Constrained Langevin Algorithms with L-mixing External Random Variables

Anonymous Author(s)
Affiliation
Address
email

Abstract

Langevin algorithms are gradient descent methods augmented with additive noise, and are widely used in Markov Chain Monte Carlo (MCMC) sampling, optimization, and learning. In recent years, the non-asymptotic analysis of Langevin algorithms for non-convex optimization learning has been extensively explored. For constrained problems with non-convex losses over compact convex domain in the case of IID data variables, Langevin algorithm achieves a deviation of $O(T^{-1/4}(\log T)^{1/2})$ from its target distribution [27]. In this paper, we obtain a deviation of $O(T^{-1/2}\log T)$ in 1-Wasserstein distance for non-convex losses with L-mixing data variables and polyhedral constraints (which are not necessarily bounded). This deviation indicates that our convergence rate is faster than those in the previous works on constrained Langevin algorithms for non-convex optimization.

1 Introduction

Langevin algorithms can be viewed as the simulation of Langevin dynamics from statistical physics [14]. They have been widely studied for Markov Chain Monte Carlo (MCMC) sampling [37], non-convex optimization [5, 23] and machine learning [42]. In the statistical community, Langevin methods are used to resolve the difficulty of exact sampling from a high dimensional distribution. For non-convex optimization, the additive noise assists the algorithms to escape from local minima and saddles. Since many modern technical challenges can be cast as sampling and optimization problems, Langevin algorithms are a potential choice for the areas of adaptive control, Deep Neural networks, reinforcement learning, time series analysis, image processing and so on [4, 10, 29].

Related Work. In recent years, the non-asymptotic analysis of Langevin algorithms has been extensively studied. The discussion below reviews theoretical studies of Langevin algorithms for MCMC sampling, optimization, and learning.

The non-asymptotic analysis of Langevin algorithms for approximate sampling began with [16, 17], with more recent relevant work given in [3, 4, 10, 13, 19, 22, 28, 30, 32–34, 41, 45]. Most works on LMC consider log-concave target distributions, though there exists some work relaxing log-concavity [10, 13, 33, 34, 41] and smoothness of the target distribution [13, 28, 34], the mainstream on LMC focuses on the unconstrained cases.

Constrained problems are less studied, but a variety of works have begun to address constraints in recent years. The work [8, 9] analyzes the case of log-concave distributions with samples constrained to a convex, compact set. Other methods derived from optimization have been introduced to handle constraints, such as mirror descent [1, 25, 26, 44] and proximal methods [7].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Pioneering work on non-asymptotic analysis of Langevin algorithms for unconstrained non-convex optimization with IID external data variables was given in [35], which was motivated by machine learning applications [42]. Since then, numerous improvements and variations on unconstrained Langevin algorithms for non-convex optimization have been reported [10–12, 21, 43].

The first analysis of Langevin algorithms for non-convex optimization with IID external variables constrained to compact convex sets is given in [27], and builds upon [8, 9]. However, the convergence rate derived in [27] is rather slow since it uses a loose result on Skorokhod problems in [40]. Recent work of [39] obtains ϵ-suboptimality guarantees in $\tilde{O}(\epsilon^{-1/3})$ steps, in contrast to $\tilde{O}(\epsilon^{-1/4})$ required in [27]. In the case of compact polyhedral sets, the work in this paper results in an improved bound of $\tilde{O}(\epsilon^{-1/2})$ steps required for ϵ-suboptimality. Here the notation $\tilde{O}$ expresses the order while hiding logarithmic factors.

Most convergence analyses for constrained non-convex optimization require no constraints or bounded constraint sets and IID external random variables or no external variables. In practice, the boundedness of constraint sets and the dependence of external variables do not always hold. The work [10] gives non-asymptotic bounds with L-mixing external variables and non-convex losses, which achieves tight performance guarantees in the unconstrained case. In contrast, our work gets a tight convergence bound (up to logarithmic factors) with the L-mixing data streams and applies to arbitrary polyhedral constraints, which may be unbounded.

Contributions. This paper focuses on the non-asymptotic analysis of constrained Langevin algorithms for a non-convex problem with L-mixing external random variables and polyhedral constraints. We show the algorithm can achieve a deviation of $O(T^{-1/2}\log T)$ from its target distribution in 1-Wasserstein distance in the polyhedral constraint and with dependent variables. The result from [10] on unconstrained Langevin algorithms with L-mixing external random variables gives a deviation of $O(T^{-1/2}(\log T)^{1/2})$, and so we see that our results match, up to a factor of $(\log T)^{1/2}$. For constrained problems, our general polyhedral assumption is not directly comparable to related work of [27], which examines compact convex constraints, and [39], which examines bounded non-convex constraints. In the cases where the domains and random variable assumptions match (i.e. bounded polyhedra with IID external random variables or no external random variables), our paper gives the tightest bounds. Indeed, the work of [27] gives a deviation of $O(T^{-1/4}(\log T)^{1/2})$ with respect to 1-Wasserstein distance from the target distribution, while [39] gives ϵ-suboptimality in $\tilde{O}(\epsilon^{-1/3})$ steps in comparison to $\tilde{O}(\epsilon^{-1/2})$ steps implied by our current results.

A key enabling result in this paper is a new quantitative bound on the deviation between Skorokhod problem solutions over polyhedra, which gives a more explicit variation of an earlier non-constructive result from [18]. Additionally, we derive a relatively simple approach to averaging out the effect of L-mixing random variables on algorithms.

2 Problem Setup

2.1 Notation and terminology

$\mathbb{R}$ denotes the set of real numbers while $\mathbb{N}$ denotes the set of non-negative integers. The Euclidean norm over $\mathbb{R}^n$ is denoted by $\|\cdot\|$.

Random variables will be denoted in bold. If $\mathbf{x}$ is a random variable, then $\mathbb{E}[\mathbf{x}]$ denotes its expected value and $\mathcal{L}(\mathbf{x})$ denotes its law. IID stands for independent, identically distributed. The indicator function is denoted by $\mathbb{1}$. If P and Q are two probability measures over $\mathbb{R}^n$, then the 1-Wasserstein distance between them with respect the Euclidean norm is denoted by $W_1(P, Q)$.

Let $\mathcal{K}$ be a convex set. (In this paper, we will assume that $\mathcal{K}$ is polyhedral with 0 in its interior.) The boundary of $\mathcal{K}$ is denoted by $\partial\mathcal{K}$. The normal cone of $\mathcal{K}$ at a point x is denoted by $N_{\mathcal{K}}(x)$. The convex projection onto $\mathcal{K}$ is denoted by $\Pi_{\mathcal{K}}$.

Let $\mathcal{Z}$ denote the domain of the external random variables $\mathbf{z}_k$.

If $\mathcal{F}$ and $\mathcal{G}$ are σ-algebras, let $\mathcal{F} \vee \mathcal{G}$ denote the σ-algebra generated by the union of $\mathcal{F}$ and $\mathcal{G}$.

The 1-Wasserstein distance is defined as below:

$$W_1(P,Q) = \inf_{\Gamma \in \mathfrak{C}(P,Q)} \int_{\mathcal{K}\times\mathcal{K}} \|x-y\| d\Gamma(x,y)$$

where $\mathfrak{C}$ is the couplings between P and Q.

2.2 Constrained Langevin algorithm

For integers k let $\hat{\mathbf{w}}_k \sim \mathcal{N}(0, I)$ be IID Gaussian random variables and let $\mathbf{z}_k$ be an L-mixing process whose properties will be described later. Assume that $\mathbf{z}_i$ is independent of $\hat{\mathbf{w}}_j$ for all $i, j \in \mathbb{N}$.

Assume that the initial value of $\mathbf{x}_0 \in \mathcal{K}$ is independent of $\mathbf{z}_i$ and $\hat{\mathbf{w}}_j$. Then the constrained Langevin algorithm has the form:

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{K}} \left(\mathbf{x}_k - \eta \nabla_x f(\mathbf{x}_k, \mathbf{z}_k) + \sqrt{\frac{2\eta}{\beta}} \hat{\mathbf{w}}_k \right), \tag{1}$$

with k an integer. Here $\eta > 0$ is the step size parameter and $\beta > 0$ is the inverse temperature parameter. In the learning context, $f(\mathbf{x}, \mathbf{z})$ is the objective function where $\mathbf{x}$ is the parameters we aim to learn and $\mathbf{z}$ is the training data.

2.3 L-mixing processes

In this paper, we assume that $\mathbf{z}_k$ is a sequence of external data variables. The class of L-mixing processes was introduced in [24] for applications in system identification and time-series analysis, and gives a means to quantitatively measure how the dependencies between the $\mathbf{z}_k$ decay over time. Formally, L-mixing requires two components: 1) M-boundedness, which specifies a global bound on the moments and 2) a measure of the decay of influence over time.

A discrete-time stochastic processes $\mathbf{z}_k$ is M-bounded if for all $m \geq 1$

$$\mathcal{M}_m(\mathbf{z}) = \sup_{k \geq 0} \mathbb{E}^{1/m} \left[\|\mathbf{z}_k\|^m \right] < \infty. \tag{2}$$

Let $\mathcal{F}_k$ be an increasing family of σ-algebras such that $\mathbf{z}_k$ is $\mathcal{F}_k$-measurable and $\mathcal{F}_k^+$ be a decreasing family of σ-algebras such that $\mathcal{F}_k$ and $\mathcal{F}_k^+$ are independent for all $k \geq 0$. Then, the process $\mathbf{z}_k$ is L-mixing with respect to $\left((\mathcal{F}_k), (\mathcal{F}_k^+) \right)$ if it is M-bounded and

$$\Psi_m(\mathbf{z}) = \sum_{\tau=0}^{\infty} \psi_m(\tau, \mathbf{z}) < \infty \tag{3a}$$

with

$$\psi_m(\tau, \mathbf{z}) = \sup_{k \geq \tau} \mathbb{E}^{1/m} \left[\left\| \mathbf{z}_k - \mathbb{E}\left[\mathbf{z}_k | \mathcal{F}_{k-\tau}^+ \right] \right\|^m \right]. \tag{3b}$$

A common example of L-mixing process is the stable autoregressive (AR) model. To be more concrete, we specifically present an AR(1) model as below:

$$\mathbf{z}_{k+1} = \alpha \mathbf{z}_k + \boldsymbol{\xi}_{k+1} \tag{4}$$

where α is a constant with $|\alpha| < 1$ and for all $k \in \mathbb{Z}$, $\boldsymbol{\xi}_k$ are IID standard Gaussian random variables and $\mathbf{z}_k \in \mathcal{Z}$. It can be observed from (4) that

$$\mathbf{z}_k = \sum_{j=0}^{\infty} \alpha^j \boldsymbol{\xi}_{k-j}. \tag{5}$$

Then, if we specify $\mathcal{F}_k = \sigma\{\boldsymbol{\xi}_i : i \leq k\}$ and $\mathcal{F}_k^+ = \sigma\{\boldsymbol{\xi}_i : i > k\}$, it can be verified that $\mathbf{z}_k$ satisfy (2) and (3) and so is an L-mixing process.

2.4 Assumptions

We assume that $\nabla_x f(x,z)$ is ℓ-Lipschitz in both x and z. In particular, this implies that $\|\nabla_x f(x_1,z) - \nabla_x f(x_2,z)\| \leq \ell\|x_1 - x_2\|$ and $\|\nabla_x f(x,z_1) - \nabla f(x,z_2)\| \leq \ell\|z_1 - z_2\|$.

We assume that $\mathbf{z}_t$ is a stationary L-mixing process, and let $\bar{f}(x) = \mathbb{E}[f(x,\mathbf{z}_t)]$ denote the function which averages $f(x,\mathbf{z}_t)$ with respect to $\mathbf{z}_t$.

Further, we assume that $\bar{f}(x)$ is μ-strongly convex outside a ball of radius $R > 0$, i.e. $(x_1 - x_2)^\top \left(\nabla\bar{f}(x_1) - \nabla\bar{f}(x_2)\right) \geq \mu\left\|x_1 - x_2\right\|^2$ for all $x_1, x_2 \in \mathcal{K}$ such that $\|x_1 - x_2\| \geq R$.

We assume that the initial second moment is bounded above as $\mathbb{E}[\|\mathbf{x}_0\|^2] \leq \varsigma < \infty$.

Throughout the paper, $\mathcal{K}$ will denote a polyhedral subset of $\mathbb{R}^n$ with 0 in its interior.

3 Main results

3.1 Convergence of the law of the iterates

For $\bar{f}$ defined above, the associated Gibbs measure is defined by:

$$\pi_{\beta\bar{f}}(A) = \frac{\int_{A\cap\mathcal{K}} e^{-\beta\bar{f}(x)}dx}{\int_{\mathcal{K}} e^{-\beta\bar{f}(x)}dx}. \tag{6}$$

The main result of this paper is stated next:

Theorem 1. *Assume that $\eta \leq \min\left\{\frac{1}{4}, \frac{\mu}{4\ell^2}\right\}$, $\mathcal{K}$ is a polyhedron with 0 in its interior, $\mathbf{x}_0 \in \mathcal{K}$, and $\mathbb{E}[\|\mathbf{x}_0\|^2] \leq \varsigma$. There are constants a, c_1, c_2, c_3, and c_4 such that the following bound holds for all integers $k \geq 4$:*

$$W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \leq (c_1 + c_2\sqrt{\varsigma})e^{-\eta a k} + (c_3 + c_4\sqrt{\varsigma})\sqrt{\eta\log(\eta^{-1})}.$$

In particular, if $\eta = \frac{\log T}{2aT}$, $T \geq 4$ and $T \geq e^{2a}$, then

$$W_1(\mathcal{L}(\mathbf{x}_T), \pi_{\beta\bar{f}}) \leq \left(c_1 + c_2\sqrt{\varsigma} + \frac{c_3 + c_4\sqrt{\varsigma}}{(2a)^{1/2}}\right)T^{-1/2}\log T.$$

Furthermore, the constants, c_1, c_2, c_3, and c_4 are $O(n)$ with respect to the dimension of $\mathbf{x}_k$, and $O(e^{\ell\beta R^2/8})$ with respect to the inverse temperature, β. And for all $\beta > 0$, $a \geq \frac{32\eta}{\mu\beta^2}e^{\frac{\beta\ell R^2}{8}}$.

The constants depend on the dimension of $\mathbf{x}_k$, n, the noise parameter, β, the Lipschitz constant, ℓ, the strong convexity constant μ, the variance bound of the initial states, ς, and some geometric properties of the polyhedron, $\mathcal{K}$.

All the constants shown in Theorem 1 and the others in the remaining of the paper are summarized in Appendix H.

3.2 Auxiliary processes for convergence analysis

Similar to the previous analyses of Langevin methods, e.g. [9, 10, 27, 35], the proof of Theorem 1 uses a collection of auxiliary processes fitting between the algorithms iterates from (1) and a stationary distribution given by (6).

The algorithm and a variation in which the $\mathbf{z}_t$ variables are averaged out are respectively given by:

$$\mathbf{x}_{t+1}^A = \Pi_{\mathcal{K}}\left(\mathbf{x}_t^A - \eta\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t) + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_t\right) \tag{7a}$$

$$\mathbf{x}_{t+1}^M = \Pi_{\mathcal{K}}\left(\mathbf{x}_t^M - \eta\nabla_x \bar{f}(\mathbf{x}_t^M) + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_t\right). \tag{7b}$$

Here $\mathbf{x}_t^A$ represents the Algorithm, while $\mathbf{x}_t^M$ represents a corresponding Mean process.

We embed the mean process in continuous time by setting $\mathbf{x}_t^M = \mathbf{x}_{\lfloor t \rfloor}^M$, where $\lfloor t \rfloor$ indicates floor function. The Gaussian noise $\hat{\mathbf{w}}_k$ can be realized as $\hat{\mathbf{w}}_k = \mathbf{w}_{k+1} - \mathbf{w}_k$ where $\mathbf{w}_t$ is a Brownian motion.

Let $\mathbf{x}_t^C$ denote a Continuous-time approximation of $\mathbf{x}_t^M$ defined by the following reflected stochastic differential equation (RSDE):

$$d\mathbf{x}_t^C = -\eta \nabla_x \bar{f}(\mathbf{x}_t^C)dt + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t - \mathbf{v}_t^C d\boldsymbol{\mu}^C(t). \tag{8}$$

Here $-\int_0^t \mathbf{v}_s^C d\boldsymbol{\mu}^C(s)$ is a bounded variation reflection process that ensures that $\mathbf{x}_t^C \in \mathcal{K}$ for all $t \geq 0$, as long as $\mathbf{x}_0^C \in \mathcal{K}$. In particular, the measure $\boldsymbol{\mu}^C$ is such that $\boldsymbol{\mu}^C([0,t])$ is finite, $\boldsymbol{\mu}^C$ supported on $\{s | \mathbf{x}_s^C \in \partial\mathcal{K}\}$, and $\mathbf{v}_s^C \in N_{\mathcal{K}}(\mathbf{x}_s^C)$ where $N_{\mathcal{K}}(x)$ is the normal cone of $\mathcal{K}$ at x. Lemma 10 in Appendix A shows that the reflection process is uniquely defined and $\mathbf{x}^C$ is the unique solution to the Skorokhod problem for the process defined by:

$$\mathbf{y}_t^C = \mathbf{x}_0^C + \sqrt{\frac{2\eta}{\beta}} \mathbf{w}_t - \eta \int_0^t \nabla_x \bar{f}(\mathbf{x}_s^C) ds. \tag{9}$$

See Appendix A for more details on the Skorokhod problem.

For compact notation, we denote the Skorokhod solution for a given trajectory, $\mathbf{y}$, by $\mathcal{S}(\mathbf{y})$. So, the fact that $\mathbf{x}^C$ is the solution to the Skorokhod problem for $\mathbf{y}^C$ will be denoted succinctly by $\mathbf{x}^C = \mathcal{S}(\mathbf{y}^C)$.

The basic idea behind the proof is to utilize the triangle inequality:

$$W_1(\mathcal{L}(\mathbf{x}_k^A), \pi_{\beta f}) \leq W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C)) + W_1(\mathcal{L}(\mathbf{x}_k^C), \pi_{\beta \bar{f}}). \tag{10}$$

and then bound each of the terms separately.

The second term is bounded by the following lemma:

Lemma 2. *Assume that $\mathbf{x}_0 \in \mathcal{K}$ and $\mathbb{E}[\|\mathbf{x}_0^C\|^2] \leq \varsigma$. There are positive constants a, c_1 and c_2 such that for all $t \geq 0$*

$$W_1(\mathcal{L}(\mathbf{x}_t^C), \pi_{\beta \bar{f}}) \leq (c_1 + c_2\sqrt{\varsigma})\, e^{-\eta a t}.$$

This result is based on an extension of the contraction results from Corollary 2 of [20] for SDEs to the case of the reflected SDEs. Appendix D steps through the methodology from [20] in order to derive a, c_1 and c_2 for our particular problem.

Most of the novel work in the paper focuses on deriving the following bound on $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C))$:

Lemma 3. *Assume that $\mathbf{x}_0^A = \mathbf{x}_0^C \in \mathcal{K}$, $\mathbb{E}[\|\mathbf{x}_0^C\|^2] \leq \varsigma$, and $\eta \leq \min\left\{\frac{1}{4}, \frac{\mu}{8\ell^2}\right\}$. Then there are positive constants c_3 and c_4 such that for all integers $k \geq 0$:*

$$W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C)) \leq (c_3 + c_4\sqrt{\varsigma})\sqrt{\eta \log(\eta^{-1})}.$$

Proof of Theorem 1 Plugging the results of Lemmas 2 and 3 into the triangle inequality bound from (10) proves the first result of the theorem. Specifically, let $\eta = \frac{\log T}{2aT}$, then

$$\begin{aligned} W_1(\mathcal{L}(\mathbf{x}_T), \pi_{\beta \bar{f}}) &\leq (c_1 + c_2\sqrt{\varsigma})T^{-1/2} + (c_3 + c_4\sqrt{\varsigma})\sqrt{\frac{\log T}{2aT}\log(\frac{2aT}{\log T})} \\ &\leq (c_1 + c_2\sqrt{\varsigma})T^{-1/2}\log T + \frac{c_3 + c_4\sqrt{\varsigma}}{(2a)^{1/2}} T^{-1/2}\log T. \end{aligned}$$

This gives the specific bound in the theorem. The last inequality utilizes the fact that $\log T > 1$ for all $T \geq 4$ and $\frac{2aT}{\log T} \leq T$ when $T \geq e^{2a}$.

Furthermore, we examine the bounds of the constants c_1, c_2, c_3, c_4 and a in Appendix H, where the dependencies of the convergence guarantee on state dimension n and the inverse temperature parameter, β can be observed directly. ■

The rest of the paper focuses on proving Lemma 3.

3.3 Proof overview for Lemma 3

This subsection describes the main ideas in the proof of Lemma 3. The results highlighted here, and proved in the appendix, cover the main novel aspects of the current work. The first novelty, captured in Lemmas 4 and 5, is a new way to bound stochastic gradient Langevin schemes with L-mixing data from a Langevin method with the data variables averaged out. The key idea is a method for examining a collection of partially averaged processes. The second novelty is a tight quantitative bound on the deviation of discretized Langevin algorithms from their continuous-time counterparts when constrained to a polyhedron. This result is based on a new quantitative bound on Skorokhod solutions over polyhedra.

First we derive time-dependent bounds (i.e. bounds that depend on k) for $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C))$. This is achieved by introducing a collection of intermediate processes and bounding their differences. Time-uniform bounds are then achieved by exploiting contractivity properties of $\mathbf{x}_t^C$.

To bound $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C))$, we first use the triangle inequality:

$$W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C)) \leq W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^M)) + W_1(\mathcal{L}(\mathbf{x}_k^M), \mathcal{L}(\mathbf{x}_k^C)). \tag{11}$$

We bound $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^M))$ via a collection of auxiliary processes in which the effect of $\mathbf{z}_k$ is partially averaged out. We bound $W_1(\mathcal{L}(\mathbf{x}_k^M), \mathcal{L}(\mathbf{x}_k^C))$ via a specialized discrete-time approximation of $\mathbf{x}_t^C$.

Now we construct the collection of partially averaged processes. Recall that $\mathbf{z}_k \in \mathcal{Z}$ is a stationary L-mixing process with respect to the σ-algebras $\mathcal{F}_k$ and $\mathcal{F}_k^+$. For $k < 0$, we set $\mathcal{F}_k = \{\emptyset, \mathcal{Z}\}$, i.e. the trivial σ-algebra. Let $\mathcal{G}_t$ be the filtration generated by the Brownian motion, $\mathbf{w}_t$.

Recall that for $k \in \mathbb{N}$, we set $\hat{\mathbf{w}}_k = \mathbf{w}_{k+1} - \mathbf{w}_k$. Define the following discrete-time processes:

$$\mathbf{x}_{k+1}^{M,s} = \Pi_{\mathcal{K}}\left(\mathbf{x}_k^{M,s} - \eta\mathbb{E}[\nabla_x f(\mathbf{x}_k^{M,s}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}_k] + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_k\right) \tag{12a}$$

$$\mathbf{x}_{k+1}^{B,s} = \Pi_{\mathcal{K}}\left(\mathbf{x}_k^{B,s} - \eta\mathbb{E}[\nabla_x f(\mathbf{x}_k^{M,s}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}_k] + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_k\right). \tag{12b}$$

Assume that all initial conditions are equal. In other words, $\mathbf{x}_0^A = \mathbf{x}_0^M = \mathbf{x}_0^{M,s} = \mathbf{x}_0^{B,s}$, for all $s \geq 0$. The iterations from (12a) define a family of algorithms in which the data variables are partially averaged, while $\mathbf{x}_k^{B,s}$ from (12b) corresponds to an auxiliary process that fits between $\mathbf{x}_k^{M,s}$ and $\mathbf{x}_k^{M,s+1}$. (Here "A" stands for algorithm, "M" stands for mean, and "B" stands for between.)

Note for $s = 0$, we have that $\mathbf{x}_k^{M,0} = \mathbf{x}_k^A$ and for $s > k$, we have that $\mathbf{x}_k^{M,s} = \mathbf{x}_k^M$. So, in order to bound $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^M))$, it suffices to bound $W_1(\mathcal{L}(\mathbf{x}_k^{M,s}), \mathcal{L}(\mathbf{x}_k^{B,s}))$ and $W_1(\mathcal{L}(\mathbf{x}_k^{B,s}), \mathcal{L}(\mathbf{x}_k^{M,s+1}))$ for all $s \geq 0$. These bounds are achieved in the following lemmas, which are proved in Appendix E.

Lemma 4. *For all $s \geq 0$ and all $k \geq 0$, the following bound holds:*

$$W_1\left(\mathcal{L}(\mathbf{x}_k^{M,s}), \mathcal{L}(\mathbf{x}_k^{B,s})\right) \leq \mathbb{E}[\|\mathbf{x}_k^{M,s} - \mathbf{x}_k^{B,s}\|] \leq 2\ell\psi_2(s, \mathbf{z})\eta\sqrt{k}.$$

Lemma 5. *For all $s \geq 0$ and all $k \geq 0$, the following bound holds*

$$W_1\left(\mathcal{L}(\mathbf{x}_k^{B,s}), \mathcal{L}(\mathbf{x}_k^{M,s+1})\right) \quad \leq \quad \mathbb{E}[\|\mathbf{x}_k^{B,s} - \mathbf{x}_k^{M,s+1}\|] \quad \leq \quad 2\ell\psi_2(s, \mathbf{z})\eta\sqrt{k}\left(e^{\eta k\ell} - 1\right).$$

Now we define the discretized approximation of $\mathbf{x}_t^C$. For any initial $\mathbf{x}_0^D \in \mathcal{K}$, we define the following iteration on the integers:

$$\mathbf{x}_{k+1}^D = \Pi_{\mathcal{K}}(\mathbf{x}_k^D + \mathbf{y}_{k+1}^C - \mathbf{y}_k^C) = \Pi_{\mathcal{K}}\left(\mathbf{x}_k^D + \int_k^{k+1} \nabla\bar{f}(\mathbf{x}_s^C)ds + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_k\right).$$

Recall that the process $\mathbf{y}^C$ is defined by (9).

Provided that $\mathbf{x}_0^D = \mathbf{x}_0^C$, we can define $\mathbf{x}^D = \mathcal{S}(\mathbf{y}^D) = \mathcal{S}(\mathcal{D}(\mathbf{y}^C))$, where $\mathcal{D}$ is the discretization operator that sets $\mathcal{D}(x)_t = x_{\lfloor t \rfloor}$ for any continuous-time trajectory x_t. Recall that $\mathcal{S}$ is the Skorokhod solution operator.

The approximation, $\mathbf{x}^D$, was utilized in [9, 27] to bound discretization errors. The next lemmas show how to bound $W_1(\mathcal{L}(\mathbf{x}_k^C), \mathcal{L}(\mathbf{x}_k^D))$ and $W_1(\mathcal{L}(\mathbf{x}_k^M), \mathcal{L}(\mathbf{x}_k^D))$, respectively. In particular, Lemma 6 is analogous to Propositions 2.4 and 3.6 of [9] and Lemma 9 of [27]. These earlier works end up with bounds of $O(\eta^{3/4}k^{1/2} + \sqrt{\eta \log k})$. It is shown in [27] that such bounds can be translated into time-uniform bounds of the form $\tilde{O}(\eta^{1/4})$. The bound from Lemma 6 is of the form $O(\eta k^{1/2} + \sqrt{\eta \log k})$, and we will see in the next subsection that this leads to a time-uniform bound of the form $\tilde{O}(\eta^{1/2})$.

Lemma 6. *Assume that $\mathcal{K}$ is a polyhedron with 0 in its interior. Assume that $\mathbf{x}_0^C = \mathbf{x}_0^D \in \mathcal{K}$ and that $\mathbb{E}[\|\mathbf{x}_0^C\|] \leq \varsigma$. There are constants, c_5, c_6 and c_7 such that for all integers $k \geq 0$, the following bound holds:*

$$W_1\left(\mathcal{L}(\mathbf{x}_k^C), \mathcal{L}(\mathbf{x}_k^D)\right) \leq \mathbb{E}\left[\|\mathbf{x}_k^C - \mathbf{x}_k^D\|\right] \leq (c_5 + c_6\sqrt{\varsigma})\,\eta\sqrt{k} + c_7\sqrt{\eta \log(4k)}.$$

Lemma 7. *Assume that $\mathcal{K}$ is a polyhedron with 0 in its interior. Assume that $\mathbf{x}_0^C = \mathbf{x}_0^D = \mathbf{x}_0^M \in \mathcal{K}$ and that $\mathbb{E}[\|\mathbf{x}_0^C\|] \leq \varsigma$. Then for all integers $k \geq 0$, the following bound holds*

$$W_1\left(\mathcal{L}(\mathbf{x}_k^M), \mathcal{L}(\mathbf{x}_k^D)\right) \leq \mathbb{E}\left[\|\mathbf{x}_k^M - \mathbf{x}_k^D\|\right] \leq \left((c_5 + c_6\sqrt{\varsigma})\,\eta\sqrt{k} + c_7\sqrt{\eta \log(4k)}\right)\left(e^{\eta\ell k} - 1\right).$$

We highlight that Lemma 6 utilizes the rather tight bounds on solutions to Skorokhod problems over a polyhedral domain shown in Theorem 9. The conclusion of such tight bounds is one of the novelties of our work. More details will be discussed in Section 4 and Appendix A.

With all of the auxiliary processes defined and their differences, we have the following lemma, which gives a time-dependent bound on $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C))$:

Lemma 8. *Assume that $\mathcal{K}$ is a polyhedron with 0 in its interior. Assume that $\mathbf{x}_0^A = \mathbf{x}_0^C \in \mathcal{K}$ and that $\mathbb{E}[\|\mathbf{x}_0^A\|] \leq \varsigma$. There are constants, c_6, c_7 and c_8, such that for all $k \geq 0$, the following bound holds:*

$$W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C)) \leq \left((c_8 + c_6\sqrt{\varsigma})\,\eta\sqrt{k} + c_7\sqrt{\eta \log(4k)}\right) e^{\eta\ell k}.$$

Proof of Lemma 8 Recalling that $\mathbf{x}_k^{M,0} = \mathbf{x}_k^A$ and $\mathbf{x}_k^{M,k+1} = \mathbf{x}_k^M$ and using the triangle inequality gives:

$$\begin{aligned}
&W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^M)) \\
&\leq \sum_{s=0}^{k} W_1(\mathcal{L}(\mathbf{x}_k^{M,s}), \mathcal{L}(\mathbf{x}_k^{M,s+1})) \\
&\leq \sum_{s=0}^{k} \left(W_1(\mathcal{L}(\mathbf{x}_k^{M,s}), \mathcal{L}(\mathbf{x}_k^{B,s})) + W_1(\mathcal{L}(\mathbf{x}_k^{B,s}), \mathcal{L}(\mathbf{x}_k^{M,s+1}))\right) \\
&\overset{\text{Lemmas 4 \& 5}}{\leq} \sum_{s=0}^{k} 2\ell\psi_2(s, \mathbf{z})\eta\sqrt{k}e^{\eta\ell k} \\
&\leq 2\ell\Psi_2(\mathbf{z})\eta\sqrt{k}e^{\eta\ell k}.
\end{aligned} \tag{13}$$

Here $\psi_2(s, \mathbf{z})$ and $\Psi_2(\mathbf{z})$ are the terms that bound the decay of probabilistic dependence between the $\mathbf{z}_k$ variables, as defined in (3).

Similarly, we bound

$$\begin{aligned}
&W_1(\mathcal{L}(\mathbf{x}_k^M), \mathcal{L}(\mathbf{x}_k^C)) \\
&\leq W_1(\mathcal{L}(\mathbf{x}_k^M), \mathcal{L}(\mathbf{x}_k^D)) + W_1(\mathcal{L}(\mathbf{x}_k^D), \mathcal{L}(\mathbf{x}_k^C)) \\
&\overset{\text{Lemmas 6 \& 7}}{\leq} \left((c_5 + c_6\sqrt{\varsigma})\eta\sqrt{k} + c_7\sqrt{\eta \log(4k)}\right) e^{\eta\ell k}.
\end{aligned} \tag{14}$$

Plugging the bounds from (13) and (14) into (11) proves the lemma, with $c_8 = c_5 + 2\ell\Psi_2(\mathbf{z})$. ■

The proof of Lemma 3 is completed by showing how the time-dependent bound from Lemma 8 can be turned into a bound that is independent of k. The technique used for this step is based on ideas from [10], and is shown in Appendix G.

4 Quantitative bounds on Skorokhod solutions over polyhedra

In this section, we present a result that enables a tight bound between the continuous-time process $\mathbf{x}_t^C$ and the discretized process $\mathbf{x}_t^M$ when constrained to the set $\mathcal{K}$ defined by:

$$\mathcal{K} = \{x | a_i^\top x \leq b_i \text{ for } i = 1, \ldots, m\}, \tag{15}$$

where a_i are unit vectors.

As discussed in Section 3.3, the bound in Lemma 6 is tighter than the corresponding results in earlier works [9, 27]. The improvement arises from the use of Theorem 9 below, which utilizes the explicit polyhedral structure of $\mathcal{K}$ to achieve a tighter bound than could be obtained for general convex constraint sets. It is a variation on an earlier result from [18]. The main distinction is that the proof in [18] is non-constructive, and so there is no way to calculate the constants, whereas the proof in Appendix A is fully constructive and the constants can be computed explicitly.

Theorem 9. *If $x = \mathcal{S}(y)$ and $x' = \mathcal{S}(y')$ are Skorokhod solutions on the polyhedral set $\mathcal{K}$, defined by (15), then there are constants, c_9 and $\alpha \in (0, 1/2]$ such that for all $t \geq 0$, the following bound holds:*

$$\sup_{0 \leq s \leq t} \|x_s - x'_s\| \leq (c_9 + 1) \sup_{0 \leq s \leq t} \|y_s - y'_s\|.$$

where

$$c_9 = 6 \left(\frac{1}{\alpha}\right)^{\mathrm{rank}(A)/2}$$

and $A = [a_1 \quad \cdots a_m]^\top$ whose rows are the $a_i^\top$ vectors.

5 Limitations

Our current work is restricted to polyhedral sets. In particular, Theorem 9 requires the polyhedral assumption, and it is unclear if Skorokhod problems satisfy similar bounds on any more general classes of constraint sets. As a result, it is unclear if our main results on projected Langevin algorithms can be extended beyond polyhedra. We also only considered constant step sizes, but in many cases decreasing or adaptive step sizes are used in practice. Finally, the dependence of the external data variables is limited to the class of L-mixing processes, which does not include all the real-world dependent data streams. Furthermore, it can be difficult to check that a data stream is L-mixing without requiring strong assumptions or knowledge about how it is generated.

6 Conclusions and future work

In this paper, we derived non-asymptotic bounds in 1-Wasserstein distance for a constrained Langevin algorithm applied to non-convex functions with dependent data streams satisfying L-mixing assumptions. Our convergence bounds match the best known bounds of the unconstrained case up to logarithmic factors, and improve on all existing bounds from the constrained case. The tighter bounds are enabled by a constructive and explicit bound on Skorokhod solutions, which builds upon an earlier non-constructive bound from [18]. The analysis of L-mixing variables followed by a comparatively simple averaging method. Future work will examine extensions beyond polyhedral domains, higher-order Langevin algorithms, alternative approaches to handling constraints, such as mirror descent, and more sophisticated step size rules. More specifically, future work will examine whether the projection step, and thus Skorkhod problems, can be circumvented by utilizing different algorithms, such as those based on proximal LMC [7]. Additionally, applications to real-world problems such as time-series analysis and adaptive control will be studied.

References

[1] Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. *arXiv preprint arXiv:2010.16212*, 2020.

[2] SV Anulova and B Sh Liptser. Diffusional approximation for processes with the normal reflection. *Theory of Probability & Its Applications*, 35(3):411–423, 1991.

[3] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

[4] Mathias Barkhagen, Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, Ying Zhang, et al. On stochastic gradient langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

[5] Vivek S Borkar and Sanjoy K Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of optimization theory and applications*, 100(3):499–513, 1999.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. *arXiv preprint arXiv:1705.08964*, 2017.

[8] Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected langevin monte carlo. *Advances in Neural Information Processing Systems*, 28:1243–1251, 2015.

[9] 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.

[10] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint arXiv:1905.13142*, 2019.

[11] Xi Chen, Simon S Du, and Xin T Tong. On stationary-point hitting time and ergodicity of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 21(68):1–41, 2020.

[12] 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.

[13] Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of langevin monte carlo from poincar\'e to log-sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

[14] William Coffey and Yu P Kalmykov. *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*, volume 27. World Scientific, 2012.

[15] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

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

[17] 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.

[18] Paul Dupuis and Hitoshi Ishii. On lipschitz continuity of the solution mapping to the skorokhod problem, with applications. *Stochastics: An International Journal of Probability and Stochastic Processes*, 35(1):31–62, 1991.

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

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

[21] 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.

[22] Murat A Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. Convergence of langevin monte carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 8151–8175. PMLR, 2022.

[23] Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in rˆd. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

[24] László Gerencsér. On a class of mixing processes. *Stochastics: An International Journal of Probability and Stochastic Processes*, 26(3):165–191, 1989.

[25] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. *Advances in Neural Information Processing Systems*, 31:2878–2887, 2018.

[26] Walid Krichene and Peter L Bartlett. Acceleration and averaging in stochastic mirror descent dynamics. *arXiv preprint arXiv:1707.06219*, 2017.

[27] Andrew Lamperski. Projected stochastic gradient langevin algorithms for constrained sampling and non-convex learning. In *Conference on Learning Theory*, pages 2891–2937. PMLR, 2021.

[28] Joseph Lehec. The langevin monte carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*, 2021.

[29] Tyler Lekang and Andrew Lamperski. Wasserstein contraction bounds on closed convex domains with applications to stochastic adaptive control. In *IEEE Conference on Decision and Control (CDC)*, 2021. Accepted, available at https://arxiv.org/abs/2109.12198.

[30] Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt (d) dimension dependence of langevin monte carlo. *arXiv preprint arXiv:2109.03839*, 2021.

[31] Torgny Lindvall and LCG Rogers. Coupling of multidimensional diffusions by reflection. *The Annals of Probability*, pages 860–872, 1986.

[32] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42): 20881–20885, 2019.

[33] Mateusz B Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *arXiv preprint arXiv:1808.07105*, 2018.

[34] Dao Nguyen, Xin Dang, and Yixin Chen. Unadjusted langevin algorithm for non-convex weakly smooth potentials. *arXiv preprint arXiv:2101.06369*, 2021.

[35] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

[36] Srinivasa Ramanujan. The lost notebook and other unpublished papers. *Bull. Am. Math. Soc*, 19:558–560, 1988.

[37] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[38] Ralph Tyrell Rockafellar. *Convex Analysis*, volume 36. Princeton University Press, 2015.

[39] Kanji Sato, Akiko Takeda, Reiichiro Kawai, and Taiji Suzuki. Convergence error analysis of reflected gradient langevin dynamics for globally optimizing non-convex constrained problems. *arXiv preprint arXiv:2203.10215*, 2022.

[40] Hiroshi Tanaka. Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Mathematical Journal*, 9(1):163–177, 1979.

[41] Xiao Wang, Qi Lei, and Ioannis Panageas. Fast convergence of langevin dynamics on manifold: Geodesics meet log-sobolev. *Advances in Neural Information Processing Systems*, 33, 2020.

[42] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[43] 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 3126–3137, 2018.

[44] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. *arXiv preprint arXiv:2002.04363*, 2020.

[45] Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pages 1152–1162. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.3 and Section 2.4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3 and Appendix A, B, C, D, E, F, G.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Background and results on Skorokhod problems

In this section, we will show that when the domain is a polyhedron, rather tight bounds on solutions to Skorokhod problems can be obtained.

A.1 Background on Skorokhod problems

Let $\mathcal{K}$ be a convex subset of $\mathbb{R}^n$ with non-empty interior. Let $y : [0, \infty) \to \mathbb{R}^n$ be a trajectory which is right-continuous with left limits and has $y_0 \in \mathcal{K}$. For each $x \in \mathbb{R}^n$, let $N_{\mathcal{K}}(x)$ be the normal cone at x. Then the functions x_t and ϕ_t solve the *Skorokhod problem* for y_t if the following conditions hold:

- $x_t = y_t + \phi_t \in \mathcal{K}$ for all $t \in [0, T)$.
- The function ϕ has the form $\phi(t) = -\int_0^t v_s d\mu(s)$, where $\|v_s\| \in \{0, 1\}$ and $v_s \in N_{\mathcal{K}}(x_s)$ for all $s \in [0, T)$, while the measure, μ, satisfies $\mu([0, T)) < \infty$ for any $T > 0$.

It can be shown that if a solution exists, it is unique. See [40]. However, existence of solutions typically relies on extra requirements beyond just convexity. For example, [40] showed the existence of solutions in the case that y is continuous and $\mathcal{K}$ is compact. Below, we will utilize results from [2] to prove existence in the case that $\mathcal{K}$ is a polyhedron. Whenever solutions are guaranteed to exist, uniqueness implies that we may view the Skorokhod solution as a mapping: $x = \mathcal{S}(y)$.

A.2 Existence of solutions over polyhedra

The following is a consequence of Theorem 4 from [2].

Lemma 10. *Let $\mathcal{K}$ be a polyhedron with non-empty interior. If y_t is a trajectory in $\mathbb{R}^n$ which is right-continuous with left-limits, then $x = \mathcal{S}(y)$ exists, is unique, and is right-continuous with left-limits.*

Proof To verify the conditions of Theorem 4 from [2], we just need to show that $\mathcal{K}$ satisfies condition β of that paper, which states that there exist constants $\epsilon > 0$ and $\bar{\delta} > 0$ such that for all $x \in \partial\mathcal{K}$, there exist $x_0 \in \mathcal{K}$ such that $\|x - x_0\| \leq \bar{\delta}$ and $\{y | \|y - x_0\| < \epsilon\} \subset \mathcal{K}$. We will show how to construct ϵ, $\bar{\delta}$, and we will see that a suitable vector, x_0, exists for any $x \in \mathcal{K}$.

Note that since $\mathcal{K}$ is a polyhedron, there are vectors $u_1, \ldots, u_p$ such that $x \in \mathcal{K}$ if and only if it can be expressed as

$$x = \sum_{i=1}^{k} \lambda_i u_i + \sum_{i=k+1}^{p} \lambda_i u_i$$

with $\lambda_i \geq 0$ for $i = 1, \ldots, p$ and $\sum_{i=1}^{k} \lambda_i = 1$. See [38]. (If $p = k$, then $\mathcal{K}$ is a compact polytope, while if $k = 0$, then $\mathcal{K}$ is a convex cone.)

Let $x^\star$ be an arbitrary point in the interior of $\mathcal{K}$ and let $\epsilon > 0$ be such that $\{y | \|y - x^\star\| < \epsilon\} \subset \mathcal{K}$.

Pick $\bar{\delta}$ such that $\|u_i - x^\star\| \leq \bar{\delta}$ for $i = 1, \ldots, k$.

For any $x = \sum_{i=1}^{p} \lambda_i u_i \in \mathcal{K}$, let $x_0 = x^\star + \sum_{i=k+1}^{p} \lambda_i u_i$. It follows that

$$\begin{aligned}
\|x - x_0\| &= \left\| \sum_{i=1}^{k} \lambda_i u_i - x^\star \right\| \\
&= \left\| \sum_{i=1}^{k} \lambda_i (u_i - x^\star) \right\| \\
&\leq \sum_{i=1}^{k} \lambda_i \|u_i - x^\star\| \\
&\leq \bar{\delta}.
\end{aligned}$$

Also, if $y \in \{y | \|y - x_0\| < \epsilon\}$, then there is a vector, v, with $\|v\| < \epsilon$, such that

$$y = x_0 + v = (x^\star + v) + \sum_{i=k+1}^{p} \lambda_i u_i.$$

Now note that $x^\star + v \in \mathcal{K}$, so there must be numbers $\lambda_i' \geq 0$ such that $\sum_{i=1}^{k} \lambda_i' = 1$ and $x^\star + v = \sum_{i=1}^{p} \lambda_i' u_i$. It follows that

$$y = \sum_{i=1}^{p} \lambda_i' u_i + \sum_{i=k+1}^{p} \lambda_i u_i = \sum_{i=1}^{k} \lambda_i' u_i + \sum_{i=k+1}^{p} (\lambda_i + \lambda_i') u_i \in \mathcal{K}.$$

■

A.3 Proof of Theorem 9

In this subsection, we provide a short proof of Theorem 9. A supporting Lemma is firstly presented to complete the proof.

The technical work in this subsection relies on some notation about the vectors defining $\mathcal{K}$ from (15). Let $A = [a_1 \quad \cdots a_m]^\top$ be the matrix whose rows are the $a_i^\top$ vectors. For $\mathcal{I} \subset \{1, \ldots, m\}$ let $A_\mathcal{I}$ be the matrix whose rows are $a_i^\top$ for $i \in \mathcal{I}$. Let $[W_\mathcal{I} \quad V_\mathcal{I}]$ be an orthogonal matrix such that $\mathcal{N}(A_\mathcal{I}) = \mathcal{R}(W_\mathcal{I})$. Here $\mathcal{N}(A_\mathcal{I})$ denotes the null space of $A_\mathcal{I}$ and $\mathcal{R}(W_\mathcal{I})$ denotes the range space of $W_\mathcal{I}$. Let $P_\mathcal{I} = W_\mathcal{I} W_\mathcal{I}^\top$, which is the orthogonal projection onto $\mathcal{N}(A_\mathcal{I})$. We will use the convention that $A_\emptyset$ is a $1 \times n$ matrix of zeros, so that $\mathcal{N}(A_\emptyset) = \mathbb{R}^n$, and thus $P_\emptyset = I$.

The following lemma is a quantitative and explicit version of Theorem 2.1 of [18]:

Lemma 11. *If $\mathcal{K}$ is a polyhedron defined by (15), then there is a compact, convex set $\mathcal{B}$ with $0 \in \operatorname{int}(\mathcal{B})$ such that if $z \in \partial\mathcal{B}$, $v \in N_\mathcal{B}(z)$, and a_j is a unit vector from (15) with $a_j^\top v \neq 0$, then*

1. $|a_j^\top z| \geq 1$
2. $\operatorname{sign}(a_j^\top z) = \operatorname{sign}(a_j^\top v)$.

Furthermore, the diameter of $\mathcal{B}$ is at most c_9, defined by

$$c_9 = 6 \left(\frac{1}{\alpha}\right)^{\operatorname{rank}(A)/2}$$

where

$$\alpha = \frac{1}{2} \min \left\{ \|P_\mathcal{I} a_j\|^2 \middle| P_\mathcal{I} a_j \neq 0,\ \mathcal{I} \subset \{1, \ldots, m\},\ j \in \{1, \ldots, m\} \right\},$$

and $\alpha \in (0, 1/2]$.

A non-constructive proof of the existence of $\mathcal{B}$ was given in [18]. While that paper shows that $\mathcal{B}$ is compact, it does not quantitatively bound its diameter. The diameter of $\mathcal{B}$ is precisely the quantity that is used to bound the difference between Skorokhod solutions.

Proof of Theorem 9. Theorem 2.2 of [18] shows that if a compact convex set with $0 \in \operatorname{int}(B)$ satisfying conditions 1 and 2 exists, then

$$\sup_{0 \leq s \leq t} \|x_s - x_s'\| \leq (\operatorname{diameter}(\mathcal{B}) + 1) \sup_{0 \leq s \leq t} \|y_s - y_s'\|.$$

The result now follows since c_9 is an upper bound on the diameter of the set $\mathcal{B}$ constructed in Lemma 11. ■

Proof of Lemma 11. We will focus on constructing a compact, convex $\mathcal{B}$ with $0 \in \operatorname{int}(\mathcal{B})$ which satisfies condition 1. Lemma 2.1 of [18] shows that condition 2 must also hold. (Note that the sign is opposite of what appears in [18], because that paper examines inward normal vectors, while we are examining outward normal vectors.)

We will find numbers $\epsilon \in (0,1)$ and $r_{\mathcal{I}} \in (0,1)$ for $\mathcal{I} \subset \{1,\dots,m\}$ such that

$$\mathcal{B} = \{x | \|P_{\mathcal{I}} x\| \le \epsilon^{-1} r_{\mathcal{I}} \text{ for } \mathcal{I} \in \{1,\dots,m\}\}$$

has the desired properties. By construction, $\mathcal{B}$ is compact and convex, $0 \in \operatorname{int}(\mathcal{B})$, and the diameter is at most $2\epsilon^{-1} r_{\emptyset} < 2\epsilon^{-1}$, since every $x \in \mathcal{B}$ satisfies $\|P_{\emptyset} x\| = \|x\| \le \epsilon^{-1} r_{\emptyset}$. Furthermore, $\mathcal{B} = \epsilon^{-1}\hat{\mathcal{B}}$, where

$$\hat{\mathcal{B}} = \{x | \|P_{\mathcal{I}} x\| \le r_{\mathcal{I}} \text{ for } \mathcal{I} \subset \{1,\dots,m\}\}.$$

A similar construction for $\mathcal{B}$ was utilized in [18]. The main distinction is that this proof will give an explicit procedure for determining the values of ϵ and $r_{\mathcal{I}}$.

Note that $z \in \hat{\mathcal{B}}$ if and only if $\epsilon^{-1} z \in \mathcal{B}$, $z \in \partial\hat{\mathcal{B}}$ if and only if $\epsilon^{-1} z \in \partial\mathcal{B}$, and $N_{\hat{\mathcal{B}}}(z) = N_{\mathcal{B}}(\epsilon^{-1} z)$. Thus, Condition 1 holds for $\mathcal{B}$ if and only if

$$z \in \partial\hat{\mathcal{B}},\ v \in N_{\hat{\mathcal{B}}}(z),\ \text{ and } a_j^\top v \neq 0 \implies |a_j^\top z| \ge \epsilon > 0. \tag{16}$$

Note that if $x \in \partial\hat{\mathcal{B}}$, then

$$\begin{aligned} N_{\hat{\mathcal{B}}}(x) &= \operatorname{cone}\{P_{\mathcal{I}} x | \|P_{\mathcal{I}} x\| = r_{\mathcal{I}}\} \\ &= \left\{ \sum_{\{\mathcal{I} | \|P_{\mathcal{I}} x\| = r_{\mathcal{I}}\}} \lambda_{\mathcal{I}} P_{\mathcal{I}} x \,\middle|\, \lambda_{\mathcal{I}} \ge 0 \right\}. \end{aligned} \tag{17}$$

See Corollary 23.8.1 of [38].

The representation in (17) implies that if $x \in \partial\hat{\mathcal{B}}$, $v \in N_{\mathcal{B}}(x)$, and $a_j^\top v \neq 0$, then there must be a set $\mathcal{I}$ such that, $\|P_{\mathcal{I}} x\| = r_{\mathcal{I}}$, $\lambda_{\mathcal{I}} > 0$, and $a_j^\top P_{\mathcal{I}} \neq 0$. We will choose ϵ such that for all $\mathcal{I}$ and j with $P_{\mathcal{I}} a_j \neq 0$, ϵ is a lower bound on the optimal value of the following (non-convex) optimization problem:

$$\begin{aligned} \min_x \quad & |a_j^\top x| && \text{(18a)} \\ \text{subject to} \quad & \|P_{\mathcal{I}} x\| \ge r_{\mathcal{I}} && \text{(18b)} \\ & \|P_{\mathcal{I}\cup\{j\}} x\| \le r_{\mathcal{I}\cup\{j\}} && \text{(18c)} \\ & \|x\| \le 1. && \text{(18d)} \end{aligned}$$

By construction, if $x \in \partial\hat{\mathcal{B}}$, $v \in N_{\mathcal{B}}(x)$, and $a_j^\top v \neq 0$, there must be some $\mathcal{I}$ such that x is feasible for (18). As a result, we must have that $|a_j^\top x| \ge \epsilon$. Thus, the implication from (16) will hold, provided that the values of $r_{\mathcal{I}}$ can be chosen so that all of the problems of the form (18) have strictly positive optimal values.

The rest of the proof proceeds as follows. First we derive conditions on $r_{\mathcal{I}}$ that ensure that the problems from (18) always have positive optimal values. Next, we compute specific values of $r_{\mathcal{I}}$ that satisfy these conditions. Finally, we use those values of $r_{\mathcal{I}}$ to compute ϵ, the desired lower bound on the optimal value of (18).

We now assume that $r_{\mathcal{I}}, r_{\mathcal{I}\cup\{j\}} \in (0,1)$ and derive sufficient conditions to make the optimal value in (18) strictly positive.

To derive the optimal value of (18), we need a few basic facts:

- If $\mathcal{I} \subset \mathcal{J}$, then $P_{\mathcal{J}} P_{\mathcal{I}} = P_{\mathcal{J}}$ and $P_{\mathcal{I}} P_{\mathcal{J}} = P_{\mathcal{J}}$.
- The matrix $\begin{bmatrix} W_{\mathcal{I}\cup\{j\}} & \frac{P_{\mathcal{I}} a_j}{\|P_{\mathcal{I}} a_j\|} & V_{\mathcal{I}} \end{bmatrix}$ is orthogonal.

First we show that $\mathcal{I} \subset \mathcal{J}$ implies that $P_{\mathcal{J}} P_{\mathcal{I}} = P_{\mathcal{J}}$. Symmetry of the projection matrices would then imply that $P_{\mathcal{I}} P_{\mathcal{J}} = P_{\mathcal{J}}$. Note that $P_{\mathcal{I}} = I - V_{\mathcal{I}} V_{\mathcal{I}}^\top$, where

$$\mathcal{R}(V_{\mathcal{I}}) = \mathcal{R}(P_{\mathcal{I}})^\perp = \mathcal{N}(A_{\mathcal{I}})^\perp \subset \mathcal{N}(A_{\mathcal{J}})^\perp = \mathcal{R}(P_{\mathcal{J}})^\perp.$$

It follows that $P_{\mathcal{J}} V_{\mathcal{I}} = 0$ and thus $P_{\mathcal{J}} P_{\mathcal{I}} = P_{\mathcal{J}}$.

Now we will show that $P_{\mathcal{I}} a_j \in \mathcal{R}(P_{\mathcal{I}}) \backslash \mathcal{R}(P_{\mathcal{I}\cup\{j\}})$. By construction, $P_{\mathcal{I}} a_j \in \mathcal{R}(P_{\mathcal{I}})$. Also, we have that $P_{\mathcal{I}\cup\{j\}} P_{\mathcal{I}} a_j = P_{\mathcal{I}\cup\{j\}} a_j = 0$, where the second equality follows because $a_j \in \mathcal{N}(A_{\mathcal{I}\cup\{j\}})^{\perp} = \mathcal{R}(P_{\mathcal{I}\cup\{j\}})^{\perp}$. Thus, we have that $P_{\mathcal{I}} a_j \neq P_{\mathcal{I}\cup\{j\}} P_{\mathcal{I}} a_j$. Now, if $P_{\mathcal{I}} a_j \in \mathcal{R}(P_{\mathcal{I}\cup\{j\}})$, then $P_{\mathcal{I}} a_j = P_{\mathcal{I}\cup\{j\}} z$ for some vector z. But then $P_{\mathcal{I}\cup\{j\}}^2 = P_{\mathcal{I}\cup\{j\}}$ would imply that $P_{\mathcal{I}\cup\{j\}} P_{\mathcal{I}} a_j = P_{\mathcal{I}\cup\{j\}} z = P_{\mathcal{I}} a_j$, which gives a contradiction. Thus, $P_{\mathcal{I}} a_j \notin \mathcal{R}(P_{\mathcal{I}\cup\{j\}})$.

Now the rank nullity theorem implies that

$$\begin{aligned} \operatorname{rank}(A_{\mathcal{I}}) &= n - \dim(\mathcal{N}(A_{\mathcal{I}})) \\ \operatorname{rank}(A_{\mathcal{I}\cup\{j\}}) &= n - \dim(\mathcal{N}(A_{\mathcal{I}\cup\{j\}})). \end{aligned}$$

Now since $A_{\mathcal{I}\cup\{j\}}$ has only one more row than $A_{\mathcal{I}}$, we must have that $\operatorname{rank}(A_{\mathcal{I}}) \leq \operatorname{rank}(A_{\mathcal{I}\cup\{j\}}) \leq \operatorname{rank}(A_{\mathcal{I}}) + 1$. Also, $\mathcal{N}(A_{\mathcal{I}\cup\{j\}}) \subset \mathcal{N}(A_{\mathcal{I}})$ by construction, and we just saw that $P_{\mathcal{I}} a_j \in \mathcal{N}(A_{\mathcal{I}}) \setminus \mathcal{N}(A_{\mathcal{I}\cup\{j\}})$, so the inclusion is strict. It follows that

$$\begin{aligned} \dim(\mathcal{N}(A_{\mathcal{I}})) &= \dim(\mathcal{R}(P_{\mathcal{I}})) \\ &= \dim(\mathcal{N}(A_{\mathcal{I}\cup\{j\}})) + 1 \\ &= \dim(\mathcal{R}(P_{\mathcal{I}\cup\{j\}})) + 1. \end{aligned} \tag{19}$$

Now, since $\mathcal{R}(W_{\mathcal{I}\cup\{j\}}) = \mathcal{N}(A_{\mathcal{I}\cup\{j\}})$, we must have that

$$\mathcal{R}\left(\begin{bmatrix} W_{\mathcal{I}\cup\{j\}} & \frac{P_{\mathcal{I}} a_j}{\|P_{\mathcal{I}} a_j\|} \end{bmatrix}\right) = \mathcal{N}(A_{\mathcal{I}}).$$

Furthermore, since $\mathcal{R}(W_{\mathcal{I}\cup\{j\}}) = \mathcal{R}(P_{\mathcal{I}\cup\{j\}})$ and $P_{\mathcal{I}\cup\{j\}} P_{\mathcal{I}} a_j = 0$, we must have that

$$\begin{bmatrix} W_{\mathcal{I}\cup\{j\}}^{\top} \\ \frac{(P_{\mathcal{I}} a_j)^{\top}}{\|P_{\mathcal{I}} a_j\|} \\ V_{\mathcal{I}}^{\top} \end{bmatrix} \begin{bmatrix} W_{\mathcal{I}\cup\{j\}} & \frac{P_{\mathcal{I}} a_j}{\|P_{\mathcal{I}} a_j\|} & V_{\mathcal{I}} \end{bmatrix} = I$$

Now we use this orthogonal matrix to perform a change of coordinates. In particular, let y_1, y_2, and y_3 be such that

$$x = W_{\mathcal{I}\cup\{j\}} y_1 + \frac{P_{\mathcal{I}} a_j}{\|P_{\mathcal{I}} a_j\|} y_2 + V_{\mathcal{I}} y_3.$$

In these new coordinates, (18) is equivalent to

$$\begin{aligned} \min_{y} \quad & |\|P_{\mathcal{I}} a_j\| y_2 + a_j^{\top} V_{\mathcal{I}} y_3| && \text{(20a)} \\ \text{subject to} \quad & \|y_1\|^2 + y_2^2 \geq r_{\mathcal{I}}^2 && \text{(20b)} \\ & \|y_1\| \leq r_{\mathcal{I}\cup\{j\}} && \text{(20c)} \\ & \|y_1\|^2 + y_2^2 + \|y_3\|^2 \leq 1. && \text{(20d)} \end{aligned}$$

The equivalence arises because

$$\begin{aligned} a_j^{\top} x &= \|P_{\mathcal{I}} a_j\| y_2 + a_j^{\top} V_{\mathcal{I}} y_3 \\ P_{\mathcal{I}} x &= W_{\mathcal{I}\cup\{j\}} y_1 + \frac{P_{\mathcal{I}} a_j}{\|P_{\mathcal{I}} a_j\|} y_2 \\ P_{\mathcal{I}\cup\{j\}} x &= W_{\mathcal{I}\cup\{j\}} y_1 \end{aligned}$$

along with orthogonality of the corresponding transformation from y to x.

If we choose $r_{\mathcal{I}} > r_{\mathcal{I}\cup\{j\}}$, then we must have

$$y_2^2 \geq r_{\mathcal{I}}^2 - \|y_1\|^2 \geq r_{\mathcal{I}}^2 - r_{\mathcal{I}\cup\{j\}}^2 > 0.$$

Now, if y is feasible, $-y$ is also feasible, and they have the same objective value in (20). So, without loss of generality, we may assume that $y_2 > 0$.

The Cauchy-Schwartz inequality, combined with (20d), implies that

$$\|P_{\mathcal{I}} a_j\| y_2 + a_j^\top V_{\mathcal{I}} y_3 \geq \|P_{\mathcal{I}} a_j\| y_2 - \|V_{\mathcal{I}}^\top a_j\| \sqrt{1 - \|y_1\|^2 - y_2^2}. \tag{21}$$

Note that this bound is achieved by setting $y_3 = -\frac{V_{\mathcal{I}}^\top a_j}{\|V_{\mathcal{I}}^\top a_j\|} \sqrt{1 - \|y_1\|^2 - y_2^2}$.

The right side of (21) is monotonically increasing in y_2. So, (20b) implies that it is minimized over y_2 by setting $y_2 = \sqrt{r_{\mathcal{I}}^2 - \|y_1\|^2}$. This leads to a lower bound of the form:

$$\|P_{\mathcal{I}} a_j\| y_2 - \|V_{\mathcal{I}}^\top a_j\| \sqrt{1 - \|y_1\|^2 - y_2^2} \geq \|P_{\mathcal{I}} a_j\| \sqrt{r_{\mathcal{I}}^2 - \|y_1\|^2} - \|V_{\mathcal{I}}^\top a_j\| \sqrt{1 - r_{\mathcal{I}}^2}.$$

The right side is now monotonically decreasing with respect to $\|y_1\|$, and so it is minimized by setting $\|y_1\| = r_{\mathcal{I}\cup\{j\}}$. This leads to the characterization:

$$\begin{aligned} &\text{Optimal Value of (20)} \\ &= \|P_{\mathcal{I}} a_j\| \sqrt{r_{\mathcal{I}}^2 - r_{\mathcal{I}\cup\{j\}}^2} - \|V_{\mathcal{I}}^\top a_j\| \sqrt{1 - r_{\mathcal{I}}^2} \\ &= \|P_{\mathcal{I}} a_j\| \sqrt{r_{\mathcal{I}}^2 - r_{\mathcal{I}\cup\{j\}}^2} - \sqrt{1 - \|P_{\mathcal{I}}^\top a_j\|^2} \sqrt{1 - r_{\mathcal{I}}^2}. \end{aligned} \tag{22}$$

The second equality follows because

$$\|V_{\mathcal{I}}^\top a_j\|^2 = a_j^\top V_{\mathcal{I}} V_{\mathcal{I}}^\top a_j = a_j^\top (I - P_{\mathcal{I}}) a_j = 1 - \|P_{\mathcal{I}} a_j\|^2.$$

Now, we have that the right side of (22) is positive if and only if:

$$\|P_{\mathcal{I}} a_j\|^2 \left(r_{\mathcal{I}}^2 - r_{\mathcal{I}\cup\{j\}}^2 \right) > \left(1 - \|P_{\mathcal{I}} a_j\|^2\right) \left(1 - r_{\mathcal{I}}^2\right) \tag{23a}$$

$$\iff r_{\mathcal{I}}^2 > 1 - \|P_{\mathcal{I}} a_j\|^2 + \|P_{\mathcal{I}} a_j\|^2 r_{\mathcal{I}\cup\{j\}}^2 \tag{23b}$$

$$\iff r_{\mathcal{I}}^2 > 1 - \|P_{\mathcal{I}} a_j\|^2 (1 - r_{\mathcal{I}\cup\{j\}}^2) \tag{23c}$$

$$\iff r_{\mathcal{I}}^2 > r_{\mathcal{I}\cup\{j\}}^2 + (1 - \|P_{\mathcal{I}} a_j\|^2)(1 - r_{\mathcal{I}\cup\{j\}}^2). \tag{23d}$$

Note that (23d) implies that $r_{\mathcal{I}} > r_{\mathcal{I}\cup\{j\}}$ holds.

Also note that any collection of $r_{\mathcal{I}}$ values in $(0, 1)$ that satisfy (23) will ensure that the corresponding set, $\hat{\mathcal{B}}$, satisfies the implication from (16). In that case, we have that $\mathcal{B}$ has the desired properties.

Now we seek a simpler, more explicit formula for the $r_{\mathcal{I}}$ values which satisfy (23). Note that (23c) implies that the right side is monotonically decreasing with respect to $\|P_{\mathcal{I}} a_j\|^2$. So, if $\alpha > 0$ is a number such that $\alpha \leq \frac{1}{2}\|P_{\mathcal{I}} a_j\|^2$ for all $\mathcal{I}$ and j with $P_{\mathcal{I}} a_j \neq 0$ we obtain a sufficient condition for (23):

$$r_{\mathcal{I}}^2 = 1 - \alpha(1 - r_{\mathcal{I}\cup\{j\}}^2) \tag{24a}$$

$$r_{\mathcal{I}}^2 = (1 - \alpha) + \alpha r_{\mathcal{I}\cup\{j\}}^2. \tag{24b}$$

Now we use (24) to derive the desired formula for $r_{\mathcal{I}}$. In particular, consider the recursion

$$x_{k+1} = (1 - \alpha) + \alpha x_k.$$

This has an explicit solution given by

$$x_k = \alpha^k x_0 + 1 - \alpha^k = 1 - \alpha^k (1 - x_0).$$

In particular, if $x_0 \in (0, 1)$, we have that $x_k \in (0, 1)$ for all $k \geq 0$.

We define $r_{\mathcal{I}}$ by fixing a value $x_0 \in (0, 1)$, which will be defined explicitly later, and setting $r_{\mathcal{I}}^2 = x_k = 1 - \alpha^k (1 - x_0)$ if $\operatorname{rank}(A) - \operatorname{rank}(A_{\mathcal{I}}) = k$.

To see that this definition satisfies (24), first note that $r_{\mathcal{I}}^2 = x_0$ for all $\mathcal{I}$ with $\operatorname{rank}(A) = \operatorname{rank}(A_{\mathcal{I}})$. Now, recall that if $P_{\mathcal{I}} a_j \neq 0$, then (19) implies that $\operatorname{rank}(A_{\mathcal{I}\cup\{j\}}) = \operatorname{rank}(A_{\mathcal{I}}) + 1$. The converse is also true: If $\operatorname{rank}(A_{\mathcal{I}\cup\{j\}}) = \operatorname{rank}(A_{\mathcal{I}}) + 1$, then we must have that $a_j \notin \mathcal{R}(A_{\mathcal{I}}^\top) = \mathcal{R}(V_{\mathcal{I}}) =$

$\mathcal{R}(P_\mathcal{I})^\perp$. It follows that $P_\mathcal{I} a_j \neq 0$. Thus, if $\text{rank}(A) - \text{rank}(A_{\mathcal{I}\cup\{j\}}) = k \geq 0$, we have that $P_\mathcal{I} a_j \neq 0$ precisely when $\text{rank}(A) - \text{rank}(A_\mathcal{I}) = k+1$. So we see that setting $r_\mathcal{I}^2 = x_{k+1} = 1 - \alpha(1 - x_k)$ gives the same value as specified in (24).

The final step in the proof requires finding a lower bound, ϵ, for the optimal value from (22). Let $r_\mathcal{I}^2 = x_k$ and $r_{\mathcal{I}\cup\{j\}}^2 = x_{k-1}$. Then we have that

$$\begin{aligned} r_\mathcal{I}^2 - r_{\mathcal{I}\cup\{j\}}^2 &= (1-\alpha)\alpha^{k-1}(1-x_0) \\ 1 - r_\mathcal{I}^2 &= \alpha\alpha^{k-1}(1-x_0). \end{aligned}$$

Also note that the right side of (22) is monotonically increasing with respect to $\|P_\mathcal{I} a_j\|^2$ and that $\|P_\mathcal{I} a_j\|^2 \geq 2\alpha$ by our choice of α. So, plugging in this lower bound gives

$$\begin{aligned} &\|P_\mathcal{I} a_j\| \sqrt{r_\mathcal{I}^2 - r_{\mathcal{I}\cup\{j\}}^2} - \sqrt{1 - \|P_\mathcal{I}^\top a_j\|^2}\sqrt{1 - r_\mathcal{I}^2} \\ &\geq \left(\sqrt{2\alpha}\sqrt{1-\alpha} - \sqrt{1-2\alpha}\sqrt{\alpha}\right)\sqrt{\alpha^{k-1}(1-x_0)} \\ &= \left(\sqrt{2-2\alpha} - \sqrt{1-2\alpha}\right)\sqrt{\alpha^k(1-x_0)} \\ &\geq \left(\sqrt{2}-1\right)\sqrt{\alpha^{\text{rank}(A)}(1-x_0)} \end{aligned}$$

The final inequality follows because $k \leq \text{rank}(A)$ and the minimum value of $\sqrt{2-2\alpha} - \sqrt{1-2\alpha}$ over $\alpha \in [0, \|P_\mathcal{I} a_j\|^2/2] \subset [0, 1/2]$ occurs at $\alpha = 0$.

To simplify the final formula for ϵ, note that $\sqrt{2} - 1 > 1/3$, and thus we can choose $x_0 \in (0,1)$ so that

$$(\sqrt{2}-1)\sqrt{1-x_0} = \frac{1}{3} \iff x_0 = 1 - \frac{1}{9\left(\sqrt{2}-1\right)^2} \approx 0.352.$$

Plugging in this value for x_0 gives the bound:

$$\text{Optimal Value of (20)} \geq \frac{1}{3}\alpha^{\frac{\text{rank}(A)}{2}} =: \epsilon$$

Now recalling that the diameter of $\mathcal{B}$ is at most $2/\epsilon$ completes the proof. ■

B Invariance of the Gibbs measure

Lemma 12. *The Gibbs measure, (6), is stationary under the dynamics of the reflected SDE from (8).*

Proof Before showing invariance of the Gibbs measure, we first remark that it is well-defined. In particular, we have that $\int_\mathcal{K} e^{-\beta \bar{f}(x)} dx < \infty$.

To see this, let $\|x\| \geq R/\theta$, where $\theta \in (0,1)$ is a number to be chosen later. Note that for $t \in [\theta, 1]$, we have that $\|\theta x\| \geq R$. So, we can use strong convexity outside a ball of radius R to show

$$\begin{aligned} \bar{f}(x) &\geq \bar{f}(0) + \int_0^1 \nabla \bar{f}(tx)^\top x dt \\ &= \bar{f}(0) + \nabla \bar{f}(0)^\top x + \int_0^\theta \left(\nabla \bar{f}(tx) - \nabla \bar{f}(0)\right)^\top x dt + \int_{1-\theta}^1 \left(\nabla \bar{f}(tx) - \nabla \bar{f}(0)\right)^\top x dt \\ &\geq \bar{f}(0) - \|\nabla \bar{f}(0)\|\|x\| - \ell\|x\|^2 \int_0^\theta t dt + \mu\|x\|^2 \int_\theta^1 t dt \\ &\geq \bar{f}(0) - \|\nabla \bar{f}(0)\|\|x\| + \frac{1}{2}\|x\|^2\left(-\ell\theta^2 + \mu(1-\theta^2)\right) \end{aligned}$$

The coefficient $-\ell^2\theta^2 + \mu(1-\theta^2)$ is positive, as long as $\theta < \sqrt{\frac{\mu}{\mu+\ell}}$. In particular, choosing $\theta^2 = \frac{1}{2}\frac{\mu}{\mu+\ell}$ gives

$$\bar{f}(x) \geq \bar{f}(0) - \|\nabla \bar{f}(0)\|\|x\| + \frac{1}{4}\mu\|x\|^2. \tag{25}$$

It follows that $Z = \int_{\mathcal{K}} e^{-\beta \bar{f}(x)} dx < \infty$.

In [27], it was shown in that the Gibbs measure is invariant under (8) when $\mathcal{K}$ is compact. We will extend the result to non-compact $\mathcal{K}$ via a limiting argument.

Let $\mathcal{K}_i = \mathcal{K} \cap \{x \in \mathbb{R}^n | \|x\|_\infty \leq i\}$. Let $Z_i = \int_{\mathcal{K}_i} e^{-\beta \bar{f}(x)} dx$. Note that $\lim_{i\to\infty} Z_i = Z$, by monotone convergence. We choose $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\} \leq i$ so that $\mathcal{K}_i$ becomes a compact polyhedron for $i \geq 1$.

Let $\mathbf{x}_t^C$ be a solution to the original form of (8) and let $\mathbf{x}_t^{C,i}$ be a solution to the RSDE from (8), with $\mathcal{K}_i$ used in place of $\mathcal{K}$. Since $\mathcal{K}_i$ is polyhedral, Lemma 10 in Appendix A shows that $\mathbf{x}_t^{C,i}$ is uniquely defined. Define the diffusion operators P and P^i by

$$(P_t g)(x) = \mathbb{E}[g(\mathbf{x}_t^C)|\mathbf{x}_0 = x]$$
$$(P_t^i g)(x) = \mathbb{E}[g(\mathbf{x}_t^{C,i})|\mathbf{x}_0 = x]$$

Let $L_2(\mathcal{K}, \pi_{\beta\bar{f}})$ be the set of functions $g : \mathcal{K} \to \mathbb{R}$ which are square integrable with respect to the measure $\pi_{\beta\bar{f}}$. We will show that $\pi_{\beta\bar{f}}$ is invariant for (8) by showing that for all $g \in L_2(\mathcal{K}, \pi_{\beta\bar{f}})$ the following equality holds for all $t \geq 0$:

$$\frac{1}{Z}\int_{\mathcal{K}} g(x)e^{-\beta\bar{f}(x)}dx = \frac{1}{Z}\int_{\mathcal{K}} (P_t g)(x)e^{-\beta\bar{f}(x)}dx. \tag{26}$$

The subset of bounded, compactly supported functions in $L_2(\mathcal{K}, \pi_{\beta\bar{f}})$ is a dense subset. Fix an arbitary bounded, compactly supported $g \in L_2(\mathcal{K}, \pi_{\beta\bar{f}})$. It suffices to show that (26) holds for g.

Lemma 19 of [27] shows that for all $i \geq 1$, the following holds:

$$\frac{1}{Z_i}\int_{\mathcal{K}_i} g(x)e^{-\beta\bar{f}(x)}dx = \frac{1}{Z_i}\int_{\mathcal{K}_i} (P_t^i g)(x)e^{-\beta\bar{f}(x)}dx. \tag{27}$$

We saw earlier that $Z_i \to Z$. Furthermore, since g is compactly supported, there is a number, m, such that $i \geq m$ implies that

$$\int_{\mathcal{K}_i} g(x)e^{-\beta\bar{f}(x)}dx = \int_{\mathcal{K}} g(x)e^{-\beta\bar{f}(x)}dx.$$

It follows that the left of (27) converges to the left of (26).

The proof will be completed if we can show that for $t > 0$,

$$\lim_{i\to\infty}\int_{\mathcal{K}_i} (P_t^i g)(x)e^{-\beta\bar{f}(x)}dx = \lim_{i\to\infty}\int_{\mathcal{K}_i} \mathbb{E}[g(\mathbf{x}_t^{C,i})|\mathbf{x}_0 = x]e^{-\beta\bar{f}(x)}dx \tag{28}$$
$$= \int_{\mathcal{K}} \mathbb{E}[g(\mathbf{x}_t^C)|\mathbf{x}_0 = x]e^{-\beta\bar{f}(x)}dx \tag{29}$$
$$= \int_{\mathcal{K}} (P_t g)(x)e^{-\beta\bar{f}(x)}dx.$$

We assumed that g was bounded, and so there is a number, b, such that $|g(x)| \leq b$ for all $x \in \mathcal{K}$. It follows from the definition of P_t and P_t^i that $|P_t g(x)| \leq b$ and $|P_t^i g(x)| \leq b$ for all t.

Fix any $t > 0$. The Brownian motion, $\mathbf{w}$, is continuous, and so for each t, $\mathbf{w}_s$ is bounded for $s \in [0, t]$. Now the form of (9) shows that $\mathbf{y}^C$, and thus $\mathbf{x}^C$ must also be continuous, and thus also bounded for $s \in [0, t]$. Thus, for each realization, we see that there is a number m such that $\mathbf{x}_s^C \in \mathcal{K}_i$ for all $s \in [0, t]$ and all $i \geq m$. Thus, we see that $\mathbf{x}_s^C = \mathbf{x}_s^{C,i}$ for $s \in [0, t]$. This argument shows that the integrand on the right of (28) converges pointwise to the integrand of (29). So, the desired equality follows by the dominated convergence theorem. ■

C Bounded variance of the processes

In this section, we derive variance bounds on all of the main processes, $\mathbb{E}[\|\mathbf{x}_k^A\|^2]$ and $\mathbb{E}[\|\mathbf{x}_t^C\|^2]$. The bound on $\mathbb{E}[\|\mathbf{x}_t^C\|^2]$ is used to prove bounds on the discretization error from $\mathbf{x}_t^M$ to $\mathbf{x}_t^C$. The bound on $\mathbb{E}[\|\mathbf{x}_k^A\|^2]$ is used to derive the time-uniform bounds on $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C))$ from Lemma 3.

C.1 Continuous-time bounds

In this section, we show that the assumption that $\bar{f}$ is strongly convex outside a ball implies that $\mathcal{V}(x) = \frac{1}{2}\|x\|^2$ can be used as a Lyapunov function for $\mathbf{x}_t^C$. In turn, we use this Lyapunov function to derive bounds on $\mathbb{E}[\|\mathbf{x}_t^C\|^2]$.

Lemma 13. *If $\bar{f}(x)$ is μ-strongly convex outside a ball with radius R, then $\mathcal{V}(x) = \frac{1}{2}x^\top x$ satisfies the following the geometric drift condition:*

$$\mathcal{A}\mathcal{V}(x) \leq -2\eta\mu\mathcal{V}(x) + c_{10}\eta.$$

Here c_{10} is defined by

$$c_{10} = (\ell + \mu)R^2 + R\|\nabla_x \bar{f}(0)\| + \frac{n}{\beta}.$$

Proof By Ito's formula, we have

$$\begin{aligned}
d\mathcal{V}(\mathbf{x}_t^C) &= \nabla_x \mathcal{V}^\top d\mathbf{x}_t^C + \frac{1}{2} d(\mathbf{x}_t^C)^\top (\nabla_x^2 \mathcal{V}) d\mathbf{x}_t^C \\
&= (\mathbf{x}_t^C)^\top (-\eta \nabla_x \bar{f}(\mathbf{x}_t^C)dt + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t - \mathbf{v}_t d\boldsymbol{\mu}_t) + \frac{1}{2} d(\mathbf{x}_t^C)^\top d\mathbf{x}_t^C \\
&= -\eta (\mathbf{x}_t^C)^\top \nabla_x \bar{f}(\mathbf{x}_t^C)dt + \sqrt{\frac{2\eta}{\beta}} (\mathbf{x}_t^C)^\top d\mathbf{w}_t - (\mathbf{x}_t^C)^\top \mathbf{v}_t d\boldsymbol{\mu}_t + \frac{\eta}{\beta}\mathrm{Tr}(d\mathbf{w}_t d\mathbf{w}_t^\top) \\
&= (-\eta (\mathbf{x}_t^C)^\top \nabla_x \bar{f}(\mathbf{x}_t^C) + \frac{n\eta}{\beta})dt + \sqrt{\frac{2\eta}{\beta}} (\mathbf{x}_t^C)^\top d\mathbf{w}_t - (\mathbf{x}_t^C)^\top \mathbf{v}_t d\boldsymbol{\mu}_t.
\end{aligned}$$

The third equality holds because $\int_0^t \mathbf{v}_s d\boldsymbol{\mu}_s$ has bounded variation. The last equality is based on the fact that $d\mathbf{w}_t d\mathbf{w}_t^\top = dt\, I$.

Since $\mathbf{v}_t \in N_{\mathcal{K}}(\mathbf{x}_t^C)$, $\boldsymbol{\mu}_t$ is a nonnegative measure, and $0 \in \mathcal{K}$, we have that $-(\mathbf{x}_t^C)^\top \mathbf{v}_t d\boldsymbol{\mu}_t \leq 0$. Thus, the generator of the Lyapunov function satisfies

$$\mathcal{A}\mathcal{V}(x) \leq -\eta x^\top \nabla_x \bar{f}(x) + \frac{n\eta}{\beta}. \tag{30}$$

If $\|x\| \geq R$, strong convexity outside a ball of radius R, along with the Cauchy-Schwartz inequality imply that

$$\begin{aligned}
x^\top \nabla_x \bar{f}(x) &= (x - 0)^\top (\nabla_x \bar{f}(x) - \nabla_x \bar{f}(0)) + x^\top \nabla_x \bar{f}(0) \\
&\geq \mu\|x\|^2 - R\|\nabla_x \bar{f}(0)\|
\end{aligned} \tag{31}$$

It follows that when $\|x\| \geq R$, we have that

$$\begin{aligned}
\mathcal{A}\mathcal{V}(x) &\leq -\eta\mu\|x\|^2 + \eta\left(R\|\nabla_x \bar{f}(0)\| + \frac{n}{\beta}\right) \\
&= -\eta 2\mu\mathcal{V}(x) + \eta\left(R\|\nabla_x \bar{f}(0)\| + \frac{n}{\beta}\right).
\end{aligned}$$

If $\|x\| \leq R$, then the Cauchy-Schwartz inequality and the Lipschitz continuity imply that

$$\begin{aligned}
-x^\top \nabla_x \bar{f}(x) &= -x^\top \left(\nabla_x \bar{f}(x) - \nabla_x \bar{f}(0) + \nabla_x \bar{f}(0)\right) \\
&\leq \|x\|\|\nabla_x \bar{f}(x) - \nabla_x \bar{f}(0)\| + R\|\nabla_x \bar{f}(0)\| \\
&\leq \ell\|x\|^2 + R\|\nabla_x \bar{f}(0)\| \\
&= -\mu\|x\|^2 + (\ell + \mu)\|x\|^2 + R\|\nabla_x \bar{f}(0)\| \\
&\leq -\mu\|x\|^2 + (\ell + \mu)R^2 + R\|\nabla_x \bar{f}(0)\|.
\end{aligned} \tag{32}$$

Note that (31) implies that (32) also holds whenever $\|x\| \geq R$. So, combining (32) with (30) shows that for all $x \in \mathcal{K}$,

$$\begin{aligned}\mathcal{A}\mathcal{V}(x) &\leq \eta\left(-\mu\|x\|^2 + (\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\|\right) + \frac{n\eta}{\beta} \\ &= -\eta 2\mu\mathcal{V}(x) + \eta\left((\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\| + \frac{n}{\beta}\right)\end{aligned}$$

■

Lemma 14. *If $\mathbb{E}[\|\mathbf{x}_0^C\|^2] \leq \varsigma$, then for all $t \geq 0$, we have that*

$$\mathbb{E}[\|\mathbf{x}_t^C\|^2] \leq \varsigma + \frac{1}{\mu}c_{10},$$

where c_{10} is defined in Lemma 13.

Proof Recall that Lyapunov generator $\mathcal{A}$ is defined as below

$$\mathcal{A}\mathcal{V}(x) = \lim_{t\downarrow 0}\mathbb{E}\left[\frac{1}{t}(\mathcal{V}(\mathbf{x}_t^C) - \mathcal{V}(\mathbf{x}_0^C))|\mathbf{x}_0^C = x\right].$$

Using Dynkin's formula and Lemma 13 gives

$$\begin{aligned}\mathbb{E}\left[\mathcal{V}(\mathbf{x}_t^C) - \mathcal{V}(\mathbf{x}_0^C)\right] &= \int_0^t \mathbb{E}\left[\mathcal{A}\mathcal{V}(\mathbf{x}_s^C)\right] ds \\ &\leq -2\eta\mu\int_0^t \mathbb{E}\left[\mathcal{V}(\mathbf{x}_s^C)\right] ds + c_{10}\eta t.\end{aligned}$$

Let $u_t = \mathbb{E}\left[\mathcal{V}(\mathbf{x}_t^C)\right]$, $u_0 = \mathbb{E}\left[\mathcal{V}(\mathbf{x}_0^C)\right]$. By Grönwall's inequality, we get

$$\begin{aligned}u_t &\leq e^{-2\eta\mu t}u_0 + \eta c_{10}\int_0^t e^{-2\eta\mu s}ds \\ &= e^{-2\eta\mu t}u_0 + \frac{c_{10}}{2\mu}\left(1 - e^{-2\eta\mu t}\right) \\ &\leq u_0 + \frac{c_{10}}{2\mu}.\end{aligned}$$

Recalling that $u_t = \frac{1}{2}\mathbb{E}[\|\mathbf{x}_t\|^2]$ and $\mathbb{E}[\|\mathbf{x}_0\|^2] \leq \varsigma$ completes the proof. ■

C.2 Discrete-time bounds

Here we derive a uniform bound on $\mathbb{E}[\|\mathbf{x}_k^A\|^2]$.

Lemma 15. *Assume that $\mathbb{E}[\|\mathbf{x}_0^A\|^2] \leq \varsigma$ and that $\eta \leq \min\left\{1, \frac{\mu}{4\ell^2}\right\}$. There is a constant, c_{11} such that for all $k \geq 0$, we have that*

$$\mathbb{E}[\|\mathbf{x}_k^A\|^2] \leq \varsigma + c_{11}.$$

The constant is given by

$$c_{11} = \frac{4}{\mu}\left(\frac{n}{\beta} + (\ell+\mu)R^2 + (2+R)\|\nabla_x \bar{f}(0)\| + \left(8\ell^2 + \frac{1}{\mu}\right)\ell^2\mathcal{M}_2(\mathbf{z})\right)$$

Proof Using non-expansiveness of the projection and then expanding the square of the norm gives:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{t+1}^A\|^2\right] &= \mathbb{E}\left[\left\|\Pi_{\mathcal{K}}\left(\mathbf{x}_t^A - \eta\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t) + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_t\right) - \Pi_{\mathcal{K}}(0)\right\|^2\right] \\ &\leq \mathbb{E}\left[\left\|\mathbf{x}_t^A - \eta\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t) + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_t\right\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{x}_t^A\|^2 + \eta^2\|\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\|^2 - 2\eta(\mathbf{x}_t^A)^\top\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\right] + \frac{2n\eta}{\beta}.\end{aligned}$$

Now we bound the term $\mathbb{E}\left[\|\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\|^2\right]$. For any $x \in \mathcal{K}$, we have that

$$\begin{aligned}\|\nabla_x f(x,z)\|^2 &= \|\nabla_x f(x,z) - \nabla_x f(0,z) + \nabla_x f(0,z)\|^2 \\ &\leq 2\|\nabla_x f(0,z)\|^2 + 2\ell^2\|x\|^2. \end{aligned} \tag{33}$$

This leads to:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{t+1}^A\|^2\right] \leq \left(1 + 2\ell^2\eta^2\right)\mathbb{E}\left[\|\mathbf{x}_t^A\|^2\right] \\ + \left(\frac{2\eta n}{\beta} + 2\eta^2\mathbb{E}[\|\nabla_x f(0,\mathbf{z}_t)\|^2]\right) - 2\eta\mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\right].\end{aligned}$$

To bound the term $\mathbb{E}[\|\nabla_x f(0,\mathbf{z}_t)\|^2]$, note that $\nabla_x \bar{f}(0) = \mathbb{E}[\nabla_x f(0, \hat{\mathbf{z}}_t)]$, where $\hat{\mathbf{z}}_t$ is identically distributed to $\mathbf{z}_t$ and independent of $\mathbf{z}_t$.

$$\begin{aligned}\mathbb{E}[\|\nabla_x f(0,\mathbf{z}_t)\|^2] &= \mathbb{E}\left[\|\nabla_x \bar{f}(0) + \nabla_x f(0,\mathbf{z}_t) - \mathbb{E}[\nabla_x f(0,\hat{\mathbf{z}}_t)]\|^2\right] \\ &\leq 2\|\nabla_x \bar{f}(0)\|^2 + 2\mathbb{E}[\|\nabla_x f(0,\mathbf{z}_t) - \mathbb{E}[\nabla_x f(0,\hat{\mathbf{z}}_t)]\|^2] \\ &\overset{\text{Jensen}}{\leq} 2\|\nabla_x \bar{f}(0)\|^2 + 2\mathbb{E}[\|\nabla_x f(0,\mathbf{z}_t) - \nabla_x f(0,\hat{\mathbf{z}}_t)\|^2] \\ &\leq 2\|\nabla_x \bar{f}(0)\| + 2\ell^2\mathbb{E}[\|\mathbf{z}_t - \hat{\mathbf{z}}_t\|^2] \\ &\leq 2\|\nabla_x \bar{f}(0)\| + 4\ell^2\mathbb{E}[\|\mathbf{z}_t\|^2 + \|\hat{\mathbf{z}}_t\|^2] \\ &\leq 2\|\nabla_x \bar{f}(0)\| + 8\ell^2\mathcal{M}_2(\mathbf{z}),\end{aligned} \tag{34}$$

where $\mathcal{M}_2(\mathbf{z})$ is a bound on $\mathbb{E}[\|\mathbf{z}_t\|^2]$ from (2).

So, we have a bound of the form

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{t+1}^A\|^2\right] \leq \left(1 + 2\ell^2\eta^2\right)\mathbb{E}\left[\|\mathbf{x}_t^A\|^2\right] \\ + \left(\frac{2\eta n}{\beta} + \eta^2\left(4\|\nabla_x \bar{f}(0)\| + 16\ell^2\mathcal{M}_2(\mathbf{z})\right)\right) - 2\eta\mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\right].\end{aligned} \tag{35}$$

To bound the inner product term, note that

$$\begin{aligned}\mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\right] &= \mathbb{E}\left[(\mathbf{x}_t^A)^\top\left(\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t) - \nabla_x f(\mathbf{x}_t^A, \hat{\mathbf{z}}_t)\right)\right] + \mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x f(\mathbf{x}_t^A, \hat{\mathbf{z}}_t)\right] \\ &= \mathbb{E}\left[(\mathbf{x}_t^A)^\top\left(\nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t) - \nabla_x f(\mathbf{x}_t^A, \hat{\mathbf{z}}_t)\right)\right] + \mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x \bar{f}(\mathbf{x}_t^A)\right].\end{aligned}$$

The second equality follows because $\hat{\mathbf{z}}_t$ is independent of $\mathbf{x}_t^A$ and identically distributed to $\mathbf{z}_t$. So, we can use the Cauchy-Schwartz inequality on the first term on the right and (32) on the second term to give:

$$\mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\right] \geq -\ell\mathbb{E}[\|\mathbf{x}_t^A\|\|\mathbf{z}_t - \hat{\mathbf{z}}_t\|] + \mu\mathbb{E}[\|\mathbf{x}_t^A\|^2] - \left((\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\|\right).$$

Using a completing-the-squares argument shows that for any numbers a and b

$$\begin{aligned}\frac{\mu}{2}a^2 - \ell ab &= \frac{\mu}{2}\left(a - \frac{\ell}{\mu}b\right)^2 - \frac{\ell^2}{2\mu}b^2 \\ &\geq -\frac{\ell^2}{2\mu}b^2.\end{aligned}$$

Setting $a = \|\mathbf{x}_t^A\|$ and $b = \|\mathbf{z}_t - \hat{\mathbf{z}}_t\|$ leads to a bound of the form

$$\begin{aligned}\mathbb{E}\left[(\mathbf{x}_t^A)^\top \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_t)\right] &\geq \frac{\mu}{2}\mathbb{E}[\|\mathbf{x}_t^A\|^2] - \frac{\ell^2}{2\mu}\mathbb{E}[\|\mathbf{z}_t - \hat{\mathbf{z}}_t\|^2] - \left((\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\|\right) \\ &\geq \frac{\mu}{2}\mathbb{E}[\|\mathbf{x}_t^A\|^2] - \frac{\ell^2}{\mu}\mathcal{M}_2(\mathbf{z}) - \left((\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\|\right).\end{aligned} \tag{36}$$

Plugging the new bounds into (35) gives

$$\mathbb{E}\left[\|\mathbf{x}_{t+1}^A\|^2\right] \leq \left(1-\mu\eta+2\ell^2\eta^2\right)\mathbb{E}\left[\|\mathbf{x}_t^A\|^2\right]$$
$$+\left(\frac{2\eta n}{\beta}+\eta^2\left(4\|\nabla_x\bar{f}(0)\|+16\ell^2\mathcal{M}_2(\mathbf{z})\right)\right)+2\eta\left(\frac{\ell^2}{\mu}\mathcal{M}_2(\mathbf{z})+\left((\ell+\mu)R^2+R\|\nabla_x\bar{f}(0)\|\right)\right)$$

Note that if $\eta \leq \frac{\mu}{4\ell^2}$, then

$$1-\mu\eta+2\ell^2\eta^2 \leq 1-\frac{\mu\eta}{2}.$$

Furthermore, if $\eta \leq 1$, we get the simplified bound:

$$\mathbb{E}\left[\|\mathbf{x}_{t+1}^A\|^2\right] \leq \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left[\|\mathbf{x}_t^A\|^2\right]$$
$$+2\eta\left(\frac{n}{\beta}+2\|\nabla_x\bar{f}(0)\|+8\ell^2\mathcal{M}_2(\mathbf{z})+\frac{\ell^2}{\mu}\mathcal{M}_2(\mathbf{z})+(\ell+\mu)R^2+R\|\nabla_x\bar{f}(0)\|\right). \quad (37)$$

Now for any $a \in [0,1)$ and any $b \geq 0$, if $u_t \geq 0$ satisfies

$$u_{t+1} \leq au_t + b$$

then

$$\begin{aligned} u_t &\leq a^t u_0 + b\sum_{k=0}^{t-1} a^k \\ &= a^t u_0 + b\frac{1-a^t}{1-a} \\ &\leq u_0 + \frac{b}{1-a} \end{aligned}$$

Applying this bound to $\mathbb{E}[\|\mathbf{x}_t^A\|^2]$ and using that $\mathbb{E}[\|\mathbf{x}_0^A\|^2] \leq \varsigma$ gives

$$\mathbb{E}[\|\mathbf{x}_t^A\|^2] \leq \varsigma + \frac{4}{\mu}\left(\frac{n}{\beta}+(\ell+\mu)R^2+(2+R)\|\nabla_x\bar{f}(0)\|+\left(8\ell^2+\frac{1}{\mu}\right)\ell^2\mathcal{M}_2(\mathbf{z})\right).$$

■

D Stochastic contraction analysis

In this Appendix, we prove Lemma 2.

D.1 Contraction for the reflected SDEs

We extend the analysis of standard SDEs from [20] to the case of reflected SDEs. The main idea of [20] is to construct a specialized metric over $\mathbb{R}^n$ and corresponding Wasserstein distance under which contraction rates can be computed. In the context of this paper, we only use Euclidean norm to construct the metric, whereas in [20], both Euclidean and a second norm were used to construct the specilized metric. Using just one norm leads to some simplifications. Our choice of reflection term in the coupling process is also slightly different, leading to further simplifications.

In the following, we firstly examine the contractivity properties of the generalized reflected SDEs and then associate the generalized process with the original process from (1).

Let $\mathcal{K}$ be a closed convex subset of $\mathbb{R}^n$ and consider a reflected stochastic differential equations of the form:

$$d\mathbf{x}_t = H(\mathbf{x}_t)dt + Gd\mathbf{w}_t - \mathbf{v}_t d\boldsymbol{\mu}(t), \quad (38)$$

where G is an invertible $n \times n$ matrix with minimum singular value $\sigma_{\min}(G)$, $\mathbf{w}_t$ is a standard Brownian motion, and $-\int_0^t \mathbf{v}_s d\boldsymbol{\mu}(s)$ is a reflection term that ensures that $\mathbf{x}_t \in \mathcal{K}$ for all $t \geq 0$. (We

are slightly abusing notation, since here $\mathbf{x}_t$ denotes the solution to a general RSDE, and is not the iterates of the original algorithm from (1).)

Following [20], we construct a function $\delta : [0, +\infty) \to \mathbb{R}$ such that $\delta(0) = 0$, $\delta'(0) = 1$, $\delta'(r) > 0$, and $\delta''(r) \leq 0$ for all $r \geq 0$. With these properties, it can be shown that $\delta(\|x - y\|)$ forms a metric over $\mathcal{K}$. The particular metric is constructed so that the dynamics are contractive with respect to the corresponding Wasserstein distance.

Assume there exists a continuous function $\kappa(r) : [0, +\infty) \to \mathbb{R}$ such that for any $x, y \in \mathbb{R}^n, x \neq y$,

$$(x - y)^\top (H(x) - H(y)) \leq \kappa(\|x - y\|)\|x - y\|^2. \tag{39}$$

Also, assume that

$$\limsup \kappa(r) < 0. \tag{40}$$

This implies that there is a postive constant, R_0, and a negative constant $\bar{\kappa}$, such that $\kappa(r) \leq \bar{\kappa} < 0$ for all $r > R_0$.

We choose

$$R_1 = \frac{R_0}{2} + \frac{1}{2}\sqrt{R_0^2 - \frac{16\sigma_{min}(G)^2 e^{h(R_0)}}{\bar{\kappa}}} > R_0,$$

and define δ via the following chain of definitions:

$$\delta(r) = \int_0^r \varphi(s)g(s)ds \tag{41a}$$

$$g(r) = 1 - \frac{\xi}{2}\int_0^{r \wedge R_1} \Phi(s)\varphi(s)^{-1}ds \tag{41b}$$

$$\xi^{-1} = \int_0^{R_1} \Phi(s)\varphi(s)^{-1}ds \tag{41c}$$

$$\Phi(r) = \int_0^r \varphi(s)ds \tag{41d}$$

$$\varphi(r) = e^{-h(r)} \tag{41e}$$

$$h(r) = \frac{1}{2\sigma_{\min}(G)^2}\int_0^r s(\kappa(s) \vee 0)ds. \tag{41f}$$

In the above definition, we use the shorthand notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

The details on the choices of R_0 and R_1 will be presented during the proof of Theorem 16 for the general reflection coupling related to (38) and Corollary 17 for the specific reflection coupling related to (8).

As discussed above, $\delta(\|x - y\|)$ is a metric. See [20] for details. The corresponding Wasserstein distance is defined by

$$W_\delta(P, Q) = \inf_{\Gamma \in \mathfrak{C}(P,Q)} \int_{\mathcal{K}\times\mathcal{K}} \delta(\|x - y\|)d\Gamma(x, y)$$

Here, $\mathfrak{C}$ is the couplings between P and Q.

To get an explicit form of the constant factor in Lemma 2, we use the following theorem, which is analogous to Corollary 2 of [20].

Theorem 16. *If $\mathbf{x}_t^1$ and $\mathbf{x}_t^2$ are two solutions to (38), then for all $0 \leq s \leq t$, their laws satisfy*

$$W_\delta(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq e^{-\tilde{a}(t-s)} W_\delta(\mathcal{L}(\mathbf{x}_s^1), \mathcal{L}(\mathbf{x}_s^2))$$

where $\tilde{a} = \xi\sigma_{\min}(G)^2$.

Proof The proof closely follows the proof of Theorem 1 from [20] with constraints handled similar to works in [27, 29]. The key is to create an explicit coupling between $\mathbf{x}_t^1$ and $\mathbf{x}_t^2$, which is known as a *reflection coupling* [31].

To define the reflection coupling, let $\boldsymbol{\tau}$ be coupling time: $\boldsymbol{\tau} = \inf\left\{t | \mathbf{x}_t^1 = \mathbf{x}_t^2\right\}$. Let $\mathbf{r}_t = \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|$, $\mathbf{u}_t = (\mathbf{x}_t^1 - \mathbf{x}_t^2)/\mathbf{r}_t$. Then the reflection coupling between $\mathbf{x}_t^1$ and $\mathbf{x}_t^2$ is defined by:

$$d\mathbf{x}_t^1 = H(\mathbf{x}_t^1)dt + Gd\mathbf{w}_t - \mathbf{v}_t^1 d\boldsymbol{\mu}^1(t) \tag{42a}$$

$$d\mathbf{x}_t^2 = H(\mathbf{x}_t^2)dt + (I - 2\mathbf{u}_t\mathbf{u}_t^\top \mathbb{1}(t < \boldsymbol{\tau}))Gd\mathbf{w}_t - \mathbf{v}_t^2 d\boldsymbol{\mu}^2(t) \tag{42b}$$

where $-\int_0^t \mathbf{v}_s^1 d\boldsymbol{\mu}^1(s)$ and $-\int_0^t \mathbf{v}_s^2 d\boldsymbol{\mu}^2(s)$ are reflection terms that ensure that $\mathbf{x}_t^1 \in \mathcal{K}$ and $\mathbf{x}_t^2 \in \mathcal{K}$ for all $t \geq 0$.

The processes from (42) define a valid coupling since $\int_0^T (I - 2\mathbf{u}_t\mathbf{u}_t^\top \mathbb{1}(t < \boldsymbol{\tau}))Gd\mathbf{w}_t$ is a Brownian motion by Lévy's characterization.

The main idea is to show that with the specially constructed metric (41), there will be a constant $\tilde{a}$ such that $e^{\tilde{a}t}\delta(\mathbf{r}_t)$ is a supermartingale. Then, the definition of W_δ and the supermartingale property shows that

$$W_\delta(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq \mathbb{E}\left[\delta(\mathbf{r}_t)\right] \leq e^{-\tilde{a}(t-s)}\mathbb{E}[\delta(\mathbf{r}_s)]$$

Since this bound holds for all couplings of the laws $\mathcal{L}(\mathbf{x}_s^1)$ and $\mathcal{L}(\mathbf{x}_s^2)$, it must hold for the optimal coupling, and so

$$W_\delta(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq e^{-\tilde{a}(t-s)} W_\delta(\mathcal{L}(\mathbf{x}_s^1), \mathcal{L}(\mathbf{x}_s^2)),$$

which is the desired conclusion.

Therefore, to complete the proof, we must show that $e^{\tilde{a}t}\delta(\mathbf{r}_t)$ is a supermartingale, which is to ensure that this process is non-increasing on average. Recall that $\boldsymbol{\tau}$ is the coupling time, so that $e^{\tilde{a}t}\delta(\mathbf{r}_t) = 0$ for $t \geq \boldsymbol{\tau}$. So we want to bound the behavior of the process for all $t < \boldsymbol{\tau}$. Specifically, it is required to show that non-martingale terms of $d\left(e^{\tilde{a}t}\delta(\mathbf{r}_t)\right)$ are non-positive. By Itô's formula, we have that

$$d\left(e^{\tilde{a}t}\delta(\mathbf{r}_t)\right) = e^{\tilde{a}t}\left(\tilde{a}\delta(r)dt + \delta'(r)d\mathbf{r}_t + \frac{1}{2}\delta''(r)(d\mathbf{r}_t)^2\right).$$

To achieve the desired differential, we have to derive the terms $d\mathbf{r}_t$ and $(d\mathbf{r}_t)^2$.

$$\begin{aligned} d\mathbf{r}_t &= \mathbf{u}_t^\top\left(d\mathbf{x}_t^1 - d\mathbf{x}_t^2\right) \\ &= \mathbf{u}_t^\top\left(\left(H(\mathbf{x}_t^1) - H(\mathbf{x}_t^2)\right)dt + 2\mathbf{u}_t\mathbf{u}_t^\top Gd\mathbf{w}_t - \mathbf{v}_t^1 d\boldsymbol{\mu}^1(t) + \mathbf{v}_t^2 d\boldsymbol{\mu}^2(t)\right) \end{aligned}$$

The above equation is simplified because $(d\mathbf{x}_t^1 - d\mathbf{x}_t^2)^\top(\nabla^2\mathbf{r}_t)(d\mathbf{x}_t^1 - d\mathbf{x}_t^2) = 0$.

Also, by assumption we have

$$(\mathbf{x}_t^1 - \mathbf{x}_t^2)^\top\left(H(\mathbf{x}_t^1) - H(\mathbf{x}_t^2)\right) \leq \kappa(\|\mathbf{x}_t^1 - \mathbf{x}_t^2\|)\|\mathbf{x}_t^1 - \mathbf{x}_t^2\|^2. \tag{43}$$

By the definition of $\mathbf{u}_t$ and the facts that $\mathbf{v}_t^1 \in N_\mathcal{K}(\mathbf{x}_t^1)$ and $\mathbf{v}_t^2 \in N_\mathcal{K}(\mathbf{x}_t^2)$ imply that $-\mathbf{u}_t^\top\mathbf{v}_t^1 d\boldsymbol{\mu}^1(t) \leq 0$ and $\mathbf{u}_t^\top\mathbf{v}_t^2 d\boldsymbol{\mu}^2(t) \leq 0$. It follows that and the assumption (39) gives

$$d\mathbf{r}_t \leq \kappa(r)rdt + 2\mathbf{u}_t^\top Gd\mathbf{w}_t.$$

Now, since the terms that were dropped in the inequality have bounded variation, we have that

$$(d\mathbf{r}_t)^2 = 4\mathbf{u}_t GG^\top\mathbf{u}_t dt \geq 4\sigma_{\min}(G)^2 dt.$$

By construction $\delta'(r) \geq 0$ and $\delta''(r) \leq 0$, and so Itô's formula gives

$$\begin{aligned} d\left(e^{\tilde{a}t}\delta(\mathbf{r}_t)\right) &\leq dte^{\tilde{a}t}\left(\tilde{a}\delta(r) + \delta'(r)\kappa(r)r + \delta''(r)2\sigma_{\min}(G)^2\right) + \mathbf{m}_t \\ &= 2\sigma_{\min}(G)^2 e^{\tilde{a}t}dt\left(\frac{\tilde{a}}{2\sigma_{\min}(G)^2}\delta(r) + \frac{k(r)r}{2\sigma_{\min}(G)^2}\delta'(r) + \delta''(r)\right) + \mathbf{m}_t, \end{aligned}$$

where $\mathbf{m}_t$ denotes a local martingale.

So it suffices to pick certain $\tilde{a}$ and R_1 to ensure that for all $r \geq 0$, the following holds:

$$\frac{\tilde{a}}{2\sigma_{\min}(G)^2}\delta(r) + \frac{\kappa(r)r}{2\sigma_{\min}(G)^2}\delta'(r) + \delta''(r) \leq 0. \tag{44}$$

Recall that

$$\begin{aligned}\delta''(r) &= \varphi'(r)g(r) + g'(r)\varphi(r) \\ &= -\frac{1}{2\sigma_{\min}(G)^2} r(\kappa(r) \vee 0)\delta'(r) - \frac{\xi}{2}\Phi(r)\mathbb{1}(r < R_1).\end{aligned}$$

So if we set $\tilde{a} = \xi\sigma_{\min}(G)^2$, then $\delta(r) \leq \Phi(r)$ implies that (44) holds for all $r < R_1$.

The remaining work is to find a sufficient condition under which (44) holds when $r \geq R_1$.

Recall that we assume that there exists $0 < R_0$, such that $k(r) < 0$ for all $r \geq R_0$. So, if we choose $R_1 > R_0$, we have for all $r \geq R_1$ that $\varphi(r) = \varphi(R_0)$. By definition, $g(r) = \frac{1}{2}$ for all $r \geq R_1$, and so we must also have $\delta'(r) = \frac{1}{2}\varphi(R_0)$.

Therefore, for $r \geq R_1$, (44) becomes

$$\frac{\tilde{a}}{2\sigma_{\min}(G)^2}\delta(r) + \frac{\kappa(r)r}{2\sigma_{\min}(G)^2}\frac{1}{2}\varphi(R_0) \leq 0.$$

So, a sufficient condition for (44) to hold when $r \geq R_1$ is given by:

$$\frac{\tilde{a}\delta(r)}{2\sigma_{\min}(G)^2} + \frac{\kappa(r)r}{2\sigma_{\min}(G)^2}\frac{1}{2}\varphi(R_0) \leq 0 \tag{46a}$$

$$\iff \tilde{a}\delta(r) + \kappa(r)r\frac{1}{2}\varphi(R_0) \leq 0 \tag{46b}$$

$$\iff \kappa(r)r\frac{1}{2}\varphi(R_0) \leq -\tilde{a}\delta(r) \tag{46c}$$

$$\iff \kappa(r)r\frac{1}{2}\varphi(R_0) \leq -\xi\sigma_{min}(G)^2\delta(r) \tag{46d}$$

$$\iff \kappa(r)r\frac{1}{2}\varphi(R_0) \leq -\frac{\sigma_{min}(G)^2}{\int_0^{R_1}\Phi(s)\varphi(s)^{-1}ds}\delta(r) \tag{46e}$$

$$\impliedby \kappa(r)r\frac{1}{2}\varphi(R_0) \leq -\frac{\sigma_{min}(G)^2}{(R_1 - R_0)\Phi(R_1)\varphi(R_0)^{-1}/2}\delta(r) \tag{46f}$$

$$\impliedby \kappa(r)r\frac{1}{2}\varphi(R_0) \leq -\frac{\sigma_{min}(G)^2}{(R_1 - R_0)\Phi(R_1)\varphi(R_0)^{-1}/2}r \tag{46g}$$

$$\iff \kappa(r) \leq -\frac{4\sigma_{min}(G)^2}{(R_1 - R_0)\Phi(R_1)} \tag{46h}$$

$$\iff (R_1 - R_0)\Phi(R_1) \geq -\frac{4\sigma_{min}(G)^2}{\kappa(r)} \tag{46i}$$

$$\impliedby (R_1 - R_0)R_1e^{-h(R_0)} \geq -\frac{4\sigma_{min}(G)^2}{\kappa(r)} \tag{46j}$$

$$\impliedby (R_1 - R_0)R_1e^{-h(R_0)} \geq -\frac{4\sigma_{min}(G)^2}{\bar{\kappa}} \tag{46k}$$

Note (46e) is implied by (46f) because for $r > R_0$, $\varphi(r) = \varphi(R_0)$, therefore, $\Phi(r) = \Phi(R_0) + \varphi(R_0)(r - R_0)$ which gives

$$\begin{aligned}\int_0^{R_1}\Phi(s)\varphi(s)^{-1}ds &\geq \int_{R_0}^{R_1}\Phi(s)\varphi(s)^{-1}ds \\ &= \int_{R_0}^{R_1}\left(\Phi(R_0) + \varphi(R_0)(s - R_0)\right)\varphi(R_0)^{-1}ds \\ &= \Phi(R_0)\varphi(R_0)^{-1}(R_1 - R_0) + \frac{(R_1 - R_0)^2}{2} \\ &\geq \frac{\Phi(R_0)\varphi(R_0)^{-1}(R_1 - R_0)}{2} + \frac{(R_1 - R_0)^2}{2} \\ &= (R_1 - R_0)\left(\Phi(R_0) + (R_1 - R_0)\varphi(R_0)\right)\varphi(R_0)^{-1}/2 \\ &= (R_1 - R_0)\Phi(R_1)\varphi(R_0)^{-1}/2.\end{aligned} \tag{47}$$

Also, (46f) is implied by (46g) because $\delta(r) < r$.

From (46i) to (46j), we use:

$$\begin{aligned}\Phi(R_1) &= \int_0^{R_1} \varphi(s)ds \\ &= \int_0^{R_1} e^{-h(s)}ds \\ &\geq \int_0^{R_1} e^{-h(R_0)}ds \\ &= R_1 e^{-h(R_0)}. \end{aligned} \tag{48}$$

The implication (46k) $\implies$ (46j) arises because of the assumption that $\kappa(r) \leq \bar{\kappa} < 0$ for all $r > R_0$. Therefore, (44) will hold all $r \geq R_1$, as long as R_1 satisfies (46k). The smallest such R_1 is given by

$$R_1 = \frac{R_0}{2} + \frac{1}{2}\sqrt{R_0^2 - \frac{16\sigma_{min}(G)^2 e^{h(R_0)}}{\bar{\kappa}}} > R_0. \tag{49}$$

■

We choose our reflection term as $(I - 2\mathbf{u}_t\mathbf{u}_t^\top \mathbb{1}(t < \boldsymbol{\tau}))Gd\mathbf{w}_t$, while [20] uses $G(I - 2\mathbf{e}_t\mathbf{e}_t^\top \mathbb{1}(t < \boldsymbol{\tau}))d\mathbf{w}_t$, with $\mathbf{e}_t = \frac{G^{-1}(\mathbf{x}_t^1 - \mathbf{x}_t^2)}{\|G^{-1}(\mathbf{x}_t^1 - \mathbf{x}_t^2)\|}$. Our form of the reflection term leads to mild simplification of some formulas.

Now we specialize the result from the previous theorem to the specific case of this paper:

Corollary 17. *If $\mathbf{x}_t^1$ and $\mathbf{x}_t^2$ are two solutions to (8), then for all $0 \leq s \leq t$, their laws satisfy*

$$W_\delta(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq e^{-\tilde{a}(t-s)} W_\delta(\mathcal{L}(\mathbf{x}_s^1), \mathcal{L}(\mathbf{x}_s^2))$$

where $\tilde{a} = \xi\frac{2\eta}{\beta}$, $R_0 = R$, and $R_1 = \frac{R}{2} + \frac{1}{2}\sqrt{R^2 + \frac{32\eta}{\mu\beta}e^{\frac{\beta\ell R^2}{8}}}$ in the construction of δ.

Proof We can see that (8) is a special case of (38) with

$$\begin{aligned} H(x) &= -\eta\nabla_x \bar{f}(x) \\ G &= \sqrt{\frac{2\eta}{\beta}}I. \end{aligned}$$

Since we assume that $\bar{f}$ is ℓ-Lipschitz and convex outside a ball with radius R, we have that (39) holds with $\kappa(s) = \eta\ell$ for $0 \leq s < R$ and $\kappa(s) = -\eta\mu$ for $s \geq R$. Therefore, we can pick $R_0 = R$ to construct the metric (41).

Now, $\sigma_{\min}(G)^2 = \frac{2\eta}{\beta}$ implies that $\tilde{a} = \xi\frac{2\eta}{\beta}$. Furthermore, the choice of $\kappa(r)$ implies that $h(R_0) = h(R) = \frac{\beta\ell R^2}{8}$.

The choice of $\kappa(r)$ also implies that $\bar{\kappa} = -\eta\mu$. Thus, the form of R_1 is given by plugging terms into (49). ■

Corollary 18. *If $\mathbf{x}_t^1$ and $\mathbf{x}_t^2$ are two solutions to (8), then for all $0 \leq s \leq t$, their laws satisfy*

$$W_1(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq 2\varphi(R)^{-1}e^{-\tilde{a}(t-s)}W_1(\mathcal{L}(\mathbf{x}_s^1), \mathcal{L}(\mathbf{x}_s^2)).$$

Proof From the special constructed of δ, we that $\delta'(r)$ is monotonically decreasing, and also $\delta'(r) = \delta(R_1)$ for all $r \geq R_1$. Furthermore, $\delta(r) = \int_0^r \delta'(s)ds \geq \delta'(r)\int_0^r ds = r\delta'(r)$. Thus, for all $r \geq 0$, the following bounds hold:

$$\delta'(R_1)r \leq \delta'(r)r \leq \delta(r) \leq r$$

These bounds are now used to relate the W_δ and W_1 distances:

$$\delta'(R_1)W_1(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq W_\delta(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq W_1(\mathcal{L}(\mathbf{x}_t), \mathcal{L}(\mathbf{y}_t)). \tag{50}$$

In particular,

$$\delta'(R_1) = \varphi(R_1)g(R_1) = \frac{1}{2}\varphi(R). \tag{51}$$

Plugging (51) into the first inequality of (50) gives

$$W_1(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq 2\varphi(R)^{-1} W_\delta(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \tag{52}$$

And combining with Corollary 17 gives

$$W_1(\mathcal{L}(\mathbf{x}_t^1), \mathcal{L}(\mathbf{x}_t^2)) \leq 2\varphi(R)^{-1} e^{-\tilde{a}(t-s)} W_\delta(\mathcal{L}(\mathbf{x}_s^1), \mathcal{L}(\mathbf{x}_s^2)) \tag{53}$$

Finally, utilizing the second inequality of (50) gives the desired result. ■

D.2 Proof of Lemma 2

In Lemma 12 of Appendix B, we showed that the Gibbs distribution, $\pi_{\beta\bar{f}}$, defined in (6) is invariant for the dynamics of $\mathbf{x}_t^C$. Thus, setting $\mathcal{L}(\mathbf{x}_t^1) = \mathcal{L}(\mathbf{x}_t^C)$ and $\mathcal{L}(\mathbf{x}_t^2) = \pi_{\beta\bar{f}}$ in Corollary 18 gives

$$W_1(\mathcal{L}(\mathbf{x}_t^C), \pi_{\beta\bar{f}}) \leq 2\varphi(R)^{-1} e^{-\tilde{a}t} W_1(\mathcal{L}(\mathbf{x}_0^C), \pi_{\beta\bar{f}}). \tag{54}$$

Let $\mathbf{y}$ be distributed according to $\pi_{\beta\bar{f}}$. For any joint distribution over $(\mathbf{x}_0^C, \mathbf{y})$ whose marginals are $\mathcal{L}(\mathbf{x}_0^C)$ and $\pi_{\beta\bar{f}}$, we have that

$$\begin{aligned}
W_1(\mathcal{L}(\mathbf{x}_0^C), \pi_{\beta\bar{f}}) &\leq \mathbb{E}[\|\mathbf{x}_0^C - \mathbf{y}\|] \\
&\leq \sqrt{\mathbb{E}[\|\mathbf{x}_0^C - \mathbf{y}\|^2]} \\
&\leq \sqrt{\mathbb{E}[2\|\mathbf{x}_0^C\|^2 + 2\|\mathbf{y}\|^2]} \\
&= \sqrt{2\mathbb{E}[\|\mathbf{x}_0^C\|^2] + 2\mathbb{E}[\|\mathbf{y}\|^2]} \\
&\leq \sqrt{2\varsigma + 2\left(\varsigma + \frac{1}{\mu}c_{10}\right)} \\
&\leq \sqrt{\frac{2}{\mu}c_{10}} + 2\sqrt{\varsigma}.
\end{aligned} \tag{55}$$

The second to last inequality uses Lemma 14.

Combining (54), (55) shows that

$$W_1(\mathcal{L}(\mathbf{x}_t^C), \pi_{\beta\bar{f}}) \leq 2\varphi(R)^{-1} e^{-\tilde{a}t} \left(\sqrt{\frac{2}{\mu}c_{10}} + 2\sqrt{\varsigma}\right).$$

Thus, the lemma is proved and the constants are given by:

$$a = \frac{2\xi}{\beta} \tag{56a}$$

$$c_1 = 2\varphi(R)^{-1}\sqrt{\frac{2}{\mu}c_{10}} \tag{56b}$$

$$c_2 = 4\varphi(R)^{-1} \tag{56c}$$

where ξ is given in (41c). ■

E Proofs of averaging lemmas

Proof of Lemma 4 Non-expansiveness of the projection and the definitions of $\mathbf{x}_t^{M,s}$ and $\mathbf{x}_t^{B,s}$ show that:

$$\begin{aligned}
&\|\mathbf{x}_{t+1}^{M,s} - \mathbf{x}_{t+1}^{B,s}\|^2 \\
&\quad \leq \left\| \mathbf{x}_t^{M,s} - \mathbf{x}_t^{B,s} + \eta \left(\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t] - \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] \right) \right\|^2 \\
&\quad = \|\mathbf{x}_t^{M,s} - \mathbf{x}_t^{B,s}\|^2 + 2\eta \left(\mathbf{x}_t^{M,s} - \mathbf{x}_t^{B,s} \right)^\top \\
&\qquad \left(\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t] - \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] \right) \\
&\qquad + \eta^2 \left\| \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t] - \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] \right\|^2 . \qquad (57)
\end{aligned}$$

We will show that the second term on the right of (57) has mean zero, and then we will bound the mean of the third term on the right of (57).

By construction, we have that $\mathbf{x}_t^{M,s}$ is $\mathcal{F}_{t-s-1} \vee \mathcal{G}_t$-measurable, while $\mathbf{x}_t^{B,s}$ is $\mathcal{F}_{t-s-2} \vee \mathcal{G}_t$-measurable. Thus, the only part of the second term on the right of (57) which is not $\mathcal{F}_{t-s-1} \vee \mathcal{G}_t$-measurable is $\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t]$. Therefore, the tower-property gives:

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbf{x}_t^{M,s} - \mathbf{x}_t^{B,s} \right)^\top \left(\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t] - \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] \right) \right] \\
&\quad = \mathbb{E}\left[\left(\mathbf{x}_t^{M,s} - \mathbf{x}_t^{B,s} \right)^\top \left(\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t] \right.\right. \\
&\qquad\qquad \left.\left. - \mathbb{E}\left[\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] \right) \middle| \mathcal{F}_{t-s-1} \vee \mathcal{G}_t \right] \right] \\
&\quad = 0.
\end{aligned}$$

Now we focus on bounding the mean of the third term on the right of (57). Recall that $\mathbf{x}_t^{M,s}$ is $\mathcal{F}_{t-s-1} \vee \mathcal{G}_t$-measurable. Furthermore, since $\mathcal{F}_{t-s}^+$ is independent of $\mathcal{F}_{t-s} \vee \mathcal{G}_t$, it must also be independent of $\mathcal{F}_{t-s-1} \vee \mathcal{G}_t$ because $\mathcal{F}_{t-s-1} \subset \mathcal{F}_{t-s}$. It follows that

$$\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] = \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t].$$

Thus, adding and subtracting $\mathbb{E}\left[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) \middle| \mathcal{F}_{t-s} \vee \mathcal{G}_t \right]$ gives

$$\begin{aligned}
&\left\| \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s-1} \vee \mathcal{G}_t] - \mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) | \mathcal{F}_{t-s} \vee \mathcal{G}_t] \right\|^2 \\
&\leq 2 \left\| \mathbb{E}\left[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) - \nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) \middle| \mathcal{F}_{t-s-1} \vee \mathcal{G}_t \right] \right\|^2 \\
&\qquad + 2 \left\| \mathbb{E}\left[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) - \nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) \middle| \mathcal{F}_{t-s} \vee \mathcal{G}_t \right] \right\|^2 . \qquad (58)
\end{aligned}$$

To bound the second term on the right of (58), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\| \mathbb{E}\left[\nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) - \nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) \middle| \mathcal{F}_{t-s} \vee \mathcal{G}_t \right] \right\|^2 \right] \\
&\quad \overset{\text{Jensen}}{\leq} \mathbb{E}\left[\left\| \nabla_x f(\mathbf{x}_t^{M,s}, \mathbf{z}_t) - \nabla_x f(\mathbf{x}_t^{M,s}, \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+]) \right\|^2 \right] \\
&\quad \overset{\text{Lipschitz}}{\leq} \ell^2 \mathbb{E}\left[\|\mathbf{z}_t - \mathbb{E}[\mathbf{z}_t | \mathcal{F}_{t-s}^+\|^2 \right] \\
&\quad \leq \ell^2 \psi_2(s, \mathbf{z})^2.
\end{aligned}$$

Here $\psi_2(s, \mathbf{z})$ was defined in (3b).

The first term on the right of (58) is bounded by analogous calculations with $\mathcal{F}_{t-s-1}$ used in place of $\mathcal{F}_{t-s}$, and gives rise to the same bound of $\ell^2\psi(s,\mathbf{z})^2$.

Plugging these bounds into (57) shows that

$$\mathbb{E}\left[\|\mathbf{x}_{t+1}^{M,s}-\mathbf{x}_{t+1}^{B,s}\|^2\right] \leq \mathbb{E}\left[\|\mathbf{x}_{t}^{M,s}-\mathbf{x}_{t}^{B,s}\|^2\right] + 4\eta^2\ell^2\psi_2(s,\mathbf{z})^2$$

Iterating (E) t times and using the the fact that $\mathbf{x}_0^{B,s}=\mathbf{x}_0^{M,s}$, shows that

$$\mathbb{E}\left[\|\mathbf{x}_{t}^{M,s}-\mathbf{x}_{t}^{B,s}\|^2\right] \leq 4\eta^2 t\ell^2\psi_2(s,\mathbf{z})^2.$$

Using the fact that

$$\mathbb{E}[\|\mathbf{x}_{t}^{M,s}-\mathbf{x}_{t}^{B,s}\|] \leq \sqrt{\mathbb{E}\left[\|\mathbf{x}_{t}^{M,s}-\mathbf{x}_{t}^{B,s}\|^2\right]}$$

gives the result. ■

Proof of Lemma 5 Non-expansiveness of the projection and the definitions of $\mathbf{x}_t^{B,s}$ and $\mathbf{x}_t^{M,s+1}$, shows that

$$\begin{aligned}
&\|\mathbf{x}_{t+1}^{B,s}-\mathbf{x}_{t+1}^{M,s+1}\| \\
&\leq \left\|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}+\eta\left(\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s+1},\mathbf{z}_t)|\mathcal{F}_{t-s-1}\vee\mathcal{G}_t]-\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s},\mathbf{z}_t)|\mathcal{F}_{t-s-1}\vee\mathcal{G}_t]\right)\right\|.
\end{aligned}$$

Let $\|\mathbf{x}\|_2=\sqrt{\mathbb{E}[\|\mathbf{x}\|^2]}$ denote the 2-norm over random vectors. The triangle inequality then implies that

$$\begin{aligned}
&\left\|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}+\eta\left(\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s+1},\mathbf{z}_t)|\mathcal{F}_{t-s-1}\vee\mathcal{G}_t]-\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s},\mathbf{z}_t)|\mathcal{F}_{t-s-1}\vee\mathcal{G}_t]\right)\right\|_2 \\
&\leq \left\|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}\right\|_2+\eta\left\|\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s+1},\mathbf{z}_t)-\nabla_x f(\mathbf{x}_t^{M,s},\mathbf{z}_t)|\mathcal{F}_{t-s-1}\vee\mathcal{G}_t]\right\|_2.
\end{aligned}$$

For any random vector, $\mathbf{x}$, and any σ-algebra, $\mathcal{F}$, Jensen's inequality followed by the tower property implies that $\mathbb{E}[\|\mathbb{E}[\mathbf{x}|\mathcal{F}]\|^2]\leq\mathbb{E}[\|\mathbf{x}\|^2]$. Applying this fact to the second term on the right of (59) and then using the Lipschitz property shows that

$$\left\|\mathbb{E}[\nabla_x f(\mathbf{x}_t^{M,s+1},\mathbf{z}_t)-\nabla_x f(\mathbf{x}_t^{M,s},\mathbf{z}_t)|\mathcal{F}_{t-s-1}\vee\mathcal{G}_t]\right\|_2 \leq \ell\|\mathbf{x}_t^{M,s+1}-\mathbf{x}_t^{M,s}\|_2.$$

Plugging this bound into (59) then adding and subtracting $\mathbf{x}_t^{B,s}$ gives:

$$\begin{aligned}
&\|\mathbf{x}_{t+1}^{B,s}-\mathbf{x}_{t+1}^{M,s+1}\|_2 \\
&\leq \|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}\|_2+\eta\ell\|\mathbf{x}_t^{M,s+1}-\mathbf{x}_t^{M,s}\|_2 \\
&\leq \|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}\|_2+\eta\ell\|\mathbf{x}_t^{M,s+1}-\mathbf{x}_t^{B,s}\|_2+\eta\ell\|\mathbf{x}_t^{B,s}-\mathbf{x}_t^{M,s}\|_2 \\
&= (1+\eta\ell)\|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}\|_2+\eta\ell\|\mathbf{x}_t^{B,s}-\mathbf{x}_t^{M,s}\|_2. \qquad (59)
\end{aligned}$$

Using the fact that $\mathbf{x}_0^{B,s}=\mathbf{x}_0^{M,s+1}$ and iterating this inequality shows that:

$$\begin{aligned}
&\|\mathbf{x}_{t}^{B,s}-\mathbf{x}_{t}^{M,s+1}\|_2 \\
&\leq \eta\ell\sum_{k=0}^{t-1}(1+\eta\ell)^k\|\mathbf{x}_{t-k}^{B,s}-\mathbf{x}_{t-k}^{M,s}\|_2 \\
&\overset{\text{Lemma 4}}{\leq} \left(2\ell\psi_2(s,\mathbf{z})\eta\sqrt{t}\right)\eta\ell\sum_{k=0}^{t-1}(1+\eta\ell)^k \\
&= \left(2\ell\psi_2(s,\mathbf{z})\eta\sqrt{t}\right)\left((1+\eta\ell)^t-1\right) \\
&\leq \left(2\ell\psi_2(s,\mathbf{z})\eta\sqrt{t}\right)\left(e^{\eta t\ell}-1\right).
\end{aligned}$$

The final inequality follows by taking logarithms and using the fact that $\log(1+\eta\ell)\leq\eta\ell$. ■

F Discretization bounds

Proof of Lemma 6 Recall that $\mathbf{y}_t^D = \mathbf{y}_{\lfloor t \rfloor}^C$ and so for all $k \in \mathbb{N}$, $\mathbf{y}_k^D = \mathbf{y}_k^C$. By the construction of Skorokhod solutions to the process $\mathbf{x}_t^C$ and $\mathbf{x}_t^D$, and using Theorem 9, we have for all $k \in \mathbb{N}$

$$\left\|\mathbf{x}_k^C - \mathbf{x}_k^D\right\| \leq (c_9 + 1) \sup_{0 \leq s \leq k} \left\|\mathbf{y}_s^C - \mathbf{y}_{\lfloor s \rfloor}^C\right\|.$$

Since

$$\mathbf{y}_t^C = \mathbf{x}_0^C - \eta \int_0^t \nabla_x \bar{f}(\mathbf{x}_s^C) ds + \sqrt{\frac{2\eta}{\beta}} \mathbf{w}_t,$$

the triangle inequality implies that

$$\left\|\mathbf{x}_k^C - \mathbf{x}_k^D\right\| \leq (c_9 + 1)\eta \sup_{s \in [0,k]} \left\| \int_{\lfloor s \rfloor}^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\| + (c_9 + 1)\sqrt{\frac{2\eta}{\beta}} \sup_{s \in [0,k]} \left\|\mathbf{w}_s - \mathbf{w}_{\lfloor s \rfloor}\right\|.$$

$\mathbb{E}\left[\sup_{s \in [0,k]} \left\|\mathbf{w}_s - \mathbf{w}_{\lfloor s \rfloor}\right\|\right]$ is upper bounded by $2n\sqrt{\log(4k)}$. See Lemma 9 in [27]. So, the remaining work is to bound the first term on the right.

Take the expectation of the first term, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s \in [0,k]} \left\| \int_{\lfloor s \rfloor}^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\|\right] \\
&= \mathbb{E}\left[\max_{i=0,\cdots,k-1} \sup_{s \in [i,i+1]} \left\| \int_i^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\|\right] \\
&\leq \mathbb{E}\left[\left(\sum_{i=0}^{k-1} \left(\sup_{s \in [i,i+1]} \left\| \int_i^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\|\right)^2\right)^{1/2}\right] \\
&\overset{\text{Jensen}}{\leq} \left(\mathbb{E}\left[\sum_{i=0}^{k-1} \left(\sup_{s \in [i,i+1]} \left\| \int_i^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\|\right)^2\right]\right)^{1/2} \\
&= \left(\sum_{i=0}^{k-1} \mathbb{E}\left[\left(\sup_{s \in [i,i+1]} \left\| \int_i^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\|\right)^2\right]\right)^{1/2}.
\end{aligned}$$

So we want to upper bound the supremum inside the expectation operation.

We can show for all $s \in [0, k]$,

$$\begin{aligned}
\left\| \int_{\lfloor s \rfloor}^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\| &\overset{\text{triangle inequality}}{\leq} \int_{\lfloor s \rfloor}^s \left\|\nabla_x \bar{f}(\mathbf{x}_\tau)\right\| d\tau \\
&\leq \int_{\lfloor s \rfloor}^{\lfloor s \rfloor + 1} \left\|\nabla_x \bar{f}(\mathbf{x}_\tau^C)\right\| d\tau \\
&\overset{\text{Jensen}}{\leq} \left(\int_{\lfloor s \rfloor}^{\lfloor s \rfloor + 1} \|\nabla_x \bar{f}(\mathbf{x}_\tau^C)\|^2 d\tau\right)^{1/2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sup_{s \in [0,k]} \left\| \int_{\lfloor s \rfloor}^s \nabla_x \bar{f}(\mathbf{x}_\tau^C) d\tau \right\|\right] &\leq \left(\sum_{i=0}^{k-1} \mathbb{E}\left[\int_i^{i+1} \|\nabla_x \bar{f}(\mathbf{x}_\tau^C)\|^2 d\tau\right]\right)^{1/2} \\
&\overset{\text{Fubini}}{=} \left(\sum_{i=0}^{k-1} \int_i^{i+1} \mathbb{E}\left[\|\nabla_x \bar{f}(\mathbf{x}_\tau^C 1)\|^2\right] d\tau\right)^{1/2}.
\end{aligned}$$

Here, we can see it suffices to bound $\mathbb{E}\left[\|\nabla_x \bar{f}(\mathbf{x}_t)\|^2\right]$.

We have assumed that $0 \in \mathcal{K}$, and so we have

$$\begin{aligned}\left\|\nabla_x \bar{f}(\mathbf{x}_t^C)\right\|^2 &= \left\|\nabla_x \bar{f}(\mathbf{x}_t^C) - \nabla_x \bar{f}(0) + \nabla_x \bar{f}(0)\right\|^2 \\ &\leq 2\left\|\nabla_x \bar{f}(\mathbf{x}_t^C) - \nabla_x \bar{f}(0)\right\|^2 + 2\left\|\nabla_x \bar{f}(0)\right\|^2 \\ &\leq 2\ell^2\left\|\mathbf{x}_t^C\right\|^2 + 2\left\|\nabla_x \bar{f}(0)\right\|^2.\end{aligned}$$

Plugging in the bound from Lemma 14 shows that

$$\begin{aligned}\mathbb{E}\left[\|\nabla_x \bar{f}(\mathbf{x}_t^C)\|^2\right] &\leq \mathbb{E}\left[2\ell^2\|\mathbf{x}_t^C\|^2 + 2\|\nabla_x \bar{f}(0)\|^2\right] \\ &= 2\ell^2\mathbb{E}\left[\|\mathbf{x}_t^C\|^2\right] + 2\|\nabla_x \bar{f}(0)\|^2 \\ &\leq 2\ell^2\left(\varsigma + \frac{1}{\mu}c_{10}\right) + 2\|\nabla_x \bar{f}(0)\|^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{s\in[0,k]}\left\|\int_{\lfloor s\rfloor}^{s}\nabla_x \bar{f}(\mathbf{x}_\tau^C)d\tau\right\|\right] &\leq \left(\sum_{i=0}^{k-1}\int_i^{i+1}\mathbb{E}\left[\|\nabla_x \bar{f}(\mathbf{x}_\tau^C)\|^2\right]d\tau\right)^{1/2} \\ &\leq \sqrt{2\ell^2\left(\varsigma + \frac{1}{\mu}c_{10}\right) + 2\|\nabla_x \bar{f}(0)\|^2}\sqrt{k} \\ &\leq \left(\sqrt{\frac{2}{\mu}\ell^2 c_{10} + 2\|\nabla_x \bar{f}(0)\|^2} + \sqrt{2\ell^2}\sqrt{\varsigma}\right)\sqrt{k}.\end{aligned}$$

Setting

$$\begin{aligned}c_5 &= (c_9+1)\sqrt{\frac{2}{\mu}\ell^2 c_{10} + 2\|\nabla_x \bar{f}(0)\|^2} \\ c_6 &= (c_9+1)\sqrt{2\ell^2} \\ c_7 &= (c_9+1)n\sqrt{\frac{8}{\beta}}\end{aligned}$$

where c_{10} is defined in Lemma 14 and c_9 is defined in Theorem 9 and combining the bound on the second supreme term gives the desired result. ■

Proof of Lemma 7 The argument of bounding $\mathbf{x}_t^M$ and $\mathbf{x}_t^D$ closely follows the proof of Lemma 10 in [27]. Recall that $\mathbf{x}_t^M$ is a discretized process and $\mathbf{x}_t^M = \mathbf{x}_{\lfloor t\rfloor}^M$. We also have $\mathbf{x}_t^M = \mathcal{S}(\mathcal{D}(\mathbf{y}_t^M))$, where $\mathbf{y}_t^M$ is defined by

$$\mathbf{y}_t^M = \mathbf{x}_0^M - \eta\int_0^t \nabla \bar{f}(\mathbf{x}_{\lfloor s\rfloor}^M)ds + \sqrt{\frac{2\eta}{\beta}}\mathbf{w}_t.$$

The intermediate process $\mathbf{x}_t^D$ satisfies $\mathbf{x}_t^D = \mathcal{S}(\mathcal{D}(\mathbf{y}_t^C))$, where

$$\mathbf{y}_t^C = \mathbf{x}_0^C - \eta\int_0^t \nabla \bar{f}(\mathbf{x}_s^C)ds + \sqrt{\frac{2\eta}{\beta}}\mathbf{w}_t.$$

So in particular,

$$\begin{aligned}\mathbf{x}_{k+1}^M &= \Pi_{\mathcal{K}}\left(\mathbf{x}_k^M + \mathbf{y}_{k+1}^M - \mathbf{y}_k^M\right) \\ &= \Pi_{\mathcal{K}}\left(\mathbf{x}_k^M - \eta\nabla \bar{f}(\mathbf{x}_k^M) + \sqrt{\frac{2\eta}{\beta}}(\mathbf{w}_{k+1} - \mathbf{w}_k)\right) \\ \mathbf{x}_{k+1}^D &= \Pi_{\mathcal{K}}\left(\mathbf{x}_k^D + \mathbf{y}_{k+1}^C - \mathbf{y}_k^C\right) \\ &= \Pi_{\mathcal{K}}\left(\mathbf{x}_k^D - \eta\int_k^{k+1}\nabla \bar{f}(\mathbf{x}_s^C)ds + \sqrt{\frac{2\eta}{\beta}}(\mathbf{w}_{k+1} - \mathbf{w}_k)\right).\end{aligned}$$

Define a difference process

$$\boldsymbol{\rho}_t = \left(\mathbf{x}_t^M + \mathbf{y}_t^M - \mathbf{y}_{\lfloor t \rfloor}^M\right) - \left(\mathbf{x}_t^D + \mathbf{y}_t^C - \mathbf{y}_{\lfloor t \rfloor}^C\right).$$

Note that at integers $k \in \mathbb{N}$, $\boldsymbol{\rho}_k = \mathbf{x}_k^M - \mathbf{x}_k^D$ and for $t \in [k, k+1)$, we have

$$\boldsymbol{\rho}_t = \left(\mathbf{x}_k^M - \mathbf{y}_k^M - \mathbf{x}_k^D + \mathbf{y}_k^D\right) + \mathbf{y}_t^M - \mathbf{y}_t^C.$$

It follows that

$$d\boldsymbol{\rho}_t = d(\mathbf{y}_t^M - \mathbf{y}_t^C) = \eta\left(\nabla \bar{f}(\mathbf{x}_t^C) - \nabla \bar{f}(\mathbf{x}_t^M)\right)$$

By construction, $\boldsymbol{\rho}_t$ is a continuous bounded variation process on the interval $[k, k+1)$. Thus, when $\boldsymbol{\rho}_t \neq 0$, we can calculate $d\|\boldsymbol{\rho}_t\|$ using the chain rule.

$$\begin{aligned}
d\left\|\boldsymbol{\rho}_t\right\| &\overset{\text{chain rule}}{=} \left(\frac{\boldsymbol{\rho}_t}{\|\boldsymbol{\rho}_t\|}\right)^\top d\boldsymbol{\rho}_t \\
&= \left(\frac{\boldsymbol{\rho}_t}{\|\boldsymbol{\rho}_t\|}\right)^\top \eta\left(\nabla \bar{f}(\mathbf{x}_t^C) - \nabla \bar{f}(\mathbf{x}_t^M)\right) dt \\
&\overset{\text{Cauchy-Schwarz}}{\leq} \eta\left\|\nabla \bar{f}(\mathbf{x}_t^C) - \nabla \bar{f}(\mathbf{x}_t^M)\right\| dt \\
&\overset{\text{Lipschitz}}{\leq} \eta\ell\left\|\mathbf{x}_t^C - \mathbf{x}_t^M\right\| dt \\
&= \eta\ell\left\|\mathbf{x}_t^C - \mathbf{x}_t^D + \mathbf{x}_t^D - \mathbf{x}_t^M\right\| dt \\
&\overset{\text{triangle}}{\leq} \eta\ell\left(\left\|\mathbf{x}_t^C - \mathbf{x}_t^D\right\| + \left\|\mathbf{x}_t^D - \mathbf{x}_t^M\right\|\right) dt.
\end{aligned}$$

To include the case that $\boldsymbol{\rho}_t = 0$, we use the Lemma 19 from [27]. The analysis is as below:

For $t \in [k, k+1)$,

$$\begin{aligned}
\|\boldsymbol{\rho}_t\| &= \|\boldsymbol{\rho}_k\| + \int_k^t d\|\boldsymbol{\rho}_t\| \\
&= \|\boldsymbol{\rho}_k\| + \lim_{\epsilon \downarrow 0} \int_k^t \mathbb{1}\left(\|\boldsymbol{\rho}_s\| \geq \epsilon\right) d\|\boldsymbol{\rho}_s\| \\
&\leq \|\boldsymbol{\rho}_k\| + \lim_{\epsilon \downarrow 0} \int_k^t \mathbb{1}\left(\|\boldsymbol{\rho}_s\| \geq \epsilon\right) \eta\ell\left(\left\|\mathbf{x}_s^C - \mathbf{x}_s^D\right\| + \left\|\mathbf{x}_s^D - \mathbf{x}_s^M\right\|\right) dt \\
&= (1 + \eta\ell)\|\boldsymbol{\rho}_k\| + \eta\ell \int_k^t \left(\left\|\mathbf{x}_s^C - \mathbf{x}_s^D\right\|\right) ds.
\end{aligned}$$

The second equality follows from Lemma 19 from [27]. The last equality holds because that $\boldsymbol{\rho}_k = \mathbf{x}_s^M - \mathbf{x}_s^D, \forall s \in [k, k+1)$.

Non-expansiveness of the convex projection implies that

$$\|\boldsymbol{\rho}_k\| = \left\|\mathbf{x}_s^M - \mathbf{x}_s^D\right\| \leq \lim_{t \uparrow k} \|\boldsymbol{\rho}_t\|. \tag{60}$$

Letting $t = k+1$ gives

$$\left\|\boldsymbol{\rho}_{k+1}\right\| \leq (1 + \eta\ell)\|\boldsymbol{\rho}_k\| + \eta\ell \int_k^{k+1} \left\|\mathbf{x}_s^C - \mathbf{x}_s^D\right\| ds.$$

Iterating this inequality, and using the assumption that $\mathbf{x}_0^M = \mathbf{x}_0^D$ gives

$$\|\boldsymbol{\rho}_k\| \leq \sum_{i=0}^{k-1} \eta\ell(1 + \eta\ell)^{k-i-1} \int_i^{i+1} \left\|\mathbf{x}_s^C - \mathbf{x}_s^D\right\| ds.$$

Taking expectation, and using Lemma 6 gives

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\rho}_k\|\right] &\leq \sum_{i=0}^{k-1} \eta\ell(1+\eta\ell)^{k-i-1} \int_i^{i+1} \left((c_5 + c_6\sqrt{\varsigma})\eta\sqrt{s} + c_7\sqrt{\eta\log(4s)}\right) ds \\
&\leq \eta\ell \left((c_5 + c_6\sqrt{\varsigma})\eta\sqrt{k} + c_7\sqrt{\eta\log(4k)}\right) \sum_{i=0}^{k-1}(1+\eta\ell)^{k-i-1} \\
&\leq \left((c_5 + c_6\sqrt{\varsigma})\eta\sqrt{k} + c_7\sqrt{\eta\log(4k)}\right) \left((1+\eta\ell)^k - 1\right) \\
&\leq \left((c_5 + c_6\sqrt{\varsigma})\eta\sqrt{k} + c_7\sqrt{\eta\log(4k)}\right) \left(e^{\eta\ell k} - 1\right).
\end{aligned}$$

The last inequality is based on the fact that $(1+\eta\ell)^k \leq e^{\eta\ell k}$ for all $\eta\ell > 0$.

Recall that for all $k \in \mathbb{N}$, $\boldsymbol{\rho}_k = \mathbf{x}_k^M - \mathbf{x}_k^D$, which gives the desired result. ■

G Conclusion of the proof of Lemma 3

This subsection uses a "switching" trick to derive a bound on $W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C))$ that is uniform in time. The essential idea is to utilize a family of processes that switch from the dynamics of $\mathbf{x}_k^A$ to the dynamics of $\mathbf{x}_k^C$, and utilize contractivity of the law of $\mathbf{x}_k^C$ to derive the uniform bounds. A similar methodology was utilized in [10].

For $s \geq 0$, let $\mathbf{x}_{s,t}^{A,C}$ be the process such that $\mathbf{x}_{s,t}^{A,C} = \mathbf{x}_t^A = \mathbf{x}_{\lfloor t \rfloor}^A$ for $t \leq s$ and for $t \geq s$, $\mathbf{x}_{s,t}^{A,C}$ follows:

$$d\mathbf{x}_{s,t}^{A,C} = -\eta\nabla_x \bar{f}(\mathbf{x}_{s,t}^{A,C})dt + \sqrt{\frac{2\eta}{\beta}}d\mathbf{w}_t - \mathbf{v}_{s,t}^{A,C} d\boldsymbol{\mu}_s^{A,C}(t).$$

In other words, $\mathbf{x}_{s,t}^{A,C}$ follows the algorithm for $t \leq s$, and then switches to the dynamics of the continuous-time approximation from (8) at $t = s$.

Now let $0 \leq s \leq \hat{s} \leq t$ where $s, \hat{s} \in \mathbb{N}$, then Corollary 18 from Appendix D shows that

$$W_1(\mathcal{L}(\mathbf{x}_{s,t}^{A,C}), \mathcal{L}(\mathbf{x}_{\hat{s},t}^{A,C})) \leq 2\varphi(R)^{-1}e^{-\tilde{a}(t-\hat{s})}W_1(\mathcal{L}(\mathbf{x}_{s,\hat{s}}^{A,C}), \mathcal{L}(\mathbf{x}_{\hat{s},\hat{s}}^{A,C})). \tag{61}$$

By starting the analysis of the processes $\mathbf{x}^A$ and $\mathbf{x}^C$ at time s, rather than time 0, Lemma 8 implies the following bound:

$$\begin{aligned}
W_1(\mathcal{L}(\mathbf{x}_{s,\hat{s}}^{A,C}), \mathcal{L}(\mathbf{x}_{\hat{s},\hat{s}}^{A,C})) &= W_1(\mathcal{L}(\mathbf{x}_{s,\hat{s}}^{A,C}), \mathcal{L}(\mathbf{x}_{\hat{s}}^A)) \\
&\leq \left(\left(c_8 + c_6\sqrt{\mathbb{E}[\|\mathbf{x}_s^A\|^2]}\right)\eta\sqrt{\hat{s}-s} + c_7\sqrt{\eta\log(4(\hat{s}-s))}\right) e^{\eta\ell(\hat{s}-s)} \\
&\leq \left(\left(c_8 + c_6\sqrt{\varsigma + c_{11}}\right)\eta\sqrt{\hat{s}-s} + c_7\sqrt{\eta\log(4(\hat{s}-s))}\right) e^{\eta\ell(\hat{s}-s)}.
\end{aligned} \tag{62}$$

The second inequality is based on Lemma 15.

Let $H = \lfloor 1/\eta \rfloor$ and $t \in [\hat{k}H, (\hat{k}+1)H)$ where $\hat{k} \in \mathbb{N}$, we have $\mathbf{x}_{0,t}^{A,C} = \mathbf{x}_t^C$ and $\mathbf{x}_{(\hat{k}+1)H,t}^{A,C} = \mathbf{x}_t^A = \mathbf{x}_{\lfloor t \rfloor}^A$. Then, the triangle inequality implies that

$$W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) \leq \sum_{i=0}^{\hat{k}} W_1(\mathcal{L}(\mathbf{x}_{iH,t}^{A,C}), \mathcal{L}(\mathbf{x}_{(i+1)H,t}^{A,C})).$$

For $i < \hat{k}$, setting $s = iH$, $\hat{s} = (i+1)H$ in (61) gives that

$$\begin{aligned}
W_1(\mathcal{L}(\mathbf{x}_{iH,t}^{A,C}), \mathcal{L}(\mathbf{x}_{(i+1)H,t}^{A,C})) &\leq 2\varphi(R)^{-1}e^{-\tilde{a}(t-(i+1)H)}W_1(\mathcal{L}(\mathbf{x}_{iH,(i+1)H}^{A,C}), \mathcal{L}(\mathbf{x}_{(i+1)H,(i+1)H}^{A,C})) \\
&\leq 2\varphi(R)^{-1}e^{-\eta a(t-(i+1)H)}g(H) \\
&\leq 2\varphi(R)^{-1}e^{-a\left(\hat{k}-i-1\right)/2)}g(\eta^{-1})
\end{aligned}$$

where

$$g(r) = \left(\left(c_8 + c_6 \sqrt{\varsigma + c_{11}} \right) \eta \sqrt{r} + c_7 \sqrt{\eta \log(4r)} \right) e^{\eta \ell r}. \tag{63}$$

The last inequality uses the facts that $1/2 \leq \eta H \leq 1$ along with monotonicity of g. The lower bound of ηH arises because $H \geq \eta^{-1} - 1$ and so $\eta H \geq 1 - \eta \geq 1/2$, since $\eta \leq 1/2$. Thus, the first $\hat{k}$ terms are bounded by:

$$\begin{aligned}\sum_{i=0}^{\hat{k}-1} W_1(\mathcal{L}(\mathbf{x}_{iH,t}^{A,C}), \mathcal{L}(\mathbf{x}_{(i+1)H,t}^{A,C})) &\leq \sum_{i=0}^{\hat{k}-1} 2\varphi(R)^{-1} e^{-a\left(\hat{k}-i-1)/2\right)} g(\eta^{-1}) \\ &\leq 2\varphi(R)^{-1} \frac{g(\eta^{-1})}{1 - e^{-a/2}}\end{aligned}$$

For $i = \hat{k}$,

$$\begin{aligned}W_1(\mathcal{L}(\mathbf{x}_{iH,t}^{A,C}), \mathcal{L}(\mathbf{x}_{(i+1)H,t}^{A,C})) &= W_1(\mathcal{L}(\mathbf{x}_{\hat{k}H,t}^{A,C}), \mathcal{L}(\mathbf{x}_t^A)) \\ &\leq g(t - \hat{k}H) \leq g(\eta^{-1})\end{aligned}$$

By triangle inequality, adding all the $\hat{k} + 1$ terms gives

$$\begin{aligned}&W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) \\ &\leq g(\eta^{-1}) \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right) \\ &\leq \left(\left(c_8 + c_6 \sqrt{\varsigma + c_{11}} \right) \eta \sqrt{\eta^{-1}} + c_7 \sqrt{\eta \log(4\eta^{-1})} \right) e^{\ell} \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right).\end{aligned} \tag{64}$$

For $\eta^{-1} \geq 4$, we have $\log(4\eta^{-1}) \leq 2\log(\eta^{-1})$, and also $\log \eta^{-1} > 1$. Thus, if $\eta \leq 1/4$, then (64) can be further upper bounded by

$$\begin{aligned}&W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) \\ &\leq \left(\left(c_8 + c_6 \sqrt{\varsigma + c_{11}} \right) \eta \sqrt{\eta^{-1} \log(\eta^{-1})} + c_7 \sqrt{2\eta \log(\eta^{-1})} \right) e^{\ell} \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right) \\ &= \left(c_8 + c_6 \sqrt{\varsigma + c_{11}} + \sqrt{2} c_7 \right) e^{\ell} \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right) \sqrt{\eta \log(\eta^{-1})} \\ &\leq \left(c_8 + c_6 \sqrt{c_{11}} + \sqrt{2} c_7 + c_6 \sqrt{\varsigma} \right) e^{\ell} \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right) \sqrt{\eta \log(\eta^{-1})}.\end{aligned}$$

So setting

$$\begin{aligned}c_3 &= \left(c_8 + c_6 \sqrt{c_{11}} + \sqrt{2} c_7 \right) e^{\ell} \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right) \\ c_4 &= c_6 e^{\ell} \left(1 + \frac{2\varphi(R)^{-1}}{1 - e^{-a/2}} \right)\end{aligned}$$

completes the proof. ■

H Bounding the constants

In this section, we summarize all the constants in Table 1. The second column of the table points to the place where these values are defined or computed. Then we show the simplified bounds of the main constants c_1, c_2, c_3, c_4, a in Theorem 1 explicitly and also discuss their dependencies on state dimension n and parameter β.

Proposition 19. *The constants c_2 and c_4 grow linearly with n. The constants c_1 and c_3 have $O(\sqrt{n})$ and $O(n)$ dependencies respectively. So overall, the dimension dependency of convergence guarantee is $O(n)$. Constants c_1, c_2, c_3, c_4 all grow exponentially with respect to $\frac{\beta \ell R^2}{8}$. And for all $\beta > 0$, $a \geq \frac{32\eta}{\mu \beta^2} e^{\frac{\beta \ell R^2}{8}}$.*

Table 1: List of constants

Constant	Definition
$a = \frac{2\xi}{\beta}$ $c_1 = 2\varphi(R)^{-1}\sqrt{\frac{2}{\mu}c_{10}}$ $c_2 = 4\varphi(R)^{-1}$	Appendix D.2 (Proof of Lemma 2)
$c_3 = \left(c_8 + c_6\sqrt{c_{11}} + \sqrt{2}c_7\right) e^{\ell}\left(1 + \frac{2\varphi(R)^{-1}}{1-e^{-a/2}}\right)$ $c_4 = c_6 e^{\ell}\left(1 + \frac{2\varphi(R)^{-1}}{1-e^{-a/2}}\right)$	Appendix G (Proof of Lemma 3)
$c_5 = (c_9+1)\sqrt{\frac{2}{\mu}\ell^2 c_{10} + 2\|\nabla_x \bar{f}(0)\|^2}$ $c_6 = (c_9+1)\sqrt{2\ell^2}$ $c_7 = (c_9+1)n\sqrt{\frac{8}{\beta}}$	Appendix F (Proof of Lemma 6)
$c_8 = c_5 + 2\ell\Psi_2(\mathbf{z})$	Section 3.3 (Proof of Lemma 8)
$c_9 = 6\left(\frac{1}{\alpha}\right)^{\mathrm{rank}(A)/2}$	Appendix A.3 (Proof of Lemma 11)
$c_{10} = (\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\| + \frac{n}{\beta}$	Appendix C.1 (Proof of Lemma 13)
$c_{11} = \frac{4}{\mu}\left(\frac{n}{\beta} + (\ell+\mu)R^2 + (2+R)\|\nabla_x \bar{f}(0)\| + \left(8\ell^2 + \frac{1}{\mu}\right)\ell^2\mathcal{M}_2(\mathbf{z})\right)$	Appendix C.2 (Proof of Lemma 15)

Proof of Proposition 19 Combining (47) and (48) gives

$$\begin{aligned}\xi &\geq (R_1 - R_0)\Phi(R_1)\varphi(R_0)^{-1}/2\\ &\geq (R_1 - R_0)R_1 e^{-h(R_0)}\varphi(R_0)^{-1}/2\\ &= (R_1 - R_0)R_1/2\end{aligned}$$

Here we choose $R_0 = R$ and $R_1 = \frac{R}{2} + \frac{1}{2}\sqrt{R^2 + \frac{32\eta}{\mu\beta}e^{\frac{\beta\ell R^2}{8}}}$ as stated in Corollary 17.

Then,

$$a = \frac{2\xi}{\beta} \geq \frac{(R_1 - R)R_1}{\beta} = \frac{32\eta}{\mu\beta^2}e^{\frac{\beta\ell R^2}{8}}.$$

We have the following inequality:

$$\begin{aligned}\frac{1}{1-e^{-a/2}} &\leq \max\left\{\frac{4}{a}, \frac{1}{1-e^{-1}}\right\}\\ &\leq \max\left\{\frac{\mu\beta^2}{8\eta}e^{-\frac{\beta\ell R^2}{8}}, \frac{1}{1-e^{-1}}\right\}\end{aligned}$$

which uses the fact that for all $y > 0$, $\frac{1}{1-e^{-y}} \leq \max\left\{\frac{2}{y}, \frac{1}{1-e^{-1}}\right\}$ and see the proof detail in [27].

So

$$1 + \frac{2\varphi(R)^{-1}}{1-e^{-a/2}} \leq 1 + \max\left\{\frac{\mu\beta^2}{4\eta}, \frac{e^{\frac{\beta\ell R^2}{8}}}{1-e^{-1}}\right\} \tag{65}$$

Without loss of generality, assume $\beta > 1$. Then, plugging the definition of ξ and φ and (65) gives

$$
\begin{aligned}
c_1 &= 2e^{\beta\ell R^2/8}\sqrt{\frac{2}{\mu}\left((\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\| + \frac{n}{\beta}\right)} \\
&\leq 2e^{\beta\ell R^2/8}\sqrt{\frac{2}{\mu}\left((\ell+\mu)R^2 + R\|\nabla_x \bar{f}(0)\| + n\right)} \\
c_2 &= 4e^{\beta\ell R^2/8} \\
c_3 &= \left(c_8 + c_6\sqrt{c_{11}} + \sqrt{2}c_7\right)e^{\ell}\left(1 + \frac{2\varphi(R)^{-1}}{1-e^{-a/2}}\right) \\
&\leq \left(c_8 + c_6\sqrt{c_{11}} + \sqrt{2}c_7\right)e^{\ell}\left(1 + \max\left\{\frac{\mu\beta^2}{4\eta}, \frac{e^{\frac{\beta\ell R^2}{8}}}{1-e^{-1}}\right\}\right) \\
&\leq p(\sqrt{n})e^{\ell}\left(1 + \max\left\{\frac{\mu\beta^2}{4\eta}, \frac{e^{\frac{\beta\ell R^2}{8}}}{1-e^{-1}}\right\}\right) \\
c_4 &= \left(6(\frac{1}{\alpha})^{\text{rank}(A)/2} + 1\right)\sqrt{2\ell^2}e^{\ell}\left(1 + \frac{2\varphi(R)^{-1}}{1-e^{-a/2}}\right) \\
&\leq \left(6(\frac{1}{\alpha})^{\text{rank}(A)/2} + 1\right)\sqrt{2\ell^2}e^{\ell}\left(1 + \max\left\{\frac{\mu\beta^2}{4\eta}, \frac{e^{\frac{\beta\ell R^2}{8}}}{1-e^{-1}}\right\}\right).
\end{aligned}
$$

For constant c_3, $p(\sqrt{n})$ is a monotonically increasing polynomial (independent of η and β). The upper bound of c_3 is derived by direct observation of the corresponding constants.

We can see neither c_2 nor c_4 depends on the state dimension, so the two constants grow linearly with n. The constant c_1 are $O(\sqrt{n})$ and c_3 are $O(n)$. As for the dependencies on β, we can see that all four constants are $O(e^{\frac{\beta\ell R^2}{8}})$.

■

I Near-optimality of Gibbs distributions

In this appendix, we prove Proposition 21 which shows that $\mathbf{x}_k$ can be near-optimal. The proof closely follows [27] and [35]. The main difference is that in our case we have to deal with the unbounded polyhedral constraint, while in [35] there is no constraint and in [27] the constraint is compact.

Firstly, we need a preliminary result shown as below.

Lemma 20. *Assume $\mathbf{x}$ is drawn according to $\pi_{\beta\bar{f}}$. There exists a positive constant c_{12} such that the following bounds hold:*

$$\mathbb{E}[\bar{f}(\mathbf{x})] \leq \min_{x\in\mathcal{K}} \bar{f}(x) + \frac{n}{\beta}\left(c_{12} + \log n\right)$$

where $c_{12} = 2\log(\varsigma + \frac{1}{\mu}c_{10}) + \frac{1}{6}\log 3 + \log 2\sqrt{\pi} - \log r_{\min}$.

Proof of Lemma 20 Recall that the probability measure $\pi_{\beta\bar{f}(A)}$ is defined by $\pi_{\beta\bar{f}(A)}(A) = \frac{\int_{A\cap\mathcal{K}} e^{-\beta\bar{f}(x)}dx}{\int_{\mathcal{K}} e^{-\beta\bar{f}(y)}dy}$.

Let $\Lambda = \int_{\mathcal{K}} e^{-\beta\bar{f}(y)}dy$ and $p(x) = \frac{e^{-\beta\bar{f}(x)}}{\Lambda}$. So $\log p(x) = -\beta\bar{f}(x) - \log\Lambda$, which implies that $\bar{f}(x) = -\frac{1}{\beta}\log p(x) - \frac{1}{\beta}\log\Lambda$. Then we have

$$
\begin{aligned}
\mathbb{E}_{\pi_{\beta\bar{f}}}[\bar{f}(\mathbf{x})] &= \int_{\mathcal{K}} \bar{f}(x)p(x)dx \\
&= -\frac{1}{\beta}\int_{\mathcal{K}} p(x)\log p(x)dx - \frac{1}{\beta}\log\Lambda. \qquad (66)
\end{aligned}
$$

We can bound the first term by maximizing the differential entropy.

Let $h(x) = -\int_{\mathcal{K}} p(x) \log p(x) dx$. Using the fact that the differential entropy with a finite moments is upper-bounded by that of a Gaussian density with the same second moment (see Theorem 8.6.5 in [15]), we have

$$h(x) \leq \frac{n}{2} \log(2\pi e \sigma^2) \leq \frac{n}{2} \log(2\pi e(\varsigma + \frac{1}{\mu} c_{10})), \tag{67}$$

where $\sigma^2 = \mathbb{E}_{\pi_{\beta \bar{f}}}[\|\mathbf{x}\|^2]$ and the second inequality uses Lemma 14.

We aim to derive the upper bound of the second term of (66).

First we show that there is a vector $x^\star \in \mathcal{K}$ which minimizes $\bar{f}$ over $\mathcal{K}$. In other words, an optimal solution exists. The bound (25) from the proof of Lemma 12 implies that $\bar{f}(x) \geq \bar{f}(0) + 1$ for all sufficiently large x. This implies that there is a compact ball, B such that if $x_n \in \mathcal{K}$ is a sequence such that $\lim_{n\to\infty} \bar{f}(x_n) = \inf_{x\in\mathcal{K}} \bar{f}(x)$, then x_n must be in $B \cap \mathcal{K}$ for all sufficiently large n. Then since $\bar{f}$ is continuous and $B \cap \mathcal{K}$ is compact, there must be a limit point $x^\star \in B \cap \mathcal{K}$ which minimizes $\bar{f}$.

Let $x^* \in \mathcal{K}$ be a minimizer. The normalizing constant can be expressed as:

$$\begin{aligned}
\log \Lambda &= \log \int_{\mathcal{K}} e^{-\beta \bar{f}(x)} dx \\
&= \log e^{-\beta \bar{f}(x^*)} \int_{\mathcal{K}} e^{\beta\left(\bar{f}(x^*) - \bar{f}(x)\right)} \\
&= -\beta \bar{f}(x^*) + \log \int_{\mathcal{K}} e^{\beta\left(\bar{f}(x^*) - \bar{f}(x)\right)} dx
\end{aligned}$$

So, to derive our desired upper bound on $-\log \Lambda$, it suffices to derive a lower bound on $\int_{\mathcal{K}} e^{\beta\left(\bar{f}(x^*) - \bar{f}(x)\right)} dx$.

We have

$$\bar{f}(x) - \bar{f}(x^*) = \int_0^1 \nabla \bar{f}(x^* + t(x - x^*))^\top (x - x^*) dt.$$

Let $y = x^* + t(x - x^*)$, $t \in [0, 1]$, then

$$\begin{aligned}
\|\nabla \bar{f}(y)\| &= \|\nabla \bar{f}(y) - \nabla \bar{f}(x^*) + \nabla \bar{f}(x^*) - \nabla \bar{f}(0) + \nabla \bar{f}(0)\| \\
&\leq \ell \|y - x^*\| + \ell \|x^*\| + \|\nabla \bar{f}(0)\| \\
&\leq \ell \|x - x^*\| t + \ell \|x^*\| + \|\nabla \bar{f}(0)\|.
\end{aligned}$$

We can show $\|x^*\|$ is upper bounded by $\max\{R, \frac{\nabla \|\bar{f}(0)\|}{\mu}\}$.

We have to find the bound for the case $\|x^*\| > R$.

The convexity outside a ball assumption gives

$$\left(\nabla \bar{f}(x^*) - \nabla \bar{f}(0)\right)^\top x^* \geq \mu \|x^*\|^2. \tag{68}$$

The optimality of x^* gives $-\nabla \bar{f}(x^*) \in N_{\mathcal{K}}(x^*)$, which is to say for all $y \in \mathcal{K}$, $-\nabla \bar{f}(x^*)^\top (y - x^*) \leq 0$. Since $0 \in \mathcal{K}$, $\nabla \bar{f}(x^*)^\top x^* \leq 0$ holds. Applying the Cauchy-Schwartz inequality to the left hand side of (68) gives

$$\|\nabla \bar{f}(0)\| \|x^*\| \geq \mu \|x^*\|^2.$$

This implies that $\|x^*\| \leq \frac{\|\nabla \bar{f}(0)\|}{\mu}$. So we can conclude that $\|x^*\| \leq \max\{R, \frac{\|\nabla \bar{f}(0)\|}{\mu}\} = c_{13}$.

Therefore,

$$\begin{aligned}\bar{f}(x) - \bar{f}(x^*) &\leq \int_0^1 \|\nabla \bar{f}(y)\| \|x - x^*\| dt \\ &\leq \frac{\ell}{2}\|x - x^*\|^2 + \left(\ell\|x^*\| + \|\nabla \bar{f}(0)\|\right)\|x - x^*\| \\ &\leq \frac{\ell}{2}\|x - x^*\|^2 + \left(\ell c_{13} + \|\nabla \bar{f}(0)\|\right)\|x - x^*\|.\end{aligned}$$

Here we set

$$\begin{aligned}& e^{\beta\left(\bar{f}(x^*) - \bar{f}(x)\right)} \geq 1/2 \\ \iff & \beta\left(\bar{f}(x^*) - \bar{f}(x)\right) \geq -\log 2 \\ \Longleftarrow & -\frac{\ell}{2}\|x - x^*\|^2 - \left(\ell c_{13} + \|\nabla \bar{f}(0)\|\right)\|x - x^*\| \geq -\frac{1}{\beta}\log 2.\end{aligned}$$

So solving the corresponding quadratic equation and taking the positive root gives an upper bound of $\|x - x^*\|$:

$$\|x - x^*\| \leq -\frac{1}{\ell}\left(\ell c_{13} + \|\nabla \bar{f}(0)\|\right) + \frac{1}{\ell}\sqrt{\left(\ell c_{13} + \|\nabla \bar{f}(0)\|\right)^2 + 2\ell\frac{1}{\beta}\log 2}.$$

So let $\epsilon = -\frac{1}{\ell}\left(\ell c_{13} + \|\nabla \bar{f}(0)\|\right) + \frac{1}{\ell}\sqrt{\left(\ell c_{13} + \|\nabla \bar{f}(0)\|\right)^2 + 2\ell\frac{1}{\beta}\log 2}$ and let $\mathcal{B}_{x^*}(\epsilon)$ be the ball of radius ϵ centered at x^*. Then we want to find a ball $\mathcal{S}$ such that

$$\int_{\mathcal{K}} e^{\beta\left(\bar{f}(x^*) - \bar{f}(x)\right)} dx \geq \frac{1}{2}\text{vol}(\mathcal{K} \cap \mathcal{B}_{x^*}(\epsilon)) \geq \frac{1}{2}\text{vol}(\mathcal{S}).$$

To find the desired ball $\mathcal{S}$, we consider the problem of finding the largest ball inscribed within $\mathcal{K} \cap \mathbf{B}_{x^\star}(\epsilon)$. This is a Chebyshev centering problem, and can be formulated as the following convex optimization problem.

$$\begin{aligned}\max_{r,y} \quad & r && \text{(69a)}\\ \text{subject to} \quad & Ay \leq b - r\mathbf{1} && \text{(69b)}\\ & \|x^* - y\| + r \leq \epsilon && \text{(69c)}\end{aligned}$$

where r and y denotes the radius and the center of the Chebyshev ball respectively. The particular form arises because the rows of A are unit vectors, and so the ball of radius r around y is inscribed in $\mathcal{K}$ if and only if (69b) holds, while this ball is contained in $\mathbf{B}_{x^\star}(\epsilon)$ if and only if (69c) holds.

We rewrite this optimization problem as:

$$\begin{aligned}\min_{r,y} \quad & -r + I_S(x^*, [r, y]^\top) && \text{(70a)}\\ \text{subject to} \quad & Ay \leq b - r\mathbf{1} && \text{(70b)}\end{aligned}$$

where $S = \{(x^*, [r, y^\top]^\top) | \|x^* - y\| + r < \epsilon\}$.

Here, I_S is defined by

$$I_S(x, [r, y^\top]^\top) = \begin{cases} +\infty & \text{if } (x, [r, y^\top]^\top) \notin S \\ 0 & \text{otherwise.}\end{cases} \tag{71}$$

Let $g(x^\star)$ denote the optimal value of (70). We will show that there is a positive constant $r_{\min} > 0$ such that $-g(x) \geq r_{\min}$ for all $x \in \mathcal{K}$. As a result, for any $x^\star$ the corresponding Chebyshev centering solutions has radius at least $r_{\min}$.

Let $F(x^*, [r, y^\top]^\top) = -r + I_s(x^*, [r, y^\top]^\top)$. We can see that F is convex in $(x^*, [r, y^\top]^\top)$ and $\text{dom } F = S$.

Let $C = \{[r, y^\top]^\top | [\mathbf{1}\ A][r, y]^\top \leq b\}$. Then the optimal value of (70) can be expressed as $g(x) = \inf_{[r,y^\top]^\top \in C} F(x, [r, y^\top]^\top)$ and $\operatorname{dom} g = \{x | \exists [r, y^\top]^\top \in C \text{ s.t. } (x, [r, y^\top]^\top) \in S\}$.

The results of Section 3.2.5 of [6] imply that if F is convex, S is convex, and $g(x) > -\infty$ for all x, then g is also convex.

If $(x, [r, y^\top]^\top) \in \operatorname{dom} F$, then

$$\begin{aligned} \|x - y\| + r \leq \epsilon &\implies r \leq \epsilon - \|x - y\| \\ &\implies -r \geq -\epsilon + \|x - y\| > -\infty. \end{aligned}$$

In particular, if there exist y, r such that $(x, [r, y^\top]^\top) \in \operatorname{dom} F$, then $\inf_{[r,y^\top]^\top \in C} F(x, [r, y^\top]^\top) \geq -\epsilon$.

There are two cases:

- If there exist $[r, y^\top]^\top \in C$ such that $(x, [r, y^\top]^\top) \in \operatorname{dom} F$, then $\inf_{[r,y^\top]^\top \in C} F(x, [r, y^\top]\top)$ is finite and bounded below.
- If there does not exist $[r, y^\top]^\top \in C$ such that $(x, [r, y^\top]^\top) \in \operatorname{dom} F$, then for all $[r, y^\top]^\top \in C$, $F(x, [r, y^\top]^\top) = +\infty$. So $g(x) = \inf_{[r,y^\top]^\top \in C} F(x, [r, y^\top]\top) = +\infty > -\infty$.

Hereby, we can conclude that for all x , $g(x) > -\infty$, so $g(x)$ is convex.

So, to found a lower bound on the inscribed radius, we want to maximize $g(x)$ over $\mathcal{K}$. Specifically, we analyze the following optimization problem

$$\max_{x \in \mathcal{K}} \qquad g(x) \tag{72a}$$

which corresponds to maximizing a convex function over a convex set.

Note that $\mathcal{K} \subset \operatorname{dom}(g)$. In particular, if $x \in \mathcal{K}$, then $(x, [0, x^\top]^\top) \in S$, which implies that $g(x) \leq 0$. Thus, $g(x) \leq 0$ for all $x \in \mathcal{K}$. Therefore, using Theorem 32.2 [38], given $\mathcal{K}$ is closed convex by our assumption and $g(x)$ is bounded above gives

$$\sup \{g(x) | x \in \mathcal{K}\} = \sup \{g(x) | x \in E\}$$

where E is a subset of $\mathcal{K}$ consisting of the extreme points of $\mathcal{K} \cap L^\perp$, where L is the linearity space of C and $L = \{x | Ax = 0\} = \mathcal{N}(A)$.

Now, we will show that E is a finite set.

Let

$$A = [U_1 \quad U_2] \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix},$$

then $\mathcal{N}(A) = L = \mathcal{R}(V_2)$ and $L^\perp = \mathcal{R}(V_1)$, and

$$K \cap L^\perp = \{V_1 Z_1 | A V_1 Z_1 \leq b\}.$$

This is a polyhedral with no lines so has a finite set of extreme points, i.e. E is finite. Then, based on the process shown in [27], we have $\operatorname{vol}(\mathcal{S}) \geq \frac{\pi^{n/2}}{\Gamma(n/2+1)} r_{\min}^n$, using the fact that a ball of radius ρ has volumn given by $\frac{\pi^{n/2}}{\Gamma(n/2+1)} \rho^n$

Then, utilizing an upper bound of Gamma function recorded in [36] shown as below:

$$\Gamma(x+1) < \sqrt{\pi} \left(\frac{x}{e}\right)^x \left(8x^3 + 4x^2 + x + \frac{1}{30}\right)^{1/6}, \ x \geq 0. \tag{73}$$

Setting $x = \frac{n}{2}$ in (73) gives:

$$\Gamma(\frac{n}{2} + 1) < \sqrt{\pi} \left(\frac{n}{2e}\right)^{\frac{n}{2}} \left(n^3 + n^2 + \frac{n}{2} + \frac{1}{30}\right)^{1/6}. \tag{74}$$

Therefore, we can find the lower bound of $\log \frac{1}{2}\text{vol}(S)$:

$$\begin{aligned}
\log \frac{1}{2}\text{vol}(S) &= \log \frac{\pi^{n/2}}{\Gamma(n/2+1)} r_{\min}^n - \log 2 \\
&> \frac{n}{2}\log \pi + n \log r_{\min} - \log \left\{ \sqrt{\pi} \left(\frac{n}{2e}\right)^{\frac{n}{2}} \left(n^3 + n^2 + \frac{n}{2} + \frac{1}{30}\right)^{1/6} \right\} - \log 2 \\
&= -\frac{1}{2}\log \pi + n \log r_{\min} + \frac{n}{2}\log(2\pi e) - \frac{n}{2}\log n - \frac{1}{6}\log\left(n^3 + n^2 + \frac{n}{2} + \frac{1}{30}\right) - \log 2 \\
&> n \log r_{\min} + \frac{n}{2}\log(2\pi e) - \frac{n}{2}\log n - \frac{1}{6}\log\left(3n^3\right) - \log(2\sqrt{\pi}) \qquad (75)
\end{aligned}$$

The last inequality holds because $n \geq 1$.

Plugging (75) and (67) in (66) gives

$$\begin{aligned}
\mathbb{E}_{\pi_{\beta \bar{f}}}[\bar{f}(x)] &< \min f(x) + \frac{n}{2\beta}\log(2\pi e(\varsigma + \frac{1}{\mu}c_{10})) \\
&\quad - \frac{1}{\beta}\left(n \log r_{\min} + \frac{n}{2}\log(2\pi e) - \frac{n}{2}\log n - \frac{1}{2}\log n - \frac{1}{6}\log 3 - \log 2\sqrt{\pi}\right) \\
&= \min f(x) + \frac{n}{2\beta}\log(\varsigma + \frac{1}{\mu}c_{10}) - \frac{1}{\beta}\left(n \log r_{\min} - \frac{n}{2}\log n - \frac{1}{2}\log n\right) \\
&\quad + \frac{1}{\beta}(\frac{1}{6}\log 3 + \log 2\sqrt{\pi}) \\
&\leq \min f(x) + \frac{n}{\beta}\left(c_{12} + \log n\right)
\end{aligned}$$

where $c_{12} = 2\log(\varsigma + \frac{1}{\mu}c_{10}) + \frac{1}{6}\log 3 + \log 2\sqrt{\pi} - \log r_{\min}$ and the last inequality holds because $n \geq 1$.

■

Proposition 21. *Let $\mathbf{x}_k$ be the iterates of the algorithms and assume $\eta \leq \frac{\mu}{3\ell^2}$ and $\mathbb{E}[\|\mathbf{x}_o\|^{2q}] < \infty$ for all $q > 1$. There exist positive constants c_{12}, c_{14}, c_{15} such that for all $q > 1$ and integers $k \geq 0$, the following bound holds:*

$$\mathbb{E}[\bar{f}(\mathbf{x}_k)] \leq \min_{x \in \mathcal{K}} \bar{f}(x) + c_{14} W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}}) + c_{15} W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}})^{\frac{2-2q}{1-2q}} + \frac{n}{\beta}\left(c_{12} + \log n\right),$$

where

$$\begin{aligned}
c_{12} &= 2\log(\varsigma + \frac{1}{\mu}c_{10}) + \frac{1}{6}\log 3 + \log 2\sqrt{\pi} - \log r_{\min} \\
c_{14} &= \|\nabla \bar{f}(0)\| + \ell \frac{2\left(\|\nabla \bar{f}(0)\| + \sqrt{\|\nabla \bar{f}(0)\|} + \sqrt{\mu \bar{f}(0)}\right)}{\mu} \\
c_{15} &= \left(\frac{2\ell}{\sqrt{\mu}} + C\right)\left(\frac{\ell}{\sqrt{\mu}(q-1)C}\right)^{\frac{2-2q}{-2q+1}}
\end{aligned}$$

Proof of Proposition 21 Let $\mathbf{x}$ be drawn according to $\pi_{\beta \bar{f}}$. Then Lemma 20 implies:

$$\begin{aligned}
\mathbb{E}[\bar{f}(\mathbf{x}_k)] &= \mathbb{E}[\bar{f}(\mathbf{x})] + \mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(\mathbf{x})] \\
&\leq \min_{x \in \mathcal{K}} \bar{f}(x) + \mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(\mathbf{x})] + \frac{n}{\beta}\left(c_{12} + \log n\right) \qquad (76)
\end{aligned}$$

So, it now suffices to bound $\mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(\mathbf{x})]$. Ideally, we would bound this term via Kantorovich duality. The problem is that $\bar{f}$ may not be globally Lipschitz. So, we must approximate it with a Lipschitz function, and then bound the gap induced by this approximation.

Namely, fix a constant $m > \bar{f}(0)$ with $m \geq$, to be chosen later. Set $g(x) = \min\{\bar{f}(x), m\}$. The inequality from (25) implies that if $\|x\| \geq \hat{R} := \frac{2\left(\|\nabla \bar{f}(0)\| + \sqrt{\|\nabla \bar{f}(0)\| + \mu(m - \bar{f}(0))}\right)}{\mu}$, then $\bar{f}(x) \geq m$. We claim that g is globally Lipschitz.

For $\|x\| \leq \hat{R}$, we have that

$$\|\nabla \bar{f}(x)\| \leq \|\nabla \bar{f}(0)\| + \ell \hat{R} =: u.$$

We will show that g is u-Lipschitz.

In the case that $f(y) \geq m$ and $f(x) \geq m$, we have $|g(x) - g(y)| = 0$, so the property holds.

Now say that $\bar{f}(x) < m$ and $\bar{f}(y) < m$. Then we must have $\|x\| \leq \hat{R}$ and $\|y\| \leq \hat{R}$. Then for all $t \in [0, 1]$, we have $\|(1-t)x + ty\| \leq \hat{R}$. It follows that

$$\begin{aligned} g(x) - g(y) &= \bar{f}(x) - \bar{f}(y) \\ &= \int_0^1 \nabla \bar{f}(x + t(y - x))^\top (y - x) dt \\ &\leq u\|x - y\|. \end{aligned}$$

Finally, consider the case that $\bar{f}(x) \geq m$ and $\bar{f}(y) < m$. Then there is some $\theta \in [0, 1]$ such that $\bar{f}(y + \theta(x - y)) = m$. Furthermore

$$\begin{aligned} |g(x) - g(y)| &= m - \bar{f}(y) \\ &= \bar{f}(y + \theta(x - y)) - \bar{f}(y) \\ &= \int_0^\theta \nabla \bar{f}(y + t(x - y))^\top (x - y) dt \\ &\leq u\|x - y\|. \end{aligned}$$

It follows that g is u-Lipschitz.

Now noting that $g(x) \leq \bar{f}(x)$ for all x gives

$$\begin{aligned} \mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(\mathbf{x})] &\leq \mathbb{E}[\bar{f}(\mathbf{x}_k) - g(\mathbf{x})] \\ &= \mathbb{E}[g(\mathbf{x}_k) - g(\mathbf{x})] + \mathbb{E}[\mathbb{1}(\bar{f}(\mathbf{x}_k) > m)(\bar{f}(\mathbf{x}_k) - m)] \\ &\leq uW_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}}) + \mathbb{E}[\mathbb{1}(\bar{f}(\mathbf{x}_k) > m)(\bar{f}(\mathbf{x}_k) - m)]. \end{aligned} \tag{77}$$

The final inequality uses Kantorovich duality. Now, it remains to bound $\mathbb{E}[\mathbb{1}(\bar{f}(\mathbf{x}_k) > m)(\bar{f}(\mathbf{x}_k) - m)]$.

Note that if $\mathbf{y}$ is a non-negative random variable, a standard identity gives that $\mathbb{E}[\mathbf{y}] = \int_0^\infty \mathbb{P}(\mathbf{y} > \epsilon) d\epsilon$. Thus, we have

$$\mathbb{E}[\mathbb{1}(\bar{f}(\mathbf{x}_k) > m)(\bar{f}(\mathbf{x}_k) - m)] = \int_0^\infty \mathbb{P}(\bar{f}(\mathbf{x}_k) - m > \epsilon) d\epsilon.$$

For all $x \in \mathcal{K}$, we have

$$\begin{aligned} \bar{f}(x) &= \bar{f}(0) - \nabla \bar{f}(0)^\top x + \int_0^1 (\nabla \bar{f}(tx) - \nabla \bar{f}(0))^\top x dt \\ &\leq \bar{f}(0) + \|\nabla \bar{f}(0)\|\|x\| + \frac{1}{2}\ell\|x\|^2 \\ &\leq \bar{f}(0) + \frac{\|\nabla \bar{f}(0)\|^2}{2\ell} + \ell\|x\|^2. \end{aligned}$$

So,

$$\begin{aligned} \bar{f}(x) - m > \epsilon \implies \bar{f}(0) + \frac{\|\nabla \bar{f}(0)\|^2}{2\ell} + \ell\|x\|^2 > m + \epsilon \\ \iff \|x\|^2 > \frac{m + \epsilon - \left(\bar{f}(0) + \frac{\|\nabla \bar{f}(0)\|^2}{2\ell}\right)}{\ell}. \end{aligned}$$

Now assume that $m/2 > \bar{f}(0) + \frac{\|\nabla \bar{f}(0)\|^2}{2\ell}$. Then the right side implies $\|x\|^2 \geq \frac{\frac{m}{2}-\epsilon}{\ell}$. It follows that for any $q > 1$, we have, via Markov's inequality and direct computation:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(\bar{f}(\mathbf{x}_k) > m)(\bar{f}(\mathbf{x}_k) - m)] &\leq \int_0^\infty \mathbb{P}\left(\|\mathbf{x}_k\|^2 > \frac{\frac{m}{2}+\epsilon}{\ell}\right) d\epsilon \\
&= \int_0^\infty \mathbb{P}\left(\|\mathbf{x}_k\|^{2q} > \left(\frac{\frac{m}{2}+\epsilon}{\ell}\right)^q\right) d\epsilon \\
&\leq \mathbb{E}[\|\mathbf{x}_k\|^{2q}] \int_0^\infty \left(\frac{\frac{m}{2}+\epsilon}{\ell}\right)^{-q} d\epsilon \\
&= \mathbb{E}[\|\mathbf{x}_k\|^{2q}] \frac{\ell^q 2^{q-1}}{(q-1)m^{q-1}}.
\end{aligned}$$

Plugging this expression into (77) and using the definition of u gives

$$\begin{aligned}
&\mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(\mathbf{x})] \\
&\leq \left(\|\nabla \bar{f}(0)\| + \ell \frac{2\left(\|\nabla \bar{f}(0)\| + \sqrt{\|\nabla \bar{f}(0)\| + \mu(m - \bar{f}(0))}\right)}{\mu}\right) W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}}) \\
&\quad + \mathbb{E}[\|\mathbf{x}_k\|^{2q}] \frac{\ell^q 2^{q-1}}{(q-1)m^{q-1}}.
\end{aligned} \tag{78}$$

We want to derive the bound of $\mathbb{E}[\|\mathbf{x}_k\|^{2q}]$.

We have

$$\|\mathbf{x}_{k+1}\|^{2q} \leq \|\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k, \mathbf{z}_k) + \sqrt{\frac{2\eta}{\beta}} \hat{\mathbf{w}}_k\|^2.$$

For notation simplicity, let $y = \frac{\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k, \mathbf{z}_k)}{\sqrt{2\eta/\beta}}$ and $w = \hat{\mathbf{w}}_k$, so the above inequality can be expressed as

$$\begin{aligned}
\|\mathbf{x}_{k+1}\|^{2q} &\leq \left(\frac{2\eta}{\beta}\right)^q \|y + w\|^{2q} \\
&= \left(\frac{2\eta}{\beta}\right)^q \left(\|y\|^2 + \|w\|^2 + 2y^\top w\right)^q \\
&= \left(\frac{2\eta}{\beta}\right)^q \sum_{k=0}^q \binom{q}{k} \left(2y^\top w\right)^{q-k} \left(\|y\|^2 + \|w\|^2\right)^k \\
&= \left(\frac{2\eta}{\beta}\right)^q \sum_{k=0}^q \binom{q}{k} \left(2y^\top w\right)^{q-k} \sum_{i=0}^k \binom{k}{i} \left(\|y\|^{2i} \|w\|^{2(k-i)}\right).
\end{aligned}$$

The last two equalities use binomial theorem. Taking expectation, we can observe that the right hand side bound can be expressed as $\left(\frac{2\eta}{\beta}\right)^q \mathbb{E}\left[\|y\|^{2q}\right] + \mathbb{E}\left[p(\|y\|)\right]$, where $p(\|y\|)$ is a polynomial in y with order strictly lower than $2q$. Note that $p(\|y\|)$ will be dominated by the higher order term if y is big. More specifically, the lower order terms will be dominated when $\mathbf{x}_k$ is big. When $\mathbf{x}_k$ is small, these terms can be bounded above by a constant.

So now we want to examine $\mathbb{E}[\|y\|^{2q}]$, which is to bound $\mathbb{E}[\|\mathbf{x}_k - \eta \nabla \bar{f}(\mathbf{x}_k)\|^{2q}]$. We have

$$\|\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^{2q} = \left(\|\mathbf{x}_k\|^2 - 2\eta \mathbf{x}_k^\top \nabla f(\mathbf{x}_k, \mathbf{z}_k) + \eta^2 \|\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^2\right)^q$$

We examine the second term:

$$\begin{aligned}
\mathbf{x}_k^\top \nabla f(\mathbf{x}_k, \mathbf{z}_k) &= \mathbf{x}_k^\top (\nabla \bar{f}(\mathbf{x}_k) - \nabla \bar{f}(0)) + \mathbf{x}_k^\top (\nabla \bar{f}(0) - \nabla \bar{f}(\mathbf{x}_k) + \nabla f(\mathbf{x}_k, \mathbf{z}_k)) \\
&\geq \mu \|\mathbf{x}_k\|^2 - (\ell + \mu) R^2 + \mathbf{x}_k^\top (\nabla \bar{f}(0) + \mathbb{E}_{\hat{\mathbf{z}}}[\nabla f(\mathbf{x}_k, \mathbf{z}_k) - \nabla f(\mathbf{x}_k, \hat{\mathbf{z}}_k)])
\end{aligned}$$

where the first term is bounded by the assumption of the strong convexity outside a ball and the detailed statement is shown below:

If $\|x\| \geq R$, then $x^\top(\nabla \bar{f}(x) - \nabla \bar{f}(0)) \geq \mu\|x\|^2$.

If $\|x\| \leq R$, then $x^\top(\nabla \bar{f}(x) - \nabla \bar{f}(0)) \geq -\ell\|x\|^2 \geq -\ell R^2$.

Therefore, we have for all $x \in \mathcal{K}$, $x^\top(\nabla \bar{f}(x) - \nabla \bar{f}(0)) \geq \mu\|x\|^2 - (\ell+\mu)R^2$.

Note here and below $\hat{\mathbf{z}}$ is independent of $\mathbf{z}$ and identically distributed.

Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^{2q}\right] \\
&= \mathbb{E}\left[\left(\|\mathbf{x}_k\|^2 - 2\eta\mathbf{x}_k^\top\nabla f(\mathbf{x}_k, \mathbf{z}_k) + \eta^2\|\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^2\right)^q\right] \\
&\leq \mathbb{E}\left[\left((1-2\mu\eta)\|\mathbf{x}_k\|^2 + 2\eta(\ell+\mu)R^2\right.\right. \\
&\qquad \left.\left. -2\eta\mathbf{x}_k^\top(\nabla\bar{f}(0) + \mathbb{E}_{\hat{\mathbf{z}}}[\nabla f(\mathbf{x}_k, \mathbf{z}_k) - \nabla f(\mathbf{x}_k, \hat{\mathbf{z}}_k)]) + \eta^2\|\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^2\right)^q\right] \\
&\leq \mathbb{E}\left[\left((1-2\mu\eta)\|\mathbf{x}_k\|^2 + 2\eta(\ell+\mu)R^2\right.\right. \\
&\qquad \left.\left. -2\eta\mathbf{x}_k^\top(\nabla\bar{f}(0) + \nabla f(\mathbf{x}_k, \mathbf{z}_k) - \nabla f(\mathbf{x}_k, \hat{\mathbf{z}}_k)) + \eta^2\|\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^2\right)^q\right] \\
&\leq \mathbb{E}\left[\left((1-2\mu\eta)\|\mathbf{x}_k\|^2 + 2\eta(\ell+\mu)R^2\right.\right. \\
&\qquad \left.\left. +2\eta\|\mathbf{x}_k\|(\|\nabla\bar{f}(0)\| + \ell(\mathbf{z}_k - \hat{\mathbf{z}}_k)) + \eta^2\|\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^2\right)^q\right]. \qquad (79)
\end{aligned}$$

The second inequality uses Jensen's inequality, and the last inequality uses Cauchy-Schwartz inequality together with ℓ-Lipschitzness of $\nabla f(x, z)$ in z.

Now we examine the last term above.

$$\begin{aligned}
\|\nabla f(x,z)\| &= \|\nabla f(x,z) - \mathbb{E}_{\hat{z}}[\nabla f(x,\hat{z})] + \mathbb{E}_{\hat{z}}[\nabla f(x,\hat{z})]\| \\
&\leq \|\mathbb{E}_{\hat{z}}[\nabla f(x,z) - \nabla f(x,\hat{z})]\| + \|\bar{f}(x)\| \\
&\leq \ell\mathbb{E}_{\hat{z}}[\|z-\hat{z}\|] + \|\nabla\bar{f}(0)\| + \ell\|x\|.
\end{aligned}$$

So

$$\|\nabla f(x,z)\|^2 \leq 3\left(\ell^2\left(\mathbb{E}_{\hat{z}}[\|z-\hat{z}\|]\right)^2 + \|\nabla\bar{f}(0)\|^2 + \ell^2\|x\|^2\right). \qquad (80)$$

Then, we can group the square terms in (79) together and simplify it:

$$\begin{aligned}
&(1-2\mu\eta)\|x\|^2 + \eta^2 3\ell^2\|x\|^2 \leq (1-\eta\mu)\|x\|^2 \\
\iff &1 - 2\mu\eta + \eta^2 3\ell^2 \leq 1 - \eta\mu \\
\iff &\eta \leq \frac{\mu}{3\ell^2}.
\end{aligned}$$

So, if $\eta \leq \frac{\mu}{3\ell^2}$, plugging (80) into (79) gives

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^{2q}\right] \leq \mathbb{E}\Big[&\Big((1-\mu\eta)\|\mathbf{x}_k\|^2 + 2\eta(\ell+\mu)R^2 \\
&-2\eta\|\mathbf{x}_k\|(\|\nabla\bar{f}(0)\| + \ell(\mathbf{z}_k - \hat{\mathbf{z}}_k)) + \eta^2 3\left(\ell^2\left(\mathbb{E}_{\hat{\mathbf{z}}}[\|\mathbf{z}_k - \hat{\mathbf{z}}_k\|]\right)^2\|\nabla\bar{f}(0)\|^2\right)\Big)^q\Big].
\end{aligned}$$

We want to further group the first and third terms above together.

Let $a = \|\mathbf{x}_k\|$, $b = \|\nabla\bar{f}(0)\| + \ell(\mathbf{z}_k - \hat{\mathbf{z}}_k)$. For all $\epsilon \geq 0$, $2ab = 2(\epsilon a)(\frac{1}{\epsilon}b) \leq (\epsilon)^2 + (\frac{1}{\epsilon}b)^2$.

Then, we have

$$\begin{aligned}
&(1-\mu\eta)\|\mathbf{x}_k\|^2 + \eta\epsilon^2\|\mathbf{x}_k\|^2 \leq (1 - \frac{\mu\eta}{2})\|\mathbf{x}_k\|^2 \\
\iff &1 - \mu\eta + \eta\epsilon^2 \leq 1 - \frac{\mu\eta}{2} \\
\iff &\epsilon \leq \sqrt{\frac{\mu}{2}}.
\end{aligned}$$

So let $\epsilon = \sqrt{\frac{\mu}{2}}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla f(\mathbf{x}_k, \mathbf{z}_k)\|^{2q}\right] &\leq \mathbb{E}\Bigg[\Bigg((1-\frac{\mu\eta}{2})\|\mathbf{x}_k\|^2 + 2\eta(\ell+\mu)R^2 \\
&\quad + \frac{2\eta}{\mu}(\|\nabla\bar{f}(0)\| + \ell(\mathbf{z}_k - \hat{\mathbf{z}}_k))^2 + \eta^2 3\left(\ell^2\left(\mathbb{E}_{\hat{\mathbf{z}}}[\|\mathbf{z}_k - \hat{\mathbf{z}}_k\|]\right)^2\|\nabla\bar{f}(0)\|^2\right)\Bigg)^q\Bigg] \\
&\leq \mathbb{E}\Bigg[\Bigg((1-\frac{\mu\eta}{2})\|\mathbf{x}_k\|^2 + 2\eta(\ell+\mu)R^2 \\
&\quad + \frac{2\eta}{\mu}(\|\nabla\bar{f}(0)\| + \ell(\mathbf{z}_k - \hat{\mathbf{z}}_k))^2 + \eta^2 3\left(\ell^2\left(\|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\right)^2\|\nabla\bar{f}(0)\|^2\right)\Bigg)^q\Bigg] \\
&= (1-\frac{\mu\eta}{2})^q\mathbb{E}[\|\mathbf{x}_k\|^{2q}] + \eta\mathbb{E}[p(\|\mathbf{x}_k\|^2, \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|^2)] \\
&\leq (1-\frac{\mu\eta}{2})\mathbb{E}[\|\mathbf{x}_k\|^{2q}] + \eta\mathbb{E}[p(\|\mathbf{x}_k\|^2, \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|^2)] \\
&\leq (1-\frac{\mu\eta}{4})\mathbb{E}[\|\mathbf{x}_k\|^{2q}] + \eta\mathbb{E}[-\frac{\mu}{4}\|\mathbf{x}\|^{2q} + p(\|\mathbf{x}_k\|^2, \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|^2)] \\
&\leq (1-\frac{\mu\eta}{4})\mathbb{E}[\|\mathbf{x}_k\|^{2q}] + \eta c_{16}.
\end{aligned}
$$

The second inequality uses Jensen's inequality twice. The polynomial $p(\|\mathbf{x}_k\|^2, \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|^2)$ is with order strictly lower than $2q$ in $\|\mathbf{x}_k\|$ and with highest order of $2q$ in $\|\mathbf{z} - \hat{\mathbf{z}}_k\|$. The last inequality is based on the claim that $\mathbb{E}[-\frac{\mu}{4}\|\mathbf{x}\|^{2q} + p(\|\mathbf{x}_k\|^2, \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|^2)]$ is bounded by a constant c_{16}, which depends on the value of q, the statistics of external random variables $\mathbf{z}$.

Therefore, we have

$$
\mathbb{E}[\|\mathbf{x}_{k+1}\|^{2q}] \leq (1-\frac{\eta\mu}{4})\mathbb{E}[\|\mathbf{x}_k\|^{2q}] + \eta c_{17}.
$$

Intuitively, the constant c_{17} represents the influence of the lower order remainders in the processes of using binomial theorem.

Now as long as $\mathbb{E}[\|\mathbf{x}_0\|^{2q}] < \infty$, we have

$$
\mathbb{E}[\|\mathbf{x}_k\|^{2q}] \leq \mathbb{E}[\|\mathbf{x}_0\|^{2q}] + c_{17}.
$$

Plugging the above result into (78) gives

$$
\begin{aligned}
&\mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(\mathbf{x})] \\
&\quad \leq \left(\|\nabla\bar{f}(0)\| + \ell\frac{2\left(\|\nabla\bar{f}(0)\| + \sqrt{\|\nabla\bar{f}(0)\| + \mu(m - \bar{f}(0))}\right)}{\mu}\right) W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \\
&\qquad + \left(\mathbb{E}[\|\mathbf{x}_0\|^{2q}] + c_{17}\right)\frac{\ell^q 2^{q-1}}{(q-1)m^{q-1}}.
\end{aligned}
$$

The remaining work is to optimize the left hand side with respect to m so that we can make a choice of the value of m which is mentioned earlier in the proof.

Let

$$
\begin{aligned}
g(m) = &\left(\|\nabla\bar{f}(0)\| + \ell\frac{2\left(\|\nabla\bar{f}(0)\| + \sqrt{\|\nabla\bar{f}(0)\|} + \sqrt{\mu m} + \sqrt{\mu\bar{f}(0)}\right)}{\mu}\right) W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \\
&+ \left(\mathbb{E}[\|\mathbf{x}_0\|^{2q}] + c_{17}\right)\frac{\ell^q 2^{q-1}}{(q-1)m^{q-1}}.
\end{aligned}
$$

We can see that $g(m)$ is an upper bound of the right hand side of (78).

Setting $g'(m) = 0$ leads to $m^* = \left(\frac{\ell W_1}{\sqrt{\mu}(q-1)C}\right)^{\frac{2}{-2q+1}}$, where $C = \left(\mathbb{E}[\|\mathbf{x}_0\|^{2q}] + c_{17}\right)\frac{\ell^q 2^{q-1}}{(q-1)}$ for notation simplicity. So

$$
\begin{aligned}
\max_{m\geq 0} g(m) &= g(m^*) \\
&\leq \left(\|\nabla \bar{f}(0)\| + \ell \frac{2\left(\|\nabla \bar{f}(0)\| + \sqrt{\|\nabla \bar{f}(0)\|} + \sqrt{\mu \bar{f}(0)}\right)}{\mu}\right) W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}}) \\
&+ \frac{2\ell}{\sqrt{\mu}}\left(\frac{\ell W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}})}{\sqrt{\mu}(q-1)C}\right)^{\frac{2-2q}{-2q+1}} + C\left(\frac{\ell W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}})}{\sqrt{\mu}(q-1)C}\right)^{\frac{2-2q}{-2q+1}} \\
&= \left(\|\nabla \bar{f}(0)\| + \ell \frac{2\left(\|\nabla \bar{f}(0)\| + \sqrt{\|\nabla \bar{f}(0)\|} + \sqrt{\mu \bar{f}(0)}\right)}{\mu}\right) W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}}) \\
&+ \left(\frac{2\ell}{\sqrt{\mu}} + C\right)\left(\frac{\ell}{\sqrt{\mu}(q-1)C}\right)^{\frac{2-2q}{-2q+1}} W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}})^{\frac{2-2q}{-2q+1}}
\end{aligned}
$$

Setting

$$
\begin{aligned}
c_{14} &= \|\nabla \bar{f}(0)\| + \ell \frac{2\left(\|\nabla \bar{f}(0)\| + \sqrt{\|\nabla \bar{f}(0)\|} + \sqrt{\mu \bar{f}(0)}\right)}{\mu} \\
c_{15} &= \left(\frac{2\ell}{\sqrt{\mu}} + \left(\mathbb{E}[\|\mathbf{x}_0\|^{2q}] + c_{17}\right)\frac{\ell^q 2^{q-1}}{(q-1)}\right)\left(\frac{\ell}{\sqrt{\mu}(q-1)\left(\mathbb{E}[\|\mathbf{x}_0\|^{2q}] + c_{17}\right)\frac{\ell^q 2^{q-1}}{(q-1)}}\right)^{\frac{2-2q}{-2q+1}}.
\end{aligned}
$$

and plugging this bound into (76) give the desired result. ■