau_k)}\xi_{\tau_k}. \end{aligned} \tag{C.42}$$

Thus,

$$\begin{aligned}
\mathcal{L}_{k\nu} &\overset{(i)}{=} \mu_k(f(r_{k\nu}) - \xi_{k\nu}) + (1-\mu_k)\cdot e^{-C_m(k\nu-\tau_k)}\cdot(f(r_{\tau_k}) - \xi_{\tau_k}) - (\sigma_{k\nu}+\phi_{k\nu}) \\
&\overset{(ii)}{\geq} \mu_k(f(r_{k\nu}) - \xi_{k\nu}) + (1-\mu_k)\cdot e^{-C_m(k\nu-\tau_k)}\cdot\left(\frac{1}{2}f(r_{k\nu}) - \xi_{k\nu} - \xi_{\tau_k} - 2\beta\right) - (\sigma_{k\nu}+\phi_{k\nu}) \\
&\overset{(iii)}{\geq} \mu_k(f(r_{k\nu}) - \xi_{k\nu}) + (1-\mu_k)\cdot\cdot\left(\frac{e^{-C_m(k\nu-\tau_k)}}{2}f(r_{k\nu}) - 2\xi_{k\nu} - 2\beta\right) - (\sigma_{k\nu}+\phi_{k\nu}) \\
&\overset{(iv)}{\geq} \mu_k(f(r_{k\nu}) - \xi_{k\nu}) + (1-\mu_k)\cdot\left(\frac{1}{200}f(r_{k\nu}) - 2\xi_{k\nu} - 2\beta\right) - (\sigma_{k\nu}+\phi_{k\nu}) \\
&\overset{(v)}{\geq} \frac{1}{200}(\mu_k\cdot f(r_{k\nu}) + (1-\mu_k)\cdot f(r_{k\nu})) - (2\xi_{k\nu}+\sigma_{k\nu}+\phi_{k\nu}) - 2\beta \\
&\overset{(vi)}{=} \frac{1}{200}(f(r_{k\nu})) - (2\xi_{k\nu}+\sigma_{k\nu}+\phi_{k\nu}) - 2\beta,
\end{aligned}$$

where (i) is by definition of $\mathcal{L}$ in Eq. (4.27). (ii) is by Eq. (C.40). (iii) is by Eq. (C.42) and the positivity of f, ξ, β. (iv) is by Eq. (C.41) and the fact that $f(r_t) \geq 0$ and $\xi_t \geq 0$ for all t. The inequalities (v) and (vi) are by algebraic manipulations.

Rearranging terms gives

$$f(r_{k\nu}) \leq 200\mathcal{L}_{k\nu} + 400\xi_{k\nu} + \sigma_{k\nu} + \phi_{k\nu} + 400\beta.$$

□

Lemma 50 *Assume that $e^{72LR^2} \geq 2$. With probability one, for all positive integers k,*

$$\mathcal{L}_{k\nu} \leq e^{-C_m\nu}\mathcal{L}_{(k-1)\nu} + (3\nu+5)\beta.$$

Applying this recursively,

$$\begin{aligned}
\mathcal{L}_{k\nu} \leq& e^{-C_mk\nu}\mathcal{L}_0 + (3\nu+5)\beta\cdot\sum_{i=0}^{k} e^{-iC_m\nu} \\
\leq& e^{-C_mk\nu}\mathcal{L}_0 + \frac{1}{1-\exp(-C_m\nu)}\cdot(3\nu+5)\beta.
\end{aligned}$$

Proof
We get the conclusion by summing the results of Lemmas 51, 52, 53 and 54. □

Below, we state the lemmas which are needed to prove Lemma 50.

Lemma 51 *Assume that $e^{72LR^2} \geq 2$. For all positive integers k, with probability 1,*

$$\mathbb{1}\{\mu_k = 1, \mu_{k-1} = 0\}\cdot\mathcal{L}_{k\nu} \leq \mathbb{1}\{\mu_k = 1, \mu_{k-1} = 0\}\cdot e^{-C_m\nu}\mathcal{L}_{(k-1)\nu} + 5\beta.$$

Proof
Given the definition of $\mathcal{L}_t$ in Eq. (4.27) we find that $\mathbb{1}\{\mu_k = 1\}\mathcal{L}_{k\nu} = \mathbb{1}\{\mu_k = 1\}f(r_{k\nu})$ and $\mathbb{1}\{\mu_{k-1} = 0\}\mathcal{L}_{(k-1)\nu} = \mathbb{1}\{\mu_{k-1} = 0\}\big(e^{-C_m((k-1)\nu-\tau_{k-1})}f(r_{\tau_{k-1}}) - \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\big)\big)$.

By the dynamics of μ_k, we can verify that

$$\begin{aligned}\mu_k = 1 \Leftrightarrow \quad & k\nu \geq \tau_k + T_{sync} \\ \Rightarrow \quad & k\nu \neq \tau_k \\ \Rightarrow \quad & \tau_k = \tau_{k-1} \\ \Rightarrow \quad & k\nu \geq \tau_{k-1} + T_{sync}.\end{aligned}$$

We can also verify that

$$\mu_{k-1} = 0 \Rightarrow (k-1)\nu < \tau_{k-1} + T_{sync}.$$

By our choice of ν, T_{sync}/ν is an integer (see comment following Eq. (4.8)), and the inequalities above imply that $k\nu = \tau_{k-1} + T_{sync}$. Thus,

$$\mathbb{1}\{\mu_k = 1, \mu_{k-1} = 0\} = \mathbb{1}\{\mu_k = 1, \mu_{k-1} = 0\} \cdot \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\}. \tag{C.43}$$

To reduce clutter, let us define $\alpha := \mathbb{1}\{\mu_k = 1, \mu_{k-1} = 0\}$ and $\alpha' := \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\}$. Hence we have,

$$\begin{aligned}\alpha \cdot \mathcal{L}_{k\nu} &\overset{(i)}{=} \alpha \cdot (f(r_{k\nu}) - \xi_{k\nu}) - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) \\ &\overset{(ii)}{=} \alpha \cdot \alpha'(f(r_{k\nu}) - \xi_{k\nu}) - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) \\ &\overset{(iii)}{\leq} \alpha \cdot \alpha' \cdot e^{-C_m T_{sync}}\big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\big) - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) + 5\beta \\ &\overset{(iv)}{=} \alpha \cdot \alpha' \cdot e^{-C_m(k\nu-\tau_{k-1})}\big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\big) - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) + 5\beta \\ &\overset{(v)}{=} \alpha \cdot \alpha' \cdot e^{-C_m\nu}e^{-C_m((k-1)\nu-\tau_{k-1})}\big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\big) - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) + 5\beta \\ &\overset{(vi)}{=} \alpha \cdot e^{-C_m\nu}e^{-C_m((k-1)\nu-\tau_{k-1})}\big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\big) - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) + 5\beta,\end{aligned} \tag{C.44}$$

where (i) is by definition of $\mathcal{L}_{k\nu}$, (ii) by Eq. (C.43), (iii) is by Lemma 34, (iv) is by the fact that $\alpha' = \mathbb{1}\{T_{sync} = k\nu - \tau_{k-1}\}$, (v) is by algebra and finally (vi) is again by Eq. (C.43).

By definition of σ_t in Eq. (4.25),

$$\begin{aligned}\alpha \cdot \sigma_{k\nu} &= \alpha \int_0^{k\nu} \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(k\nu-s)} \cdot 4r_s ds \\ &\overset{(i)}{=} \alpha \int_0^{(k-1)\nu} \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(k\nu-s)} \cdot 4r_s ds \\ &= \alpha e^{-C_m\nu}\sigma_{(k-1)\nu},\end{aligned} \tag{C.45}$$

where (i) is because $\alpha = 1$ implies that $\mu_{\lfloor \frac{s}{\nu} \rfloor} = \mu_{k-1} = 0$ for all $s \in [(k-1)\nu, k\nu)$.

Similarly, by the definition of ϕ_t in Eq. (4.26),

$$\begin{aligned}
&\alpha \cdot \phi_{k\nu} \\
&= \alpha \int_0^{k\nu} \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(k\nu - s)} f'(r_s) q'(\|z_s + w_s\|_2) \left\langle \frac{z_s + w_s}{\|z_s + w_s\|_2}, 4\sqrt{\frac{c_\kappa}{L}} \left(\gamma_s \gamma_s^T dB_s + \frac{1}{2} \bar{\gamma}_s \bar{\gamma}_s^T dA_s \right) \right\rangle \\
&\overset{(i)}{=} \alpha \int_0^{(k-1)\nu} \mu_{\lfloor \frac{s}{\nu} \rfloor} \cdot e^{-C_m(k\nu - s)} f'(r_s) q'(\|z_s + w_s\|_2) \left\langle \frac{z_s + w_s}{\|z_s + w_s\|_2}, 4\sqrt{\frac{c_\kappa}{L}} \left(\gamma_s \gamma_s^T dB_s + \frac{1}{2} \bar{\gamma}_s \bar{\gamma}_s^T dA_s \right) \right\rangle \\
&= \alpha e^{-C_m \nu} \phi_{(k-1)\nu}, \qquad \text{(C.46)}
\end{aligned}$$

where (i) is again because $\alpha = 1$ implies that $\mu_{\lfloor \frac{s}{\nu} \rfloor} = \mu_{k-1} = 0$ for all $s \in [(k-1)\nu, k\nu)$.

Combining these results,

$$\begin{aligned}
\alpha \mathcal{L}_{k\nu} &\overset{(i)}{\leq} \alpha \cdot e^{-C_m \nu} e^{-C_m((k-1)\nu - \tau_{k-1})} \big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}} \big) \\
&\quad - \alpha \cdot (\sigma_{k\nu} + \phi_{k\nu}) + 5\beta \\
&\overset{(ii)}{=} \alpha \cdot e^{-C_m \nu} e^{-C_m((k-1)\nu - \tau_{k-1})} \big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}} \big) \\
&\quad - \alpha e^{-C_m \nu} \cdot \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu} \big) + 5\beta \\
&\overset{(iii)}{=} \alpha \cdot e^{-C_m \nu} \cdot \mathcal{L}_{(k-1)\nu} + 5\beta,
\end{aligned}$$

where (i) is by Eq. (C.44) and (ii) is by Eq. (C.45) and Eq. (C.46). Inequality (iii) is by the definition of $\mathcal{L}_t$ in Eq. (4.27), and because $\mathbb{1}\{\mu_{k-1} = 0\}\mathcal{L}_{(k-1)\nu} = \mathbb{1}\{\mu_{k-1} = 0\}\big(e^{-C_m((k-1)\nu - \tau_{k-1})} f(r_{\tau_{k-1}}) - \big(\sigma_{(k-$ as noted in the beginning of the proof. □

Lemma 52 *For all positive integers k, with probability one,*

$$\mathbb{1}\{\mu_k = 0, \mu_{k-1} = 0\} \cdot \mathcal{L}_{k\nu} \leq \mathbb{1}\{\mu_k = 0, \mu_{k-1} = 0\} \cdot e^{-C_m \nu} \mathcal{L}_{(k-1)\nu} + 5\beta.$$

Proof
Define α_1, α_2 and α_3 to be indicators for the following events:

$$\alpha_1 := \mathbb{1}\{\mu_k = 0, \mu_{k-1} = 0\}, \alpha_2 := \mathbb{1}\{k\nu = \tau_k\} \text{ and } \alpha_3 := \mathbb{1}\{k\nu = \tau_{k-1} + T_{sync}\}.$$

By the definition of the Lyapunov function in Eq. (4.27) we find that

$$\begin{aligned}
&\alpha_1 \cdot \mathcal{L}_{k\nu} = \alpha_1 \cdot \big(e^{-C_m(k\nu - \tau_k)} (f(r_{\tau_k}) - \xi_{\tau_k}) - (\sigma_{k\nu} + \phi_{k\nu}) \big), \quad \text{and,} \\
&\alpha_1 \cdot \mathcal{L}_{(k-1)\nu} = \alpha_1 \cdot \big(e^{-C_m((k-1)\nu - \tau_{k-1})} \big(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}} \big) - \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu} \big) \big). \qquad \text{(C.47)}
\end{aligned}$$

We now consider two cases: when $k\nu = \tau_k$ and when $k\nu \neq \tau_k$ and prove the result in both of these cases.

Case 1: $k\nu = \tau_k$
From the definition of τ_t in Eq. (4.17), we know that $k\nu = \tau_k \Rightarrow k\nu - \tau_{k-1} \geq T_{sync}$. Additionally, $\mu_{k-1} = 0 \Rightarrow (k-1)\nu - \tau_{k-1} < T_{sync}$. By our choice of ν; T_{sync}/ν is an integer

(immediately below (4.8)). Thus it must be that $k\nu = \tau_{k-1} + T_{sync}$. Hence we have shown that

$$\alpha_1 \cdot \alpha_2 = \alpha_1 \cdot \alpha_2 \cdot \alpha_3. \tag{C.48}$$

Thus,

$$\begin{aligned}
&\alpha_1 \cdot \alpha_2 \cdot \mathcal{L}_{k\nu} \\
&\overset{(i)}{=} \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \left(e^{-C_m(k\nu-\tau_k)}(f(r_{\tau_k}) - \xi_{\tau_k}) - (\sigma_{\tau_k} + \phi_{\tau_k})\right) \\
&\overset{(ii)}{=} \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot ((f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})) \\
&\overset{(iii)}{\leq} \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \left(e^{-C_m(k\nu-T_{sync})} \cdot \left(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\right) - (\sigma_{k\nu} + \phi_{k\nu})\right) + 5\beta \\
&\overset{(iv)}{=} \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \left(e^{-C_m(k\nu-T_{sync})} \cdot \left(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\right) - e^{-C_m\nu}\left(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\right)\right) + 5\beta \\
&\overset{(v)}{=} \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \left(e^{-C_m\nu}\mathcal{L}\left(\theta_{(k-1)\nu}\right)\right) + 5\beta \\
&\overset{(vi)}{=} \alpha_1 \cdot \alpha_2 \cdot \left(e^{-C_m\nu}\mathcal{L}\left(\theta_{(k-1)\nu}\right)\right) + 5\beta,
\end{aligned}$$

where (i) is by Eq. (C.48), (ii) is because $\alpha_2 = 1$ implies $\tau_k = k\nu$, (iii) is by Lemma 34. Inequality (iv) is because $\alpha_1 = 1$ implies $\mu_{k-1} = 0$, we can thus verify from Eq. (4.25) and Eq. (4.26) that $\alpha_1 \cdot (\sigma_{k\nu} + \phi_{k\nu}) = \alpha_1 \cdot e^{-C_m\nu}\left(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\right)$ (the detailed proof is identical to proof of Eq. (C.45) and (C.46), and is not repeated here). (v) follows by our expression for $\mathcal{L}_{(k-1)\nu}$ in Eq. (C.47) and (vi) is again by Eq. (C.48).

Case 2: $k\nu \neq \tau_k$

In this case, by the definition of τ_t (in Eq. (4.17)) that $\tau_k = \tau_{k-1}$. Thus,

$$\begin{aligned}
&\alpha_1 \cdot (1-\alpha_2) \cdot \mathcal{L}_{k\nu} \\
&\overset{(i)}{=} \alpha_1 \cdot (1-\alpha_2) \cdot e^{-C_m(k\nu-\tau_k)}(f(r_{\tau_k}) - \xi_{\tau_k}) - \alpha_1 \cdot (1-\alpha_2) \cdot (\sigma_{k\nu} + \phi_{k\nu}) \\
&\overset{(ii)}{=} \alpha_1 \cdot (1-\alpha_2) \cdot e^{-C_m(k\nu-\tau_{k-1})}\left(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\right) - \alpha_1 \cdot (1-\alpha_2) \cdot (\sigma_{k\nu} + \phi_{k\nu}) \\
&\overset{(iii)}{=} \alpha_1 \cdot (1-\alpha_2) \cdot e^{-C_m(k\nu-\tau_{k-1})}\left(f(r_{\tau_{k-1}}) - \xi_{\tau_{k-1}}\right) - \alpha_1 \cdot (1-\alpha_2) \cdot e^{-C_m\nu}\left(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\right) \\
&\overset{(iv)}{=} \alpha_1 \cdot (1-\alpha_2) \cdot e^{-C_m\nu}\mathcal{L}_{(k-1)\nu},
\end{aligned}$$

where (i) is by the expression for $\mathcal{L}_{k\nu}$ in Eq. (C.47), (ii) is because $\tau_k = \tau_{k-1}$. Inequality (iii) is because $\alpha_1 \cdot (\sigma_{k\nu} + \phi_{k\nu}) = \alpha_1 \cdot e^{-C_m\nu}(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu})$. The proof of this fact is identical to proof of inequalities Eqs. (C.45) and (C.46), and is not repeated here. Finally (iv) is by pulling out a factor of $e^{-C_m\nu}$, and then using the equality in Eq. (C.47).

Therefore, summing the two cases, we get our conclusion that

$$\mathbb{1}\{\mu_k = 0, \mu_{k-1} = 0\} \cdot \mathcal{L}_{k\nu} \leq \mathbb{1}\{\mu_k = 0, \mu_{k-1} = 0\} \cdot e^{-C_m\nu}\mathcal{L}_{(k-1)\nu} + 5\beta.$$

□

Lemma 53 *For all positive integers k, with probability 1,*

$$\mathbb{1}\{\mu_k = 1, \mu_{k-1} = 1\} \cdot \mathcal{L}_{k\nu} \leq \mathbb{1}\{\mu_k = 1, \mu_{k-1} = 1\} \cdot e^{-C_m \nu} \mathcal{L}_{(k-1)\nu} + 5\beta\nu.$$

Proof
Let α denote the indicator of the following event, $\alpha := \mathbb{1}\{\mu_k = 1, \mu_{k-1} = 1\}$. By the definition of our Lyapunov function (see Eq. (4.27)) that

$$\begin{aligned} \alpha \cdot \mathcal{L}_{k\nu} &= \alpha \cdot ((f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})), \quad \text{and,} \\ \alpha \cdot \mathcal{L}_{(k-1)\nu} &= \alpha \cdot \big(\big(f(r_{(k-1)\nu}) - \xi_{(k-1)\nu}\big) - \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\big)\big). \end{aligned} \tag{C.49}$$

Thus we have,

$$\begin{aligned} \alpha \cdot \mathcal{L}_{k\nu} &\overset{(i)}{=} \alpha \cdot ((f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})) \\ &\overset{(ii)}{=} \alpha \cdot (\mu_k (f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})) \\ &\overset{(iii)}{\leq} \alpha \cdot \big(e^{-C_m \nu} \mu_k \cdot \big(f(r_{(k-1)\nu}) - \xi_{(k-1)\nu}\big) - \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\big)\big) + 5\beta\nu \\ &\overset{(iv)}{=} \alpha \cdot \big(e^{-C_m \nu} \cdot \big(f(r_{(k-1)\nu}) - \xi_{(k-1)\nu}\big) - \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\big)\big) + 5\beta\nu \\ &\overset{(v)}{=} \alpha \cdot e^{-C_m \nu} \mathcal{L}_{(k-1)\nu}, \end{aligned}$$

where (i) is by Eq. (C.49), (ii) is because $\alpha = \alpha \cdot \mu_k$, (iii) is by Lemma 33, (iv) is again because $\alpha = \alpha \cdot \mu_k$ and (v) is again by Eq. (C.49). □

Lemma 54 *For all positive integers k, with probability 1,*

$$\mathbb{1}\{\mu_k = 0, \mu_{k-1} = 1\} \cdot \mathcal{L}_{k\nu} \leq \mathbb{1}\{\mu_k = 0, \mu_{k-1} = 1\} \cdot e^{-C_m \nu} \mathcal{L}_{(k-1)\nu} + 5\beta\nu.$$

Proof
Let $\alpha := \mathbb{1}\{\mu_k = 0, \mu_{k-1} = 1\}$. We can verify using the definition of the Lyapunov function in Eq. (4.27) that:

$$\begin{aligned} \alpha \cdot \mathcal{L}_{k\nu} &= \alpha \cdot \big(e^{-C_m(k\nu - \tau_k)}(f(r_{\tau_k}) - \xi_{\tau_k}) - (\sigma_{k\nu} + \phi_{k\nu})\big) \quad \text{and,} \\ \alpha \cdot \mathcal{L}_{(k-1)\nu} &= \alpha \cdot \big(\big(f(r_{(k-1)\nu}) - \xi_{(k-1)\nu}\big) - \big(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\big)\big). \end{aligned} \tag{C.50}$$

Additionally, we can verify from Eq. (4.18) that $\mu_k = 0$ implies that $k\nu - T_{sync} < \tau_k$ and that $\mu_{k-1} = 1$ implies thta $(k-1)\nu - T_{sync} \geq \tau_{k-1}$. Putting this together, we get

$$\tau_k > k\nu - T_{sync} > (k-1)\nu - T_{sync} \geq \tau_{k-1}.$$

Thus $\tau_k > \tau_{k-1}$. From the definition of μ_t (in Eq. (4.18)), we see that τ_k is either equal to τ_{k-1} or is equal to $k\nu$, so that it must be that

$$\tau_k = k\nu,$$

when $\alpha = 1$. In particular, this implies that

$$\begin{aligned}
\alpha \cdot \mathcal{L}_{k\nu} &\overset{(i)}{=} \alpha \cdot \left(e^{-C_m(k\nu-\tau_k)}(f(r_{\tau_k}) - \xi_{\tau_k}) - (\sigma_{k\nu} + \phi_{k\nu})\right) \\
&\overset{(ii)}{=} \alpha \cdot (\mu_k \cdot (f(r_{k\nu}) - \xi_{k\nu}) - (\sigma_{k\nu} + \phi_{k\nu})) \\
&\overset{(iii)}{\leq} \alpha \cdot \left(e^{-C_m\nu}\mu_k \cdot \left(f(r_{(k-1)\nu}) - \xi_{(k-1)\nu}\right) - \left(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\right)\right) + 5\beta\nu \\
&\overset{(iv)}{\leq} \alpha \cdot \left(e^{-C_m\nu} \cdot \left(f(r_{(k-1)\nu}) - \xi_{(k-1)\nu}\right) - \left(\sigma_{(k-1)\nu} + \phi_{(k-1)\nu}\right)\right) + 5\beta\nu \\
&\overset{(v)}{=} \alpha \cdot e^{-C_m\nu}\mathcal{L}_{(k-1)\nu} + 5\beta\nu,
\end{aligned}$$

where (i) is by Eq. (C.50), (ii) is by $\alpha \cdot \mu_k = \alpha$ and because $\alpha = \alpha \cdot \mathbb{1}\{\tau_k = k\nu\}$, (iii) is by Lemma 33, (iv) is again by $\alpha \cdot \mu_k = \alpha$ and finally (v) is by Eq. (C.50). □

C.5 Properties of f

Lemma 55 *Assume that $e^{2\alpha_f \mathcal{R}_f^2} \geq 2\frac{1}{2}$. The function f defined in Eq. (4.22) has the following properties.*

(F1) $f(0) = 0$, $f'(0) = 1$.

(F2) For all $r \geq 0$, $\frac{1}{2}e^{-2\alpha_f \mathcal{R}_f^2} \leq \frac{1}{2}\psi(r) \leq f'(r) \leq 1$.

(F3) For all $r \geq 0$, $\frac{1}{2}e^{-2\alpha_f \mathcal{R}_f^2} r \leq \frac{1}{2}\Psi(r) \leq f(r) \leq \Psi(r) \leq r$.

(F4) For all $0 < r \leq \mathcal{R}_f$, $f''(r) + \alpha_f r f'(r) \leq -\frac{e^{-2\alpha_f \mathcal{R}_f^2}}{4\mathcal{R}_f^2} f(r)$

(F5) For all $r > 0$, f'' *is defined,* $f''(r) \leq 0$, *and* $f''(r) = 0$ *when* $r > 2\mathcal{R}_f$.

(F6) If $2\alpha_f \mathcal{R}_f^2 \geq \ln 2$, *for any* $0.5 < s < 1$, $f(sr) \leq \exp\left(-\frac{1-s}{4}e^{-2\alpha_f \mathcal{R}_f^2}\right) f(r)$.

(F7) For $r > 0$, $|f''(r)| \leq 4\alpha_f \mathcal{R}_f + \frac{4}{\mathcal{R}_f}$

Proof
We refer to definitions of the functions ψ, Ψ, g in Eq. (D.19) and the definition of f in Eq. (4.22).

(F1) $f(0) = 0$ and $f'(0) = 1$ by the definition of f and ψ.

(F2),(F3) are verified from the definitions, noting that $\frac{1}{2} \leq g(r) \leq 1$ and $e^{-2\alpha_f \mathcal{R}_f^2} \leq \psi(2\mathcal{R}_f) \leq \psi(r) \leq \psi(0)$.

(F4) To prove this property first we observe that $f'(r) = \psi(r)g(r)$ so

$$f''(r) = \psi'(r)g(r) + \psi(r)g'(r).$$

By the definition of ψ, $\psi'(r) = -2\alpha_f r\psi(r)$ if $r < \mathcal{R}_f$, thus

$$\begin{aligned} f''(r) + 2\alpha_f r f'(r) &= -2\alpha_f r\psi(r)g(r) + \psi(r)g'(r) + 2\alpha_f r f'(r) \\ &= \psi(r)g'(r) \\ &= -\frac{1}{2}\frac{h(r)\Psi(r)}{\int_0^\infty h(s)\frac{\Psi(s)}{\psi(s)}ds} \\ &\overset{(i)}{\leq} -\frac{1}{2}\frac{f(r)}{\int_0^\infty h(s)\frac{\Psi(s)}{\psi(s)}ds} \\ &\overset{(ii)}{\leq} -\frac{e^{-2\alpha_f\mathcal{R}_f^2}}{4\mathcal{R}_f^2}f(r), \end{aligned}$$

where (i) is because $f(r) \leq \Psi(r)$ and $h(r) = 1$ for $r \leq \mathcal{R}_f$.

(ii) is because $f(r) \geq 0$ and

$$\int_0^\infty h(s)\frac{\Psi(s)}{\psi(s)}ds = \int_0^{2\mathcal{R}_f} h(s)\frac{\Psi(s)}{\psi(s)}ds \leq \int_0^{2\mathcal{R}_f} \frac{2s}{e^{-2\alpha_f\mathcal{R}_f^2}}ds \leq 4\mathcal{R}_f^2 e^{2\alpha_f\mathcal{R}_f^2}.$$

The first inequality above is by (F2), (F3) and the definition of $h(s)$.

(F5) $f''(r) \leq 0$ follows from its expression $f''(r) = \psi'(r)g(r) + \psi(r)g'(r)$, and the fact that $\psi(r) \geq 0$ from (F2), $g(r) \geq 1/2$, $g'(r) \leq 0$ and $\psi'(r) \leq 0$ for all r. For $r > 2\mathcal{R}_f$, $\psi'(r) = g'(r) = 0$, so in that case $f''(r) = \psi'(r)g(r) + \psi(r)g'(r) = 0$.

(F6) For any $0 < c < 1$,

$$f((1+c)r) = f(r) + \int_r^{(1+c)r} f'(s)ds \geq f(r) + cr \cdot \frac{1}{2}e^{-2\alpha_f\mathcal{R}_f^2} \geq \left(1 + \frac{c}{2}e^{-2\alpha_f\mathcal{R}_f^2}\right) f(r),$$

where the first inequality follows from (F2), and the second inequality follows from (F3). Under the assumption that $e^{-2\alpha_f\mathcal{R}_f^2} \leq \frac{1}{2}$, and using the inequality $1 + x \geq e^{x/2}$ for all $x \in [0, 1/2]$, we get $1 + (c/2)e^{-2\alpha_f\mathcal{R}_f^2} \geq e^{(c/4)e^{-2\alpha_f\mathcal{R}_f^2}}$.

Thus, for any $s \in (1/2, 1)$, let $r' := sr$, so that $r = \frac{1}{s}r' = \big(1 + \big(\frac{1}{s} - 1\big)\big)r'$ Applying the above with $c = \frac{1}{s} - 1$, we get

$$\begin{aligned} f(sr) = f(r') &\leq \frac{1}{1 + \frac{c}{2}\exp\left(-2\alpha_f \mathcal{R}_f^2\right)} f((1+c)r') \\ &= \frac{1}{1 + \frac{c}{2}\exp\left(-2\alpha_f \mathcal{R}_f^2\right)} f(r) \\ &\leq \exp\left(-\frac{c}{4} e^{-2\alpha_f \mathcal{R}_f^2}\right) f(r) \\ &= \exp\left(-\frac{1/s - 1}{4} e^{-2\alpha_f \mathcal{R}_f^2}\right) f(r) \leq \exp\left(-\frac{1-s}{4} e^{-2\alpha_f \mathcal{R}_f^2}\right) f(r). \end{aligned}$$

where we use the fact that $-\frac{1-s}{s} \leq -(1-s)$.

(F7) Recall that

$$f''(r) = \psi'(r)g(r) + \psi(r)g'(r)$$

Thus

$$\begin{aligned} |f''(r)| &\leq |\psi'(r)g(r)| + |\psi(r)g'(r)| \\ &\leq 2\alpha_f r h(r) + |\psi(r)g'(r)| \end{aligned}$$

From our definition of $h(r)$, we know that $rh(r) \leq 2\mathcal{R}_f$. In addition, since $\psi(r)$ is monotonically decreasing, $\Psi(r) = \int_0^r \psi(s)ds \geq r\psi(r)$, so that

$$\frac{\Psi(r)}{\psi(r)} \geq r. \tag{C.51}$$

Thus $\Psi(r)/r \geq r$ for all r. On the other hand, using the fact that $\psi(s) \leq 1$,

$$\Psi(r) = \int_0^r \psi(s)ds \leq r. \tag{C.52}$$

Combining the previous expressions,

$$\begin{aligned} |\psi(r)\nu'(r)| &= \left| \frac{1}{2} \frac{h(r)\Psi(r)}{\int_0^{4\mathcal{R}_f} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \right| \\ &\leq \left| \frac{1}{2} \frac{2\mathcal{R}_f}{\int_0^{\mathcal{R}_f} \frac{\Psi(s)}{\psi(s)} ds} \right| \\ &\leq \left| \frac{1}{2} \frac{2\mathcal{R}_f}{\int_0^{\mathcal{R}_f} s ds} \right| \\ &\leq \frac{4}{\mathcal{R}_f}, \end{aligned}$$

where the first inequality is by the definition of $h(r) = 1$ for $r \leq \mathcal{R}_f$ and $h(r) = 0$ for $r \geq 2\mathcal{R}_f$, and the second-to-last inequality is by (C.51).

Put together, we get

$$|f''(r)| \leq 4\alpha_f \mathcal{R}_f + \frac{4}{\mathcal{R}_f}.$$

□

C.6 Bounding moments

To bound the discretization error it is necessary to bound the moments of the random variables x_t, u_t and y_t, v_t. The main results of this section are Lemma 56 (which bounds the moments of x_t and u_t) and Lemma 57 (which bounds the moments of y_t and v_t).

Lemma 56 *For $\delta \leq 2^{-10} c_\kappa$, and for all $t \geq 0$,*

$$\mathbb{E}\left[\|x_t\|_2^8 + \|x_t + u_t\|_2^8\right] \leq 2^{70}\left(R^2 + \frac{d}{m}\right)^4.$$

Lemma 57 *For all $t \geq 0$,*

$$\mathbb{E}\left[\|y_t\|_2^8 + \|y_t + v_t\|_2^8\right] \leq 2^{66}\left(R^2 + \frac{d}{m}\right)^4.$$

C.6.1 Proof of Lemma 56

Let us consider the Lyapunov function $l(x_t, u_t) := \left(\|x_t\|_2^2 + \|x_t + u_t\|_2^2 - 4R^2\right)_+^4$.

By calculating the derivaties of l we can verify that:

$$\begin{aligned}
\nabla_x l(x_t, u_t) &= 8l(x_t, u_t)^{3/4}(x_t) \\
\nabla_u l(x_t, u_t) &= 8l(x_t, u_t)^{3/4}(x_t + u_t) \\
\nabla_u^2 l(x_t, u_t) &= 8l(x_t, u_t)^{3/4} I + 24 l(x_t, u_t)^{2/4}(x_t + u_t)(x_t + u_t)^T.
\end{aligned}$$

The following are two useful inequalities which we will use in this proof:

$$\begin{aligned}
\|x\|_2^2 + \|x + u\|_2^2 &\leq l(x, u)^{1/4} + 4R^2 \\
\|x\|_2^2 + \|x + u\|_2^2 &\geq l(x, u)^{1/4}.
\end{aligned} \tag{C.53}$$

Recall from the dynamics defined in Eq. (4.11) and Eq. (4.12) that

$$\begin{aligned}
dx_t =& u_t dt \\
du_t =& -2u_t - \frac{c_\kappa}{L}\nabla U(x_{\lfloor \frac{t}{\delta} \rfloor \delta})dt + 2\sqrt{\frac{c_\kappa}{L}} dB_t.
\end{aligned}$$

Thus by studying the evolution of the Lyapunov function $l(x_t, u_t)$ we have:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\left[l(x_t,u_t)\right] &= \mathbb{E}\left[8l(x_t,u_t)^{3/4}\Big(\langle x_t,u_t\rangle + \Big\langle x_t+u_t, -u_t - \frac{c_\kappa}{L}\nabla U(x_{\lfloor \frac{t}{\delta}\rfloor \delta})\Big\rangle\Big)\right] \\
&\quad + \mathbb{E}\left[\frac{16c_\kappa}{L}\big(l(x_t,u_t)^{3/4}d + 3l(x_t,u_t)^{2/4}\|x_t+u_t\|_2^2\big)\right] \\
&= \mathbb{E}\left[\underbrace{8l(x_t,u_t)^{3/4}\Big(\langle x_t,u_t\rangle + \Big\langle x_t+u_t, -u_t - \frac{c_\kappa}{L}\nabla U(x_t)\Big\rangle\Big)}_{=:\spadesuit}\right] \\
&\qquad + \mathbb{E}\left[\underbrace{8\cdot\frac{c_\kappa}{L}\cdot l(x_t,u_t)^{3/4}\Big(\Big\langle x_t+u_t, \nabla U(x_t) - \nabla U(x_{\lfloor \frac{t}{\delta}\rfloor \delta})\Big\rangle\Big)}_{=:\heartsuit}\right] \\
&\qquad + \mathbb{E}\left[\underbrace{\frac{16c_\kappa}{L}\big(l(x_t,u_t)^{3/4}d + 3l(x_t,u_t)^{2/4}\|x_t+u_t\|_2^2\big)}_{=:\clubsuit}\right].
\end{aligned}$$

We will bound the three terms separately. We begin by bounding $\spadesuit$:

$$\begin{aligned}
\spadesuit &= 8l(x_t,u_t)^{3/4}\Big(\langle x_t,u_t\rangle + \Big\langle x_t+u_t, -u_t - \frac{c_\kappa}{L}\nabla U(x_t)\Big\rangle\Big) \\
&\overset{(i)}{\leq} -c_\kappa^2 l(x_t,u_t)^{3/4}\big(\|x_t\|_2^2 + \|x_t+u_t\|_2^2\big) \\
&\overset{(ii)}{\leq} -c_\kappa^2 l(x_t,u_t),
\end{aligned}$$

where (i) is by invoking Lemma 59, and (ii) is by Eq. (C.53). Next consider the term $\heartsuit$:

$$
\begin{aligned}
\heartsuit &= 8 \cdot \frac{c_\kappa}{L} \cdot l(x_t, u_t)^{3/4} \Big(\Big\langle x_t + u_t, \nabla U(x_t) - \nabla U(x_{\lfloor \frac{t}{\delta} \rfloor \delta}) \Big\rangle \Big) \\
&\overset{(i)}{\leq} 8 c_\kappa l(x_t, u_t)^{3/4} \|x_t + u_t\|_2 \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2 \\
&\overset{(ii)}{\leq} 8 c_\kappa l(x_t, u_t)^{3/4} \big(l(x_t, u_t)^{1/8} + 2R \big) \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2 \\
&\overset{(iii)}{\leq} 8 c_\kappa l(x_t, u_t)^{7/8} \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2 + 16 c_\kappa l(x_t, u_t)^{3/4} R \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2 \\
&\overset{(iv)}{\leq} \frac{c_\kappa^2}{8} l(x_t, u_t) + \frac{2^{32}}{c_\kappa^6} \left(\Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2^8 \right) + \frac{c_\kappa^2}{8} l(x_t, u_t) + \frac{2^{28}}{c_\kappa^2} \left(R^4 \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2^4 \right) \\
&\overset{(v)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + \frac{2^{32}}{c_\kappa^6} \left(\Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2^8 \right) + 2^{28} c_\kappa^2 R^8 + \frac{2^{28}}{c_\kappa^6} \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2^8 \\
&\overset{(vi)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + 2^{28} c_\kappa^2 R^8 + \frac{2^{33}}{c_\kappa^6} \Big\| x_t - x_{\lfloor \frac{t}{\delta} \rfloor \delta} \Big\|_2^8 \\
&\overset{(vii)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + 2^{28} c_\kappa^2 R^8 + \frac{2^{33}}{c_\kappa^6} \left\| \int_{\lfloor \frac{t}{\delta} \rfloor \delta}^t u_s ds \right\|_2^8 \\
&\overset{(viii)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + 2^{28} c_\kappa^2 R^8 + \frac{2^{33}}{c_\kappa^6} \left(\left(t - \left\lfloor \frac{t}{\delta} \right\rfloor \delta \right)^7 \int_{\lfloor \frac{t}{\delta} \rfloor \delta}^t \|u_s\|_2^8 ds \right) \\
&\overset{(ix)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + 2^{28} c_\kappa^2 R^8 + \frac{2^{33}}{c_\kappa^6} \left(\delta^7 \int_{\lfloor \frac{t}{\delta} \rfloor \delta}^t \|u_s\|_2^8 ds \right),
\end{aligned}
$$

where (i) is by Cauchy-Schwarz and Assumption (A1), (ii) is by Eq. (C.53), (iii) is again by Eq. (C.53), (iv) is by Young's inequality, (v) is again by Young's inequality, (vi) follows by an algebraic manipulation, (vii) is by the dynamics defined in Eq. (4.11), $(viii)$ is by Jensen's inequality and finally (ix) is because $t - \lfloor \frac{t}{\delta} \rfloor \delta \leq \delta$. Also:

$$
\begin{aligned}
\clubsuit &= \frac{16 c_\kappa}{L} \big(l(x_t, u_t)^{3/4} d + 3 l(x_t, u_t)^{2/4} \|x_t + u_t\|_2^2 \big) \\
&\overset{(i)}{\leq} \frac{16 c_\kappa}{L} \big(l(x_t, u_t)^{3/4} d + 3 l(x_t, u_t)^{3/4} + 12 l(x_t, u_t)^{2/4} R^2 \big) \\
&\overset{(ii)}{\leq} \frac{c_\kappa^2}{16} l(x_t, u_t) + \frac{2^{28}}{c_\kappa^2 L^4} d^4 + \frac{c_\kappa^2}{16} l(x_t, u_t) + \frac{2^{36}}{c_\kappa^2 L^4} + \frac{c_\kappa^2}{16} l(x_t, u_t) + \frac{2^{16} c_\kappa^2 R^4}{L^2} \\
&\overset{(iii)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + 2^{29} c_\kappa^2 \left(\frac{d^4}{m^4} + \frac{R^4}{L^2} \right) \\
&\overset{(iv)}{\leq} \frac{c_\kappa^2}{4} l(x_t, u_t) + 2^{30} c_\kappa^2 \left(\frac{d^4}{m^4} + R^8 \right),
\end{aligned}
$$

where (i) is by Eq. (C.53), (ii) is by Young's inequality, (iii) follows by definition of c_κ in Eq. (4.6) and (iv) is by Young's inequality, and because $m \leq L$.

Putting together the upper bounds on $\spadesuit, \heartsuit, \clubsuit$:

$$
\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[l(x_t,u_t)\right] = \spadesuit + \heartsuit + \clubsuit \\
&\leq \mathbb{E}\left[-c_\kappa^2 l(x_t,u_t) + 2^{28}c_\kappa^2 R^8 + \frac{2^{33}}{c_\kappa^6}\left(\delta^7 \int_{\lfloor \frac{t}{\delta}\rfloor \delta}^{t} \|u_s\|_2^8 ds\right) + \frac{c_\kappa^2}{4} l(x_t,u_t) + 2^{30}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right)\right] \\
&\leq -\frac{c_\kappa^2}{2}\mathbb{E}\left[l(x_t,u_t)\right] + 2^{33}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right) + \frac{2^{33}}{c_\kappa^6}\delta^7 \int_{\lfloor \frac{t}{\delta}\rfloor \delta}^{t} \mathbb{E}\left[\|u_s\|_2^8\right] ds \\
&\overset{(i)}{\leq} -\frac{c_\kappa^2}{2}\mathbb{E}\left[l(x_t,u_t)\right] + 2^{33}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right) + \frac{2^{33}}{c_\kappa^6}\delta^8\left(1.1\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta}\rfloor \delta}\right\|_2^8\right)\right] + 2\left(\frac{d}{m}\right)^4\right) \\
&\overset{(ii)}{\leq} -\frac{c_\kappa^2}{2}\mathbb{E}\left[l(x_t,u_t)\right] + 2^{33}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right) + \frac{c_\kappa^2}{8}\left(\mathbb{E}\left[l\left(x_{\lfloor \frac{t}{\delta}\rfloor \delta}, u_{\lfloor \frac{t}{\delta}\rfloor \delta}\right)\right] + R^8 + \left(\frac{d}{m}\right)^4\right) \\
&\leq -\frac{c_\kappa^2}{2}\mathbb{E}\left[l(x_t,u_t)\right] + 2^{34}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right) + \frac{c_\kappa^2}{8}\mathbb{E}\left[l\left(x_{\lfloor \frac{t}{\delta}\rfloor \delta}, u_{\lfloor \frac{t}{\delta}\rfloor \delta}\right)\right], \qquad \text{(C.54)}
\end{aligned}
$$

where (i) is by Lemma 58, and (ii) is by Eq. (C.53) and Eq. (4.6) along with some algebra.

Consider an arbitrary positive interger k. By Grönwall's Lemma applied over $s \in [k\delta, (k+1)\delta)$,

$$
\begin{aligned}
&\mathbb{E}\left[l(x_{(k+1)\delta}, u_{(k+1)\delta}\right] \\
&\quad \leq e^{-\frac{c_\kappa^2}{2}\delta}\mathbb{E}\left[l(x_{k\delta},u_{k\delta})\right] + \delta \cdot \left(2^{34}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right) + \frac{c_\kappa^2}{8}\mathbb{E}\left[l\left(x_{\lfloor \frac{t}{\delta}\rfloor \delta}, u_{\lfloor \frac{t}{\delta}\rfloor \delta}\right)\right]\right) \\
&\quad \overset{(i)}{\leq} \left(1 - \frac{c_\kappa^2\delta}{4}\right)\mathbb{E}\left[l(x_{k\delta},u_{k\delta})\right] + \delta \cdot \left(2^{34}c_\kappa^2\left(\frac{d^4}{m^4}+R^8\right) + \frac{c_\kappa^2}{8}\mathbb{E}\left[l\left(x_{\lfloor \frac{t}{\delta}\rfloor \delta}, u_{\lfloor \frac{t}{\delta}\rfloor \delta}\right)\right]\right) \\
&\quad \overset{(ii)}{\leq} e^{-\frac{c_\kappa^2\delta}{8}}\mathbb{E}\left[l(x_{k\delta},u_{k\delta})\right] + 2^{34}c_\kappa^2\delta\left(2^{34}\left(\frac{d^4}{m^4}+R^8\right)\right),
\end{aligned}
$$

where (i) and (ii) use the fact that $c_\kappa^2\delta \leq \frac{1}{10}$, along with $1-a \leq e^{-a} \leq 1 - \frac{a}{2}$ for $|a| \leq \frac{1}{10}$.

Applying the above recursively, using the geometric sum, and Eq. (4.10), we show that for all positive integers k,

$$\mathbb{E}\left[l(x_{k\delta},u_{k\delta})\right] \leq 2^{38}\left(\frac{d^4}{m^4}+R^8\right).$$

For an arbitrary $t \geq 0$, we can similarly verify using the above result, Eq. (C.54), and Grönwall's Lemma that

$$\mathbb{E}\left[l(x_t,u_t)\right] \leq 2^{39}\left(\frac{d^4}{m^4}+R^8\right).$$

This completes the proof of the lemma.

We now state and prove some auxillary lemmas that were useful in the proof above.

Lemma 58 *Assume that $\delta \leq \frac{1}{1000}$. Then for all $t \geq 0$,*

$$\mathbb{E}\left[\|x_t\|_2^8 + \|u_t\|_2^8\right] \leq 1.1\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right)\right] + 2\left(\frac{d}{m}\right)^4.$$

Proof

From the stochastic dynamics defined in Eq. (4.11), Eq. (4.12), Eq. (4.13) and Eq. (4.14), we can verify that

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\left[\left(\|x_t\|_2^8 + \|u_t\|_2^8\right)\right] &\overset{(i)}{=} \mathbb{E}\left[8\|x_t\|_2^6 \langle x_t, u_t\rangle + 8\|u_t\|_2^6 \left\langle u_t, -2u_t - \frac{c_\kappa}{L}\nabla U(x_{\lfloor \frac{t}{\delta} \rfloor \delta})\right\rangle\right] \\
&\quad + \mathbb{E}\left[\frac{8c_\kappa d}{L}\|u_t\|_2^6 + \frac{48c_\kappa}{L}\|u_t\|_2^6\right] \\
&\overset{(ii)}{\leq} 8\mathbb{E}\left[\|x_t\|_2^8 + \|u_t\|_2^8 + \|u_t\|_2^8 + c_\kappa^8\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right] + \mathbb{E}\left[\frac{d}{m}\|u_t\|_2^6\right] \\
&\overset{(iii)}{\leq} 64\mathbb{E}\left[\|x_t\|_2^8 + \|u_t\|_2^8\right] + \mathbb{E}\left[\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right] + \left(\frac{d}{m}\right)^4,
\end{aligned}$$

where (i) is by Itô's Lemma, (ii) is by Assumption (A1), Young's inequality and by the definition of c_κ in Eq. (4.6), and (iii) is again by Young's inequality and definition of c_κ.

Consider an arbitrary $t \geq 0$, and let $k := \lfloor \frac{t}{\delta} \rfloor$. Then for all $s \in [k\delta, (k+1)\delta)$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left(\|x_t\|_2^8 + \|u_t\|_2^8\right)\right] \\
&\quad \leq e^{64(s-k\delta)}\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right)\right] + \left(e^{64(s-k\delta)} - 1\right)\left(\mathbb{E}\left[\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right] + \left(\frac{d}{m}\right)^4\right) \\
&\quad \leq (1 + 128\delta)\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right)\right] + 128\delta\left(\frac{d}{m}\right)^4 \\
&\quad \leq 1.1\mathbb{E}\left[\left(\left\|x_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8 + \left\|u_{\lfloor \frac{t}{\delta} \rfloor \delta}\right\|_2^8\right)\right] + 2\frac{d}{m},
\end{aligned}$$

where the final two inequalities are both by our assumption that $\delta \leq \frac{1}{1000}$. □

Lemma 59 *For (x_t, u_t) satisfying $\|x_t\|_2^2 + \|x_t + u_t\|_2^2 \geq 4R^2$,*

$$\langle x_t, u_t\rangle + \left\langle x_t + u_t, -u_t - \frac{c_\kappa}{L}\nabla U(x_t)\right\rangle \leq -\frac{c_\kappa^2}{6}\left(\|x_t\|_2^2 + \|x_t + u_t\|_2^2\right).$$

Proof
We first verify that

$$\begin{aligned}
&\langle x_t, u_t\rangle + \Big\langle x_t + u_t, -u_t - \frac{c_\kappa}{L}\nabla U(x_{i\nu})\Big\rangle \\
&\quad = -\|u_t\|_2^2 - \frac{c_\kappa}{L}\langle x_t, \nabla U(x_t)\rangle - \Big\langle u_t, \frac{c_\kappa}{L}\nabla U(x_t)\Big\rangle \\
&\quad = -\|u_t\|_2^2 - \frac{c_\kappa}{L}\langle x_t, \nabla U(x_t)\rangle - \frac{c_\kappa}{L}\langle u_t, \nabla U(x_t)\rangle \\
&\quad = -\|u_t\|_2^2 - \frac{c_\kappa}{L}\langle x_t, \nabla U(x_t)\rangle + \frac{1}{2}\|u_t\|_2^2 + \frac{c_\kappa^2}{2L^2}\|\nabla U(x_t)\|_2^2 - \frac{1}{2}\Big\|u_t + \frac{c_\kappa}{L}\nabla U(x_t)\Big\|_2^2 \\
&\quad \le -\frac{1}{2}\|u_t\|_2^2 - \frac{c_\kappa}{L}\langle x_t, \nabla U(x_t)\rangle + \frac{c_\kappa^2}{2}\|x_t\|_2^2 =: \spadesuit
\end{aligned} \tag{C.55}$$

Now consider two cases:
Case 1: ($\|x_t\|_2 \le R$) By Young's inequality we get that,

$$\|x_t + u_t\|_2^2 \le 11\|u_t\|_2^2 + 1.1\|x_t\|_2^2.$$

Furthermore, by our assumption that $\|x_t\|_2^2 + \|x_t + u_t\|_2^2 \ge 4R^2$,

$$\begin{aligned}
11\|u_t\|_2^2 \ge& \|x_t + u_t\|_2^2 - 1.1\|x_t\|_2^2 \\
\ge& 4R^2 - 2.1R^2 \\
\ge& 1.9R^2 \\
\ge& 1.9\|x_t\|_2^2.
\end{aligned} \tag{C.56}$$

Thus in this case $\|u_t\|_2^2 \ge \frac{1}{10}R^2$, and $\spadesuit$ can be upper bounded by

$$\begin{aligned}
\spadesuit =& -\frac{1}{2}\|u_t\|_2^2 - \Big\langle x_t, \frac{c_\kappa}{L}\nabla U(x_t)\Big\rangle + \frac{c^2}{2}\|x_t\|_2^2 \\
\overset{(i)}{\le}& -\frac{1}{2}\|u_t\|_2^2 + c_\kappa\|x_t\|_2^2 + \frac{c_\kappa^2}{2}\|x_t\|_2^2 \\
\overset{(ii)}{\le}& -\frac{1}{2}\|u_t\|_2^2 + 2c_\kappa\|x_t\|_2^2 \\
\overset{(iii)}{\le}& -\frac{1}{4}\|u_t\|_2^2 \\
\overset{(iv)}{\le}& -\frac{1}{160}\big(\|x_t\|_2^2 + \|x_t + u_t\|_2^2\big),
\end{aligned}$$

where (i) is by L-Lipschitz of ∇U and Cauchy-Schwarz, (ii) and (iii) are because $c_\kappa := \frac{1}{1000\kappa} \le \frac{1}{1000}$ and by Eq. (C.56), the (iv) is because

$$\begin{aligned}
\|x_t\|_2^2 + \|x_t + u_t\|_2^2 \le& 3\|x_t\|_2^2 + 2\|u_t\|_2^2 \\
\le& 30\|u_t\|_2 + 2\|u_t\|_2 \\
\le& 40\|u_t\|_2^2,
\end{aligned}$$

where the second inequality is by again by Eq. (C.56).

Case 2: ($\|x_t\|_2 \geq R$)

By Assumption (A3), $-\frac{c_\kappa}{L}\langle x_t, \nabla_t\rangle \leq -\frac{c_\kappa}{\kappa}\|x_t\|_2^2$. Thus we can upper bound ♠ as follows:

$$\begin{aligned}
\spadesuit =& -\frac{1}{2}\|u_t\|_2^2 - \frac{c_\kappa}{L}\langle x_t, \nabla_t\rangle + \frac{c^2}{2}\|x_t\|_2^2 \\
\leq& -\frac{1}{2}\|u_t\|_2^2 - c_\kappa^2\|x_t\|_2^2 + \frac{c^2}{2}\|x_t\|_2^2 \\
\leq& -\|u_t\|_2^2 - \frac{c_\kappa^2}{2}\|x_t\|_2^2 \\
\leq& -\frac{c_\kappa^2}{3}\big(\|x_t\|_2^2 + \|x_t+u_t\|_2^2\big).
\end{aligned}$$

Putting the previous two results together, and using Young's inequality:

$$\begin{aligned}
\spadesuit \leq& -\frac{c_\kappa^2}{3}\big(\|x_t\|_2^2 + \|x_t+u_t\|_2^2\big) \\
\leq& -\frac{c_\kappa^2}{3}\big(\|x_t\|_2^2 + \|x_t+u_t\|_2^2\big) \\
\leq& -\frac{c_\kappa^2}{6}\big(\|x_t\|_2^2 + \|x_t+u_t\|_2^2\big).
\end{aligned}$$

□

C.6.2 Proof of Lemma 57

Let us consider the Lyapunov function $l(y_t, v_t) := \big(\|y_t\|_2^2 + \|y_t+v_t\|_2^2 - 4R^2\big)_+^4$.

By calculating its derivatives we can verify that

$$\begin{aligned}
\nabla_x l(y_t, v_t) &= 8l(y_t, v_t)^{3/4}(y_t) \\
\nabla_u l(y_t, v_t) &= 8l(y_t, v_t)^{3/4}(y_t + v_t) \\
\nabla_u^2 l(y_t, v_t) &= 8l(y_t, v_t)^{3/4} I + 24 l(y_t, v_t)^{2/4}(y_t+v_t)(y_t+v_t)^T.
\end{aligned}$$

Recall the dynamics of the variables y_t and v_t,

$$\begin{aligned}
dy_t =& v_t dt \\
dv_t =& -2v_t - \frac{c_\kappa}{L}\nabla U(x_{y_t})dt + 2\sqrt{\frac{c_\kappa}{L}}dB_t.
\end{aligned}$$

By Itô's lemma we can study the time evolution of this Lyapunov function:

$$\begin{aligned}
dl(y_t, v_t) &= 8l(y_t, v_t)^{3/4}\Big(\langle y_t, v_t\rangle + \Big\langle y_t + v_t, -v_t - \frac{c_\kappa}{L}\nabla U(y_t)\Big\rangle\Big)dt \\
&\quad + \frac{16c_\kappa}{L}\big(l(y_t, v_t)^{3/4}d + l(y_t, v_t)^{2/4}\|y_t + v_t\|_2^2\big)dt \\
&\quad\quad + 16\sqrt{\frac{c_\kappa}{L}}l(y_t, v_t)^{3/4}(\langle y_t, v_t\rangle + \langle y_t + v_t, dB_t\rangle) \\
&\overset{(i)}{\leq} 8l(y_t, v_t)^{3/4}\bigg(-\frac{c_\kappa^2}{6}\big(\|y_t\|_2^2 + \|y_t + v_t\|_2^2\big)\bigg)dt \\
&\quad + \frac{16c_\kappa}{L}\big(l(y_t, v_t)^{3/4}d + l(y_t, v_t)^{2/4}\|y_t + v_t\|_2^2\big)dt \\
&\quad\quad + 16\sqrt{\frac{c_\kappa}{L}}l(y_t, v_t)^{3/4}(\langle y_t, v_t\rangle + \langle y_t + v_t, dB_t\rangle) \\
&\overset{(ii)}{\leq} -c_\kappa^2 l(y_t, v_t)dt \\
&\quad + \frac{32c_\kappa}{L}\big(l(y_t, v_t)^{3/4}d\big)dt + \frac{64c_\kappa}{L}\big(l(y_t, v_t)^{2/4}R^2\big)dt \\
&\quad\quad + 16\sqrt{\frac{c_\kappa}{L}}l(y_t, v_t)^{3/4}(\langle y_t, v_t\rangle + \langle y_t + v_t, dB_t\rangle) \\
&\leq -c_\kappa^2 l(y_t, v_t)dt \\
&\quad + \frac{c_\kappa^2}{8}l(y_t, v_t)dt + \frac{2^{25}d^4}{c_\kappa^2 L^4}dt + \frac{c_\kappa^2}{8}l(y_t, v_t)dt + \frac{2^{16}R^4}{L^2} \\
&\quad\quad + 16\sqrt{\frac{c_\kappa}{L}}l(y_t, v_t)^{3/4}(\langle y_t, v_t\rangle + \langle y_t + v_t, dB_t\rangle) \\
&\leq -\frac{c_\kappa^2}{2}l(y_t, v_t)dt + 2^{26}\bigg(\frac{d^4}{c_\kappa^2 L^4} + c^2R^8\bigg)dt \\
&\quad + 16\sqrt{\frac{c_\kappa}{L}}l(y_t, v_t)^{3/4}(\langle y_t, v_t\rangle + \langle y_t + v_t, dB_t\rangle),
\end{aligned}$$

where (i) can be proved by an argument similar to the proof of Lemma 59, and is omitted, while (ii) follows because

$$\|y + v\|_2^2 \leq l(y, v)^{1/4} + 4R^2, \quad \text{and,} \quad \|y\|_2^2 + \|x + u\|_2^2 \geq l(y, v)^{1/4}$$

by the definition of $l(x, u)$. Taking expectations on both sides, the term involving the Brownian motion, dB_t, goes to zero. Note also that (y_t, v_t) is distributed according to the invariant distribution p^* for all $t \geq 0$, therefore,

$$\begin{aligned}
0 = \frac{d}{dt}\mathbb{E}\left[l(y_t, v_t)\right] &\leq -\frac{c_\kappa^2}{2}\mathbb{E}\left[l(y_t, v_t)\right] + 2^{26}\bigg(\frac{d^4}{c_\kappa^2 L^4} + c^2R^8\bigg) \\
&\leq -\frac{c_\kappa^2}{2}\mathbb{E}\left[l(y_t, v_t)\right] + 2^{26}c_\kappa^2\bigg(10^{12}\frac{d^4}{m^4} + R^8\bigg)
\end{aligned}$$

Thus

$$\mathbb{E}\left[l(y_t, v_t)\right] \leq 2^{27}\left(10^{12}\frac{d^4}{m^4} + R^8\right) \leq 2^{66}\left(\frac{d}{m} + R^2\right)^4.$$

This completes the proof of the lemma.

We now state and prove some auxillary lemmas that were useful in the proof above.

Lemma 60 *Let x_t be evolved according to the dynamics in Eq.* (C.3)*. Then for all $t \geq 0$,*

$$\mathbb{E}\left[\|x_t\|_2^2\right] \leq 2\left(R^2 + \frac{d}{m}\right).$$

Proof
Let $\theta_k \sim \mathcal{N}(0, I)$ then we have,

$$\begin{aligned}
\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\} \cdot \left\|x_{(k+1)\delta}\right\|_2^2 &= \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\} \cdot \left\|x_{k\delta} - \delta\nabla U(x_k) + \sqrt{2\delta}\theta_k\right\|_2^2 \\
&\leq \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\} \cdot \|x_{k\delta} - \delta\nabla U(x_k)\|_2^2 \\
&\quad + \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\} \cdot \left\|\sqrt{2\delta}\theta_k\right\|_2^2 \\
&\quad + \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\} \cdot 2\left\langle x_{k\delta} - \delta\nabla U(x_k), \sqrt{2\delta}\theta_k\right\rangle.
\end{aligned}$$

Consider two cases:

If $\|x_{k\delta}\|_2 \geq R$, then

$$\begin{aligned}
\|x_{k\delta} - \delta\nabla U(x_k)\|_2^2 &= \|x_{k\delta}\|_2^2 - \langle x_{k\delta}, 2\delta\nabla U(x_k)\rangle + \delta^2\|\nabla U(x_k)\|_2^2 \\
&\overset{(i)}{\leq} (1 - 2\delta m)\|x_{k\delta}\|_2^2 + \delta^2\|\nabla U(x_k)\|_2^2 \\
&\overset{(ii)}{\leq} \left(1 - 2\delta m + \delta^2 L^2\right)\|x_{k\delta}\|_2^2 \\
&\overset{(iii)}{\leq} (1 - \delta m)\|x_{k\delta}\|_2^2,
\end{aligned}$$

where (i) is by Assumption (A3), (ii) is by Assumption (A1), and (iii) is by our assumption that $\delta \leq \frac{1}{\kappa L}$.

While I=if $\|x_{k\delta}\|_2 \leq R$, then

$$\begin{aligned}
\|x_{k\delta} - \delta\nabla U(x_k)\|_2^2 &= \|x_{k\delta}\|_2^2 - \langle x_{k\delta}, 2\delta\nabla U(x_k)\rangle + \delta^2\|\nabla U(x_k)\|_2^2 \\
&\overset{(i)}{\leq} \left(1 + 2\delta L + \delta^2 L^2\right)\|x_k\|_2^2 \\
&\overset{(ii)}{\leq} (1 + 3\delta L)\|x_k\|_2^2,
\end{aligned}$$

where (i) is by Assumption (A1), and (ii) is by our assumption that $\delta \leq \frac{1}{\kappa L}$.

Thus for both cases we have,

$$\begin{aligned}\|x_{k\delta} - \delta\nabla U(x_k)\|_2^2 \leq & \mathbb{1}\left\{\|x_{k\delta}\|_2 \geq R\right\}(1-\delta m)\|x_{k\delta}\|_2^2 + \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}(1+3\delta L)\|x_k\|_2^2 \\ \leq & \|x_{k\delta}\|_2^2 - \delta m\|x_{k\delta}\|_2^2 + \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\cdot(3\delta L + \delta m)\|x_{k\delta}\|_2^2.\end{aligned}$$

Thus we have:

$$\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\|x_{k\delta} - \delta\nabla U(x_k)\|_2^2 \leq \mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}(1-\delta m)\|x_{k\delta}\|_2^2.$$

By taking expectations with respect to the Brownian motion we get,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\left\|x_{(k+1)\delta}\right\|_2^2\right] \leq & (1-\delta m)\mathbb{E}\left[\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\|x_{k\delta}\|_2^2\right] \\ & + \mathbb{E}\left[\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\left\|\sqrt{2\delta}\theta_k\right\|_2^2\right] \\ \leq & (1-\delta m)\mathbb{E}\left[\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\|x_{k\delta}\|_2^2\right] + 2\delta d.\end{aligned}$$

Applying this inequality recursively over k steps we arrive at,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\left\|x_{(k+1)\delta}\right\|_2^2\right] \leq & e^{-\delta m}\mathbb{E}\left[\mathbb{1}\left\{\|x_{k\delta}\|_2 \leq R\right\}\left\|x_{(k+1)\delta}\right\|_2^2\right] + \frac{2\delta d}{\delta m} \\ \leq & \frac{2d}{m}.\end{aligned}$$

Thus we get that,

$$\mathbb{E}\left[\left\|x_{(k+1)\delta}\right\|_2^2\right] \leq 2\left(R^2 + \frac{d}{m}\right).$$

□

Lemma 61 *Let $y \sim p^*(y) \propto e^{-U(y)}$. Then*

$$\mathbb{E}\left[\|y\|_2^8\right] \leq 2^{20}\left(\frac{d^4}{m^4} + R^8\right).$$

Proof
Let $l(y) := \left(\|y\|_2^2 - R^2\right)_+^4$. We calculate derivatives and verify that

$$\begin{aligned}\nabla l(y) &= 8l(y)^{3/4}\cdot y \\ \nabla^2 l(y) &= 48l(y)^{2/4}\cdot yy^T + 8l(y)^{3/4}I,\end{aligned}$$

where I is the identity matrix. By Itô's Lemma:

$$dl(y_t) = \left\langle\nabla l(y_t), -\nabla U(y_t)\right\rangle dt + \left\langle\nabla l(y_t), \sqrt{2}dB_t\right\rangle + \frac{1}{2}\mathrm{tr}\left(\nabla^2 l(y_t)\right). \tag{C.57}$$

We start by analyzing the first term,

$$\begin{aligned}\langle \nabla l(y_t), -\nabla U(y_t)\rangle &= \left\langle 8l(y_t)^{3/4}\cdot y_t, -\nabla U(y_t)\right\rangle \\ &\overset{(i)}{\leq} \mathbb{1}\left\{\|y_t\|_2 \geq R\right\}\left(8l(y_t)^{3/4}\right)\left(-m\|y_t\|_2^2\right) \\ &\qquad + \mathbb{1}\left\{\|y_t\|_2 < R\right\}\left\langle 8l(y_t)^{3/4}y, -\nabla U(y_t)\right\rangle \\ &\overset{(ii)}{=} \left(8l(y_t)^{3/4}\right)\left(-m\|y_t\|_2^2\right) \\ &\leq -8ml(y_t),\end{aligned}$$

where (i) is by Assumption (A3), and, (ii) is because $\mathbb{1}\left\{\|y_t\|_2 < R\right\}\cdot l(y_t) = 0$ and $\mathbb{1}\left\{\|y_t\|_2 \geq R\right\}\cdot l(y_t) = l(y_t)$ by definition of $l(y_t)$.

Consider the other term on the right-hand side of Eq. (C.57):

$$\begin{aligned}\operatorname{tr}\left(\nabla^2 l(y_t)\right) &= 48l(y_t)^{2/4}\|y\|_2 + 8l(y_t)^{3/4}d \\ &\overset{(i)}{\leq} 48l(y_t)^{3/4} + 8l(y_t)^{3/4}d + 48l(y_t)^{2/4}R^2 \\ &\leq 64l(y_t)^{3/4}d + 48l(y_t)^{2/4}R^2 \\ &\overset{(ii)}{\leq} 2ml(y_t) + 2^{21}\frac{d^4}{m^3} + 2ml(y_t) + 2^{11}\frac{R^4}{m} \\ &\overset{(iii)}{\leq} 4ml(y_t) + 2^{22}\left(\frac{d^4}{m^3} + mR^8\right),\end{aligned}$$

where (i) is by definition of $l(y)$, while (ii) and (iii) are by Young's inequality.

Put together into Eq. (C.57) and taking expectations,

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[l(y_t)\right] &\leq -8m\mathbb{E}\left[l(y_t)\right] + 4ml(y_t) + 2^{22}\left(\frac{d^4}{m^3} + mR^8\right) \\ &\leq -4m\mathbb{E}\left[l(y_t)\right] + 2^{22}\left(\frac{d^4}{m^3} + mR^8\right).\end{aligned}$$

Since $y_t \sim p^*$ for all t, $\frac{d}{dt}\mathbb{E}\left[l(y_t)\right] = 0$, thus we get,

$$\mathbb{E}\left[l(y_t)\right] \leq 2^{20}\left(\frac{d^4}{m^4} + R^8\right).$$

□

C.7 Existence of Coupling

Proof of Lemma 15

We prove the existence of a unique strong solution for $(x_t, u_t, y_t, v_t, \tau_{\lfloor \frac{t}{\nu}\rfloor})$ inductively: Let

k be an arbitrary nonnegative integer, and suppose that the lemma statement holds for all $s \in [0, k\nu]$. We show that the lemma statement holds for all $s \in [0, (k+1)\nu]$.

First, we can verify that for $t \in [k\nu, (k+1)\nu)$,

$$\tau_{\lfloor \frac{t}{\nu} \rfloor} = \tau_k,$$

that is, $\tau_{\lfloor \frac{t}{\nu} \rfloor}$ is a constant, and so $\mu_{\lfloor \frac{t}{\nu} \rfloor} = \mu_k$ is also a constant.

Next, we find that for $t \in [k\nu, (k+1)\nu)$, the following is algebraically equivalent to dynamics described by Eqs.(4.11)–(4.14):

$$\begin{aligned}
dx_t &= u_t dt, \\
du_t &= -2u_t dt - \frac{c_\kappa}{L} \nabla U\Big(x_{\lfloor \frac{t}{\delta} \rfloor \delta}\Big) dt + 2\sqrt{\frac{c_\kappa}{L}} dB_t, \\
dy_t &= v_t dt, \\
dv_t &= -2v_t - \frac{c_\kappa}{L} \nabla U(y_t) dt + 2\sqrt{\frac{c_\kappa}{L}} dB_t - \mu_k \cdot \left(4\sqrt{\frac{c_\kappa}{L}} \gamma_t \gamma_t^T dB_t + 2\sqrt{\frac{c_\kappa}{L}} \bar{\gamma}_t \bar{\gamma}_t^T dA_t \right),
\end{aligned}$$

where we use the fact that μ_t takes on a constant value over $t \in [k\nu, (k+1)\nu)$.

We proceed by applying Theorem 5.2.1 of [66], which states that if the following holds:

1. $\mathbb{E}\left[\|x_{k\nu}\|_2^2 + \|y_{k\nu}\|_2^2 + \|u_{k\nu}\|_2^2 + \|v_{k\nu}\|_2^2\right] < \infty$.
2. For all $x, y \in \mathbb{R}^d$, $\|\nabla U(x) - \nabla U(y)\|_2 \leq D\|x - y\|_2$ for some constant $D > 0$.
3. For all $(x, y, u, v), (x', y', u', v')$,

   $$\left\|\gamma\gamma^T - \gamma'\gamma'^T\right\|_2 + \left\|\bar{\gamma}\bar{\gamma}_t'^T - \bar{\gamma}_t'\bar{\gamma}'^T\right\|_2 \leq D(\|x - x'\|_2 + \|y - y'\|_2 + \|u - u'\|_2 + \|v - v'\|_2),$$

   for some constant D (where γ and $\bar{\gamma}$ are functions of (x, y, u, v), as defined in Eq. (4.15), similarly for γ', $\bar{\gamma}'$ and (x', y', u', v')),

then there is a solution (x_t, y_t, u_t, v_t) for $t \in [k\nu, (k+1)\nu]$ with the properties:

(a) (x_t, y_t, u_t, v_t) is unique and t-continuous with probability one.

(b) (x_t, y_t, u_t, v_t) is adapted to the filtration $\mathcal{F}_t$ generated by $(x_{k\nu}, y_{k\nu}, u_{k\nu}, v_{k\nu})$ and dB_t and dA_t for $t \in [k\nu, (k+1)\nu)$.

(c) $\int_0^T \mathbb{E}\left[\|x_t\|_2^2 + \|y_t\|_2^2 + \|u_t\|_2^2 + \|v_t\|_2^2\right] dt < \infty$.

We can verify the first condition holds by using Lemma 56 and Lemma 57. Condition 2 holds due to our smoothness assumption, Assumption (A1).

We can verify that Condition 3 also holds using the argument below:

From the definition of $\mathcal{M}$ in Eq. (4.15), we know that $|\mathcal{M}(r)'| \leq \frac{1}{2}|\sin(r \cdot 2\pi/\beta)| \cdot \frac{2\pi}{\beta} \leq \frac{\pi}{\beta}$.

By definition of γ_t in Eq. (4.15),

$$\gamma_t \gamma_t^T := \mathcal{M}(\|z_t + w_t\|_2) \cdot \frac{(z_t + w_t)(z_t + w_t)^T}{\|z_t + w_t\|_2^2}.$$

To simplify notation, consider an arbitrary $x, y \in \mathbb{R}^d$, and assume wlog that $\|x\|_2 \leq \|y\|_2$. We will bound

$$\left\| \mathcal{M}(\|x\|_2) \frac{xx^T}{\|x\|_2^2} - \mathcal{M}(\|y\|_2) \frac{yy^T}{\|y\|_2^2} \right\|_2 \leq D\|x - y\|_2,$$

for some D, which implies condition 3.

By the triangle inequality,

$$\begin{aligned}
&\left\| \mathcal{M}(\|x\|_2) \frac{xx^T}{\|x\|_2^2} - \mathcal{M}(\|y\|_2) \frac{yy^T}{\|y\|_2^2} \right\|_2 \\
&\leq \mathcal{M}(\|x\|_2) \left\| \frac{xx^T}{\|x\|_2^2} - \frac{yy^T}{\|y\|_2^2} \right\|_2 + \left\| \frac{yy^T}{\|y\|_2^2} \right\|_2 |\mathcal{M}(\|x\|_2) - \mathcal{M}(\|y\|_2)| \\
&\leq \mathcal{M}(\|x\|_2) \left\| \frac{xx^T}{\|x\|_2^2} - \frac{yy^T}{\|y\|_2^2} \right\|_2 + |\mathcal{M}(\|x\|_2) - \mathcal{M}(\|y\|_2)|. \qquad \text{(C.58)}
\end{aligned}$$

The second term can be bounded as

$$|\mathcal{M}(\|x\|_2) - \mathcal{M}(\|y\|_2)| \leq \frac{\pi}{\beta} |\|x\|_2 - \|y\|_2|_2 \leq \frac{\pi}{\beta} \|x - y\|_2,$$

where we use the upper bound we established on $|\mathcal{M}'(r)|$.

To bound the first term, we consider two cases:

If $\|x\|_2 \leq \beta/2$, $\mathcal{M}(\|x\|_2) = 0$ and we are done.

If $\|x\|_2 \geq \beta/2$, we verify that the transformation $T(x) = \frac{x}{\|x\|_2}$ has Jacobian $\nabla T(x) = \frac{1}{\|x\|_2}\left(I - \frac{xx^T}{\|x\|_2}\right)$, so that $\|\nabla T(x)\|_2 \leq \frac{1}{\|x\|_2}$. By our earlier assumption that $\|x\|_2 \leq \|y\|_2$, we know that $\|ax + (1-a)y\|_2 \geq \beta/2$ for all $a \in [0, 1]$. Therefore,

$$\left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 = \|T(x) - T(y)\|_2 \leq \frac{1}{\|x\|_2} \|x - y\|_2 \leq \frac{2}{\beta}.$$

By the triangle inequality and some algebra, we obtain:

$$\begin{aligned}
&\left\| \frac{xx^T}{\|x\|_2^2} - \frac{yy^T}{\|y\|_2^2} \right\|_2 \\
&\leq \left\| \frac{x}{\|x\|_2} + \frac{y}{\|y\|_2} \right\|_2 \left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 \\
&\leq 2 \left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 \\
&\leq \frac{4}{\beta} \|x - y\|_2,
\end{aligned}$$

where the first two inequalities are due to the triangle inequality. Combined with the fact that $\mathcal{M}(r) \leq 1$ for all r, we can bound Eq. (C.58) by $\frac{8}{\beta}\|x - y\|_2$.

A similar argument can be used to show that $\bar{\gamma}_t$ is Lipschitz. Let $\mathcal{N}(x) := \left(1 - (1 - 2\mathcal{M}(\|x\|_2))^2\right)^{1/2}$. Then we verify that

$$\mathcal{N}(r) := \begin{cases} 1, & \text{for } r \in [\beta, \infty) \\ \sin\left(r \cdot \frac{2\pi}{\beta}\right), & \text{for } r \in [\beta/2, \beta] \\ 0, & \text{for } r \in [0, \beta/2] \end{cases}$$

$$\bar{\gamma}_t := (\mathcal{N}(\|z_t + w_t\|_2))^{1/2} \frac{z_t + w_t}{\|z_t + w_t\|_2}.$$

The proof is almost identical to the proof of (C.58), so we omit it, but highlight two crucial facts:

$$\begin{aligned}
&1.\ \mathcal{N}(r) \in [0, 1] && \text{for all } r \\
&2.\ |\mathcal{N}'(r)| \leq \frac{2\pi}{\beta} && \text{for all } r.
\end{aligned}$$

Thus we find that Condition 3 is satisfied with $D = \frac{16}{\beta}$, and in turn show that (a)-(c) hold for $t \in [k\nu, (k+1)\nu]$. From Eq. (4.17) we know that $\tau_{(k+1)\nu}$ is a function of $(x_{(k+1)\nu}, u_{(k+1)\nu}, y_{(k+1)\nu}, v_{(k+1)\nu}, \tau_k)$. Thus we have shown the existence of a unique solution $(x_t, y_t, u_t, v_t, \tau_{\lfloor \frac{t}{\nu} \rfloor})$ for $t \in [k\nu, (k+1)\nu]$, where (x_t, y_t, u_t, v_t) is t-continuous.

The proof of the lemma now follows by induction over k. □

Lemma 62 *Let B_t and A_t be two independent Brownian motions, and let $\mathcal{F}_t$ be the σ-algebra generated by B_s, A_s; $s \leq t$, and (x_0, u_0, y_0, v_0).*

For all $t \geq 0$, the stochastic process ϕ_t defined in Eqs. (4.24) has a unique solution such that ϕ_t is t-continuous with probability one, and satisfies the following, for all $s \geq 0$:

1. *ϕ_t is adapted to the filtration $\mathcal{F}_s$.*

2. $\mathbb{E}\left[\|\phi_t\|_2^2\right] \leq \infty$.

Proof
The proof is almost identical to that of Lemma 15. The main additional requirement is showing that there exists a constant D such that for any (x, y, u, v) and (x', y', u', v'),

$$\left\|\nabla_w f(r)\gamma\gamma^T - \nabla_{w'} f(r')\gamma'\gamma'^T\right\|_2, \tag{C.59}$$

with γ (resp γ') being a function of (x, y, u, v) (resp γ') as defined in (4.15). and r being a function of (x, y, u, v) as defined in (4.19). In the proof of Lemma 15, we already showed that $\gamma\gamma^T$ and $\gamma'\gamma'^T$ are uniformly bounded and lipschitz, thus it is sufficient to show that

$$\begin{aligned} &1.\ \|\nabla_w f(r) - \nabla_{w'} f(r')\|_2 \leq D\|x - x'\|_2 + \|y - y'\|_2 + \|u - u'\|_2 + \|v - v'\|_2 \\ &2.\ \|\nabla_w f(r)\|_2 \leq D. \end{aligned} \tag{C.60}$$

The second point is easy to verify:

$$\nabla_w f(r) = f'(r)\nabla_w r = f'(r)(\nabla_w \ell(z + w))$$

Thus $\|\nabla_w f(r)\|_2 \leq 1$ using item (F2) of Lemma 55 and item 2 of Lemma 31.

To prove the first point, we verify that

$$\nabla_w^2 f(r) = f''(r)(\nabla_w \ell(z + w)) + f'(r)\big(\nabla_w^2 \ell(z + w)\big).$$

Using item (F7) of Lemma 55 and item

$$\left\|\nabla_w^2 f(r)\right\|_2 \leq 4\alpha_f \mathcal{R}_f + \frac{4}{\mathcal{R}_f} + \frac{8}{\beta};$$

this implies C.60 which in turn implies (C.59). Note that $\|w - w'\|_2 \leq \|u - u'\|_2 + \|v - v'\|$. □

C.8 Coupling and Discretization

Proof of Lemma 14
Let us define

$$\bar{B}_t := \int_0^t \left(dB_t - \mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\} \cdot \left(2\gamma_t\gamma_t^T dB_t + \bar{\gamma}_t\bar{\gamma}_t^T dA_t\right)\right).$$

We will show that $\bar{B}_t$ is a Brownian motion by using Levy's characterization. The conclusion then follows immediately from the dynamics defined in Eq. (3.1).

Since B_t and A_t are Brownian motions, $\bar{B}_t$ is also a continuous martingale with respect to the filtration $\mathcal{F}_t$. Further the quadratic variation of $\bar{B}_t$ over an interval $[s, s']$ is

$$\int_s^{s'} \underbrace{\left(I - 2\mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\}\gamma_t\gamma_t^T\right)^2 + \mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\}\left(\bar{\gamma}_t\bar{\gamma}_t^T\right)^2}_{=:\spadesuit} dt.$$

If $\mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\} = 0$, then the above is clearly the identity matrix – I.

If, on the other hand, $\mathbb{1}\left\{k\nu \geq \tau_{\lfloor \frac{t}{\nu} \rfloor} + T_{sync}\right\} = 1$, define $c_t := z_t + w_t$; then by the definition of γ_t and $\bar{\gamma}_t$ in Eq. (4.15), we find that

$$\begin{aligned}
\spadesuit &= \left(I - 2\mathcal{M}(\|c_t\|_2)\frac{c_t c_t^T}{\|c_t\|_2^2}\right)^2 + \left(1 - (1 - 2\mathcal{M}(\|c_t\|_2))^2\right)\frac{c_t c_t^T}{\|c_t\|_2^2} \\
&\overset{(i)}{=} I - \frac{c_t c_t^T}{\|c_t\|_2^2} + (1 - 2\mathcal{M}(\|c_t\|_2))^2\frac{c_t c_t^T}{\|c_t\|_2^2} + \left(1 - (1 - 2\mathcal{M}(\|c_t\|_2))^2\right)\frac{c_t c_t^T}{\|c_t\|_2^2} \\
&= I,
\end{aligned}$$

where (i) follows by the eigenvalue decomposition of the matrix $\left(I - 2\mathcal{M}(\|c_t\|_2)\frac{c_t c_t^T}{\|c_t\|_2^2}\right)^2$.

Thus the quadratic variation of $\bar{B}_t$ over the interval $[s, s]$ is $(s' - s)I$, thus satisfying Levy's characterization of a Brownian motion.

□

Proof of Lemma 32
Using similar steps as Lemma 14, we can verify that

$$\bar{B}_t := \int_0^t \sqrt{2} dB_t - 2\sqrt{2}\gamma_t \gamma_t^T dB_t + \sqrt{2}\bar{\gamma}_t \bar{\gamma}_t^T dA_t.$$

is a Brownian motion. The proof follows immediately. □

Lemma 63 *Given $(x_{k\delta}, u_{k\delta})$, the solution (x_t, u_t), for $t \in (k\delta, (k+1)\delta]$, of the discrete underdamped Langevin diffusion defined by the dynamics in Eq. (4.7) is*

$$\begin{aligned}
u_t &= u_{k\delta} e^{-2(t-k\delta)} - \frac{c_\kappa}{L}\left(\int_{k\delta}^t e^{-2(t-s)}\nabla U(x_{k\delta}) ds\right) + \sqrt{\frac{4c_\kappa}{L}}\int_{k\delta}^t e^{-2(t-s)} dB_s \qquad \text{(C.61)} \\
x_t &= x_{k\delta} + \int_{k\delta}^t u_s ds.
\end{aligned}$$

Proof
It can be easily verified that the above expressions have the correct initial values $(x_{k\delta}, u_{k\delta})$. By taking derivatives, one can also verify that they satisfy the stochastic differential equations in Eq. (4.7). □

Lemma 64 *Conditioned on $(x_{k\delta}, u_{k\delta})$, the solution $(x_{(k+1)\delta}, u_{(k+1)\delta})$ of Eq. (4.7) is a Gaussian with mean,*

$$\begin{aligned}
\mathbb{E}\left[u_{(k+1)\delta}\right] &= u_{k\delta} e^{-2\delta} - \frac{c_\kappa}{2L}(1 - e^{-2\delta})\nabla f(x_{k\delta}) \\
\mathbb{E}\left[x_{(k+1)\delta}\right] &= x_{k\delta} + \frac{1}{2}(1 - e^{-2\delta})u_{k\delta} - \frac{c_\kappa}{2L}\left(\delta - \frac{1}{2}\left(1 - e^{-2\delta}\right)\right)\nabla U(x_{k\delta}),
\end{aligned}$$

and covariance,

$$\mathbb{E}\left[\left(x_{(k+1)\delta} - \mathbb{E}\left[x_{(k+1)\delta}\right]\right)\left(x_{(k+1)\delta} - \mathbb{E}\left[x_{(k+1)\delta}\right]\right)^\top\right] = \frac{c_\kappa}{L}\left[\delta - \frac{1}{4}e^{-4\delta} - \frac{3}{4} + e^{-2\delta}\right] \cdot I_{d\times d}$$
$$\mathbb{E}\left[\left(u_{(k+1)\delta} - \mathbb{E}\left[u_{(k+1)\delta}\right]\right)\left(u_{(k+1)\delta} - \mathbb{E}\left[u_{(k+1)\delta}\right]\right)^\top\right] = \frac{c_\kappa}{L}(1 - e^{-4\delta}) \cdot I_{d\times d}$$
$$\mathbb{E}\left[\left(x_{(k+1)\delta} - \mathbb{E}\left[x_{(k+1)\delta}\right]\right)\left(u_{(k+1)\delta} - \mathbb{E}\left[u_{(k+1)\delta}\right]\right)^\top\right] = \frac{c_\kappa}{2L}\left[1 + e^{-4\delta} - 2e^{-2\delta}\right] \cdot I_{d\times d}.$$

Proof
Consider some $t \in [k\delta, (k+1)\delta)$.

It follows from the definition of Brownian motion that the distribution of (x_t, u_t) is a $2d$-dimensional Gaussian distribution. We will compute its moments below, using the expression in Lemma 63. Computation of the conditional means is straightforward, as we can simply ignore the zero-mean Brownian motion terms:

$$\mathbb{E}\left[u_t\right] = u_{k\delta}e^{-2(t-k\delta)} - \frac{c_\kappa}{2L}(1 - e^{-2(t-k\delta)})\nabla U(x_{k\delta}) \tag{C.62}$$
$$\mathbb{E}\left[x_t\right] = x_{k\delta} + \frac{1}{2}(1 - e^{-2(t-k\delta)})u_{k\delta} - \frac{c_\kappa}{2L}\left(t - k\delta - \frac{1}{2}\left(1 - e^{-2(t-k\delta)}\right)\right)\nabla U(x_{k\delta}). \tag{C.63}$$

The conditional variance for u_t only involves the Brownian motion term:

$$\begin{aligned}\mathbb{E}\left[\left(u_t - \mathbb{E}\left[u_t\right]\right)\left(u_t - \mathbb{E}\left[u_t\right]\right)^\top\right] &= \frac{4c_\kappa}{L}\mathbb{E}\left[\left(\int_{k\delta}^t e^{-2(t-s)}dB_s\right)\left(\int_{k\delta}^t e^{-2(s-t)}dB_s\right)^\top\right] \\ &= \frac{4c_\kappa}{L}\left(\int_{k\delta}^t e^{-4(t-s)}ds\right) \cdot I_{d\times d} \\ &= \frac{c_\kappa}{L}(1 - e^{-4(t-k\delta)}) \cdot I_{d\times d}.\end{aligned}$$

The Brownian motion term for x_t is given by

$$\begin{aligned}\sqrt{\frac{4c_\kappa}{L}}\int_{k\delta}^t\left(\int_{k\delta}^r e^{-2(r-s)}dB_s\right)dr &= \sqrt{\frac{4c_\kappa}{L}}\int_{k\delta}^t e^{2s}\left(\int_s^t e^{-2r}dr\right)dB_s \\ &= \sqrt{\frac{c_\kappa}{L}}\int_{k\delta}^t\left(1 - e^{-2(t-s)}\right)dB_s.\end{aligned}$$

Here the second equality follows by Fubini's theorem. The conditional covariance for x_t now follows as

$$\begin{aligned}\mathbb{E}\left[\left(x_t - \mathbb{E}\left[x_t\right]\right)\left(x_t - \mathbb{E}\left[x_t\right]\right)^\top\right] &= \frac{c_\kappa}{L}\mathbb{E}\left[\left(\int_{k\delta}^t\left(1 - e^{-2(t-s)}\right)dB_s\right)\left(\int_{k\delta}^t\left(1 - e^{-2(t-s)}\right)dB_s\right)^\top\right] \\ &= \frac{c_\kappa}{L}\left[\int_{k\delta}^t\left(1 - e^{-2(t-s)}\right)^2 ds\right] \cdot I_{d\times d} \\ &= \frac{c_\kappa}{L}\left[t - k\delta - \frac{1}{4}e^{-4(t-k\delta)} - \frac{3}{4} + e^{-2(t-k\delta)}\right] \cdot I_{d\times d}.\end{aligned}$$

Finally we compute the cross-covariance between x_t and u_t,

$$\begin{aligned}
\mathbb{E}\left[(x_t - \mathbb{E}[x_t])(u_t - \mathbb{E}[u_t])^\top\right] &= \frac{2c_\kappa}{L}\mathbb{E}\left[\left(\int_{k\delta}^{t}\left(1 - e^{-2(t-s)}\right)dB_s\right)\left(\int_{k\delta}^{t} e^{-2(t-s)}dB_s\right)^\top\right] \\
&= \frac{2c_\kappa}{L}\left[\int_{k\delta}^{t}(1 - e^{-2(t-s)})(e^{-2(t-s)})ds\right] \cdot I_{d\times d} \\
&= \frac{c_\kappa}{2L}\left[1 + e^{-4(t-k\delta)} - 2e^{-2(t-k\delta)}\right] \cdot I_{d\times d}.
\end{aligned}$$

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 D

Proofs for Chapter 5

D.1 Proofs for Convergence under Gaussian Noise (Theorem 10)

D.1.1 Proof Overview

The main proof of Theorem 10 is contained in Appendix D.1.4.

Here, we outline the steps of our proof:

1. In Appendix D.1.2, we construct a coupling between (5.3) and (5.2) over a single step (i.e. for $t \in [k\delta, (k+1)\delta]$, for some k and δ).

2. Appendix D.1.3, we prove Lemma 65, which shows that under the coupling constructed in Step 1, a Lyapunov function $f(x_T - y_T)$ contracts exponentially with rate λ, plus a discretization error term. The function f is defined in Appendix D.5, and sandwiches $\|x_T - y_T\|_2$. In Corollary 66, we apply the results of Lemma 65 recursively over multiple steps to give a bound on $f(x_{k\delta} - y_{k\delta})$ for all k, and for sufficiently small δ.

3. Finally, in Appendix D.1.4, we prove Theorem 10 by applying the results of Corollary 66, together with the fact that $f(z)$ upper bounds $\|z\|_2$ up to a constant factor.

D.1.2 A coupling construction

In this subsection, we will study the evolution of (5.3) and (5.2) over a small time interval. Specifically, we will study

$$dx_t = -\nabla U(x_t)dt + M(x_t)dB_t \tag{D.1}$$
$$dy_t = -\nabla U(y_0)dt + M(y_0)dB_t \tag{D.2}$$

One can verify that (D.1) is equivalent to (5.3), and (D.2) is equivalent to a single step of (5.2) (i.e. over an interval $t \leq \delta$).

We first give the explicit coupling between (D.1) and (D.2): (A similar coupling in the continuous-time setting is first seen in [39] in their proof of contraction of (5.3).)

Given arbirary (x_0, y_0), define (x_t, y_t) using the following coupled SDE:

$$\begin{aligned} x_t =& x_0 + \int_0^t -\nabla U(x_s)ds + \int_0^t c_m dV_s + \int_0^t N(x_s)dW_s \\ y_t =& y_0 + \int_0^t -\nabla U(y_0)dt + \int_0^t c_m\big(I - 2\gamma_s\gamma_s^T\big)dV_s + \int_0^t N(y_0)dW_s, \end{aligned} \tag{D.3}$$

where dV_t and dW_t are two independent standard Brownian motion, and

$$\gamma_t := \frac{x_t - y_t}{\|x_t - y\|_2} \cdot \mathbb{1}\left\{\|x_t - y_t\|_2 \in [2\varepsilon, \mathcal{R}_q)\right\}. \tag{D.4}$$

By Lemma 70, we show that (D.1) has the same distribution as x_t in (D.3), and (D.2) has the same distribution as y_t in (D.3). Thus, for any t, the process (x_t, y_t) defined by (D.3) is a valid coupling for (D.1) and (D.2).

D.1.3 One step contraction

Lemma 65 *Let f be as defined in Lemma 82 with parameters ε satisfying $\varepsilon \le \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$. Let x_t and y_t be as defined in* (D.3). *If we assume that $\mathbb{E}\left[\|y_0\|_2^2\right] \le 8(R^2 + \beta^2/m)$ and $T \le \min\left\{\frac{\varepsilon^2}{\beta^2}, \frac{\varepsilon}{6L\sqrt{R^2+\beta^2/m}}\right\}$, then*

$$\mathbb{E}\left[f(x_T - y_T)\right] \le e^{-\lambda T}\mathbb{E}\left[f(x_0 - y_0)\right] + 3T(L + L_N^2)\varepsilon.$$

Remark 14 *For ease of reference: m, L, L_R, R are from Assumption A, c_m, β are from Assumption B, $\alpha_q, \mathcal{R}_q, L_N, \lambda$ are defined in* (5.7).

Proof of Lemma 65

For notational convenience, for the rest of this proof, let us define $z_t := x_t - y_t$ and $\nabla_t := \nabla U(x_t) - \nabla U(y_t)$, $\Delta_t := \nabla U(y_0) - \nabla U(y_t)$ $N_t := N(x_t) - N(y_t)$.

It follows from (D.3) that

$$dz_t = -\nabla_t dt + \Delta_t dt + 2c_m\gamma_t\gamma_t^T dV_t + (N_t + N(y_t) - N(y_0))dW_t. \tag{D.5}$$

Using Ito's Lemma, the dynamics of $f(z_t)$ is given by

$$\begin{aligned}
&df(z_t)\\
=&\langle \nabla f(z_t), dz_t\rangle + 2c_m^2 \mathrm{tr}\big(\nabla^2 f(z_t)\big(\gamma_t\gamma_t^T\big)\big)dt + \frac{1}{2}\mathrm{tr}\big(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\big)dt\\
=&\underbrace{-\langle \nabla f(z_t), \nabla_t\rangle}_{①}\, dt + \underbrace{\langle \nabla f(z_t), \Delta_t\rangle}_{②}\, dt + \underbrace{\big\langle \nabla f(z_t), 2c_m\gamma_t\gamma_t^T dV_t + (N_t + N(y_t) - N(y_0))dW_t\big\rangle}_{③}\\
&+ \underbrace{2c_m^2 \mathrm{tr}\big(\nabla^2 f(z_t)\big(\gamma_t\gamma_t^T\big)\big)}_{④}\, dt + \underbrace{\frac{1}{2}\mathrm{tr}\big(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\big)}_{⑤}\, dt. \qquad \text{(D.6)}
\end{aligned}$$

③ goes to 0 when we take expectation, so we will focus on ①, ②, ④, ⑤. We will consider 3 cases

Case 1: $\|z_t\|_2 \le 2\varepsilon$

From item 1(c) of Lemma 82, $\|\nabla f(z)\|_2 \le 1$. Using Assumption A.1, $\|\nabla_t\| \le L\|z_t\|_2$, so that

$$① \le \|\nabla_t\|_2 \le L\|z_t\|_2 \le 2L\varepsilon.$$

Also by Cauchy Schwarz,

$$② = \langle \nabla f(z_t), \Delta_t\rangle \le \|\Delta_t\|_2 \le L\|y_t - y_0\|_2$$

Since $\gamma_t = 0$ in this case by definition in (D.4), ④ $= 0$.

Using Lemma 82.2.c. $\|\nabla^2 f(z_t)\|_2 \le \frac{2}{\varepsilon}$, so that

$$\begin{aligned}
⑤ \le& \frac{1}{\varepsilon}\Big(\mathrm{tr}\big(N_t^2 + N(y_t) - N(y_0)\big)^2\Big)\\
\le& \frac{2}{\varepsilon}\big(\mathrm{tr}\big(N_t^2\big) + \mathrm{tr}\big((N(y_t) - N(y_0))^2\big)\big)\\
\le& \frac{2L_N^2}{\varepsilon}\big(\|z_t\|_2^2 + \|y_t - y_0\|_2^2\big)\\
\le& 4L_N^2\varepsilon + \frac{2L_N^2}{\varepsilon}\|y_t - y_0\|_2^2,
\end{aligned}$$

where the second inequality is by Young's inequality, the third inequality is by item 2 of Lemma 80, the fourth inequality is by our assumption that $\|z_t\|_2 \le 2\varepsilon$.

Summing these,

$$① + ② + ④ + ⑤ \le 4\big(L + L_N^2\big)\varepsilon + L\|y_t - y_0\|_2 + \frac{2L_N^2}{\varepsilon}\|y_t - y_0\|_2^2.$$

Case 2: $\|z_t\|_2 \in (2\varepsilon, \mathcal{R}_q)$

In this case, $\gamma_t = \frac{z_t}{\|z_t\|_2}$. Let q be as defined in (D.20) and g be as defined in Lemma 84. By

items 1(b) and 2(b) of Lemma 82 and items 1(b) and 2(b) of Lemma 84,

$$\begin{aligned}\nabla f(z_t) =& q'(g(z_t))\nabla g(z_t)\\ =& q'(g(z_t))\frac{z_t}{\|z_t\|_2}\\ \nabla^2 f(z_t) =& q''(g(z_t))\nabla g(z_t)\nabla g(z_t)^T + q'(g(z_t))\nabla^2 g(z_t)\\ =& q''(g(z_t))\frac{z_t z_t^T}{\|z_t\|_2^2} + q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right).\end{aligned}$$

Once again, by Assumption A.3,

$$① \le q'(g(z_t))\|\nabla_t\|_2 \le q'(g(z_t)) \cdot L_R \cdot \|z_t\|_2 \le L \cdot q'(g(z_t))g(z_t) + 2L\varepsilon,$$

where the last inequality uses Lemma 84.4. We can also verify that

$$② \le L\|y_t - y_0\|_2.$$

Using the expression for $\nabla^2 f(z_t)$,

$$④ = 2c_m^2 \mathrm{tr}\big(\nabla^2 f(z_t)\gamma_t\gamma_t^T\big) = 2c_m^2 \cdot q''(g(z_t)).$$

Finally,

$$\begin{aligned}⑤ =& \frac{1}{2}\mathrm{tr}\big(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\big)\\ =& \frac{1}{2}\mathrm{tr}\left(\left(q''(g(z_t))\frac{z_t z_t^T}{\|z_t\|_2^2} + q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)\right)(N_t + N(y_t) - N(y_0))^2\right)\\ \le& \frac{1}{2}\mathrm{tr}\left(\left(q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)\right)(N_t + N(y_t) - N(y_0))^2\right)\\ \le& \frac{q'(g(z_t))}{\|z_t\|_2} \cdot \big(\mathrm{tr}\big(N_t^2\big) + \mathrm{tr}\big((N(y_t) - N(y_0))^2\big)\big)\\ \le& q'(g(z_t)) \cdot L_N^2\|z_t\|_2 + \frac{L_N^2\|y_t - y_0\|_2^2}{2\varepsilon}\\ \le& q'(g(z_t)) \cdot L_N^2 g(z_t) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\varepsilon} + 2L_N^2\varepsilon.\end{aligned}$$

The above uses multiples times the fact that $0 \le q' \le 1$ and $q'' \le 0$ (proven in items 3 and 4 of Lemma 85). The second inequality is by Young's inequality, the third inequality is by item 2 of Lemma 80, the fourth inequality uses item 4 of Lemma 84.

Summing these,

$$\begin{aligned}
①+②+④+⑤ \leq & \big(L_R + L_N^2\big)q'(g(z_t))g(z_t) + 2c_m^2 q''(g(z_t)) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\varepsilon} + 2\big(L + L_N^2\big)\varepsilon \\
\leq & -\frac{2c_m^2 \exp\left(-\frac{7\alpha_q \mathcal{R}_q{}^2}{3}\right)}{32\mathcal{R}_q{}^2} q(g(z_t)) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\varepsilon} + 2\big(L + L_N^2\big)\varepsilon \\
\leq & -\lambda q(g(z_t)) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\varepsilon} + 2(L + L_N^2)\varepsilon \\
= & -\lambda f(z_t) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\varepsilon} + 2(L + L_N^2)\varepsilon + L\|y_t - y_0\|_2,
\end{aligned}$$

where the last inequality follows from Lemma 85.1. and the definition of λ in (5.7).

Case 3: $\|z_t\|_2 \geq \mathcal{R}_q$

In this case, $\gamma_t = 0$. Similar to case 2,

$$\nabla f(z_t) = q'(g(z_t))\frac{z_t}{\|z_t\|_2}.$$

Thus by Assumption A.3,

$$\begin{aligned}
① = & \left\langle q'(g(z_t))\frac{z_t}{\|z_t\|_2}, -\nabla_t \right\rangle \\
\leq & -mq'(g(z_t))\|z_t\|_2,
\end{aligned}$$

where the inequality is by Assumption A.3.

For identical reasons as in Case 1, $② \leq L_R\|y_t - y_0\|_2$, and $④ = 0$. Finally,

$$\begin{aligned}
⑤ = & \frac{1}{2}\mathrm{tr}\big(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\big) \\
= & \frac{1}{2}\mathrm{tr}\left(\left(q''(g(z_t))\frac{z_t z_t^T}{\|z_t\|_2^2} + q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)\right)(N_t + N(y_t) - N(y_0))^2\right) \\
\leq & \frac{1}{2}\mathrm{tr}\left(\left(q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)\right)(N_t + N(y_t) - N(y_0))^2\right) \\
\leq & \frac{q'(g(z_t))}{\|z_t\|_2} \cdot \big(\mathrm{tr}\big(N_t^2\big) + \mathrm{tr}\big((N(y_t) - N(y_0))^2\big)\big),
\end{aligned}$$

where the first inequality is because $q'' \leq 0$ from item 4 of Lemma 85, the second inequality is by Young's inequality. (These steps are identical to Case 2). Continuing from above, and using item 2 and 3 of Lemma 80,

$$\begin{aligned}
⑤ \leq & q'(g(z_t)) \cdot \left(\frac{8\beta^2 L_N}{c_m} + \frac{L_N^2\|y_t - y_0\|_2^2}{\varepsilon}\right) \\
\leq & q'(g(z_t)) \cdot \left(\frac{m}{2}\|z_t\|_2\right) + q'(g(z_t)) \cdot \left(\frac{L_N^2\|y_t - y_0\|_2^2}{\varepsilon}\right),
\end{aligned}$$

where the second inequality is by our definition of $\mathcal{R}_q$ in the lemma statement, which ensures that $\frac{8\beta^2 L_N}{c_m} \leq \frac{m}{2}\mathcal{R}_q \leq \frac{m}{2}\|z_t\|_2$.

Thus

$$\begin{aligned}
&①+②+④+⑤\\
\leq& -mq'(g(z_t))\|z_t\|_2 + L_R\|y_t-y_0\|_2 + \frac{m}{2}q'(g(z_t))\|z_t\|_2 + q'(g(z_t))\cdot\left(\frac{L_N^2\|y_t-y_0\|_2^2}{\varepsilon}\right)\\
\leq& -\frac{m}{2}q'(g(z_t))\|z_t\|_2 + \frac{L_N^2}{\varepsilon}\|y_t-y_0\|_2^2 + L\|y_t-y_0\|_2\\
\leq& -\lambda f(z_t) + \frac{L_N^2}{\varepsilon}\|y_t-y_0\|_2^2 + L\|y_t-y_0\|_2,
\end{aligned}$$

where the second inequality uses $q' \leq 1$ from item 3 of Lemma 85, the third inequality uses our definition of λ in (5.7).

Combining the three cases, (D.6) can be upper bounded with probability 1:

$$\begin{aligned}
df(z_t) \leq& -\lambda f(z_t) + \frac{L_N^2}{\varepsilon}\|y_t-y_0\|_2^2 + L\|y_t-y_0\|_2\\
&+\left\langle \nabla f(z_t), 2c_m\gamma_t\gamma_t^T dV_t + (N_t + N(y_t) - N(y_0))dW_t\right\rangle.
\end{aligned}$$

To simplify notation, let us define $G_t \in \mathbb{R}^{1\times 2d}$ as $G_t := \left[\nabla f(z_t)^T 2c_m\gamma_t\gamma_t^T, \nabla f(z_t)^T(N_t + N(y_t) - N(y_0))\right]$, and let A_t be a $2d$-dimensional Brownian motion from concatenating $A_t = \begin{bmatrix} V_t \\ W_t \end{bmatrix}$. Thus

$$df(z_t) \leq -\lambda f(z_t)dt + \left(\frac{L_N^2}{\varepsilon}\|y_t-y_0\|_2^2 + L\|y_t-y_0\|_2\right) + G_t dA_t.$$

We will study the Lyapunov function

$$\mathcal{L}_t := f(z_t) - \int_0^t e^{-\lambda(t-s)}\left(\frac{L_N^2}{\varepsilon}\|y_s-y_0\|_2^2 + L\|y_s-y_0\|_2\right)ds - \int_0^t e^{-\lambda(t-s)}G_s dA_s.$$

By taking derivatives, we see that

$$\begin{aligned}
d\mathcal{L}_t \leq& -\lambda f(z_t)dt + \left(\frac{L_N^2}{\varepsilon}\|y_t-y_0\|_2^2 + L\|y_t-y_0\|_2\right)dt + G_t dA_t\\
&+\lambda\left(\int_0^t e^{-\lambda(t-s)}\left(\frac{L_N^2}{\varepsilon}\|y_s-y_0\|_2^2 + L\|y_s-y_0\|_2\right)ds\right)dt\\
&-\left(\frac{L_N^2}{\varepsilon}\|y_t-y_0\|_2^2 + L\|y_t-y_0\|_2\right)dt\\
&+\lambda\left(\int_0^t e^{-\lambda(t-s)}G_s dA_s\right)dt - G_t dA_t\\
=& -\lambda\mathcal{L}_t dt.
\end{aligned}$$

We can then apply Gronwall's Lemma to $\mathcal{L}_t$, so that

$$\mathcal{L}_T \leq e^{-\lambda T}\mathcal{L}_0,$$

which is equivalent to

$$f(z_T) - \int_0^T e^{-\lambda(T-s)}\left(\frac{L_N^2}{\varepsilon}\|y_s - y_0\|_2^2 + L\|y_s - y_0\|_2\right)ds - \int_0^T e^{-\lambda(t-s)}G_s dA_s \leq e^{-\lambda T}f(z_0).$$

Observe that G_s is measurable wrt the natural filtration generated by A_s, so that $\int_0^T e^{-\lambda(T-s)}G_s dA_s$ is a martingale. Thus taking expectations,

$$\mathbb{E}\left[f(z_T)\right] \leq e^{-\lambda T}\mathbb{E}\left[f(z_0)\right] + \int_0^T \frac{L_N^2}{\varepsilon}\mathbb{E}\left[\|y_s - y_0\|_2^2\right] + L\mathbb{E}\left[\|y_s - y_0\|_2\right]ds.$$

By Lemma 75, $\mathbb{E}\left[\|y_t - y_0\|_2^2\right] \leq t^2L^2\mathbb{E}\left[\|y_0\|_2^2\right] + t\beta^2$, so that

$$\int_0^T \frac{L_N^2}{\varepsilon}\mathbb{E}\left[\|y_s - y_0\|_2^2\right]ds \leq \frac{T^3L_N^2L^2}{\varepsilon}\mathbb{E}\left[\|y_0\|_2^2\right] + \frac{T^2L_N^2}{\varepsilon}\beta^2$$
$$L\mathbb{E}\left[\|y_s - y_0\|_2\right] \leq T^2L^2\sqrt{\mathbb{E}\left[\|y_0\|_2^2\right]} + T^{3/2}L\beta.$$

Furthermore, using our assumption in the lemma statement that $T \leq \min\left\{\frac{\varepsilon^2}{\beta^2}, \frac{\varepsilon}{6L\sqrt{R^2+\beta^2/m}}\right\}$ and $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$, we can verify that

$$\int_0^T \frac{L_N^2}{\varepsilon}\mathbb{E}\left[\|y_s - y_0\|_2^2\right]ds \leq \frac{1}{4}TL_N^2\varepsilon + TL_N^2\varepsilon$$
$$L\mathbb{E}\left[\|y_s - y_0\|_2\right] \leq \frac{1}{2}TL\varepsilon + TL\varepsilon.$$

Combining the above gives

$$\mathbb{E}\left[f(z_T)\right] \leq e^{-\lambda T}\mathbb{E}\left[f(z_0)\right] + 3T\big(L + L_N^2\big)\varepsilon.$$

□

Corollary 66 *Let f be as defined in Lemma 82 with parameter ε satisfying $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q\mathcal{R}_q{}^2+1}$.*

Let $\delta \leq \min\left\{\frac{\varepsilon^2}{\beta^2}, \frac{\varepsilon}{8L\sqrt{R^2+\beta^2/m}}\right\}$, and let $\bar{x}_t$ and $\bar{y}_t$ have dynamics as defined in (5.3) *and* (5.2) *respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{y}_0\|_2^2\right] \leq R^2 + \beta^2/m$. Then there exists a coupling between $\bar{x}_t$ and $\bar{y}_t$ such that*

$$\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{y}_{i\delta})\right] \leq e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{6}{\lambda}\big(L + L_N^2\big)\varepsilon.$$

Proof of Corollary 66
From Lemma 71 and 72, our initial conditions imply that for all t, $\mathbb{E}\left[\|\bar{x}_t\|_2^2\right] \leq 6\left(R^2 + \frac{\beta^2}{m}\right)$ and $\mathbb{E}\left[\|\bar{y}_{k\delta}\|_2^2\right] \leq 8\left(R^2 + \frac{\beta^2}{m}\right)$.

Consider an arbitrary k, and for $t \in [k\delta, (k+1)\delta)$, define

$$x_t := \bar{x}_{k\delta+t} \quad \text{and} \quad y_t := \bar{y}_{k\delta+t}.$$

Under this definition, x_t and y_t have dynamics described in (D.1) and (D.2). Thus the coupling in (D.3), which describes a coupling between x_t and y_t, equivalently describes a coupling between $\bar{x}_t$ and $\bar{y}_t$ over $t \in [k\delta, (k+1)\delta)$.

We now apply Lemma 65. Given our assumed bound on δ and our proven bounds on $\mathbb{E}\left[\|\bar{x}_t\|_2^2\right]$ and $\mathbb{E}\left[\|\bar{y}_t\|_2^2\right]$,

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{x}_{(k+1)\delta} - \bar{y}_{(k+1)\delta})\right] \\
=&\mathbb{E}\left[f(x_\delta - y_\delta)\right] \\
\leq& e^{-\lambda\delta}\mathbb{E}\left[f(x_0 - y_0)\right] + 6\delta(L + L_N^2)\varepsilon \\
=& e^{-\lambda\delta}\mathbb{E}\left[f(\bar{x}_{k\delta} - \bar{y}_{k\delta})\right] + 6\delta(L + L_N^2)\varepsilon.
\end{aligned}$$

Applying the above recursively gives, for any i

$$\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{y}_{i\delta})\right] \leq e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\varepsilon.$$

□

D.1.4 Proof of Theorem 10

For ease of reference, we re-state Theorem 10 below as Theorem 13 below. We make a minor notational change: using the letters $\bar{x}_t$ and $\bar{y}_t$ in Theorem 13, instead of the letters x_t and y_t in Theorem 10. This is to avoid some notation conflicts in the proof.

Theorem 13 (Equivalent to Theorem 10) *Let $\bar{x}_t$ and $\bar{y}_t$ have dynamics as defined in (5.3) and (5.2) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{y}_0\|_2^2\right] \leq R^2 + \beta^2/m$. Let $\hat{\varepsilon}$ be a target accuracy satisfying $\hat{\varepsilon} \leq \left(\frac{16\left(L+L_N^2\right)}{\lambda}\right) \cdot \exp\left(7\alpha_q\mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q{\mathcal{R}_q}^2+1}$. Let δ be a step size satisfying*

$$\delta \leq \min\left\{ \begin{array}{l} \frac{\lambda^2\hat{\varepsilon}^2}{512\beta^2\left(L^2+L_N^4\right)\exp\left(\frac{14\alpha_q{\mathcal{R}_q}^2}{3}\right)} \\ \frac{2\lambda\hat{\varepsilon}}{\left(L^2+L_N^4\right)\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\sqrt{R^2+\beta^2/m}} \end{array} \right. .$$

If we assume that $\bar{x}_0 = \bar{y}_0$, then there exists a coupling between $\bar{x}_t$ and $\bar{y}_t$ such that for any k,

$$\mathbb{E}\left[\|\bar{x}_{k\delta} - \bar{y}_{k\delta}\|_2\right] \leq \hat{\varepsilon}.$$

Alternatively, if we assume $k \geq \frac{3\alpha_q {\mathcal{R}_q}^2}{\delta} \log \frac{R^2+\beta^2/m}{\hat{\varepsilon}}$, then

$$W_1(p^*, p^y_{k\delta}) \leq 2\hat{\varepsilon},$$

where $p^y_t := \mathsf{Law}(\bar{y}_t)$.

Proof of Theorem 13

Let $\varepsilon := \frac{\lambda}{16(L+L_N^2)} \exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\hat{\varepsilon}$. Let f be defined as in Lemma 82 with the parameter ε.

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{y}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{y}_{i\delta})\right] + 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\varepsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{6}{\lambda}\big(L + L_N^2\big)\varepsilon\right) + 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\varepsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{16(L + L_N^2)}{\lambda}\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\cdot\varepsilon \qquad \text{(D.7)} \\
=& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \hat{\varepsilon},
\end{aligned}$$

where the first inequality is by item 4 of Lemma 82, the second inequality is by Corollary 66 (notice that δ satisfies the requirement on T in Theorem 10, for the given ε). The third inequality uses the fact that $1 \leq L/m \leq \frac{\left(L+L_N^2\right)}{\lambda}$.

The first claim follows from substituting $\bar{x}_0 = \bar{y}_0$ into (D.7), so that the first term is 0, and using the definition of ε, so that the second term is 0.

For the second claim, let $\bar{x}_0 \sim p^*$, the invariant distribution of (5.3). From Lemma 71, we know that $\bar{x}_0$ satisfies the required initial conditions in this Lemma. Continuing from (D.7),

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{y}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\mathbb{E}\left[\|\bar{x}_0\|_2^2 + \|\bar{y}_0\|_2^2\right] + \frac{6}{\lambda}\big(L + L_N^2\big)\varepsilon\right) + \varepsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\big(2e^{-\lambda i\delta}\big(R^2 + \beta^2/m\big)\big) + \frac{16}{\lambda}\exp\left(2\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\big(L + L_N^2\big)\varepsilon \\
=& 4\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\big(e^{-\lambda i\delta}\big(R^2 + \beta^2/m\big)\big) + \hat{\varepsilon}.
\end{aligned}$$

By our assumption that $i \geq \frac{1}{\delta}\cdot 3\alpha_q {\mathcal{R}_q}^2 \log\frac{R^2+\beta^2/m}{\hat{\varepsilon}}$, the first term is also bounded by $\hat{\varepsilon}$, and this proves our second claim. $\square$

D.1.5 Simulating the SDE

One can verify that the SDE in (5.2) can be simulated (at discrete time intervals) as follows:

$$y_{(k+1)\delta} = y_{k\delta} - \delta \nabla U(y_{k\delta}) + \sqrt{\delta} M(y_{k\delta}) \theta_k,$$

where $\theta_k \sim \mathcal{N}(0, I)$. This however requires access to $M(y_{k,\delta})$, which may be difficult to compute.

If for any y, one is able to draw samples from some distribution p_y such that

1. $\mathbb{E}_{\xi \sim p_y}[\xi] = 0$
2. $\mathbb{E}_{\xi \sim p_y}\left[\xi \xi^T\right] = M(y)$
3. $\|\xi\|_2 \leq \beta$ almost surely, for some β,

then one might sample a noise that is δ close to $M(y_{k\delta})\theta_k$ through Theorem 15.

Specifically, if one draws n samples $\xi_1 ... \xi_n \overset{iid}{\sim} p_y$, and let $S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i$, Theorem 15 guarantees that $W_2(S_n, M(y)\theta) \leq \left(6\sqrt{d}\beta\sqrt{\log n}\right)/\sqrt{n}$. We remark that the proof of Theorem 10 can be modified to accommodate for this sampling error. The number of samples needed to achieve ε accuracy will be on the order of $n \approxeq O(\delta\varepsilon)^{-2} = O(\varepsilon^{-6})$.

D.2 Proofs for Convergence under Non-Gaussian Noise (Theorem 11)

D.2.1 Proof Overview

The main proof of Theorem 11 is contained in Appendix D.2.4.

Here, we outline the steps of our proof:

1. In Appendix D.2.2, we construct a coupling between (5.3) and (5.1) over an epoch which consists of an interval $[k\delta, (k+n)\delta)$ for some k. The coupling in (D.2.2) consists of four processes (x_t, y_t, v_t, w_t), where y_t and v_t are auxiliary processes used in defining the coupling. Notably, the process (x_t, y_t) has the same distribution over the epoch as (D.3).

2. In Appendix D.2.3, we prove Lemma 67 and Lemma 68, which, combined with Lemma 65 from Appendix D.1.3, show that under the coupling constructed in Step 1, a Lyapunov function $f(x_T - w_T)$ contracts exponentially with rate λ, plus a discretization error term. In Corollary 69, we apply the results of Lemma 65, Lemma 67 and Lemma 68 recursively over multiple steps to give a bound on $f(x_{k\delta} - w_{k\delta})$ for all k, and for sufficiently small δ.

3. Finally, in Appendix D.2.4, we prove Theorem 11 by applying the results of Corollary 69, together with the fact that $f(z)$ upper bounds $\|z\|_2$ up to a constant.

D.2.2 Constructing a Coupling

In this sectopn, we construct a coupling between (5.1) and (5.3), given arbitrary initialization (x_0, w_0). We will consider a finite time $T = n\delta$, which we will refer to as an *epoch*.

1. Let V_t and W_t be two independent Brownian motion.
2. Using V_t and W_t, define
$$x_t = x_0 + \int_0^t -\nabla U(x_s)ds + \int_0^t c_m dV_s + \int_0^t N(w_0)dW_s. \tag{D.8}$$
3. Using the same V_t and W_t in (D.8), we will define y_t as
$$y_t = w_0 + \int_0^t -\nabla U(w_0)ds + \int_0^t c_m\big(I - 2\gamma_s\gamma_s^t\big)dV_s + \int_0^T N(x_s)dW_s, \tag{D.9}$$
where $\gamma_t := \frac{x_t - y_t}{\|x_t - y_t\|_2} \cdot \mathbb{1}\left\{\|x_t - y_t\|_2 \in [2\varepsilon, \mathcal{R}_q)\right\}$. The coupling (x_t, y_t) defined in (D.8) and (D.9) is identical to the coupling in (D.3) (with $y_0 = w_0$).
4. We now define a process $v_{k\delta}$ for $k = 0...n$:
$$v_{k\delta} = w_0 + \sum_{i=0}^{k-1} -\delta\nabla U(w_0) + \sqrt{\delta}\sum_{i=0}^{k-1}\xi(w_0, \eta_i), \tag{D.10}$$
where marginally, the variables $(\eta_0...\eta_{n-1})$ are drawn $i.i.d$ from the same distribution as in (5.1).

   Notice that $y_T - w_0 - T\nabla U(w_0) = \int_0^T c_m dB_t + \int_0^T N(w_0)dW_t$, so that $\mathsf{Law}(y_T - w_0 - T\nabla U(w_0)) = \mathcal{N}(0, TM(w_0)^2)$. Notice also that $v_T - w_0 - T\nabla U(w_0) = \sqrt{\delta}\sum_{i=0}^{n-1}\xi(w_0, \eta_i)$. By Corollary 88, $W_2(y_T - w_0 - T\nabla U(w_0), v_T - w_0 - T\nabla U(w_0)) = 6\sqrt{d\delta}\beta\sqrt{\log n}$. Let the joint distribution between (D.10) and (D.9) be the one induced by the optimal coupling between $y_T - w_0 - T\nabla U(w_0)$ and $v_T - w_0 - T\nabla U(w_0)$, so that
$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\|y_T - v_T\|_2^2\right]} \\
=&\sqrt{\mathbb{E}\left[\|y_T - T\nabla U(w_0) - v_T + T\nabla U(w_0)\|_2^2\right]} \\
=&W_2(y_T - w_0 - T\nabla U(w_0), v_T - w_0 - T\nabla U(w_0)) \\
\leq&6\sqrt{d\delta}\beta\sqrt{\log n},
\end{aligned} \tag{D.11}$$
where the last inequality is by Corollary 88.

5. Given the sequence $(\eta_0 ... \eta_{n-1})$ from (D.10), we can define

$$w_{k\delta} = w_0 + \sum_{i=0}^{k-1} -\delta \nabla U(w_{i\delta}) + \sqrt{\delta} \sum_{i=0}^{k-1} \xi(w_{i\delta}, \eta_i), \tag{D.12}$$

specifically, $(w_0 ... w_{n\delta})$ in (D.12) and $(v_0 ... v_{n\delta})$ in (D.10) are coupled through the shared $(\eta_0 ... \eta_{n-1})$ variables.

For convenience, we will let $v_t := v_{i\delta}$ and $w_t := w_{i\delta}$, where i is the unique integer satisfying $t \in [i\delta, (i+1)\delta)$.

We can verify that, marginally, the process x_t in (D.8) has the same distribution as (5.3), using the proof as Lemma 70. It is also straightforward to verify that $w_{k\delta}$, as defined in (D.12), has the same marginal distribution as (5.1), due to the definition of η_i in (D.10).

D.2.3 One Epoch Contraction

In Lemma 67, we prove a discretization error bound between $f(x_T - y_T)$ and $f(x_T - v_T)$, for the coupling defined in (D.8), (D.9) and (D.10).

In Lemma 68, we prove a discretization error bound between $f(x_T - v_T)$ and $f(x_T - w_T)$, for the coupling defined in (D.8), (D.10) and (D.12).

Lemma 67 *Let f be as defined in Lemma 82 with parameter ε satisfying $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$. Let x_t, y_t and v_t be as defined in* (D.8), (D.9), (D.10). *Let n be any integer and δ be any step size, and let $T := n\delta$.*

If $\mathbb{E}\left[\|x_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$, $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$ and $T \leq \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}\right\}$ and

$$\delta \leq \min\left\{ \frac{T\varepsilon^2 L}{36 d\beta^2 \log\left(\frac{36 d\beta^2}{\varepsilon^2 L}\right)}, \frac{T\varepsilon^4 L^2}{2^{14} d\beta^4 \log\left(\frac{2^{14} d\beta^4}{\varepsilon^4 L^2}\right)} \right\},$$

Then

$$\mathbb{E}\left[f(x_T - v_T)\right] - \mathbb{E}\left[f(x_T - y_T)\right] \leq 4TL\varepsilon,$$

Proof

By Taylor's Theorem,

$$\begin{aligned}
&\mathbb{E}\left[f(x_T - v_T)\right] \\
=&\mathbb{E}\left[f(x_T - y_T) + \langle \nabla f(x_T - y_T), y_T - v_T\rangle\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - y_T + s(y_T - v_T)), (y_T - v_T)(y_T - v_T)^T \right\rangle ds dt\right] \\
=&\mathbb{E}\left[f(x_T - y_T) + \underbrace{\langle \nabla f(x_0 - y_0), y_T - v_T\rangle}_{①} + \underbrace{\langle \nabla f(x_T - y_T) - \nabla f(x_0 - y_0), y_T - v_T\rangle}_{②}\right] \\
&\quad + \mathbb{E}\left[\underbrace{\int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - y_T + s(y_T - v_T)), (y_T - v_T)(y_T - v_T)^T \right\rangle ds dt}_{③}\right].
\end{aligned}$$

We will bound each of the terms above separately.

$$\begin{aligned}
&\mathbb{E}\left[①\right] \\
=&\mathbb{E}\left[\langle \nabla f(x_0 - y_0), y_T - v_T\rangle\right] \\
=&\mathbb{E}\left[\langle \nabla f(x_0 - y_0), n\delta \nabla U(y_0) - n\delta \nabla U(v_0)\rangle\right] \\
&\quad + \mathbb{E}\left[\left\langle \nabla f(x_0 - y_0), \int_0^T -\nabla U(w_0) dt + \int_0^T c_m dV_t + \int_0^T N(w_0) dW_t + \sum_{i=0}^{n-1} \sqrt{\delta}\xi(v_0, \eta_i)\right\rangle\right] \\
=&\mathbb{E}\left[\langle \nabla f(x_0 - y_0), n\delta \nabla U(y_0) - n\delta \nabla U(v_0)\rangle\right] \\
=&0,
\end{aligned}$$

where the third equality is because $\int_0^T dB_t$, $\int_0^T dW_t$ and $\sum_{k=1}^T \xi(v_0, \eta_i)$ have zero mean conditioned on the information at time 0, and the fourth equality is because $y_0 = v_0$ by definition in (D.9) and (D.10).

$$\begin{aligned}
&\mathbb{E}\left[②\right] \\
=&\mathbb{E}\left[\langle \nabla f(x_T - y_T) - \nabla f(x_0 - y_0), y_T - v_T\rangle\right] \\
\leq&\sqrt{\mathbb{E}\left[\|\nabla f(x_T - y_T) - \nabla f(x_0 - y_0)\|_2^2\right]} \sqrt{\mathbb{E}\left[\|y_T - v_T\|_2^2\right]} \\
\leq&\frac{2}{\varepsilon}\sqrt{2\mathbb{E}\left[\|x_T - x_0\|_2^2 + \|y_T - y_0\|_2^2\right]} \sqrt{\mathbb{E}\left[\|y_T - v_T\|_2^2\right]} \\
\leq&\frac{2}{\varepsilon}\sqrt{(32T\beta^2 + 4T\beta^2)} \cdot \left(6\sqrt{d\delta}\beta \log n\right) \\
\leq&\frac{128}{\varepsilon}\sqrt{T}\beta^2 \cdot \left(\sqrt{d\delta} \log n\right),
\end{aligned}$$

where the second inequality is by $\|\nabla^2 f\|_2 \leq \frac{2}{\varepsilon}$ from item 2(c) of Lemma 82 and Young's inequality. The third inequality is by Lemma 74 and Lemma 75 and (D.11).

Finally, we can bound

$$\begin{aligned}
&\mathbb{E}\left[③\right] \\
\leq &\int_0^1 \int_0^s \mathbb{E}\left[\left\|\nabla^2 f(x_T - y_T + s(y_T - v_T))\right\|_2 \|y_T - v_T\|_2^2\right] dsdt \\
\leq &\frac{2}{\varepsilon}\mathbb{E}\left[\|y_T - v_T\|_2^2\right] \\
\leq &\frac{72d\delta\beta^2 \log^2 n}{\varepsilon},
\end{aligned}$$

where the second inequality is by $\|\nabla^2 f\|_2 \leq \frac{2}{\varepsilon}$ from item 2(c) of Lemma 82, the third inequality is by (D.11).

Summing these 3 terms,

$$\begin{aligned}
&\mathbb{E}\left[f(x_T - v_T) - f(x_T - y_T)\right] \\
\leq &\frac{128}{\varepsilon}\sqrt{T}\beta^2 \cdot \left(\sqrt{d\delta}\sqrt{\log n}\right) + \frac{36d\delta\beta^2 \log n}{\varepsilon} \\
= &\frac{128}{\varepsilon}\sqrt{T}\beta^2 \cdot \left(\sqrt{d\delta}\sqrt{\log \frac{T}{\delta}}\right) + \frac{36d\delta\beta^2 \log \frac{T}{\delta}}{\varepsilon}.
\end{aligned}$$

Let us bound the first term. We apply Lemma 89 (with $x = \frac{T}{\delta}$ and $c = \frac{\varepsilon^4}{2^{14}d\beta^4}$), which shows that

$$\frac{T}{\delta} \geq \frac{2^{14}d\beta^4}{\varepsilon^4}\log\left(\frac{2^{14}d\beta^4}{\varepsilon^4 L^2}\right) \quad \Rightarrow \quad \frac{T}{\delta}\frac{1}{\log\frac{T}{\delta}} \geq \frac{2^{14}d\beta^4}{\varepsilon^4 L^2} \quad \Leftrightarrow \quad \frac{128}{\varepsilon}\sqrt{T}\beta^2 \cdot \left(\sqrt{d\delta}\log\frac{T}{\delta}\right) \leq TL\varepsilon.$$

For the second term, we can again apply Lemma 89 ($x = \frac{T}{\delta}$ and $c = \frac{\varepsilon^2 L}{36d\beta^2}$), which shows that

$$\frac{T}{\delta} \geq \frac{36d\beta^2}{\varepsilon^2 L}\log\left(\frac{36d\beta^2}{\varepsilon^2 L}\right) \quad \Rightarrow \quad \frac{T}{\delta}\frac{1}{\log\frac{T}{\delta}} \geq \frac{36d\beta^2}{\varepsilon^2 L} \quad \Rightarrow \quad \frac{36d\delta\beta^2\log\frac{T}{\delta}}{\varepsilon} \leq TL\varepsilon.$$

The above imply that

$$\mathbb{E}\left[f(x_T - v_T) - f(x_T - y_T)\right] \leq 2TL\varepsilon.$$

□

Lemma 68 *Let f be as defined in Lemma 82 with parameter ε satisfying $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$. Let x_t, v_t and w_t be as defined in* (D.8), (D.10), (D.12). *Let n be an integer and δ be a step size, and let $T := n\delta$.*

If we assume that $\mathbb{E}\left[\|x_0\|_2^2\right]$, $\mathbb{E}\left[\|v_0\|_2^2\right]$, *and* $\mathbb{E}\left[\|w_0\|_2^2\right]$ *are each upper bounded by* $8(R^2+\beta^2/m)$ *and that* $T \leq \min\left\{\frac{1}{16L}, \frac{\varepsilon}{32\sqrt{L}\beta}, \frac{\varepsilon^2}{128\beta^2}, \frac{\varepsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$, *then*

$$\mathbb{E}\left[f(x_T - w_T)\right] - \mathbb{E}\left[f(x_T - v_T)\right] \leq 4T(L + L_N^2)\varepsilon.$$

Remark 15 *For sufficiently small* ε, *our assumption on* T *boils down to* $T = o(\varepsilon^4)$.

Proof
First, we can verify using Taylor's theorem that for any x, y,

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \int_0^s \left\langle \nabla^2 f(x + s(y-x)), (y-x)(y-x)^T \right\rangle ds dt, \tag{D.13}$$

$$\nabla f(y) = \nabla f(x) + \left\langle \nabla^2 f(x), y - x\right\rangle + \int_0^1 \int_0^s \left\langle \nabla^3 f(x + s(y-x)), (y-x)(y-x)^T \right\rangle ds dt. \tag{D.14}$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[f(x_T - w_T)\right] \\
=&\mathbb{E}\left[f(x_T - v_T) + \langle \nabla f(x_T - v_T), v_T - w_T\rangle\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - v_T + s(v_T - w_T)), (v_T - w_T)(v_T - w_T)^T \right\rangle ds dt\right] \\
=&\mathbb{E}\left[f(x_T - v_T) + \underbrace{\langle \nabla f(x_0 - v_0), v_T - w_T\rangle}_{①} + \underbrace{\langle \nabla f(x_T - v_T) - \nabla f(x_0 - v_0), v_T - w_T\rangle}_{②}\right] \\
&\quad + \mathbb{E}\left[\underbrace{\int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - v_T + s(v_T - w_T)), (v_T - w_T)(v_T - w_T)^T \right\rangle ds dt}_{③}\right].
\end{aligned}$$

Recall from (D.10) and (D.12) that

$$\begin{aligned}
v_{n\delta} =& w_0 + \sum_{i=0}^{n-1} \delta \nabla U(w_0) + \sqrt{\delta} \sum_{i=0}^{n-1} \xi(w_0, \eta_i), \\
w_{n\delta} =& w_0 + \sum_{i=0}^{n-1} \delta \nabla U(w_{i\delta}) + \sqrt{\delta} \sum_{i=0}^{n-1} \xi(w_{i\delta}, \eta_i).
\end{aligned}$$

Note that conditioned on the randomness up to time 0, $\mathbb{E}\left[\sum_{i=0}^{n-1}\xi(w_0,\eta_i)\right]=\mathbb{E}\left[\sum_{i=0}^{n-1}\xi(w_{i\delta},\eta_i)\right]=0$, so that

$$\begin{aligned}
&\mathbb{E}\left[①\right]\\
=&\mathbb{E}\left[\langle\nabla f(x_0-v_0),v_T-w_T\rangle\right]\\
=&\delta\mathbb{E}\left[\left\langle\nabla f(x_0-v_0),\sum_{i=0}^{n-1}\nabla U(w_0)-\nabla U(w_{i\delta})\right\rangle\right]\\
&\quad+\sqrt{\delta}\mathbb{E}\left[\left\langle\nabla f(x_0-v_0),\sum_{i=0}^{n-1}\xi(w_0,\eta_i)-\sum_{i=0}^{n-1}\xi(w_{i\delta},\eta_i)\right\rangle\right]\\
=&\delta\mathbb{E}\left[\left\langle\nabla f(x_0-v_0),\sum_{i=0}^{n-1}\nabla U(w_0)-\nabla U(w_{i\delta})\right\rangle\right]\\
\le&\delta\sum_{i=0}^{n-1}L\mathbb{E}\left[\|w_0-w_{i\delta}\|_2\right]\\
\le&TL\sqrt{32T\beta^2}\le 8T^{3/2}L\beta,
\end{aligned}$$

where the third equality is becayse $\xi(\cdot,\eta_i)$ has 0 mean conditioned on the randomness at time 0, and the second inequality is by Lemma 77.

Next,

$$\begin{aligned}
&\mathbb{E}\left[②\right]\\
=&\mathbb{E}\left[\langle\nabla f(x_T-v_T)-\nabla f(x_0-v_0),v_T-w_T\rangle\right]\\
\le&\mathbb{E}\left[\|\nabla f(x_T-v_T)-\nabla f(x_0-v_0)\|_2\|v_T-w_T\|\right]\\
\le&\frac{4}{\varepsilon}\sqrt{\mathbb{E}\left[\|x_T-x_0\|_2^2+\|v_T-v_0\|_2^2\right]}\cdot\sqrt{\mathbb{E}\left[\|v_T-w_T\|_2^2\right]}\\
\le&\frac{4}{\varepsilon}\sqrt{16T\beta^2+2T\beta^2}\cdot\sqrt{32\big(T^2L^2+TL_\xi^2\big)T\beta^2}\\
\le&\frac{128}{\varepsilon}T\beta^2\Big(\sqrt{T}L_\xi+TL\Big),
\end{aligned}$$

where the second inequality is because $\|\nabla^2 f\|_2\le\frac{2}{\varepsilon}$ from item 2(c) of Lemma 82 and by Young's inequality. The third inequality is by Lemma 74, Lemma 76 and Lemma 78.

Finally,

$$\begin{aligned}
&\mathbb{E}\left[\textcircled{3}\right]\\
=&\mathbb{E}\left[\int_0^1\int_0^s \left\langle \nabla^2 f(x_T - v_T + s(v_T - w_T)), (v_T - w_T)(v_T - w_T)^T\right\rangle ds dt\right]\\
\leq&\int_0^1\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(x_T - v_T + s(v_T - w_T))\right\|_2 \|v_T - w_T\|_2^2\right] ds\\
\leq&\frac{1}{\varepsilon}\mathbb{E}\left[\|v_T - w_T\|_2^2\right]\\
\leq&\frac{32}{\varepsilon}\big(T^2L^2 + TL_\xi^2\big)T\beta^2,
\end{aligned}$$

where the second inequality is because $\|\nabla^2 f\|_2 \leq \frac{2}{\varepsilon}$ from item 2(c) of Lemma 82 and by Young's inequality. The third inequality is by Lemma 78.

Summing the above,

$$\begin{aligned}
&\mathbb{E}\left[f(x_T - w_T) - f(x_T - v_T)\right]\\
\leq&8T^{3/2}L\beta + \frac{128}{\varepsilon}T\beta^2\Big(\sqrt{T}L_\xi + TL\Big) + \frac{32}{\varepsilon}\big(T^2L^2 + TL_\xi^2\big)T\beta^2\\
\leq&T^{3/2}\varepsilon,
\end{aligned}$$

where the last inequality is by our assumption on T, specifically,

$$\begin{aligned}
&T \leq \frac{\varepsilon^2}{128\beta^2} \Rightarrow T^{3/2}L\beta \leq TL\varepsilon\\
&T \leq \frac{\varepsilon^2}{128\beta^2} \Rightarrow \frac{128}{\varepsilon}T^2L\beta^2 \leq TL\varepsilon\\
&T \leq \frac{\varepsilon}{32\sqrt{L}\beta} \Rightarrow \frac{32}{\varepsilon}(T^3L^2\beta^2) \leq TL\varepsilon\\
&T \leq \frac{\varepsilon^4 L_N^2}{2^{14}\beta^2 c_m^2} \Rightarrow \frac{128}{\varepsilon}T^{3/2}\beta^2 L_\xi \leq TL_N^2\varepsilon\\
&T \leq \frac{\varepsilon^2}{128\beta^2} \Rightarrow T \leq \frac{\varepsilon^2}{128c_m^2} \Rightarrow \frac{32}{\varepsilon}T^2L_\xi^2\beta^2 \leq TL_N^2\varepsilon,
\end{aligned}$$

where the last line uses the fact that $\beta \geq c_m^2$.

□

Corollary 69 *Let f be as defined in Lemma 82 with parameter ε satisfying $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2+1}$. Let $T = \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}, \frac{\varepsilon}{32\sqrt{L}\beta}, \frac{\varepsilon^2}{128\beta^2}, \frac{\varepsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$ and let $\delta \leq \min\left\{\frac{T\varepsilon^2 L}{36d\beta^2\log\left(\frac{36d\beta^2}{\varepsilon^2 L}\right)}, \frac{T\varepsilon^4 L^2}{2^{14}d\beta^4\log\left(\frac{2^{14}d\beta^4}{\varepsilon^4 L^2}\right)}\right\}$. Assume additionally that $n = T/\delta$ is an integer.*

Let $\bar{x}_t$ and $\bar{w}_t$ have dynamics as defined in (5.3) *and* (5.2) *respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{w}_0\|_2^2\right] \leq R^2 + \beta^2/m$. Then there exists a coupling between $\bar{x}_t$ and $\bar{w}_t$ such that*

$$\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{w}_{i\delta})\right] \leq e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{6}{\lambda}\big(L + L_N^2\big)\varepsilon$$

Proof
From Lemma 71 and 73, our initial conditions imply that for all t, $\mathbb{E}\left[\|\bar{x}_t\|_2^2\right] \leq 6\Big(R^2 + \frac{\beta^2}{m}\Big)$ and $\mathbb{E}\left[\|\bar{w}_{k\delta}\|_2^2\right] \leq 8\Big(R^2 + \frac{\beta^2}{m}\Big)$.

Consider an arbitrary k, and for $t \in [0, T)$, define

$$x_t := \bar{x}_{kT+t} \quad \text{and} \quad w_t := \bar{w}_{kT+t} \tag{D.15}$$

Notice that as described above, x_t and w_t have dynamics described in (5.3) and (5.1). Let x_t, w_t have joint distribution as described in (D.8) and (D.12), and let (y_t, v_t) be the processes defined in (D.9) and (D.10). Notice that the joint distribution between x_t and w_t equivalently describes a coupling between $\bar{x}_t$ and $\bar{w}_t$ over $t \in [kT, (k+1)T)$.

First, notice that the processes (D.8) and (D.9) have the same distribution as (D.3). We can thus apply Lemma 65:

$$\mathbb{E}\left[f(x_T - y_T)\right] \leq e^{-\lambda T}\mathbb{E}\left[f(x_0 - y_0)\right] + 6T(L + L_N^2)\varepsilon$$

By Lemma 67,

$$\mathbb{E}\left[f(x_T - v_T)\right] - \mathbb{E}\left[f(x_T - y_T)\right] \leq 4TL\varepsilon$$

By Lemma 68,

$$\mathbb{E}\left[f(x_T - w_T)\right] - \mathbb{E}\left[f(x_T - v_T)\right] \leq 4T(L + L_N^2)\varepsilon$$

Summing the above three equations,

$$\mathbb{E}\left[f(x_T - w_T)\right] \leq e^{-\lambda\delta}\mathbb{E}\left[f(x_0 - w_0)\right] + 14T(L + L_N^2)$$

Where we use the fact that $y_0 = w_0$ by construction in (D.9).

Recalling (D.15), this is equivalent to

$$\mathbb{E}\left[f(\bar{x}_{(k+1)T} - \bar{w}_{(k+1)T})\right] \leq e^{-\lambda\delta}\mathbb{E}\left[f(\bar{x}_{kT} - \bar{w}_{kT})\right] + 14T(L + L_N^2)$$

Applying the above recursively gives, for any i

$$\mathbb{E}\left[f(\bar{x}_{iT} - \bar{w}_{iT})\right] \leq e^{-\lambda iT}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{14}{\lambda}\big(L + L_N^2\big)\varepsilon$$

□

D.2.4 Proof of Theorem 11

For ease of reference, we re-state Theorem 11 below as Theorem 14 below. We make a minor notational change: using the letters $\bar{x}_t$ and $\bar{y}_t$ in Theorem 14, instead of the letters x_t and y_t in Theorem 11. This is to avoid some notation conflicts in the proof.

Theorem 14 (Equivalent to Theorem 11) *Let $\bar{x}_t$ and w_t have dynamics as defined in (5.3) and (5.1) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2+\beta^2/m$ and $\mathbb{E}\left[\|\bar{w}_0\|_2^2\right] \leq R^2+\beta^2/m$. Let $\hat{\varepsilon}$ be a target accuracy satisfying $\hat{\varepsilon} \leq \left(\frac{16\left(L+L_N^2\right)}{\lambda}\right) \cdot \exp\left(7\alpha_q\mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q{\mathcal{R}_q}^2+1}$. Let $\varepsilon := \frac{\lambda}{16(L+L_N^2)} \exp\left(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\hat{\varepsilon}$. Let*

$$T := \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}, \frac{\varepsilon}{32\sqrt{L}\beta}, \frac{\varepsilon^2}{128\beta^2}, \frac{\varepsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$$

and let δ be a step size satisfying

$$\delta \leq \min\left\{\frac{T\varepsilon^2 L}{36d\beta^2\log\left(\frac{36d\beta^2}{\varepsilon^2 L}\right)}, \frac{T\varepsilon^4 L^2}{2^{14}d\beta^4\log\left(\frac{2^{14}d\beta^4}{\varepsilon^4 L^2}\right)}\right\}.$$

If we assume that $\bar{x}_0 = \bar{w}_0$, then there exists a coupling between $\bar{x}_t$ and $\bar{w}_t$ such that for any k,

$$\mathbb{E}\left[\|\bar{x}_{k\delta} - \bar{w}_{k\delta}\|_2\right] \leq \hat{\varepsilon}.$$

Alternatively, if we assume that $k \geq \frac{3\alpha_q{\mathcal{R}_q}^2}{\delta} \cdot \log\frac{R^2+\beta^2/m}{\hat{\varepsilon}}$, then

$$W_1(p^*, p_{k\delta}^w) \leq 2\hat{\varepsilon},$$

where $p_t^w := \mathsf{Law}(\bar{w}_t)$.

Proof of Theorem 14
Let f be defined as in Lemma 82 with parameter ε.

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{w}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{w}_{i\delta})\right] + 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\varepsilon \\
\leq& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{6}{\lambda}\left(L+L_N^2\right)\varepsilon\right) + 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\varepsilon \\
\leq& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{16(L+L_N^2)}{\lambda}\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\cdot\varepsilon \qquad \text{(D.16)} \\
=& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \hat{\varepsilon},
\end{aligned}$$

where the first inequality is by item 4 of Lemma 82, the second inequality is by Corollary 69 (notice that δ satisfies the requirement on T in Theorem 10, for the given ε). The third inequality uses the fact that $1 \leq L/m \leq \frac{\left(L+L_N^2\right)}{\lambda}$.

The first claim follows from substituting $\bar{x}_0 = \bar{w}_0$ into (D.16), so that the first term is 0, and using the definition of ε, so that the second term is 0.

For the second claim, let $\bar{x}_0 \sim p^*$, the invariant distribution of (5.3). From Lemma 71, we know that $\bar{x}_0$ satisfies the required initial conditions in this Lemma. Continuing from (D.16),

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{w}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\mathbb{E}\left[\|\bar{x}_0\|_2^2 + \|\bar{w}_0\|_2^2\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\varepsilon\right) + \varepsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\left(R^2 + \beta^2/m\right)\right) + \frac{16}{\lambda}\exp\left(2\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(L + L_N^2\right)\varepsilon \\
=& 4\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\left(R^2 + \beta^2/m\right)\right) + \hat{\varepsilon}.
\end{aligned}$$

By our assumption that $i \geq \frac{1}{\delta} \cdot 3\alpha_q {\mathcal{R}_q}^2 \log\frac{R^2+\beta^2/m}{\hat{\varepsilon}}$, the first term is also bounded by $\hat{\varepsilon}$, and this proves our second claim. □

D.3 Coupling Properties

Lemma 70 *Consider the coupled (x_t, y_t) in* (D.3)*. Let p_t denote the distribution of x_t, and q_t denote the distribution of y_t. Let p'_t and q'_t denote the distributions of* (D.1) *and* (D.2)*.*

If $p_0 = p'_0$ and $q_0 = q'_0$, then $p_t = p'_t$ and $q_t = q'_t$ for all t.

Proof

Consider the coupling in (D.3), reproduced below for ease of reference:

$$\begin{aligned}
x_t =& x_0 + \int_0^t -\nabla U(x_s)ds + \int_0^t c_m dV_s + \int_0^t N(x_s)dW_s \\
y_t =& y_0 + \int_0^t -\nabla U(y_0)dt + \int_0^t c_m\left(I - 2\gamma_s\gamma_s^T\right)dV_s + \int_0^t N(y_0)dW_s.
\end{aligned}$$

Let us define the stochastic process $A_t := \int_0^t M(x_s)^{-1}c_m dV_s + \int_0^t M(x_s)^{-1}N(x_s)dW_s$. We can verify using Levy's characterization that A_t is a standard Brownian motion: first, since V_t and W_t are Brownian motions, and $N(x)$ is differentiable with bounded derivatives, we know that A_t has continuous sample paths. We now verify that $A_t^i A_t^j - \mathbb{1}\{i=j\}t$ is a martingale.

Notice that $dA_t = c_m dV_t + M(x_s)^{-1}N(x_s)dW_s$. Then

$$\begin{aligned}
dA_t^i A_t^j =& dA_t^T\left(e_i e_j^T\right)A_t \\
=& A_t\left(e_i e_j^T\right)\left(c_m dV_t + M(x_s)^{-1}N(x_s)dW_s\right)^T + \left(c_m dV_t + M(x_s)^{-1}N(x_s)dW_s\right)\left(e_j e_i^T\right)a_t^T \\
&+ \frac{1}{2}\mathrm{tr}\left(\left(e_i e_j^T + e_j e_i^T\right)\left(c_m^2 M(x_s)^{-2} + M(x_s)^{-1}N(x_s)^2 M(x_s)^{-1}\right)\right)dt,
\end{aligned}$$

where the second inequality is by Ito's Lemma applied to $f(A_t) = A_t^T e_j e_j^T A_t$. Taking expectations,

$$\begin{aligned}
d\mathbb{E}\left[A_t^i A_t^j\right] =& \mathbb{E}\left[\frac{1}{2}\mathrm{tr}\Big(\big(e_i e_j^T + e_j e_i^T\big)\Big(c_m^2 M(x_s)^{-2} + M(x_s)^{-1}N(x_s)N(x_s)^T\big(M(x_s)^{-1}\big)^T\Big)\Big)\right] dt \\
=& \mathbb{E}\left[\frac{1}{2}\mathrm{tr}\Big(\big(e_i e_j^T + e_j e_i^T\big)\big(M(x_s)^{-1}\big(c_m^2 I + N(x_s)^2\big)M(x_s)^{-1}\big)\Big)\right] dt \\
=& \mathbb{E}\left[\frac{1}{2}\mathrm{tr}\Big(\big(e_i e_j^T + e_j e_i^T\big)\big(M(x_s)^{-1}\big(M(x_s)^2\big)M(x_s)^{-1}\big)\Big)\right] dt \\
=& \mathbb{E}\left[\frac{1}{2}\mathrm{tr}\Big(\big(e_i e_j^T + e_j e_i^T\big)\Big)\right] dt \\
=& \mathbb{1}\left\{i = j\right\} dt.
\end{aligned}$$

This verifies that $A_t^i A_t^j - \mathbb{1}\left\{i = j\right\} t$ is a martingale, and hence by Levy's characterization, A_t is a standard Brownian motion. In turn, we verify that by definition of A_t,

$$\begin{aligned}
x_t =& x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t c_m dV_s + \int_0^t N(x_s) dW_s \\
=& x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t M(x_s)\big(M(x_s)^{-1}(c_m dV_s + N(x_s) dW_s)\big) \\
=& x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t M(x_s) dA_s
\end{aligned}$$

Since we showed that A_t is a standard Brownian motion, we verify that x_t as defined in (D.3) has the same distribution as (5.3).

On the other hand, we can verify that $A'_t := \int_0^T (I - 2\gamma_s\gamma_s^T) V_s$ is a standard Brownian motion by the reflection principle. Thus

$$\int_0^t c_m\big(I - 2\gamma_s\gamma_s^T\big) dV_s + \int_0^t N(y_0) dW_s \sim \mathcal{N}(0, \big(c_m^2 I + N(y_0)^2\big)) = \mathcal{N}(0, M(y_0)^2)$$

where the equality is by definition of N in (5.6).

It follows immediately that y_t in (D.3) has the same distribution as y_t in (5.2).

□

D.3.1 Energy Bounds

Lemma 71 *Consider x_t as defined in* (5.3)*. If x_0 satisfies $\mathbb{E}\left[\|x_0\|_2^2\right] \leq R^2 + \frac{\beta^2}{m}$, then Then for all t,*

$$\mathbb{E}\left[\|x_t\|_2^2\right] \leq 6\left(R^2 + \frac{\beta^2}{m}\right).$$

We can also show that

$$\mathbb{E}_{p^*}\left[\|x\|_2^2\right] \leq 4\bigg(R^2 + \frac{\beta^2}{m}\bigg).$$

Proof
We consider the potential function $a(x) = (\|x\|_2 - R)_+^2$ We verify that

$$\begin{aligned}\nabla a(x) =& (\|x\|_2 - R)_+ \frac{x}{\|x\|_2} \\ \nabla^2 a(x) =& \mathbb{1}\left\{\|x\|_2 \geq R\right\}\frac{xx^T}{\|x\|_2^2} + \frac{(\|x\|_2 - R)_+}{\|x\|_2}\bigg(I - \frac{xx^T}{\|x\|_2^2}\bigg).\end{aligned}$$

Observe that

1. $\|\nabla^2 a(x)\|_2 \leq 2\mathbb{1}\left\{\|x\|_2 \geq R\right\} \leq 2$
2. $\langle \nabla a(x), -\nabla U(x)\rangle \leq -ma(x)$. This can be verified by considering 2 cases. If $\|x\|_2 \leq R$, then $\nabla a(x) = 0$ and $a(x) = 0$. If $\|x\|_2 \geq R$, then by Assumption A,
$$\langle \nabla a(x), -\nabla U(x)\rangle \leq -m(\|x\|_2 - R)_+\|w\|_2 \leq -m(\|x\|_2 - R)_+^2 = -m \cdot a(x).$$
3. $a(x) \geq \frac{1}{2}\|x\|_2^2 - 2R^2$. One can first verify that $a(x) \geq (\|x\|_2 - R)^2 - R^2$. Next, by Young's inequality, $(\|x\|_2 - R)^2 = \|x\|_2^2 + R^2 - 2\|x\|_2 R \geq \|x\|_2^2 + R^2 - \frac{1}{2}\|x\|_2^2 - 2R^2 = \frac{1}{2}\|x\|_2^2 - R^2$.

Therefore,

$$\begin{aligned}&\frac{d}{dt}\mathbb{E}\left[a(x_t)\right] = \mathbb{E}\left[\langle \nabla a(x_t), -\nabla U(x_t)dt\rangle\right] + \frac{1}{2}\mathbb{E}\left[\mathrm{tr}\big(M(x_t)^2\nabla^2 a(x)\big)\right] \leq -m\mathbb{E}\left[a(x_t)\right] + \beta^2 \\ \Rightarrow\quad &\frac{d}{dt}\bigg(\mathbb{E}\left[a(x_t)\right] - \frac{\beta^2}{m}\bigg) \leq -m\bigg(\mathbb{E}\left[a(x_t)\right] - \frac{\beta^2}{m}\bigg) \\ \Rightarrow\quad &\frac{d}{dt}\bigg(\mathbb{E}\left[a(x_t)\right] - R^2 - \frac{\beta^2}{m}\bigg) \leq -m\bigg(\mathbb{E}\left[a(x_t)\right] - R^2 - \frac{\beta^2}{m}\bigg).\end{aligned}$$

Thus if $\mathbb{E}\left[\|x_0\|_2^2\right] \leq R^2 + \frac{\beta^2}{m}$, then $\mathbb{E}\left[a(x_0)\right] \leq R^2 - \frac{\beta^2}{m}$, then $\bigg(\mathbb{E}\left[a(x_0)\right] - R^2 - \frac{\beta^2}{m}\bigg) \leq 0$, and $\bigg(\mathbb{E}\left[a(x_t)\right] - R^2 + \frac{\beta^2}{m}\bigg) \leq e^{-mt} \cdot 0 \leq 0$ for all t. This implies that, for all t,

$$\mathbb{E}\left[\|x_t\|_2^2\right] \leq \mathbb{E}\left[2a(x_t) + 4R^2\right] \leq 6\bigg(R^2 + \frac{\beta^2}{m}\bigg).$$

For our second claim that $\mathbb{E}_{p^*}\left[\|x\|_2^2\right] \leq R^2 + \frac{\beta^2}{m}$, we can use the fact that if $x_0 \sim p^*$, then $\mathbb{E}\left[a(x_t)\right]$ does not change as p^* is invariant, so that

$$0 = \frac{d}{dt}\mathbb{E}\left[a(x_t)\right] \leq -m\mathbb{E}\left[a(x_t)\right] + \beta^2.$$

Thus

$$\mathbb{E}\left[a(x_t)\right] \leq \frac{\beta^2}{m}.$$

Again,

$$\mathbb{E}_{p^*}\left[\|x\|_2^2\right] = \mathbb{E}\left[\|x_t\|_2^2\right] \leq 2\mathbb{E}\left[a(x_t)\right] + 4R^2 \leq 4\left(R^2 + \frac{\beta^2}{m}\right).$$

□

Lemma 72 *Let the sequence $y_{k\delta}$ be as defined in* (5.1)*. Assuming that $\delta \leq m/(16L^2)$ and $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 2\left(R^2 + \frac{\beta^2}{m}\right)$ Then for all k,*

$$\mathbb{E}\left[\|y_{k\delta}\|_2^2\right] \leq 8\left(R^2 + \frac{\beta^2}{m}\right).$$

Proof
Let $a(w) := (\|w\|_2 - R)_+^2$. We can verify that

$$\begin{aligned}
\nabla a(w) =& (\|w\|_2 - R)_+ \frac{w}{\|w\|_2} \\
\nabla^2 a(w) =& \mathbb{1}\left\{\|w\|_2 \geq R\right\}\frac{ww^T}{\|w\|_2^2} + (\|w\|_2 - R)_+ \frac{1}{\|w\|_2}\left(I - \frac{ww^T}{\|w\|_2^2}\right)
\end{aligned}$$

Observe that

1. $\|\nabla^2 a(w)\|_2 \leq 2\mathbb{1}\left\{\|w\|_2 \geq R\right\} \leq 2$
2. $\langle \nabla a(w), -\nabla U(w)\rangle \leq -ma(w)$.
3. $a(w) \geq \frac{1}{2}\|w\|_2^2 - 2R^2$.

The proofs are identical to the proof at the start of Lemma 73, so we omit them here.

Using Taylor's Theorem, and taking expectation of $y_{(k+1)\delta}$ conditioned on $y_{k\delta}$,

$$\begin{aligned}
&\mathbb{E}\left[a(y_{(k+1)\delta})\right] \\
=&\mathbb{E}\left[a(y_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(y_{k\delta}), y_{(k+1)\delta} - y_{k\delta}\right\rangle\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \int_0^t \left\langle \nabla^2 a(y_{k\delta} + s(y_{(k+1)\delta} - y_{k\delta}), (y_{(k+1)\delta} - y_{k\delta})(y_{(k+1)\delta} - y_{k\delta})^T\right\rangle dtds\right] \\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(y_{k\delta}), y_{(k+1)\delta} - y_{k\delta}\right\rangle\right] + \mathbb{E}\left[\left\|(y_{(k+1)\delta} - y_{k\delta})\right\|_2^2 ds\right] \\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] + \mathbb{E}\left[\langle \nabla a(y_{k\delta}), -\delta\nabla U(y_{k\delta})\rangle\right] + 2\delta^2\|\nabla U(y_{k\delta})\|_2^2 + 2\delta\mathbb{E}\left[\mathrm{tr}\big(M(y_{k\delta})^2\big)\right] \\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] - m\delta\mathbb{E}\left[a(y_{k\delta})\right] + 2\delta^2\mathbb{E}\left[\|\nabla U(y_{k\delta})\|_2^2\right] + 2\delta\mathbb{E}\left[\mathrm{tr}\big(M(y_{k\delta})^2\big)\right] \\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] - m\delta\mathbb{E}\left[a(y_{k\delta})\right] + 2\delta^2 L^2\mathbb{E}\left[\|y_{k\delta}\|_2^2\right] + 2\delta\beta^2 \\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] - m\delta\mathbb{E}\left[a(y_{k\delta})\right] + 4\delta^2 L^2\mathbb{E}\left[a(y_{k\delta})\right] + 8\delta^2 L^2 R^2 + 2\delta\beta^2 \\
\leq&(1 - m\delta/2)\mathbb{E}\left[a(y_{k\delta})\right] + m\delta R^2 + 2\delta\beta^2,
\end{aligned}$$

where the first inequality uses the upper bound on $\|\nabla^2 a(y)\|_2$ above, the second inequality uses the fact that $y_{(k+1)\delta} \sim \mathcal{N}(y_{k\delta} - \delta\nabla U(y_{k\delta}), \delta M(y_{k\delta})^2)$, the third inequality uses claim 2. at the start of this proof, the fourth inequality uses item 2 of Assumption B. The fifth inequality uses claim 3. above, the sixth inequality uses our assumption that $\delta \leq \frac{m}{16L^2}$.

Taking expectation wrt $y_{k\delta}$,

$$\begin{aligned} & \mathbb{E}\left[a(y_{(k+1)\delta})\right] \leq \mathbb{E}\left[a(y_k)\right] - m\delta\big(\mathbb{E}\left[a(y_{k\delta})\right] - 2R^2 + 2\beta^2/m\big) \\ \Rightarrow \quad & \mathbb{E}\left[a(y_{(k+1)\delta})\right] - (2R^2/2 + 2\beta^2/m) \leq (1 - m\delta)\big(\mathbb{E}\left[a(y_{k\delta})\right] - (2R^2 + 2\beta^2/m\big). \end{aligned}$$

Thus, if $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 2R^2 + 2\beta^2/m$, then $\mathbb{E}\left[a(y_0)\right] - (2R^2 + 2\beta^2/m) \leq 0$, then $\mathbb{E}\left[a(y_{k\delta})\right] - (2R^2 + 2\beta^2/m) \leq 0$ for all k, which implies that

$$\mathbb{E}\left[\|y_{k\delta}\|_2^2\right] \leq 2\mathbb{E}\left[a(y_{k\delta})\right] + 4R^2 \leq 8\big(R^2 + \beta^2/m\big)$$

for all k. □

Lemma 73 *Let the sequence $w_{k\delta}$ be as defined in* (5.1)*. Assuming that $\delta \leq m/(16L^2)$ and $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 2\Big(R^2 + \frac{\beta^2}{m}\Big)$ Then for all k,*

$$\mathbb{E}\left[\|w_{k\delta}\|_2^2\right] \leq 8\bigg(R^2 + \frac{\beta^2}{m}\bigg).$$

Proof
The proof is almost identical to that of Lemma 72. Let $a(w) := (\|w\|_2 - R)_+^2$. We can verify that

$$\begin{aligned} \nabla a(w) =& (\|w\|_2 - R)_+ \frac{w}{\|w\|_2} \\ \nabla^2 a(y) =& \mathbb{1}\left\{\|w\|_2 \geq R\right\}\frac{ww^T}{\|w\|_2^2} + (\|w\|_2 - R)_+ \frac{1}{\|w\|_2}\bigg(I - \frac{ww^T}{\|w\|_2^2}\bigg). \end{aligned}$$

Observe that

1. $\|\nabla^2 a(w)\|_2 \leq 2\mathbb{1}\left\{\|w\|_2 \geq R\right\} \leq 2$
2. $\langle \nabla a(w), -\nabla U(w)\rangle \leq -ma(w)$.
3. $a(w) \geq \frac{1}{2}\|w\|_2^2 - 2R^2$.

The proofs are identical to the proof at the start of Lemma 73, so we omit them here.

Using Taylor's Theorem, and taking expectation of $w_{(k+1)\delta}$ conditioned on $w_{k\delta}$,

$$\begin{aligned}
&\mathbb{E}\left[a(w_{(k+1)\delta})\right] \\
=&\mathbb{E}\left[a(w_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(w_{k\delta}), w_{(k+1)\delta} - w_{k\delta}\right\rangle\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \int_0^t \left\langle \nabla^2 a(w_{k\delta} + s(w_{(k+1)\delta} - w_{k\delta}), (w_{(k+1)\delta} - w_{k\delta})(w_{(k+1)\delta} - w_{k\delta})^T\right\rangle dtds\right] \\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(w_{k\delta}), w_{(k+1)\delta} - w_{k\delta}\right\rangle\right] + \mathbb{E}\left[\left\|(w_{(k+1)\delta} - w_{k\delta})\right\|_2^2 ds\right] \\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] + \mathbb{E}\left[\langle \nabla a(w_{k\delta}), -\delta \nabla U(w_{k\delta})\rangle\right] + 2\delta^2 \|\nabla U(w_{k\delta})\|_2^2 + 2\delta \mathbb{E}\left[\|\xi(w_{k\delta}, \eta_k)\|_2^2\right] \\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] - m\delta \mathbb{E}\left[a(w_{k\delta})\right] + 2\delta^2 \mathbb{E}\left[\|\nabla U(w_{k\delta})\|_2^2\right] + 2\delta \mathbb{E}\left[\|\xi(w_{k\delta}, \eta_k)\|_2^2\right] \\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] - m\delta \mathbb{E}\left[a(w_{k\delta})\right] + 2\delta^2 L^2 \mathbb{E}\left[\|w_{k\delta}\|_2^2\right] + 2\delta\beta^2 \\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] - m\delta \mathbb{E}\left[a(w_{k\delta})\right] + 2\delta^2 L^2 a(w_{k\delta}) + 2\delta^2 L^2 R^2 + 2\delta\beta^2 \\
\leq&(1 - m\delta/2) a(w_{k\delta}) + m\delta R^2 + 2\delta\beta^2,
\end{aligned}$$

where the first inequality uses the upper bound on $\|\nabla^2 a(y)\|_2$ above, the second inequality uses the fact that $w_{(k+1)\delta} = (y_{k\delta} - \delta \nabla U(y_{k\delta}) = \xi(w_{k\delta}, \eta_k))$, and $\mathbb{E}\left[\xi(w_{k\delta}, \eta_k) | w_{k\delta}\right] = 0$, the third inequality uses claim 2. at the start of this proof, the fourth inequality uses item 2 of Assumption B. The fifth inequality uses claim 3. above, and the sixth inequality uses our assumption that $\delta \leq \frac{m}{16L^2}$.

Taking expectation wrt $w_{k\delta}$,

$$\begin{aligned}
&\mathbb{E}\left[a(w_{(k+1)\delta})\right] \leq \mathbb{E}\left[a(w_k)\right] - m\delta\big(\mathbb{E}\left[a(w_{k\delta})\right] - 2R^2 + 2\beta^2/m\big) \\
\Rightarrow \quad &\mathbb{E}\left[a(w_{(k+1)\delta})\right] - (2R^2/2 + 2\beta^2/m) \leq (1 - m\delta)\big(\mathbb{E}\left[a(w_{k\delta})\right] - (2R^2 + 2\beta^2/m)\big).
\end{aligned}$$

Thus, if $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 2R^2 + 2\beta^2/m$, then $\mathbb{E}\left[a(w_0)\right] - (2R^2 + 2\beta^2/m) \leq 0$, then $\mathbb{E}\left[a(w_{k\delta})\right] - (2R^2 + 2\beta^2/m) \leq 0$ for all k, which implies that

$$\mathbb{E}\left[\|w_{k\delta}\|_2^2\right] \leq 2\mathbb{E}\left[a(w_{k\delta})\right] + 4R^2 \leq 8\big(R^2 + \beta^2/m\big)$$

for all k. □

D.3.2 Divergence Bounds

Lemma 74 *Let x_t be as defined in* (D.1) *(or equivalently* (D.3) *or* (D.8)*), initialized at x_0. Then for any $T \leq \frac{1}{16L}$,*

$$\mathbb{E}\left[\|x_T - x_0\|_2^2\right] \leq 8\big(T\beta^2 + T^2 L^2 \mathbb{E}\left[\|x_0\|_2^2\right]\big).$$

If we additionally assume that $\mathbb{E}\left[\|x_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$ and $T \leq \frac{\beta^2}{8L^2(R^2 + \beta^2/m)}$, then

$$\mathbb{E}\left[\|x_T - x_0\|_2^2\right] \leq 16T\beta^2.$$

Proof
By Ito's Lemma,

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[\|x_t\|_2^2\right] \\
=&2\mathbb{E}\left[\langle \nabla U(x_t), x_t - x_0\rangle\right] + \mathbb{E}\left[\mathrm{tr}\big(M(x_t)^2\big)\right] \\
\leq& 2L\mathbb{E}\left[\|x_t\|_2\|x_t - x_0\|_2\right] + \beta^2 \\
\leq& 2L\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + 2L\mathbb{E}\left[\|x_0\|_2\|x_t - x_0\|_2\right] + \beta^2 \\
\leq& 2L\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + L^2T\mathbb{E}\left[\|x_0\|_2^2\right] + \frac{1}{T}\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + \beta^2 \\
\leq& \frac{2}{T}\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + \big(L^2T\mathbb{E}\left[\|x_0\|_2^2\right] + \beta^2\big),
\end{aligned}$$

where the first inequality is by item 1 of Assumption A and item 2 of Assumption B, the second inequality is by triangle inequality, the third inequality is by Young's inequality, the last inequality is by our assumption on T.

Applying Gronwall's inequality for $t \in [0, T]$,

$$\begin{aligned}
&\big(\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + L^2T^2\mathbb{E}\left[\|x_0\|_2^2\right] + T\beta^2\big) \\
\leq& e^2\big(\mathbb{E}\left[\|x_0 - x_0\|\right] + L^2T^2\mathbb{E}\left[\|x_0\|_2^2\right] + T\beta^2\big) \\
\leq& 8L^2T^2\mathbb{E}\left[\|x_0\|_2^2\right] + T\beta^2.
\end{aligned}$$

This concludes our proof. □

Lemma 75 *Let y_t be as defined in* (D.2) *(or equivalently* (D.3) *or* (D.8)*), initialized at y_0. Then for any T,*

$$\mathbb{E}\left[\|y_T - y_0\|_2^2\right] \leq T^2L^2\mathbb{E}\left[\|y_0\|_2^2\right] + T\beta^2$$

If we additionally assume that $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|y_T - y_0\|_2^2\right] \leq 2T\beta^2.$$

Proof
Notice from the definition in (D.2) that $y_T - y_0 \sim \mathcal{N}(-T\nabla U(y_0), TM(y_0)^2)$, the conclusion immediately follows from item 1 of Assumption A and item 2 of Assumption B, and the fact that

$$\mathrm{tr}\big(M(x)^2\big) = \mathrm{tr}\big(\mathbb{E}\left[\xi(x,\eta)\xi(x,\eta)^T\right]\big) = \mathbb{E}\left[\|\xi(x,\eta)\|_2^2\right].$$

□

Lemma 76 *Let v_t be as defined in* (D.10)*, initialized at v_0. Then for any $T = n\delta$,*

$$\mathbb{E}\left[\|v_T - v_0\|_2^2\right] \leq T^2 L^2 \mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2.$$

If we additionally assume that $\mathbb{E}\left[\|v_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|v_T - v_0\|_2^2\right] \leq 2T\beta^2.$$

Proof
From (D.10),

$$v_T - v_0 = -T\nabla U(v_0) + \sqrt{\delta}\sum_{i=0}^{n-1}\xi(v_0, \eta_i).$$

Conditioned on the randomness up to time i, $\mathbb{E}\left[\xi(v_0, \eta_{i+1})\right] = 0$. Thus

$$\begin{aligned}
&\mathbb{E}\left[\|v_T - v_0\|_2^2\right] \\
=&T^2\mathbb{E}\left[\|\nabla U(v_0)\|_2^2\right] + \delta\sum_{i=0}^{n-1}\mathbb{E}\left[\|\xi(v_0, \eta_i)\|_2^2\right] \\
\leq&T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2,
\end{aligned}$$

where the inequality is by item 1 of Assumption A and item 2 of Assumption B. □

Lemma 77 *Let w_t be as defined in* (D.12)*, initialized at w_0. Then for any $T = n\delta$ such that $T \leq \frac{1}{2L}$,*

$$\mathbb{E}\left[\|w_T - w_0\|_2^2\right] \leq 16\big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\big).$$

If we additionally assume that $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|w_T - w_0\|_2^2\right] \leq 32T\beta^2$$

.

Proof

$$\begin{aligned}
&\mathbb{E}\left[\left\|w_{(k+1)\delta} - w_0\right\|_2^2\right] \\
=&\mathbb{E}\left[\left\|w_{k\delta} - \delta\nabla U(w_{k\delta}) + \sqrt{\delta}\xi(w_{k\delta}, \eta_k) - w_0\right\|_2^2\right] \\
=&\mathbb{E}\left[\|w_{k\delta} - \delta\nabla U(w_{k\delta}) - w_0\|_2^2\right] + \delta\mathbb{E}\left[\|\xi(w_{k\delta}, \eta_k)\|_2^2\right] \qquad \text{(D.17)}
\end{aligned}$$

We can bound $\delta\mathbb{E}\left[\|\xi(w_{k\delta},\eta_k)\|_2^2\right] \leq \delta\beta^2$ by item 2 of Assumption B.

$$\begin{aligned}
&\mathbb{E}\left[\|w_{k\delta} - \delta\nabla U(w_{k\delta}) - w_0\|_2^2\right] \\
\leq&\mathbb{E}\left[\left(\|w_{k\delta} - w_0 - \delta(\nabla U(w_{k\delta}) - \nabla U(w_0))\|_2 + \delta\|\nabla U(w_0)\|_2\right)^2\right] \\
\leq&\left(1+\frac{1}{n}\right)\mathbb{E}\left[\|w_{k\delta} - w_0 - \delta(\nabla U(w_{k\delta}) - \nabla U(w_0))\|_2^2\right] \\
&\quad + (1+n)\delta^2\mathbb{E}\left[\|\nabla U(w_0)\|_2^2\right] \\
\leq&\left(1+\frac{1}{n}\right)(1+\delta L)^2\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] \\
\leq&e^{1/n+2\delta L}\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right],
\end{aligned}$$

where the first inequality is by triangle inequality, the second inequality is by Young's inequality, the third inequality is by item 1 of Assumption A.

Inserting the above into (D.17) gives

$$\mathbb{E}\left[\left\|w_{(k+1)\delta} - w_0\right\|_2^2\right] \leq e^{1/n+2\delta L}\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + \delta\beta^2.$$

Applying the above recursively for $k = 1...n$, we see that

$$\begin{aligned}
&\mathbb{E}\left[\|w_{n\delta} - w_0\|_2^2\right] \\
\leq&\sum_{k=0}^{n-1} e^{(n-k)\cdot(1/n+2\delta L)} \cdot \left(2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + \delta\beta^2\right) \\
\leq&16\left(n^2\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + n\delta\beta^2\right) \\
=&16\left(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\right).
\end{aligned}$$

□

D.3.3 Discretization Bounds

Lemma 78 *Let $v_{k\delta}$ and $w_{k\delta}$ be as defined in* (D.10) *and* (D.12)*. Then for any δ, n, such that $T := n\delta \leq \frac{1}{16L}$,*

$$\mathbb{E}\left[\|v_T - w_T\|_2^2\right] \leq 8\left(2T^2L^2\left(T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\right) + TL_\xi^2\left(16\left(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\right)\right)\right)$$

If we additionally assume that $\mathbb{E}\left[\|v_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$, $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 8(R^2 + \beta^2/m)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|v_T - w_T\|_2^2\right] \leq 32\left(T^2L^2 + TL_\xi^2\right)T\beta^2.$$

Proof
Using the fact that conditioned on the randomness up to step k, $\mathbb{E}\left[\xi(v_0, \eta_{k+1}) - \xi(w_{k\delta}, \eta_{k+1})\right] = 0$, we can show that for any $k \leq n$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{(k+1)\delta} - w_{(k+1)\delta}\right\|_2^2\right] \\
=&\mathbb{E}\left[\left\|v_{k\delta} - \delta\nabla U(v_0) - w_{k\delta} + \delta\nabla U(w_{k\delta}) + \sqrt{\delta}\xi(w_0, \eta_k) - \sqrt{\delta}\xi(w_{k\delta}, \eta_k)\right\|_2^2\right] \\
=&\mathbb{E}\left[\|v_{k\delta} - \delta\nabla U(v_0) - w_{k\delta} + \delta\nabla U(w_{k\delta})\|_2^2\right] + \delta\mathbb{E}\left[\|\xi(w_0, \eta_k) - \xi(w_{k\delta}, \eta_k)\|_2^2\right] \qquad \text{(D.18)}
\end{aligned}$$

Using Assumption B, and Lemma 76, we can bound

$$\begin{aligned}
&\delta\mathbb{E}\left[\|\xi(w_0, \eta_k) - \xi(w_{k\delta}, \eta_k)\|_2^2\right] \\
\leq&\delta L_\xi^2\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] \\
\leq&\delta L_\xi^2\big(16\big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\big)\big)
\end{aligned}$$

We can also bound

$$\begin{aligned}
&\mathbb{E}\left[\|v_{k\delta} - \delta\nabla U(v_0) - w_{k\delta} + \delta\nabla U(w_{k\delta})\|_2^2\right] \\
\leq&\left(1 + \frac{1}{n}\right)\mathbb{E}\left[\|v_{k\delta} - \delta\nabla U(v_{k\delta}) - w_{k\delta} + \delta\nabla U(w_{k\delta})\|_2^2\right] + (1+n)\delta^2\mathbb{E}\left[\|\nabla U(v_{k\delta}) - \nabla U(v_0)\|_2^2\right] \\
\leq&\left(1 + \frac{1}{n}\right)(1+\delta L)^2\mathbb{E}\left[\|v_{k\delta} - w_{k\delta}\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|v_{k\delta} - v_0\|_2^2\right] \\
\leq&e^{1/n+2\delta L}E\|v_{k\delta} - w_{k\delta}\|_2^2 + 2n\delta^2L^2\mathbb{E}\left[\|v_{k\delta} - v_0\|_2^2\right] \\
\leq&e^{1/n+2\delta L}E\|v_{k\delta} - w_{k\delta}\|_2^2 + 2n\delta^2L^2\big(T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\big),
\end{aligned}$$

where the first inequality is by Young's inequality and the second inequality is by item 1 of Assumption A, the fourth inequality uses Lemma 76.

Substituting the above two equation blocks into (D.18), and applying recursively for $k = 0...n-1$ gives

$$\begin{aligned}
&\mathbb{E}\left[\|v_T - w_T\|_2^2\right] \\
=&\mathbb{E}\left[\|v_{n\delta} - w_{n\delta}\|_2^2\right] \\
\leq&e^{1+2n\delta L}\big(2n^2\delta^2L^2\big(T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\big) + n\delta L_\xi^2\big(16\big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\big)\big)\big) \\
\leq&8\big(2T^2L^2\big(T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\big) + TL_\xi^2\big(16\big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\big)\big)\big).
\end{aligned}$$

The last inequality is by noting that $T = n\delta \leq \frac{1}{4L}$. □

D.4 Regularity of M and N

Lemma 79

1. $tr\big(M(x)^2\big) \le \beta^2$
2. $tr\big((M(x)^2 - M(y)^2)^2\big) \le 16\beta^2 L_\xi^2 \|x-y\|_2^2$
3. $tr\big((M(x)^2 - M(y)^2)^2\big) \le 32\beta^3 L_\xi \|x-y\|_2$

Proof
In this proof, we will use the fact that $\xi(\cdot,\eta)$ is L_ξ-Lipschitz from Assumption B.

The first property is easy to see:

$$\begin{aligned}
&\mathrm{tr}\big(M(x)^2\big)\\
=&\mathrm{tr}\big(\mathbb{E}_\eta\left[\xi(x,\eta)\xi(x,\eta)^T\right]\big)\\
=&\mathbb{E}_\eta\left[\mathrm{tr}\big(\xi(x,\eta)\xi(x,\eta)^T\big)\right]\\
=&\mathbb{E}_\eta\left[\|\xi(x,\eta)\|_2^2\right]\\
\le&\beta^2.
\end{aligned}$$

We now prove the second and third claims. Consider a fixed x and fixed y, let $u_\eta := \xi(x,\eta)$, $v_\eta := \xi(y,\eta)$. Then

$$\begin{aligned}
&\mathrm{tr}\Big(\big(M(x)^2 - M(y)^2\big)^2\Big)\\
=&\mathrm{tr}\Big(\big(\mathbb{E}_\eta\left[u_\eta u_\eta^T - v_\eta v_\eta^T\right]\big)^2\Big)\\
=&\mathrm{tr}\big(\mathbb{E}_{\eta,\eta'}\left[\big(u_\eta u_\eta^T - v_\eta v_\eta^T\big)\big(u_{\eta'} u_{\eta'}^T - v_{\eta'} v_{\eta'}^T\big)\right]\big)\\
=&\mathbb{E}_{\eta,\eta'}\left[\mathrm{tr}\big(\big(u_\eta u_\eta^T - v_\eta v_\eta^T\big)\big(u_{\eta'} u_{\eta'}^T - v_{\eta'} v_{\eta'}^T\big)\big)\right].
\end{aligned}$$

For any fixed η and η', let's further simplify notation by letting u, u', v, v' denote $u_\eta, u_{\eta'}, v_\eta, v_{\eta'}$. Thus

$$\begin{aligned}
&\mathrm{tr}\big(\big(uu^T - vv^T\big)\big(u'u'^T - v'v'^T\big)\big)\\
=&\mathrm{tr}\big(\big((u-v)v^T + v(u-v)^T + (u-v)(u-v)^T\big)\big((u'-v')v'^T + v'(u'-v')^T + (u'-v')(u'-v')^T\big)\big)\\
=&\mathrm{tr}\big((u-v)v^T(u'-v')v'^T\big) + \mathrm{tr}\big((u-v)v^Tv'(u'-v')^T\big) + \mathrm{tr}\big((u-v)v^T(u'-v')(u'-v')^T\big)\\
&+ \mathrm{tr}\big(v(u-v)^T(u'-v')v'^T\big) + \mathrm{tr}\big(v(u-v)^Tv'(u'-v')^T\big) + \mathrm{tr}\big(v(u-v)^T(u'-v')(u'-v')^T\big)\\
&+ \mathrm{tr}\big((u-v)(u-v)^T(u'-v')v'^T\big) + \mathrm{tr}\big((u-v)(u-v)^Tv'(u'-v')^T\big)\\
&+ \mathrm{tr}\big((u-v)(u-v)^T(u'-v')(u'-v')^T\big)\\
\le&\min\left\{16\beta^2L_\xi^2\|x-y\|_2^2, 32\beta^3L_\xi\|x-y\|_2\right\},
\end{aligned}$$

where the last inequality uses Assumption B.2 and B.3; in particular, $\|v\|_2 \le \beta$ and $\|u-v\|_2 \le \min\{2\beta, L_\xi\|x-y\|_2\}$. This proves 2. and 3. of the lemma statement. □

Lemma 80 *Let $N(x)$ be as defined in* (5.6) *and L_N be as defined in* (5.7)*. Then*

1. $tr\big(N(x)^2\big) \leq \beta^2$
2. $tr\big((N(x) - N(y))^2\big) \leq L_N^2 \|x - y\|_2^2$
3. $tr\big((N(x) - N(y))^2\big) \leq \dfrac{8\beta^2}{c_m} \cdot L_N \|x - y\|_2.$

Proof of Lemma 80
The first inequality holds because $N(x)^2 := M(x)^2 - c_m^2 I$, and then applying Lemma 79.1, and the fact that $\mathrm{tr}(M(x)^2 - c_m^2 I) \leq \mathrm{tr}(M(x)^2)$ by Assumption B.4.

The second inequality is an immediate consequence of Lemma 81, Lemma 79.2, and the fact that $\lambda_{min}(N(x)^2) = \lambda_{min}(M(x)^2 - c_m^2) \geq c_m^2$ by Assumption B.4.

The proof for the third inequality is similar to the second inequality, and follows from Lemma 79 and Lemma 81.

□

Lemma 81 (Simplified version of Lemma 1 from [35]) *Let A, B be positive definite matrices. Then*

$$tr\left(\left(\sqrt{A} - \sqrt{B}\right)^2\right) \leq tr\big((A - B)^2 A^{-1}\big)$$

D.5 Defining f and related inequalities

In this section, we define the Lyapunov function f which is central to the proof of our main results. Here, we give an overview of the various functions defined in this section:

1. $g(z) : \mathbb{R}^d \to \mathbb{R}^+$: A smoothed version of $\|z\|_2$, with bounded derivatives up to third order.
2. $q(r) : \mathbb{R}^+ \to \mathbb{R}^+$: A concave potential function, similar to the one defined in [32], which has bounded derivatives up to third order everywhere except at $r = 0$.
3. $f(z) = q(g(z)) : \mathbb{R}^d \to \mathbb{R}^+$, a concave function which upper and lower bounds $\|z\|_2$ within a constant factor, has bounded derivatives up to third order everywhere.

Lemma 82 (Properties of f) *Let ε satisfy $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$. We define the function*

$$f(z) := q(g(z))$$

Where q is as defined in (D.20) *Appendix D.5.1, and g is as defined in Lemma 84 (with parameter ε). Then*

1. *a)* $\nabla f(z) = q'(g(z)) \cdot \nabla g(z)$

b) For $\|z\|_2 \geq 2\varepsilon$, $\nabla f(z) = q'(g(z))\frac{z}{\|z\|_2}$

c) For all z, $\|\nabla f(z)\|_2 \leq 1$.

2. a) $\nabla^2 f(z) = q''(g(z))\nabla g(z)\nabla g(z)^T + q'(g(z))\nabla^2 g(z)$

b) For $r \geq 2\varepsilon$, $\nabla^2 f(z) = q''(g(z))\frac{zz^T}{\|z\|_2^2} + q'(g(z))\frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right)$

c) For all z, $\|\nabla^2 f(z)\|_2 \leq \frac{2}{\varepsilon}$

d) For all z, v, $v^T\nabla^2 f(z)v \leq \frac{q'(g(z))}{\|z\|_2}$

3. For any z, $\|\nabla^3 f(z)\|_2 \leq \frac{9}{\varepsilon^2}$

4. For any z, $f(z) \in \left[\frac{1}{2}\exp\left(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)g(\|z\|_2), g(\|z\|_2)\right] \in \left[\frac{1}{2}\exp\left(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)(\|z\|_2 - 2\varepsilon), \|z\|_2\right]$

Proof of Lemma 82

1. a) chain rule
   b) Use definition of $\nabla g(z)$ from Lemma 84.
   c) By definition, $\nabla f(z) = q'(g(z))\nabla g(z)$. From Lemma 85, $|q'(g(z))| \leq 1$. By definition, $\nabla g(z) = h'(\|z\|_2)\frac{z}{\|z\|_2}$. Our conclusion follows from $h' \leq 1$ using item 2 of Lemma 83.

2. a) chain rule
   b) by item 2 b) of Lemma 84
   c) by item 1 c) and item 2 d) of Lemma 84, and item 3 and item 4 of Lemma 85, and our assumption that $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q + {\mathcal{R}_q}^2 + 1}$.
   d) by item 4 of Lemma 85), and items 2 c) and 2 d) of Lemma 84, and our expression for $\nabla^2 f(z)$ established in item 2 a).

3. It can be verified that

$$\begin{aligned}\nabla^3 f(z) =& q'''(g(z)) \cdot \nabla g(z)^{\otimes 3} + q''(g(z))\nabla g(z) \bigotimes \nabla^2 g(z) + q''(g(z))\nabla^2 g(z) \bigotimes \nabla g(z) \\ &+ q''(g(z))\nabla g(z) \bigotimes \nabla^2 g(z) + q'(g(z))\nabla^3 g(z)\end{aligned}$$

Thus

$$\begin{aligned}\left\|\nabla^3 f(z)\right\|_2 \leq& |q'''(g(z))|\|\nabla g(z)\|_2^3 + 3q''(g(z))\|\nabla g(z)\|_2\left\|\nabla^2 g(z)\right\|_2 + q'(g(z))\left\|\nabla^3 g(z)\right\| \\ \leq& 5\left(\alpha_q + \frac{1}{{\mathcal{R}_q}^2}\right)\left(\alpha_q{\mathcal{R}_q}^2 + 1\right) + 3\left(\frac{5\alpha_q\mathcal{R}_q}{4} + \frac{4}{\mathcal{R}_q}\right)\cdot\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2} \\ \leq& \frac{9}{\varepsilon^2}\end{aligned}$$

Where the first inequality uses Lemma 85 and Lemma 84, and the second inequality assumes that $\varepsilon \leq \frac{\mathcal{R}_q}{\alpha_q {\mathcal{R}_q}^2+1}$

4.

$$f(z) \in \left[\frac{1}{2}\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right) g(\|z\|_2), g(\|z\|_2)\right] \in \left[\frac{1}{2}\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)(\|z\|_2 - 2\varepsilon), \|z\|_2\right]$$

The first containment is by Lemma 85.2.: $\frac{1}{2}\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right) \cdot g(z) \leq q(g(z)) \leq g(z)$. THe second containment is by Lemma 84.4: $g(\|z\|_2) \in [\|z\|_2 - 2\varepsilon, \|z\|_2]$.

□

Lemma 83 (Properties of h) *Given a parameter ε, define*

$$h(r) := \begin{cases} \frac{r^3}{6\varepsilon^2}, & \text{for } r \in [0, \varepsilon] \\ \frac{\varepsilon}{6} + \frac{r-\varepsilon}{2} + \frac{(r-\varepsilon)^2}{2\varepsilon} - \frac{(r-\varepsilon)^3}{6\varepsilon^2}, & \text{for } r \in [\varepsilon, 2\varepsilon] \\ r, & \text{for } r \geq 2\varepsilon \end{cases}$$

1. The derivatives of h are as follows:

$$h'(r) = \begin{cases} \frac{r^2}{2\varepsilon^2}, & \text{for } r \in [0, \varepsilon] \\ \frac{1}{2} + \frac{r-\varepsilon}{\varepsilon} - \frac{(r-\varepsilon)^2}{2\varepsilon^2}, & \text{for } r \in [\varepsilon, 2\varepsilon] \\ 1, & \text{for } r \geq 2\varepsilon \end{cases}$$

$$h''(r) = \begin{cases} \frac{r}{\varepsilon^2}, & \text{for } r \in [0, \varepsilon] \\ \frac{1}{\varepsilon} - \frac{r-\varepsilon}{\varepsilon^2}, & \text{for } r \in [\varepsilon, 2\varepsilon] \\ 0, & \text{for } r \geq 2\varepsilon \end{cases}$$

$$h'''(r) = \begin{cases} \frac{1}{\varepsilon^2}, & \text{for } r \in [0, \varepsilon] \\ -\frac{1}{\varepsilon^2}, & \text{for } r \in [\varepsilon, 2\varepsilon] \\ 0, & \text{for } r \geq 2\varepsilon \end{cases}$$

2. a) h' is positive, motonically increasing.
b) $h'(0) = 0$, $h'(r) = 1$ for $r \geq \varepsilon$
c) $\frac{h'(r)}{r} \leq \min\left\{\frac{1}{\varepsilon}, \frac{1}{r}\right\}$ for all r

3. a) $h''(r)$ is positive

   b) $h''(r) = 0$ *for* $r = 0$ *and* $r \geq 2\varepsilon$
   c) $h''(r) \leq \frac{1}{\varepsilon}$
   d) $\frac{h''(r)}{r} \leq \frac{1}{\varepsilon^2}$
4. $|h'''(r)| \leq \frac{1}{\varepsilon^2}$
5. $r - 2\varepsilon \leq h(r) \leq r$

Proof of Lemma 83
The claims can all be verified with simple algebra. □

Lemma 84 (Properties of g) *Given a parameter ε, let us define*

$$g(z) := h(\|z\|_2)$$

Where h is as defined in Lemma 83 (using parameter ε). Then

1. *a)* $\nabla g(z) = h'(\|z\|_2)\frac{z}{\|z\|_2}$
   b) For $\|z\|_2 \geq 2\varepsilon$, $\nabla g(z) = \frac{z}{\|z\|_2}$.
   c) For any $\|z\|_2$, $\|\nabla g(z)\|_2 \leq 1$
2. *a)* $\nabla^2 g(z) = h''(\|z\|_2)\frac{zz^T}{\|z\|_2^2} + h'(\|z\|_2)\frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right)$
   b) For $\|z\|_2 \geq 2\varepsilon$, $\nabla^2 g(z) = \frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right)$.
   c) For $\|z\|_2 \geq 2\varepsilon$, $\|\nabla^2 g(z)\|_2 = \frac{1}{\|z\|_2}$
   d) For all z, $\|\nabla^2 g(z)\|_2 \leq \frac{1}{\varepsilon}$
3. $\|\nabla^3 g(z)\|_2 \leq \frac{5}{\varepsilon^2}$
4. $\|z\|_2 - 2\varepsilon \leq g(z) \leq \|z\|_2$.

Proof of Lemma 84
All the properties can be verified with algebra. We provide a proof for 3. since it is a bit involved.

Let us define the functions $\kappa^1(z) = \nabla(\|z\|_2), \kappa^2(z) = \nabla^2(\|z\|_2), \kappa^3(z) = \nabla^3(\|z\|_2)$. Specifically,

$$\begin{aligned}
\kappa^1(z) =& \frac{z}{\|z\|_2} \\
\kappa^2(z) =& \frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right) \\
\kappa^3(z) =& -\frac{1}{\|z\|_2^2}\frac{z}{\|z\|_2} \bigotimes \left(I - \frac{zz^T}{\|z\|_2^2}\right) + \frac{1}{\|z\|_2}\left(\frac{z}{\|z\|_2} \bigotimes \kappa^2(z) + \kappa^2(z) \bigotimes \frac{z}{\|z\|_2}\right)
\end{aligned}$$

It can be verified that

$$\left\|\kappa^2(z)\right\|_2 = \frac{1}{\|z\|_2}$$
$$\left\|\kappa^3(z)\right\|_2 = \frac{1}{\|z\|_2^2}$$

It can be verified that $\nabla^2 g(z)$ has the following form:

$$\nabla^3 g(z) = h'''(\|z\|_2)\big(\kappa^1(z)\big)^{\bigotimes 3} + h''(\|z\|_2)\kappa^1(z) \bigotimes \kappa^2(z) + h''(\|z\|_2)\kappa^2(z) \bigotimes \kappa^1(z) + h'(\|z\|_2)\kappa^3(z) + h''(\|z\|_2)\kappa^1(z) \bigotimes \kappa^2(z)$$

Thus

$$\left\|\nabla^3 g(z)\right\|_2 \leq |h'''(\|z\|_2)| + 3\frac{h''(\|z\|_2)}{\|z\|_2} + \frac{h'(\|z\|_2)}{\|z\|_2^2} \leq \frac{5}{\varepsilon^2}$$

Where we use properties of h from Lemma 83.
The last claim follows immediately from Lemma 83.4. □

D.5.1 Defining q

In this section, we define the function q that is used in Lemma 82. Our construction is a slight modification to the original construction in [32].

Let α_q and $\mathcal{R}_q$ be as defined in (5.7). We begin by defining auxiliary functions $\psi(r)$, $\Psi(r)$ and $\nu(r)$, all from $\mathbb{R}^+$ to $\mathbb{R}$:

$$\psi(r) := e^{-\alpha_q \tau(r)}, \qquad \Psi(r) := \int_0^r \psi(s)ds, \qquad \nu(r) := 1 - \frac{1}{2}\frac{\int_0^r \frac{\mu(s)\Psi(s)}{\psi(s)}ds}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)}ds}, \tag{D.19}$$

Where $\tau(r)$ and $\mu(r)$ are as defined in Lemma 86 and Lemma 87 with $\mathcal{R} = \mathcal{R}_q$.
Finally we define q as

$$q(r) := \int_0^r \psi(s)\nu(s)ds. \tag{D.20}$$

We now state some useful properties of the distance function q.

Lemma 85 *The function q defined in* (D.20) *has the following properties.*

1. *For all $r \leq \mathcal{R}_q$, $q''(r) + \alpha_q q'(r) \cdot r \leq -\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{32{\mathcal{R}_q}^2} q(r)$*
2. *For all r, $\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{2} \cdot r \leq q(r) \leq r$*

3. *For all* r, $\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{2} \leq q'(r) \leq 1$

4. *For all* r, $q''(r) \leq 0$ *and* $|q''(r)| \leq \left(\frac{5\alpha_q \mathcal{R}_q}{4} + \frac{4}{\mathcal{R}_q}\right)$

5. *For all* r, $|q'''(r)| \leq 5\alpha_q + 2\alpha_q\left(\alpha_q {\mathcal{R}_q}^2 + 1\right) + \frac{2(\alpha_q {\mathcal{R}_q}^2+1)}{{\mathcal{R}_q}^2}$

Proof of Lemma 85

Proof of 1. It can be verified that

$$\begin{aligned}
\psi'(r) =& \psi(r)(-\alpha_q \tau'(r)) \\
\psi''(r) =& \psi(r)\left(\left(\alpha_q \tau'(r)\right)^2 + \alpha_q \tau''(r)\right) \\
\nu'(r) =& -\frac{1}{2}\frac{\frac{\mu(r)\Psi(r)}{\psi(r)}}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds}
\end{aligned}$$

For $r \in [0, \mathcal{R}_q]$, $\tau'(r) = r$, so that $\psi'(r) = \psi(r)(-\alpha_q r)$. Thus

$$\begin{aligned}
q'(r) =& \psi(r)\nu(r) \\
q''(r) =& \psi'(r)\nu(r) + \psi(r)\nu'(r) \\
=& \psi(r)\nu(r)(-\alpha_q r) + \psi(r)\nu'(r) \\
=& -\alpha_q r \nu'(r) + \psi(r)\nu'(r) \\
q''(r) + \alpha_q r q'(r) =& \psi(r)\nu'(r) \\
=& -\frac{1}{2}\frac{\mu(r)\Psi(r)}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \\
=& -\frac{1}{2}\frac{\Psi(r)}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds}
\end{aligned}$$

Where the last equality is by definition of $\mu(r)$ in Lemma 87 and the fact that $r \leq \mathcal{R}_q$.

We can upper bound

$$\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds \leq \int_0^{4\mathcal{R}_q} \frac{\Psi(s)}{\psi(s)} ds \leq \frac{\int_0^{4\mathcal{R}_q} s ds}{\psi(4\mathcal{R}_q)} = \frac{16{\mathcal{R}_q}^2}{\psi(4\mathcal{R}_q)} \leq 16{\mathcal{R}_q}^2 \cdot \exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)$$

Where the first inequality is by Lemma 87, the second inequality is by the fact that $\psi(s)$ is monotonically decreasing, the third inequality is by Lemma 86.

Thus

$$\begin{aligned} q''(r) + \alpha_q r q'(r) \leq & -\frac{1}{2}\left(\frac{\exp\left(-\frac{7\alpha_q \mathcal{R}_q{}^2}{3}\right)}{16\mathcal{R}_q{}^2}\right)\Psi(r) \\ \leq & -\frac{\exp\left(-\frac{7\alpha_q \mathcal{R}_q{}^2}{3}\right)}{32\mathcal{R}_q{}^2} q(r) \end{aligned}$$

Where the last inequality is by $\Psi(r) \geq q(r)$.

Proof of 2. Notice first that $\nu(r) \geq \frac{1}{2}$ for all r. Thus

$$\begin{aligned} q(r) := & \int_0^r \psi(s)\nu(s)ds \\ \geq & \frac{1}{2}\int_0^r \psi(s)ds \\ \geq & \frac{\exp\left(-\frac{7\alpha_q \mathcal{R}_q{}^2}{3}\right)}{2} \cdot r \end{aligned}$$

Where the last inequality is by Lemma 86.

Proof of 3. By definition of f, $q'(r) = \psi(r)\nu(r)$, and

$$\frac{\exp\left(-\frac{7\alpha_q \mathcal{R}_q{}^2}{3}\right)}{2} \leq \psi(r)\nu(r) \leq 1$$

Where we use Lemma 86 and the fact that $\nu(r) \in [1/2, 1]$

Proof of 4. Recall that

$$q''(r) = \psi'(r)\nu(r) + \psi(r)\nu'(r)$$

That $q'' \leq 0$ can immediately be verified from the definitions of ψ and ν.

Thus

$$\begin{aligned} |q''(r)| \leq & |\psi'(r)\nu(r)| + |\psi(r)\nu'(r)| \\ \leq & \alpha_q \tau'(r) + |\psi(r)\nu'(r)| \end{aligned}$$

From Lemma 86, we can upperbound $\tau'(r) \leq \frac{5\mathcal{R}_q}{4}$. In addition, $\Psi(r) = \int_0^r \psi(s) \geq r\psi(r)$, so that

$$\frac{\Psi(r)}{\psi(r)} \geq r \tag{D.21}$$

(Recall again that $\psi(s)$ is monotonically decreasing). Thus $\Psi(r)/r \geq r$ for all r. In addition, using the fact that $\psi(r) \leq 1$,

$$\Psi(r) = \int_0^r \psi(s)ds \leq r \tag{D.22}$$

Combining the previous expressions,

$$\begin{aligned}
|\psi(r)\nu'(r)| =& \left| \frac{1}{2} \frac{\mu(r)\Psi(r)}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \right| \\
\leq& \left| \frac{1}{2} \frac{\mu(r) r}{\int_0^{\mathcal{R}_q} \frac{\Psi(s)}{\psi(s)} ds} \right| \\
\leq& \left| \frac{1}{2} \frac{4\mathcal{R}_q}{\int_0^{\mathcal{R}_q} s ds} \right| \\
\leq& \frac{4}{\mathcal{R}_q}
\end{aligned}$$

Where the first inequality are by definition of $\mu(r)$ and (D.22), and the second inequality is by (D.21) and the fact that $\mu(r) = 0$ for $r \geq 4\mathcal{R}_q$. Combining with our bound on $\psi'(r)\nu(r)$ gives the desired bound.

Proof of 5.

$$q'''(r) = \psi''(r)\nu(r) + 2\psi'(r)\nu'(r) + \psi(r)\nu''(r)$$

We first bound the middle term:

$$\begin{aligned}
|\psi'(r)\nu' r)| =& |\psi(r)(\alpha_q \tau'(r))\nu' r)| \\
\leq& \alpha_q |\tau'(r)||\psi(r)\nu' r)| \\
\leq& \frac{5\alpha_q \mathcal{R}_q}{4} \cdot \frac{4}{\mathcal{R}_q} \\
\leq& 5\alpha_q
\end{aligned}$$

Where the second last line follows form Lemma 86 and our proof of 4..

Next,

$$\psi''(r) = \psi(r)\big(\alpha_q^2 \tau'(r)^2 - \alpha_q \tau''(r)\big)$$

Thus applying Lemma 86.1 and Lemma 86.3,

$$|\psi''(r)\nu(r)| \leq 2\alpha_q^2 {\mathcal{R}_q}^2 + \alpha_q$$

Finally,

$$\nu''(r) = \frac{1}{2\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \cdot \frac{d}{dr} \mu(r)\Psi(r)/\psi(r)$$

Expanding the numerator,

$$\begin{aligned}\frac{d}{dr}\frac{\mu(r)\Psi(r)}{\psi(r)} =& \mu'(r)\frac{\Psi(r)}{\psi(r)} + \mu(r) - \mu(r)\frac{\Psi(r)\psi'(r)}{\psi(r)^2} \\ =& \mu'(r)\frac{\Psi(r)}{\psi(r)} + \mu(r) + \mu(r)\frac{\Psi(r)\psi(r)\alpha_q\tau'(r)}{\psi(r)^2}\end{aligned}$$

Thus

$$\psi(r)\nu''(r) = \frac{1}{2\int_0^{4\mathcal{R}_q}\frac{\mu(s)\Psi(s)}{\psi(s)}ds}\cdot(\mu'(r)\Psi(r) + \mu(r)\psi(r) + \mu(r)\Psi(r)\alpha_q\tau'(r))$$

Using the same argument as from the proof of 4., we can bound

$$\begin{aligned}\frac{1}{2\int_0^{4\mathcal{R}_q}\frac{\mu(s)\Psi(s)}{\psi(s)}ds} \leq& \frac{1}{2\int_0^{\mathcal{R}_q} sds} \\ \leq& \frac{1}{{\mathcal{R}_q}^2}\end{aligned}$$

Finally, from Lemma 87, $|\mu'(r)| \leq \frac{\pi}{6\mathcal{R}_q}$, so

$$\begin{aligned}|\psi(r)\nu''(r)| \leq& \frac{\pi/6 + 1 + 5\alpha_q{\mathcal{R}_q}^2/4}{{\mathcal{R}_q}^2} \\ \leq& \frac{2(\alpha_q{\mathcal{R}_q}^2 + 1)}{{\mathcal{R}_q}^2}\end{aligned}$$

□

Lemma 86 *Let $\tau(r) : [0,\infty) \to \mathbb{R}$ be defined as*

$$\tau(r) = \begin{cases} \frac{r^2}{2}, & \text{for } r \leq \mathcal{R} \\ \frac{\mathcal{R}^2}{2} + \mathcal{R}(r-\mathcal{R}) + \frac{(r-\mathcal{R})^2}{2} - \frac{(r-\mathcal{R})^3}{3\mathcal{R}}, & \text{for } r \in [\mathcal{R}, 2\mathcal{R}] \\ \frac{5\mathcal{R}^2}{3} + \mathcal{R}(r-2\mathcal{R}) - \frac{(r-2\mathcal{R})^2}{2} + \frac{(r-2\mathcal{R})^3}{12\mathcal{R}}, & \text{for } r \in [2\mathcal{R}, 4\mathcal{R}] \\ \frac{7\mathcal{R}^2}{3}, & \text{for } r \geq 4\mathcal{R}] \end{cases}$$

Then

1. *$\tau'(r) \in [0, \frac{5\mathcal{R}}{4}]$, with maxima at $r = \frac{3\mathcal{R}}{2}$. $\tau'(r) = 0$ for $r \in \{0\}\bigcup[4\mathcal{R}, \infty)$*
2. *As a consequence of 1, $\tau(r)$ is monotonically increasing*
3. *$\tau''(r) \in [-1, 1]$*

Proof of Lemma 86
We provide the derivatives of τ below. The claims in the Lemma can then be immediately verified.

$$\tau'(r) = \begin{cases} r, & \text{for } r \leq \mathcal{R} \\ \mathcal{R} + (r - \mathcal{R}) - \frac{(r-\mathcal{R})^2}{\mathcal{R}}, & \text{for } r \in [\mathcal{R}, 2\mathcal{R}] \\ \mathcal{R} - (r - 2\mathcal{R}) + \frac{(r-2\mathcal{R})^2}{4\mathcal{R}}, & \text{for } r \in [2\mathcal{R}, 4\mathcal{R}] \\ 0, & \text{for } r \geq 4\mathcal{R}] \end{cases}$$

$$\tau''(r) = \begin{cases} 1, & \text{for } r \leq \mathcal{R} \\ 1 - \frac{2(r-\mathcal{R})}{\mathcal{R}}, & \text{for } r \in [\mathcal{R}, 2\mathcal{R}] \\ -1 + \frac{r-2\mathcal{R}}{2\mathcal{R}}, & \text{for } r \in [2\mathcal{R}, 4\mathcal{R}] \\ 0, & \text{for } r \geq 4\mathcal{R}] \end{cases}$$

□

Lemma 87 *Let*

$$\mu(r) := \begin{cases} 1, & \textit{for } r \leq \mathcal{R} \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{\pi(r-\mathcal{R})}{3\mathcal{R}}\right), & \textit{for } r \in [\mathcal{R}, 4\mathcal{R}] \\ 0, & \textit{for } r \geq 4\mathcal{R} \end{cases}$$

Then

$$\mu'(r) := \begin{cases} 0, & \textit{for } r \leq \mathcal{R} \\ -\frac{\pi}{6\mathcal{R}}\sin\left(\frac{\pi(r-\mathcal{R})}{\mathcal{R}}\right), & \textit{for } r \in [\mathcal{R}, 4\mathcal{R}] \\ 0, & \textit{for } r \geq 4\mathcal{R} \end{cases}$$

Furthermore, $\mu'(r) \in [-\frac{\pi}{6\mathcal{R}}, 0]$

This lemma can be easily verified by algebra.

D.6 Miscellaneous

The following theorem, taken from [35], establishes a quantitative CLT.

Theorem 15 *Let $X_1...X_n$ be random vectors with mean 0, covariance Σ, and $\|X_i\| \leq \beta$ almost surely for each i. Let $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$, and let Z be a Gaussian with covariance Σ, then*

$$W_2(S_n, Z) \leq \frac{6\sqrt{d}\beta\sqrt{\log n}}{\sqrt{n}}.$$

Corollary 88 *Let $X_1...X_n$ be random vectors with mean 0, covariance Σ, and $\|X_i\| \leq \beta$ almost surely for each i. let Y be a Gaussian with covariance $n\Sigma$. Then*

$$W_2\left(\sum_i X_i, Y\right) \leq 6\sqrt{d}\beta\sqrt{\log n}.$$

This is simply taking the result of Theorem 15 and scaling the inequality by $\sqrt{n}$ on both sides.

The following lemma is taken from [19] and included here for completeness.

Lemma 89 *For any $c > 0$, $x > 3\max\left\{\frac{1}{c}\log\frac{1}{c}, 0\right\}$, the inequality*

$$\frac{1}{c}\log(x) \leq x$$

holds.

Proof
We will consider two cases:

Case 1: If $c \geq \frac{1}{e}$, then the inequality

$$\log(x) \leq cx$$

is true for all x.

Case 2: $c \leq \frac{1}{e}$.

In this case, we consider the Lambert W function, defined as the inverse of $f(x) = xe^x$. We will particularly pay attention to W_{-1} which is the lower branch of W.

We can lower bound $W_{-1}(-c)$ using Theorem 1 from [13]:

$$\forall u > 0, \quad W_{-1}(-e^{-u-1}) > -u - \sqrt{2u} - 1$$

$$\text{equivalently} \quad \forall c \in (0, 1/e), \quad -W_{-1}(-c) < \log\left(\frac{1}{c}\right) + 1 + \sqrt{2\left(\log\left(\frac{1}{c}\right) - 1\right)} - 1$$

$$= \log\left(\frac{1}{c}\right) + \sqrt{2\left(\log\left(\frac{1}{c}\right) - 1\right)}$$

$$\leq 3\log\frac{1}{c}$$

Thus by our assumption,

$$\begin{aligned} x &\geq 3 \cdot \frac{1}{c} \log\left(\frac{1}{c}\right) \\ \Rightarrow x &\geq \frac{1}{c}(-W_{-1}(-c)) \end{aligned}$$

then $W_{-1}(-c)$ is defined, so

$$\begin{aligned} x &\geq \frac{1}{c} \max\{-W_{-1}(-c), 1\} \\ \Rightarrow (-cx)e^{-cx} &\geq -c \\ \Rightarrow xe^{-cx} &\leq 1 \\ \Rightarrow \log(x) &\leq cx \end{aligned}$$

The first implication is justified as follows: $W_{-1}^{-1} : [-\frac{1}{\varepsilon}, \infty) \to (-\infty, -1)$ is monotonically decreasing. Thus its inverse $W_{-1}^{-1}(y) = ye^y$, defined over the domain $(-\infty, -1)$ is also monotonically decreasing. By our assumption, $-cx \leq -3 \log \frac{1}{c} \leq -3$, thus $-cx \in (-\infty, -1]$, thus applying W_{-1}^{-1} to both sides gives us the first implication. $\square$