Projected Stochastic Gradient Langevin Algorithms for Constrained Sampling and Non-Convex Learning

Andrew Lamperski ALAMPERS@UMN.EDU
200 Union St. Se, Keller Hall 4-174, Minneapolis, MN 55455, USA

Abstract

Langevin algorithms are gradient descent methods with additive noise. They have been used for decades in Markov Chain Monte Carlo (MCMC) sampling, optimization, and learning. Their convergence properties for unconstrained non-convex optimization and learning problems have been studied widely in the last few years. Other work has examined projected Langevin algorithms for sampling from log-concave distributions restricted to convex compact sets. For learning and optimization, log-concave distributions correspond to convex losses. In this paper, we analyze the case of non-convex losses with compact convex constraint sets and IID external data variables. We term the resulting method the projected stochastic gradient Langevin algorithm (PSGLA). We show the algorithm achieves a deviation of $O(T^{-1/4}(\log T)^{1/2})$ from its target distribution in 1-Wasserstein distance. For optimization and learning, we show that the algorithm achieves ϵ-suboptimal solutions, on average, provided that it is run for a time that is polynomial in ϵ^{-1} and slightly super-exponential in the problem dimension.

Keywords: Langevin Methods, Stochastic Gradient Algorithms, Non-Convex Learning, Non-Asymptotic Analysis, Markov Chain Monte Carlo Sampling

1. Introduction

Langevin dynamics originate in the study of statistical physics Coffey and Kalmykov (2012), and have a long history of applications to Markov Chain Monte Carlo (MCMC) sampling Roberts et al. (1996), non-convex optimization Gelfand and Mitter (1991); Borkar and Mitter (1999), and machine learning Welling and Teh (2011). Langevin algorithms amount to gradient descent augmented with additive Gaussian noise. This additive noise enables the algorithms to escape saddles and local minima. For optimization and learning, this enables the algorithms to find near optimal solutions even when the losses are non-convex. For sampling, Langevin algorithms give a simple approach to produce samples that converge to target distributions which are not log-concave.

Related Work. A large amount of progress on the non-asymptotic analysis of Langevin algorithms has been reported in recent years. This work has two main streams: 1) unconstrained non-convex problems and 2) constrained convex problems. These works will be reviewed below.

The bulk of the recent work on non-asymptotic analysis of Langevin algorithms has examined unconstrained problems Raginsky et al. (2017); Majka et al. (2020); Fehrman et al. (2020); Chen et al. (2020); Erdogdu et al. (2018); Durmus et al. (2017); Chau et al. (2019); Xu et al. (2018); Cheng et al. (2018); Ma et al. (2019). The basic algorithm in the unconstrained case has the form:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_x f(\mathbf{x}_k, \mathbf{z}_k) + \sqrt{\frac{2\eta}{\beta}} \mathbf{w}_k,$$

where $\mathbf{x}_k$ is the decision variable, $\mathbf{z}_k$ are external random variables, $\mathbf{w}_k$ is Gaussian noise, and η and β are parameters. In a learning context, $\mathbf{z}_k$ correspond to data, $\mathbf{x}_k$ are parameters of a model, and $f(x, z)$ is a loss function that describes how well the model parameters fit the data. With no Gaussian noise, $\mathbf{w}_k$, this algorithm reduces to stochastic gradient descent.

A breakthrough was achieved in Raginsky et al. (2017), which gave non-asymptotic bounds in the case that $f(x, z)$ is non-convex in x and $\mathbf{z}_k$ are independent identically distributed (IID). A wide number of improvements and variations on these results have since been obtained in works such as Majka et al. (2020); Fehrman et al. (2020); Chen et al. (2020); Erdogdu et al. (2018); Durmus et al. (2017); Chau et al. (2019); Xu et al. (2018); Cheng et al. (2018); Ma et al. (2019). In particular, Chau et al. (2019) achieves tighter performance guarantees and extends to the case that $\mathbf{z}_k$ is a mixing process.

For problems with constraints, most existing work focuses convex losses over compact convex constraint sets with no external variables $\mathbf{z}_k$. Most closely related to our work is that of Bubeck et al. (2015, 2018) which augments the Langevin algorithm with a projection onto the constraint set. Proximal-type algorithms were examined Brosse et al. (2017). Variations on mirror descent were examined in Ahn and Chewi (2020); Hsieh et al. (2018); Zhang et al. (2020); Krichene and Bartlett (2017).

Recent work of Wang et al. (2020) examines Langevin dynamics on Riemannian manifolds. In this case, the losses may be non-convex, but still there are no external variables, $\mathbf{z}_k$. It utilizes results from diffusion theory to give convergence with respect to Kullback-Liebler (KL) divergence. Many of the ideas in that paper could likely be translated to the current setting. However, such KL divergence bounds become degenerate if the algorithm is initialized as a constant value, e.g. $\mathbf{x}_0 = 0$. In contrast, our work focuses on bounds in the 1-Wasserstien distance, which gives well-defined bounds as long as the initialization is feasible for the constraints.

Contributions. This paper gives non-asymptotic convergence bounds for Langevin algorithms for problems that are constrained to a compact convex set. In particular, we examine a generalized version of the algorithm examined in Bubeck et al. (2018, 2015). As discussed above, the existing works on constrained Langevin methods (aside from the Riemannian manifold results of Wang et al. (2020)) focus on convex loss functions, and none consider external random variables. This paper examines the case of non-convex losses with IID external randomness. For the purpose of sampling, it is shown that after T steps, the error from the target in the 1-Wasserstein is of $O(T^{-1/4}(\log T)^{1/2})$. For optimization and learning, this bound is used to show that the algorithm can achieve a suboptimality of ϵ in a number of steps that is polynomial in ϵ and slightly superexponential in the dimension of $\mathbf{x}_k$. To derive the bounds, a novel result on contractions for reflected stochastic differential equations is derived.

2. 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 $W_1(P, Q)$.

Throughout the paper, $\mathcal{K}$ will denote a compact convex subset of $\mathbb{R}^n$ of diameter D such that a ball of radius $r > 0$ around the origin is contained in $\mathcal{K}$. 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 $\Pi_{\mathcal{K}}$.

2.2. The Project Stochastic Gradient 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 IID random variables whose properties will be described later. Assume that $\mathbf{z}_i$ and $\hat{\mathbf{w}}_j$ are independent 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 projected stochastic gradient 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 a noise parameter.

Let $\bar{f}(x) = \mathbb{E}[f(x, \mathbf{z})]$, where the expectation is over $\mathbf{z}$, which has the same distribution as $\mathbf{z}_k$. We will assume that $\nabla_x f(x, \mathbf{z}) - \nabla_x \bar{f}(x)$ are uniformly sub-Gaussian for each $x \in \mathbb{R}^n$. That is, there is a number $\sigma > 0$ such that for all $\alpha \in \mathbb{R}^n$, the following bound holds:

$$\mathbb{E}\left[\exp\left(\alpha^\top \left(\nabla_x f(x, \mathbf{z}) - \nabla_x \bar{f}(x) \right) \right) \right] \leq e^{\sigma^2 \|\alpha\|^2 / 2}. \tag{2}$$

The uniform sub-Gaussian property holds under the following conditions:

- **Gradient Noise:** $\nabla_x f(x, \mathbf{z}) = \nabla_x \bar{f}(x) + \mathbf{z}$ with $\mathbf{z}$ sub-Gaussian.
- **Lipschitz Gradients and Strongly Log-Concave z:** $\nabla_x f(x, z)$ is Lipschitz in z and $\mathbf{z}$ has a density of the form $e^{-U(z)}$ with $\nabla^2 U(z) \succeq \kappa I$ for all z. Here $\kappa > 0$ and the inequality is with respect to the positive semidefinite partial order. (See Theorem 5.2.15 of Vershynin (2018).)
- **Convex Gradients and Bounded z:** Each component $\frac{\partial f(x,z)}{\partial x_i}$ is convex in z and $\mathbf{z}$ is bounded with independent components. (See Theorem 3.24 of Wainwright (2019).)

For learning, the last two conditions are the most useful, since they give general classes of losses and variables for which the method can be applied. In particular, the second case applies to many common scenarios. It includes Gaussian $\mathbf{z}$ as a special case, and it can be applied to neural networks with smooth activation functions. A variety of more specialized cases in which the sub-Gaussian condition holds are presented in Chapter 5 of Vershynin (2018). Future work will relax the uniform sub-Gaussian assumption and the requirement of IID $\mathbf{z}_k$.

We assume that for each z, $\nabla_x f(x, z)$ is ℓ-Lipschitz in x, i.e. $\|\nabla_x f(x_1, z) - \nabla_x f(x_2, z)\| \leq \ell \|x_1 - x_2\|$. The mean function, $\bar{f}$, is assumed to be u-smooth, so that $\|\nabla_x \bar{f}(x)\| \leq u$ for all $x \in \mathcal{K}$. The assumptions on $\bar{f}$ imply that we can have $u \leq \|\nabla_x \bar{f}(0)\| + \ell D$ and that $\bar{f}$ is u-Lipschitz.

In Bubeck et al. (2018), the case with $\bar{f}$ is convex and no $\mathbf{z}_k$ variables is studied. It is shown that by choosing the step size, η, appropriately, the law of $\mathbf{x}_k$ is given approximately given by $\pi_{\beta \bar{f}}$,

which is defined by

$$\pi_{\beta\bar{f}}(A) = \frac{\int_A e^{-\beta\bar{f}(x)}dx}{\int_K e^{-\beta\bar{f}(x)}dx}. \tag{3}$$

In this paper, we will bound the convergence of (1) to (3) in the case of non-convex f with external random variables $\mathbf{z}_k$.

3. Main Results

3.1. Convergence of the Law of the Iterates

The following is the main result of the paper. It is proved in Subsection 3.4.

Theorem 1 *Assume that $\eta \leq 1/2$. There are positive constants a, c_1 and c_2 such that for all integers $k \geq 4$, the following bound holds:*

$$W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \leq c_1 e^{-\eta a k} + c_2(\eta \log k)^{1/4}$$

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

$$W_1(\mathcal{L}(\mathbf{x}_T), \pi_{\beta\bar{f}}) \leq \left(c_1 + \frac{c_2}{(4a)^{1/4}}\right) T^{-1/4}(\log T)^{1/2}.$$

The constants depend on the dimension of $\mathbf{x}_k$, n, the noise parameter, β, the Lipschitz constant, ℓ, the diameter, D, the size of the inscribed ball r, and the smoothness constant, u. The specific form of the constants will be derived in the proof. For applications, it is useful to know how the constants depend on the dimension, n, and the noise parameter, β. The result below indicates that the algorithm exhibits two distinct regimes in which convergence is fast and slow, respectively. It is proved in Appendix C.

Proposition 2 *The constants c_1 and c_2 grow linearly with n. If $D^2\ell\beta < 8$, then we can set $a = \frac{4}{D^2\beta} \geq \frac{\ell}{2}$, while c_1 and c_2 grow polynomially with respect to $\left(1 - \frac{D^2\ell\beta}{8}\right)^{-2}$ and $\beta^{-1/4}$. In general, we have a positive constant c_3 and a monotonically increasing polynomial p (independent of η and β) such that for all $\beta > 0$, the following bounds hold:*

$$a \geq c_3\beta \exp\left(-\frac{D^2\ell\beta}{4}\right)$$
$$\max\{c_1, c_2\} \leq p(\beta^{-1/4}) \exp\left(\frac{3D^2\ell\beta}{4}\right).$$

3.2. Application to Optimization and Learning

The following result shows that the $\mathbf{x}_k$ can be made arbitrarily near optimal, but the required time may be slightly super-exponential with respect to problem dimension, n. It is proved in Appendix D.

Proposition 3 *Assume that $\eta \leq 1/2$. There is a positive constant, c_5 such that for all $k \geq 4$, the following bound holds:*

$$\mathbb{E}[\bar{f}(\mathbf{x}_k)] \leq \min_{x \in \mathcal{K}} \bar{f}(x) + uW_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta \bar{f}}) + \frac{n \log(c_5 \max\{1, \beta\})}{\beta}.$$

In particular, given any $\rho > 4$ and any $\zeta > 1$, there are choices of η, β, and T, along with positive numbers $m(\rho, \zeta)$ and $\alpha(\rho, \zeta)$, such that for any suboptimality level, $\epsilon > 0$, the following implication holds:

$$T \geq \frac{m(\rho, \zeta)}{\epsilon^\rho} \exp\left(\alpha(\rho, \zeta) n^\zeta\right) \implies \mathbb{E}[\bar{f}(\mathbf{x}_T)] \leq \min_{x \in \mathcal{K}} \bar{f}(x) + \epsilon.$$

3.3. The Auxiliary Processes Used for the Main Bound

Similar to other analyses of Langevin methods, e.g. Raginsky et al. (2017); Bubeck et al. (2018); Chau et al. (2019), the proof of Theorem 1 utilizes a collection of auxiliary stochastic processes that fit between the algorithm iterates from (1) and a stationary Markov process with state distribution given by (3).

We will embed the iterates of the algorithm into continuous time by setting $\mathbf{x}_t^A = \mathbf{x}_{\lfloor t \rfloor}$. The A superscript is used to highlight the connection between this process and the algorithm. The Gaussian variables $\hat{\mathbf{w}}_k$ can be realized as $\hat{\mathbf{w}}_k = \mathbf{w}_{k+1} - \mathbf{w}_k$ where $\mathbf{w}_t$ is a Brownian motion.

We will let $\mathbf{x}_t^C$ be a continuous approximation of $\mathbf{x}_t^A$ and we will let $\mathbf{x}_t^M$ be a variation on the process $\mathbf{x}_t^C$ in which averages out the effect of the $\mathbf{z}_k$ variables. The proof will proceed by showing that the law of $\mathbf{x}_t^M$ converges exponentially to (3), that $\mathbf{x}_t^C$ has a similar law to $\mathbf{x}_t^M$, and that $\mathbf{x}_t^A$ has a similar law to $\mathbf{x}_t^C$. Below we make these statements more precise.

The continuous approximation of the algorithm is defined by the following reflected stochastic differential equation (RSDE):

$$d\mathbf{x}_t^C = -\eta \nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor}) dt + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t - \mathbf{v}_t^C d\boldsymbol{\mu}^C(t). \tag{4}$$

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. Under these conditions, the reflection process is uniquely defined and $\mathbf{x}^C$ is the unique solution to the Skorokhod problem for the process defined by:

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

See Appendix E for more details on the Skorokhod problem.

For compact notation, we denote the Skorkohod solution for 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 averaged version of $\mathbf{x}_t^C$, denoted by $\mathbf{x}_t^M$, where the M corresponds to "mean", is defined by:

$$d\mathbf{x}_t^M = -\eta \nabla_x \bar{f}(\mathbf{x}_t^M) dt + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t - \mathbf{v}_t^M d\boldsymbol{\mu}^M(t). \tag{5}$$

Again $-\int_0^t \mathbf{v}_s^M d\boldsymbol{\mu}^M(s)$ is the unique reflection process that ensures that $\mathbf{x}_t^M \in \mathcal{K}$ for all t whenever $\mathbf{x}_0^M \in \mathcal{K}$. By construction, $\mathbf{x}_t^M$ satisfies the Skorokhod problem for the continuous process defined by

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

See Appendix E for more details on the Skorokhod problem.

The following lemmas describe the relationships between the laws all of these processes. They are proved in Sections 4, 6, 7 respectively.

Lemma 4 *There are positive constants c_1 and a such that for all $t \geq 0$*

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

Lemma 5 *Assume that $\mathbf{x}_0^A = \mathbf{x}_0^C \in \mathcal{K}$ and $\eta \leq 1/2$. There is a positive constant, c_6, such that for all $t \geq 4$,*

$$W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) \leq c_6 \left(\eta \log t\right)^{1/4}.$$

Lemma 6 *Assume that $\mathbf{x}_0^M = \mathbf{x}_0^C \in \mathcal{K}$ and $\eta \leq 1/2$. There is a positive constant, c_7 such that for all $t \geq 0$,*

$$W_1(\mathcal{L}(\mathbf{x}_t^M), \mathcal{L}(\mathbf{x}_t^C)) \leq c_7 \eta^{1/4}.$$

Most of the rest of the paper focuses on proving these lemmas. Assuming that these lemmas hold, the main result now has a short proof, which we describe next.

3.4. Proof of Theorem 1

Recall that $\mathbf{x}_k^A = \mathbf{x}_k$ for all integers $k \in \mathbb{N}$. Assume that $\mathbf{x}_0 = \mathbf{x}_0^A = \mathbf{x}_0^C = \mathbf{x}_0^M$. The triangle inequality followed by Lemmas 4, 5, and 6 shows that

$$\begin{aligned} W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta f}) &\leq W_1(\mathcal{L}(\mathbf{x}_k^A), \mathcal{L}(\mathbf{x}_k^C)) + W_1(\mathcal{L}(\mathbf{x}_k^C), \mathcal{L}(\mathbf{x}_k^M)) + W_1(\mathcal{L}(\mathbf{x}_k^M), \pi_{\beta f}) \\ &\leq c_1 e^{-\eta a k} + c_6 \left(\eta \log k\right)^{1/4} + c_7 \eta^{1/4}. \end{aligned}$$

The result now follows by noting that $\log k \geq 1$ for $k \geq 4$ setting $c_2 = c_6 + c_7$. The specific bound when $\eta = \frac{\log T}{4aT}$ arises from direct computation. ■

4. Contractions for the Reflected SDEs

In this section, we will show how the laws of the processes $\mathbf{x}_t^C$ and $\mathbf{x}_t^M$ are contractive with respect to a specially constructed Wasserstein distance. By relating this specially constructed distance with W_1 we will prove Lemma 4 which states that $\mathcal{L}(\mathbf{x}_t^M)$ converges to $\pi_{\beta f}$ exponentially with respect to W_1. The contraction will be derived by an extension of the reflection coupling argument of Eberle (2016) to the case of reflected SDEs with external randomness (from $\mathbf{z}_k$). This result may be of independent interest.

Proposition 7 *There are positive constants a and c_8 such that for any two solutions, $\mathbf{x}_t^{C,1}$ and $\mathbf{x}_t^{C,2}$ to the continuous-time RSDE, (4), their laws converge according to*

$$W_1(\mathcal{L}(\mathbf{x}_t^{C,1}), \mathcal{L}(\mathbf{x}_t^{C,2})) \leq c_8 e^{-\eta a t} W_1(\mathcal{L}(\mathbf{x}_0^{C,1}), \mathcal{L}(\mathbf{x}_0^{C,2})) \tag{6}$$

To define the constants, let the natural frequency and damping ratio be given by

$$\omega_N = \frac{\sqrt{a\beta}}{2} \quad \textit{and} \quad \xi = \frac{D\ell}{4}\sqrt{\frac{\beta}{a}}. \tag{7}$$

The constants can always be set to

$$a = \frac{D^2\ell^2\beta}{16}\left(1 - \tanh^2\left(\frac{D^2\ell\beta}{8}\right)\right)$$
$$c_8 = \frac{e^{D\omega_N\xi}}{\cosh(D\omega_N\sqrt{\xi^2-1}) - \frac{\xi}{\sqrt{\xi^2-1}}\sinh(D\omega_N\sqrt{1-\xi^2})}$$

When $D^2\ell\beta < 8$, a larger decay constant, a, can be defined by setting

$$a = \frac{4}{D^2\beta}$$
$$c_8 = \frac{e^{D\omega_N\xi}}{\cos(D\omega_N\sqrt{1-\xi^2}) - \frac{\xi}{\sqrt{1-\xi^2}}\sin(D\omega_N\sqrt{1-\xi^2})}.$$

Proof We will follow the main idea behind Eberle (2016). We will correlate the solutions using reflection coupling, and then construct a distance function, h, from the coupling. Then h will be used to construct a Wasserstein distance for which the laws $\mathcal{L}(\mathbf{x}_t^{C,1})$ and $\mathcal{L}(\mathbf{x}_t^{C,2})$ converge exponentially. The desired bound is found by comparing this auxiliary distance to the classical W_1 distance.

Let $\boldsymbol{\rho}_t = \mathbf{x}_t^{C,1} - \mathbf{x}_t^{C,2}$, $\mathbf{u}_t = \boldsymbol{\rho}_t/\|\boldsymbol{\rho}_t\|$ and $\boldsymbol{\tau} = \inf\{t | \mathbf{x}_t^{C,1} = \mathbf{x}_t^{C,2}\}$. Note that $\boldsymbol{\tau}$. The *reflection coupling* between $\mathbf{x}_t^{C,1}$ and $\mathbf{x}_t^{C,2}$ is defined by:

$$d\mathbf{x}_t^{C,1} = -\eta\nabla_x f(\mathbf{x}_t^{C,1}, \mathbf{z}_{\lfloor t \rfloor}) + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t - \mathbf{v}_t^{C,1} d\boldsymbol{\mu}^{C,1}(t) \tag{8a}$$

$$d\mathbf{x}_t^{C,2} = -\eta\nabla_x f(\mathbf{x}_t^{C,2}, \mathbf{z}_{\lfloor t \rfloor}) + \sqrt{\frac{2\eta}{\beta}}(I - 2\mathbf{u}_t\mathbf{u}_t^\top \mathbb{1}(t < \boldsymbol{\tau}))d\mathbf{w}_t - \mathbf{v}_t^{C,2} d\boldsymbol{\mu}^{C,2}(t). \tag{8b}$$

Here I is the $n \times n$ identity matrix. Also, $\boldsymbol{\varphi}_t^1 = -\int_0^t \mathbf{v}_s^{C,1} d\boldsymbol{\mu}^{C,1}(s)$ and $\boldsymbol{\varphi}_t^2 = -\int_0^t \mathbf{v}_s^{C,2} d\boldsymbol{\mu}^{C,2}(s)$ are the unique projection processes that ensure that respective Skorkhod problem solutions, $\mathbf{x}_t^{C,1}$ and $\mathbf{x}_t^{C,2}$, remain in $\mathcal{K}$.

The processes from (8) define a valid coupling because $\int_0^T (I - 2\mathbf{u}_s\mathbf{u}_s^\top \mathbb{1}(s < \boldsymbol{\tau}))d\mathbf{w}_s$ is a Brownian motion. Furthermore, for $t \geq \boldsymbol{\tau}$, we have that $\mathbf{x}_t^{C,1} = \mathbf{x}_t^{C,2}$.

Analogous to Eberle (2016), we aim to construct a function $h : [0, D] \to \mathbb{R}$ such that $h(0) = 0$, $h'(0) = 1$, $h'(x) > 0$, and $h''(x) < 0$ and a constant $a > 0$ such that $e^{\eta a t} h(\|\mathbf{z}_t\|)$ is a supermartingale. One simplifying assumption for the construction is that we only need to define h over the compact set $[0, D]$, while Eberle (2016) requires h to be defined over $[0, \infty)$. This is due to the fact

that our solutions will be contained in $\mathcal{K}$ which has diameter D, while in Eberle (2016) the solutions are unconstrained.

Now we will describe why the construction of such an h proves the lemma. The supermartingale property will ensure that

$$\mathbb{E}[h(\|\boldsymbol{\rho}_t\|)] \leq e^{-\eta a t}\mathbb{E}[h(\|\boldsymbol{\rho}_0\|)] \tag{9}$$

Let W_h denote that 1-Wasserstein distance corresponding to the function $d(x,y) = h(\|x-y\|)$ for $x, y \in \mathcal{K}$. In other words, if P and Q are probability distributions on $\mathcal{K}$ and $C(P,Q)$ is the set of couplings between P and Q, then

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

By the hypotheses on h, $d(x,y) = d(y,x) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$. Thus, W_h is a valid Wasserstein distance.

Assume that Γ_0 is an optimal coupling of the initial laws $C(\mathcal{L}(\mathbf{x}_0^{C,1}), \mathcal{L}(\mathbf{x}_0^{C,2}))$ so that

$$W_h(\mathcal{L}(\mathbf{x}_0^{C,1}), \mathcal{L}(\mathbf{x}_0^{C,2})) = \int_{\mathcal{K}\times\mathcal{K}} h(\|x-y\|) d\Gamma_0(x,y).$$

Such a coupling exists by Theorem 4.1 of Villani (2008). Then using this initial coupling on the right of (9) and minimizing over all couplings of the dynamics on the left shows that

$$W_h(\mathcal{L}(\mathbf{x}_t^{C,1}), \mathcal{L}(\mathbf{x}_t^{C,2})) \leq \mathbb{E}[h(\|\boldsymbol{\rho}_t\|)] \leq e^{-\eta a t} W_h(\mathcal{L}(\mathbf{x}_0^{C,1}), \mathcal{L}(\mathbf{x}_0^{C,1})). \tag{10}$$

In other words, the law of the continuous-time RSDE is contractive with respect to W_h.

By the assumptions that $h(0) = 0$, $h'(0) = 1$, $h'(x) > 0$, and $h''(x) < 0$, we have that for all $x \in [0, D]$:

$$h'(D)x \leq h(x) \leq x.$$

It then follows from the definition of W_h and W_1 that for all probability measures P and Q over $\mathcal{K}$ that

$$h'(D) W_1(P,Q) \leq W_h(P,Q) \leq W_1(P,Q).$$

Combining these inequalities with (10) gives (6) with $c_8 = h'(D)^{-1}$.

Now we will construct h. The restriction of h to the domain of $[0, D]$, along with the Lipschitz bound on $\nabla_x f$ will enable an explicit construction of h as the solution to a simple harmonic oscillator problem. This is in contrast to the more abstract construction in terms of integrals from Eberle (2016).

To ensure that $e^{\eta a t} h(\|\boldsymbol{\rho}_t\|)$ is a supermartingale, we must ensure that this process is non-increasing on average. Recall that $\boldsymbol{\tau}$ is the coupling time so that $e^{\eta a t} h(\|\boldsymbol{\rho}_t\|) = 0$ for $t \geq \boldsymbol{\tau}$. So, it suffices to bound the behavior of the process for all $t < \boldsymbol{\tau}$. In this case, we require that non-martingale terms of $d\left(e^{\eta a t} h(\|\boldsymbol{\rho}_t\|)\right)$ are non-positive. By Itô's formula we have that

$$d(e^{\eta a t} h(\|\boldsymbol{\rho}_t\|)) = e^{\eta a t} \eta a h(\|\boldsymbol{\rho}_t\|) dt + e^{\eta a t} h'(\|\boldsymbol{\rho}_t\|) d\|\boldsymbol{\rho}_t\| + \frac{1}{2} e^{\eta a t} h''(\|\boldsymbol{\rho}_t\|)(d\|\boldsymbol{\rho}_t\|)^2. \tag{11}$$

Thus, the desired differential is computed from $d\|\boldsymbol{\rho}_t\|$ and $(d\|\boldsymbol{\rho}_t\|)^2$. So, our next goal is to derive these terms.

Let $\mathbf{b}_t$ be the one-dimensional Brownian motion defined by $d\mathbf{b}_t = \mathbf{u}_t^\top d\mathbf{w}_t$. Then for $t < \boldsymbol{\tau}$, $d\boldsymbol{\rho}_t$ can be expressed as

$$d\boldsymbol{\rho}_t = \eta(\nabla_x f(\mathbf{x}_t^{C,2}, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x f(\mathbf{x}_t^{C,1}, \mathbf{z}_{\lfloor t \rfloor}))dt + \sqrt{\frac{8\eta}{\beta}}\mathbf{u}_t d\mathbf{b}_t + \mathbf{v}_t^{C,2} d\boldsymbol{\mu}^{C,2}(t) - \mathbf{v}_t^{C,1} d\boldsymbol{\mu}^{C,1}(t). \quad (12)$$

Since $\boldsymbol{\varphi}_t^1$ and $\boldsymbol{\varphi}_t^2$ are bounded variation processes, the quadratic terms are given by

$$(d\boldsymbol{\rho}_t)(d\boldsymbol{\rho}_t)^\top = \frac{8\eta}{\beta}\mathbf{u}_t\mathbf{u}_t^\top dt. \quad (13)$$

If $u = \rho/\|\rho\|$ and $\rho \neq 0$, then the gradient and Hessian of $\|\rho\|$ are given by

$$\nabla\|\rho\| = \|\rho\|^{-1}\rho = u \quad \text{and} \quad \nabla^2\|\rho\| = \|\rho\|^{-1}I - \|\rho\|^{-1}uu^\top. \quad (14)$$

Plugging (12), (13), and (14) into Itô's formula and simplifying gives

$$\begin{aligned} d\|\boldsymbol{\rho}_t\| &= \eta\mathbf{u}_t^\top(\nabla_x f(\mathbf{x}_t^{C,2}, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x f(\mathbf{x}_t^{C,1}, \mathbf{z}_{\lfloor t \rfloor}))dt + \sqrt{\frac{8\eta}{\beta}}d\mathbf{b}_t \\ &\quad + \mathbf{u}_t^\top \mathbf{v}_t^{C,2} d\boldsymbol{\mu}^{C,2}(t) - \mathbf{u}_t^\top \mathbf{v}_t^{C,1} d\boldsymbol{\mu}^{C,1}(t) \\ &\leq \eta\ell D dt + \sqrt{\frac{8\eta}{\beta}}d\mathbf{b}_t. \end{aligned} \quad (15)$$

The simplification in the equality arises because $(d\boldsymbol{\rho}_t)^\top(\nabla^2\|\boldsymbol{\rho}_t\|)(d\boldsymbol{\rho}_t) = 0$. The inequality uses two simplifications. The first term on the right arises due to the Lipschitz bound on $\nabla_x f$ and the diameter bound on $\mathcal{K}$. The other terms can be removed since $\mathbf{x}_t^{C,1}$ and $\mathbf{x}_t^{C,2}$ are both in $\mathcal{K}$, so that $\mathbf{v}_t^2 \in N_\mathcal{K}(\mathbf{x}_t^{C,2})$ implies that $(\mathbf{x}_t^{C,1} - \mathbf{x}_t^{C,2})^\top \mathbf{v}_t^2 \leq 0$. Likewise, $\mathbf{v}_t^1 \in N_\mathcal{K}(\mathbf{x}_t^{C,1})$ implies that $-(\mathbf{x}_t^{C,1} - \mathbf{x}_t^{C,2})^\top \mathbf{v}_t^1 \leq 0$. Then since $\boldsymbol{\mu}^1$ and $\boldsymbol{\mu}^2$ are non-negative measures, the corresponding terms are non-positive.

Note that we also have that $(d\|\boldsymbol{\rho}_t\|)^2 = \frac{8\eta}{\beta}dt$. Plugging the bounds for $d\|\boldsymbol{\rho}_t\|$ and $(d\|\boldsymbol{\rho}_t\|)^2$ into (11) gives

$$d(e^{\eta a t}h(\|\boldsymbol{\rho}_t\|)) \leq \frac{4\eta}{\beta}e^{\eta a t}\left(\frac{a\beta}{4}h(\|\boldsymbol{\rho}_t\|) + \frac{D\ell\beta}{4}h'(\|\boldsymbol{\rho}_t\|) + h''(\|\boldsymbol{\rho}_t\|)\right)dt + \sqrt{\frac{8\eta}{\beta}}e^{\eta a t}h'(\|\boldsymbol{\rho}_t\|)d\mathbf{b}_t. \quad (16)$$

Thus, we see that a sufficient condition for $e^{\eta a t}h(\|\boldsymbol{\rho}_t\|)$ to be a supermartingale is that

$$\frac{a\beta}{4}h(x) + \frac{D\ell\beta}{4}h'(x) + h''(x) = 0 \quad (17)$$

for all $x \in [0, D]$. This is precisely the simple harmonic oscillator equation for natural frequency and damping ratio defined by:

$$\omega_N^2 = \frac{a\beta}{4} \quad \text{and} \quad 2\xi\omega_N = \frac{D\ell\beta}{4}.$$

For any positive a, the simple harmonic oscillator has a solution with $h(0) = 0$, and $h'(0) = 1$. Lemma 13 from Appendix A gives explicit values of a that lead to h with $h'(x) > 0$ and $h''(x) < 0$ for all $x \in D$, and gives explicit expressions for $c_8 = (h'(D))$ in these cases. The result follows by plugging in these values. ■

Note that the function $\bar{f}(x)$ satisfies all of the same assumptions that $f(x, z)$ does, with the further property that it is independent of z. As a result, Proposition 7 applies to $\mathbf{x}_t^M$ as well. We can use this fact to prove the exponential convergence with respect to W_1 result from 4.

Proof of Lemma 4. Lemma 18 from Appendix F implies that $\pi_{\beta\bar{f}}$ is invariant with respect to the dynamics of the process $\mathbf{x}^M$.

Now, apply Proposition 7 to $\mathbf{x}^M = \mathbf{x}^{M,1}$ and $\mathbf{x}^{M,2}$ such that $\mathcal{L}(\mathbf{x}_0^{M,2}) = \pi_{\beta\bar{f}}$ to give

$$W_1(\mathcal{L}(\mathbf{x}_t^M), \pi_{\beta\bar{f}}) \leq c_8 e^{-\eta a t} W_1(\mathcal{L}(\mathbf{x}_0^M), \pi_{\beta\bar{f}}) \leq c_8 D e^{-\eta a t}.$$

The specific form from the lemma arises because in this case $\mathcal{L}(\mathbf{x}_t^{M,2}) = \pi_{\beta\bar{f}}$ for all $t \geq 0$, and also that $W_1(\mathcal{L}(\mathbf{x}_0^M), \pi_{\beta\bar{f}}) \leq D$, since $\mathcal{K}$ has diameter D. Setting $c_1 = c_8 D$ gives the result. ■

5. A Switching Argument for Uniform Bounds

The following lemma, which is based on a method from Chau et al. (2019), is useful for deriving W_1 bounds from $\mathcal{L}(\mathbf{x}_t^C)$ that hold uniformly over time. It is proved in Appendix B.

Lemma 8 *Assume that $\eta \leq 1/2$. Let $\hat{\mathbf{x}}$ be a process such that for all $0 \leq s \leq t$, if $\hat{\mathbf{x}}_s = \mathbf{x}_s^C$ then $W_1(\mathcal{L}(\hat{\mathbf{x}}_t), \mathcal{L}(\mathbf{x}_t^C)) \leq g(t - s)$, where g is a monotonically increasing function. If $\hat{\mathbf{x}}_0 = \mathbf{x}_0^C$, then for all $t \geq 0$, we have that*

$$W(\mathcal{L}(\hat{\mathbf{x}}_t), \mathcal{L}(\mathbf{x}_t^C)) \leq g(\eta^{-1}) \left(1 + \frac{c_8}{1 - e^{-a/2}}\right)$$

We will refer to this as the "switching lemma", as the proof follows by constructing a sequence of processes that switch from the dynamics of $\hat{\mathbf{x}}$ to the dynamics of $\mathbf{x}^C$.

6. Bounding the Algorithm from the Continuous RSDE

The goal of this section is to prove Lemma 5, which states that the law of the algorithm, $\mathbf{x}_t^A$, is close to the law of the continuous reflected SDE, $\mathbf{x}_t^C$. To derive this bound, we introduce an intermediate process $\mathbf{x}^D$, and show that its law is close to that of both $\mathbf{x}^C$ and $\mathbf{x}^A$.

Recall the process $\mathbf{y}^C$ defined in Subsection 3.3. For any initial $\mathbf{x}_0^D \in \mathcal{K}$, we define the following iteration on the integers:

$$\mathbf{x}_{k+1}^D = \Pi_{\mathcal{K}}(\mathbf{x}_k^D + \mathbf{y}_{k+1}^C - \mathbf{y}_k^C),$$

and set $\mathbf{x}_t^D = \mathbf{x}_{\lfloor t \rfloor}^D$ for all $t \in \mathbb{R}$.

The process, $\mathbf{x}^D$, can also be interpreted as a Skorokhod solution. Indeed, let $\mathcal{D}$ be the discretization operator that sets $\mathcal{D}(x)_t = x_{\lfloor t \rfloor}$ for any continuous-time trajectory, x_t. Then, provided that $\mathbf{x}_0^D = \mathbf{x}_0^C$, we have that $\mathbf{x}^D = \mathcal{S}(\mathcal{D}(\mathbf{y}^C))$. Recall that $\mathcal{S}$ corresponds to the Skorokhod solution. See Appendix E for a more detailed explanation of this construction.

The following lemmas give the specific bounds on the differences between $\mathcal{L}(\mathbf{x}_t^C)$ and $\mathcal{L}(\mathbf{x}_t^D)$, and between $\mathcal{L}(\mathbf{x}_t^A)$ and $\mathcal{L}(\mathbf{x}_t^D)$, respectively. They are proved in Appendix B.

Lemma 9 *Assume that $\mathbf{x}_0^D = \mathbf{x}_0^C$ and $\eta \leq 1$. There are constants, c_9 and c_{10} such that for all $t \geq 0$, the following bound holds:*

$$W_1(\mathcal{L}(\mathbf{x}_t^C), \mathcal{L}(\mathbf{x}_t^D)) \leq \mathbb{E}\left[\|\mathbf{x}_t^C - \mathbf{x}_t^D\|\right] \leq (\eta \log(4\max\{1,t\}))^{1/4} \left(c_9\sqrt{\eta t} + c_{10}\right)$$

The constants are given by:

$$c_9 = \sqrt{2\left(\frac{u+\ell D}{2r} + \frac{n\sigma}{\sqrt{2}r} + \frac{2n\sqrt{2}}{r\sqrt{\beta}}\right)\left(\frac{n}{\beta} + Du + 2Dn\sigma\right)}$$

$$c_{10} = \sqrt{2\left(Du + 2n\sigma + \frac{n}{\beta}\right)} + D\sqrt{\frac{u+\ell D}{2r} + \frac{n\sigma}{\sqrt{2}r} + \frac{2n\sqrt{2}}{r\sqrt{\beta}}}$$

Lemma 10 *Assume that $\mathbf{x}_0^A = \mathbf{x}_0^D$ and $\eta \leq 1$. Then for all $t \geq 0$:*

$$W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^D)) \leq (\eta \log(4\max\{1,t\}))^{1/4} \left(c_9\sqrt{\eta t} + c_{10}\right)\left((1+\eta\ell)^t - 1\right)$$

Now Lemma 5 can be proved by combining Lemmas 8, 9, and 10.

Proof of Lemma 5 Using the triangle inequality and Lemmas 9 and 10 gives

$$\begin{aligned} W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) &\leq W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^D)) + W(\mathcal{L}(\mathbf{x}_t^D), \mathcal{L}(\mathbf{x}_t^C)) \\ &\leq (\eta \log(4\max\{1,t\}))^{1/4} \left(c_9\sqrt{\eta t} + c_{10}\right)(1+\eta\ell)^t. \end{aligned}$$

Now we will utilize the switching trick from Lemma 8 to simplify the bound. Define $g : [0,t] \to \mathbb{R}$ by $g(s) = (\eta \log(4\max\{1,t\}))^{1/4} \left(c_9\sqrt{\eta s} + c_{10}\right)(1+\eta\ell)^s$. Then applying Lemma 8 using the bound from g gives the desired bound:

$$\begin{aligned} &W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) \\ &\quad\leq (\eta \log(4\max\{1,t\}))^{1/4} (c_9 + c_{10})(1+\eta\ell)^{1/\eta}\left(1 + \frac{c_8}{1-e^{-a/2}}\right) \\ &\quad\leq (\eta \log(4\max\{1,t\}))^{1/4} (c_9 + c_{10}) e^{\ell}\left(1 + \frac{c_8}{1-e^{-a/2}}\right) \end{aligned}$$

The second inequality uses the fact that for all $\eta > 0$,

$$(1+\eta\ell)^{1/\eta} \leq e^{\ell} \iff \frac{\log(1+\eta\ell)}{\eta} \leq \ell$$

where the right inequality holds due to concavity of the logarithm.

Now, for $t \geq 4$ we have $\log(4t) \leq 2\log(t)$. So, setting

$$c_6 = 2^{1/4}(c_9 + c_{10}) e^{\ell}\left(1 + \frac{c_8}{1-e^{-a/2}}\right)$$

gives the bound $W_1(\mathcal{L}(\mathbf{x}_t^A), \mathcal{L}(\mathbf{x}_t^C)) \leq c_6(\eta \log t)^{1/4}$. ■

7. Averaging Out the External Variables

Now we show that the dynamics of the continuous reflected SDE, $\mathbf{x}^C$, and its averaged version, $\mathbf{x}^M$, have similar laws. In particular, we will prove Lemma 6. The general strategy is similar to that of Section 6. Namely, we devise a new process, $\mathbf{x}^B$ that fits "between" $\mathbf{x}^C$ and $\mathbf{x}^M$. Then the desired bound is given by showing that $\mathcal{L}(\mathbf{x}_t^M)$ is close to $\mathcal{L}(\mathbf{x}_t^B)$ and that $\mathcal{L}(\mathbf{x}_t^B)$ is close to $\mathcal{L}(\mathbf{x}_t^C)$.

The new process is defined by $\mathbf{x}^B = \mathcal{S}(\mathbf{y}^B)$ where

$$\mathbf{y}_t^B = \mathbf{x}_0^B + \sqrt{\frac{2\eta}{\beta}}\mathbf{w}_t - \eta \int_0^t \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_{\lfloor s \rfloor})ds. \tag{18}$$

So, we see that $\mathbf{x}^B$ has similar dynamics to $\mathbf{x}^C$, but $\mathbf{x}^M$ is used in place of $\mathbf{x}^C$ in the drift term.

The lemmas describing the relations between $\mathcal{L}(\mathbf{x}_t^M)$ and $\mathcal{L}(\mathbf{x}_t^B)$ and between $\mathcal{L}(\mathbf{x}_t^C)$ and $\mathcal{L}(\mathbf{x}_t^B)$ are stated below. They are proved in Appendix B.

Lemma 11 *Assume that $\mathbf{x}_0^M = \mathbf{x}_0^B$ and that $\eta \leq 1$. Then is a positive constants, c_{11}, c_{12}, abd c_{13} such that for all $t \geq 0$,*

$$W_1(\mathcal{L}(\mathbf{x}_t^B), \mathcal{L}(\mathbf{x}_t^M)) \leq \mathbb{E}\left[\|\mathbf{x}_t^B - \mathbf{x}_t^M\|\right] \leq c_{11}\eta t^{1/2} + c_{12}\eta^{1/2}t^{1/4} + c_{13}\eta t^{3/4}$$

The constants are given by:

$$c_{11} = 2\sigma\sqrt{n}$$
$$c_{12} = \sqrt{\frac{64n\sigma D\sqrt{2\pi}}{r}}$$
$$c_{13} = \sqrt{\frac{128n\sigma\sqrt{2\pi}}{r}\left(\frac{n}{\beta} + Du + 2Dn\sigma\right)}$$

Lemma 12 *Assume that $\mathbf{x}_0^C = \mathbf{x}_0^B$ and $\eta \leq 1$. Then for all $t \geq 0$,*

$$W_1(\mathcal{L}(\mathbf{x}_t^B), \mathcal{L}(\mathbf{x}_t^C)) \leq \left(c_{11}\eta t^{1/2} + c_{12}\eta^{1/2}t^{1/4} + c_{13}\eta t^{3/4}\right)(e^{\eta\ell t} - 1).$$

Proof of Lemma 6 Using the triangle inequality along with Lemmas 11 and 12 shows that

$$\begin{aligned} W_1(\mathcal{L}(\mathbf{x}_t^M), \mathcal{L}(\mathbf{x}_t^C)) &\leq W_1(\mathcal{L}(\mathbf{x}_t^M), \mathcal{L}(\mathbf{x}_t^B)) + W_1(\mathcal{L}(\mathbf{x}_t^B), \mathcal{L}(\mathbf{x}_t^C)) \\ &\leq \left(c_{11}\eta t^{1/2} + c_{12}\eta^{1/2}t^{1/4} + c_{13}\eta t^{3/4}\right)e^{\eta\ell t} \end{aligned}$$

Using Lemma 8 along with the fact that $\eta^{1/2} \leq \eta^{1/4}$ gives $W_1(\mathcal{L}(\mathbf{x}_t^M), \mathcal{L}(\mathbf{x}_t^C)) \leq c_7\eta^{1/4}$ with

$$c_7 = (c_{11} + c_{12} + c_{13})e^{\ell}\left(1 + \frac{c_8}{1 - e^{-a/2}}\right).$$

■

8. Conclusions and Future Work

In this paper, we have given non-asymptotic bounds for a projected stochastic gradient Langevin algorithm applied to non-convex functions with IID external random variables. In particular, we demonstrated convergence of sampling methods with respect to the 1-Wasserstein distance and showed how the sampling results can be utilized for non-convex learning. The results were derived using a novel approach contraction analysis for reflected SDEs. The contraction analysis utilizes connections with simple harmonic oscillator problems to get explicit contraction rate bounds. Future work will include a variety of extensions. The assumption of a compact convex domain, $\mathcal{K}$, can likely be relaxed to a general class of non-convex non-compact domains. This would require use of more general analysis of Skorokhod problems as in Lions and Sznitman (1984) along with a dissipativity condition to ensure that the reflected SDEs remain contractive. The assumptions that $\mathbf{z}_k$ are IID and that $\nabla_x f(x, \mathbf{z}_k)$ is sub-Gaussian will also be relaxed in future work. The eventual goal will be to use the method for problems in time-series analysis, control, and reinforcement learning.

Acknowledgments

The author would like to thank Tyler Lekang, Jonah Roux, Suneel Sheikh, Chuck Hisamoto, and Michael Schmit for helpful discussions. The author acknowledges funding from NASA STTR 19-1-T4.03-3451 and NSF CMMI 1727096

References

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

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.

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.

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.

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.

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.

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.

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.

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.

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

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

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

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.

Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21, 2020.

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

David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*. Springer, 1998.

Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

J Michael Harrison and Ruth J Williams. Multidimensional reflected brownian motions having exponential stationary distributions. *The Annals of Probability*, pages 115–137, 1987.

Mark Herbster and Manfred K Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1(Sep):281–309, 2001.

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

Olav Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, 2002.

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

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *Cambridge University Press (preprint)*, 2019.

John M Lee. *Introduction to Smooth Manifolds*. Springer, 2013.

Pierre-Louis Lions and Alain-Sol Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984.

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.

Mateusz B Majka, Aleksandar Mijatović, Łukasz Szpruch, et al. Nonasymptotic bounds for sampling algorithms without log-concavity. *Annals of Applied Probability*, 30(4):1534–1581, 2020.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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

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

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

Leszek Słomiński. Euler's approximations of solutions of sdes with reflecting boundary. *Stochastic processes and their applications*, 94(2):317–337, 2001.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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.

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.

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.

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

Appendix A. Bounds on Simple Harmonic Oscillator Coefficients.

Lemma 13 *Consider the simple harmonic oscillator*

$$\omega_N^2 h(x) + 2\xi\omega_N h'(x) + h''(x) = 0$$

with

$$\omega_N = \frac{\sqrt{a\beta}}{2} \quad \textit{and} \quad \xi = \frac{D\ell}{4}\sqrt{\frac{\beta}{a}}.$$

and boundary condition $h(0) = 0$ *and* $h'(0) = 1$.

For any positive values of D, ℓ, *and* β *if* a *is set to*

$$a = \frac{D^2\ell^2\beta}{16}\left(1 - \tanh^2\left(\frac{D^2\ell\beta}{8}\right)\right)$$

then $h'(x) > 0$ *and* $h''(x) < 0$ *for all* $x \in [0, D]$ *and*

$$(h'(D))^{-1} = \frac{e^{D\omega_N\xi}}{\cosh(D\omega_N\sqrt{\xi^2-1}) - \frac{\xi}{\sqrt{\xi^2-1}}\sinh(D\omega_N\sqrt{1-\xi^2})}$$

If $D^2\ell\beta < 8$*, then* a *can be set to* $a = \frac{4}{D^2\beta}$ *and in this case* $h'(x) > 0$ *and* $h''(x) < 0$ *for all* $x \in [0, D]$ *and*

$$(h'(D))^{-1} = \frac{e^{D\omega_N\xi}}{\cos(D\omega_N\sqrt{1-\xi^2}) - \frac{\xi}{\sqrt{1-\xi^2}}\sin(D\omega_N\sqrt{1-\xi^2})}.$$

Proof We will tune a to ensure that $h'(x) > 0$ for all $x \in [0, D]$. Then since $h(x) \geq 0$ for $x \in [0, D]$, the simple harmonic oscillator equation implies that $h''(x) < 0$ for $x \in [0, D]$.

We will consider the underdamped case with $\xi < 1$ and the overdamped case with $\xi > 1$. We will see that for any collection of parameters, a can be chosen to give an overdamped solution with the desired properties. However, when $D^2\ell\beta < 8$, a larger a can be chosen which gives rise to an underdamped solution with the desired properties.

First we consider the underdamped case. The expression for ξ from (7) shows that

$$\xi^2 < 1 \iff \frac{D^2\ell^2\beta}{16} < a. \tag{19}$$

Now we will try to maximize a while ensuring that $h'(x) > 0$. Standard methods from linear differential equations show that h and its derivative are given by:

$$h(x) = e^{-x\omega_N\xi}\frac{\sin(x\omega_N\sqrt{1-\xi^2})}{\omega_N\sqrt{1-\xi^2}}$$
$$h'(x) = \frac{e^{-x\omega_N\xi}}{\sqrt{1-\xi^2}}(\sqrt{1-\xi^2}\cos(x\omega_N\sqrt{1-\xi^2}) - \xi\sin(x\omega_N\sqrt{1-\xi^2})).$$

The smallest $x > 0$ such that $h'(x) = 0$ is the smallest $x > 0$ such that

$$(\cos(x\omega_N\sqrt{1-\xi^2}), \sin(x\omega_N\sqrt{1-\xi^2})) = (\xi, \sqrt{1-\xi^2}).$$

Using the fact that $\sin'(\theta) < 1$ for $\theta \neq 2\pi k$, we have that

$$\sin(D\omega_N\sqrt{1-\xi^2}) < D\omega_N\sqrt{1-\xi^2}.$$

So, if we choose $\omega_N \leq D^{-1}$ we will have $\sin(x\omega_N\sqrt{1-\xi^2}) < \sqrt{1-\xi^2}$ and thus $h''(x) > 0$ for all $x \in [0, D]$. Plugging in the expression for ω_N from (7) shows that a must satisfy

$$a \leq \frac{4}{D^2\beta}. \tag{20}$$

Comparing (19) and (20) shows that a suitable a can only be chosen when $D^2\ell\beta < 8$. The a from the lemma statement is chosen by taking the largest possible value. Note that by construction, a satisfies (19) and so $\xi < 1$ in this case.

Now we consider the overdamped case, so that $\xi^2 > 1$. In this case, standard methods from linear differential equations show that h and its derivative are given by:

$$h(x) = e^{-x\omega_N\xi}\frac{\sinh(x\omega_N\sqrt{\xi^2-1})}{\omega_N\sqrt{\xi^2-1}}$$
$$h'(x) = \frac{e^{-x\omega_N\xi}}{\sqrt{\xi^2-1}}(\sqrt{\xi^2-1}\cosh(x\omega_N\sqrt{\xi^2-1}) - \xi\sinh(x\omega_N\sqrt{\xi^2-1}))$$

Thus $h'(x) = 0$ precisely when $\tanh(x\omega_N\sqrt{\xi^2-1}) = \frac{\sqrt{\xi^2-1}}{\xi}$. Since $\tanh$ is monotonically increasing, if $\tanh(D\omega_N\sqrt{\xi^2-1}) < \frac{\sqrt{\xi^2-1}}{\xi}$, then we will have that $h'(x) > 0$ for all $x \in [0, D]$.

Plugging in the expressions for ω_N and ξ gives for all $a > 0$

$$\tanh(D\omega_N\sqrt{\xi^2-1}) = \tanh\left(\frac{D}{2}\sqrt{\frac{D^2\ell^2\beta^2}{16} - a\beta}\right) < \tanh\left(\frac{D^2\ell\beta}{8}\right).$$

So to ensure that $h'(x) > 0$ for all $x \in [0, D]$, it suffices to choose a so that $\frac{\sqrt{\xi^2-1}}{\xi}$ achieves the bound on the right. In particular, after some algebra we find that

$$a = \frac{D^2\ell^2\beta}{16}\left(1 - \tanh^2\left(\frac{D^2\ell\beta}{8}\right)\right) > 0.$$

Plugging this expression into the definition of ξ shows that $\xi^2 > 1$, and so the oscillator is indeed overdamped. Thus, h is well-defined and has all the desired properties, so the proof is complete. ■

Appendix B. Proofs of Supporting Lemmas

The proofs below use the following notation from Rockafellar (2015). Let $\gamma(x|\mathcal{K})$ denote the gauge function:

$$\gamma(x|\mathcal{K}) = \inf\{t > 0 | x \in t\mathcal{K}\}$$

and let $\gamma^*(x|\mathcal{K})$ be the support function:

$$\delta^\star(x|\mathcal{K}) = \sup\{y^\top x | y \in \mathcal{K}\}.$$

By the assumption on $\mathcal{K}$, it follows that $\gamma(x|\mathcal{K}) \leq r^{-1}\|x\|$.

Proof of Lemma 8 Consider the family of switching processes $\hat{\mathbf{x}}^C_{s,t}$ be the process such that $\hat{\mathbf{x}}^C_{s,t} = \hat{\mathbf{x}}_t$ for $t \le s$ and then for $t \ge s$, the dynamics of $\hat{\mathbf{x}}^C_{s,t}$ follow (4), the definition of $\mathbf{x}^C_t$.

Let $H = \lfloor 1/\eta \rfloor$ and assume that $t \in [kH, (k+1)H)$. It follows that $\hat{\mathbf{x}}^C_{0,t} = \mathbf{x}^C_t$ and $\hat{\mathbf{x}}^C_{(k+1)H,t} = \hat{\mathbf{x}}_t$. The triangle inequality then implies that

$$W_1(\mathcal{L}(\mathbf{x}^C_t), \mathcal{L}(\hat{\mathbf{x}}_t)) \le \sum_{i=0}^{k} W_1(\mathcal{L}(\hat{\mathbf{x}}^C_{iH,t}), \mathcal{L}(\hat{\mathbf{x}}^C_{(i+1)H,t}))$$

For $i < k$, using Proposition 7, followed by the hypothesis gives

$$\begin{aligned} W_1(\mathcal{L}(\hat{\mathbf{x}}^C_{iH,t}), \mathcal{L}(\hat{\mathbf{x}}^C_{(i+1)H,t})) &\le c_8 e^{-\eta a(t-(i+1)H)} W(\mathcal{L}(\hat{\mathbf{x}}^C_{iH,(i+1)H}), \mathcal{L}(\hat{\mathbf{x}}_{(i+1)H})) \\ &\le c_8 e^{-\eta a(t-(i+1)H)} g(H) \\ &\le c_8 e^{-a(k-i-1)/2} g(\eta^{-1}) \end{aligned}$$

The final inequality uses the facts that $\frac{1}{2} \le \eta H \le 1$ along with monotonicity of g. The lower bound on ηH arises because $H \ge \eta^{-1} - 1$ and so $\eta H \ge 1 - \eta \ge 1/2$, since $\eta \le 1/2$.

It follows that the first k terms of the sum can be bounded by:

$$\begin{aligned} \sum_{i=0}^{k-1} W_1(\mathcal{L}(\hat{\mathbf{x}}^C_{iH,t}), \mathcal{L}(\hat{\mathbf{x}}^C_{(i+1)H,t})) &\le \sum_{i=0}^{k-1} c_8 e^{-a(k-i-1)/2} g(1) \\ &\le \frac{c_8 g(1)}{1 - e^{-a/2}} \end{aligned}$$

For $i = k$ the hypothesis gives

$$\begin{aligned} W_1(\mathcal{L}(\hat{\mathbf{x}}^C_{iH,t}), \mathcal{L}(\hat{\mathbf{x}}^C_{(i+1)H,t})) &= W_1(\mathcal{L}(\hat{\mathbf{x}}^C_{kH,t}), \mathcal{L}(\hat{\mathbf{x}}_t)) \\ &\le g(t - kH) \le g(\eta^{-1}) \end{aligned}$$

Adding this to the bound from the first k terms gives the result. ■

Proof of Lemma 9 The basic idea follows arguments from Bubeck et al. (2018). However, we must deviate from the method to account for the extra randomness due to $\mathbf{z}_i$. First note that since $\mathbf{x}^D_t = \mathbf{x}^D_{\lfloor t \rfloor}$, we get the following triangle inequality bound:

$$\begin{aligned} \|\mathbf{x}^C_t - \mathbf{x}^D_t\| &= \|\mathbf{x}^C_t - \mathbf{x}^C_{\lfloor t \rfloor} + \mathbf{x}^C_{\lfloor t \rfloor} - \mathbf{x}^D_{\lfloor t \rfloor}\| &(21)\\ &\le \|\mathbf{x}^C_t - \mathbf{x}^C_{\lfloor t \rfloor}\| + \|\mathbf{x}^C_{\lfloor t \rfloor} - \mathbf{x}^D_{\lfloor t \rfloor}\|. &(22) \end{aligned}$$

For simpler notation, set $\lfloor t \rfloor = k$.

The first term can be estimated directly via Itô's rule:

$$\begin{aligned} &d\|\mathbf{x}^C_t - \mathbf{x}^C_k\|^2 \\ &\quad = 2(\mathbf{x}^C_t - \mathbf{x}^C_k)^\top \left(-\eta \nabla_x f(\mathbf{x}^C_t, \mathbf{z}_{\lfloor t \rfloor}) dt - \mathbf{v}_s d\boldsymbol{\mu}(s) + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t \right) + \frac{2\eta n}{\beta} dt. \quad (23) \end{aligned}$$

Note that the inner products between the state difference and gradient terms can be bounded by:

$$\begin{aligned}&(\mathbf{x}_t^C - \mathbf{x}_k^C)^\top \nabla_x \bar{f}(\mathbf{x}_t^C) + (\mathbf{x}_t^C - \mathbf{x}_k^C)^\top \left(\nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x \bar{f}(\mathbf{x}_t^C)\right) \\ &\quad \leq Du + D\|\nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x \bar{f}(\mathbf{x}_t^C)\| \end{aligned} \tag{24}$$

The uniform sub-Gaussian assumption implies that the mean of the term on the right is bounded above by $2n\sigma$. Indeed, if $\mathbf{g}$ is a sub-Gaussian vector in $\mathbb{R}^n$ with sub-gaussian parameter σ, then

$$\mathbb{E}\left[\|\mathbf{g}\|\right] \leq \sum_{i=1}^n \mathbb{E}[|e_i^\top \mathbf{g}|] = \sum_{i=1}^n \mathbb{E}[\max\{e_i^\top \mathbf{g}, -e_i^\top \mathbf{g}\}] \leq n\sigma\sqrt{2\log(2)} \leq 2n\sigma.$$

Here e_i are the standard basis vectors of $\mathbb{R}^n$. The third inequality is based on a standard bounding method for sub-Gaussian variables. See exercise 2.21 of Wainwright (2019). Additionally, we will work out these details for bounding a different maximum of sub-Gaussian variables.

Thus, by integrating (23), taking expectations, and noting that $t - k \leq 1$, we can conclude that

$$\mathbb{E}\left[\|\mathbf{x}_t^C - \mathbf{x}_k^C\|^2\right] \leq 2\eta\left(\frac{n}{\beta} + uD + 2n\sigma\right)$$

Thus, an elementary Cauchy-Schwarz bound gives that

$$\mathbb{E}\left[\|\mathbf{x}_t^C - \mathbf{x}_k^C\|\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{x}_t^C - \mathbf{x}_k^C\|^2\right]} \leq \sqrt{2\eta\left(Du + 2n\sigma + \frac{n}{\beta}\right)}. \tag{25}$$

The rest of the proof bounds $\mathbb{E}[\|\mathbf{x}_k^C - \mathbf{x}_k^D\|]$. When $k = 0$, this term is 0, so we focus on the $k \geq 1$ case. Recall that $\mathbf{x}_t^C$ solves the Skorokhod problem for $\mathbf{y}_t^C$ and $\mathbf{x}_t^D$ solves the Skorkhod problem for $\mathcal{D}(\mathbf{y}^C)_t = \mathbf{y}_{\lfloor t \rfloor}^C$. Let $\boldsymbol{\varphi}_t^D = -\int_0^t \mathbf{v}_t^D d\boldsymbol{\mu}^D(t)$ be the unique projection process such that $\mathbf{x}_t^D = \mathbf{y}_{\lfloor t \rfloor}^C + \boldsymbol{\varphi}_t^D$. Then, Lemma 2.2 of Tanaka et al. (1979) implies that

$$\begin{aligned}&\|\mathbf{x}_k^C - \mathbf{x}_k^D\|^2 \\ &\quad \leq \|\mathbf{y}_k^C - \mathbf{y}_k^C\|^2 + 2\int_0^k (\mathbf{y}_k^C - \mathbf{y}_k^C - \mathbf{y}_s^C + \mathbf{y}_{\lfloor s \rfloor}^C)^\top (\mathbf{v}_s^D d\boldsymbol{\mu}^D(s) - \mathbf{v}_s^C d\boldsymbol{\mu}^C(s)) \\ &\quad = 2\int_0^k (\mathbf{y}_{\lfloor s \rfloor}^C - \mathbf{y}_s^C)^\top (\mathbf{v}_s^D d\boldsymbol{\mu}^D(s) - \mathbf{v}_s d\boldsymbol{\mu}(s))\end{aligned}$$

Note that for any integer, i, $\mathbf{y}_{\lfloor s \rfloor}^C$ is constant for $s \in (i, i+1)$. It follows that the measure, $\boldsymbol{\mu}^D$ is supported on the integers. However, the integrand is zero on the integers, so we arrive at the simplified bound:

$$\|\mathbf{x}_k^C - \mathbf{x}_k^D\|^2 \leq 2\int_0^k (\mathbf{y}_s^C - \mathbf{y}_{\lfloor s \rfloor}^C)^\top \mathbf{v}_s^C d\boldsymbol{\mu}^C(s).$$

Now, the elementary inequality $x^\top y \leq \gamma(x|\mathcal{K})\delta^*(x|\mathcal{K})$ followed by Hölder's inequality gives:

$$\begin{aligned}\|\mathbf{x}_k^C - \mathbf{x}_k^D\|^2 &\leq 2\int_0^k \gamma(\mathbf{y}_s^C - \mathbf{y}_{\lfloor s \rfloor}^C|\mathcal{K})\delta^\star(\mathbf{v}_s|\mathcal{K}) d\boldsymbol{\mu}(s) \\ &\leq 2\left(\sup_{s\in[0,k]} \gamma(\mathbf{y}_s^C - \mathbf{y}_{\lfloor s \rfloor}^C|\mathcal{K})\right)\int_0^k \delta^\star(\mathbf{v}_s|\mathcal{K}) d\boldsymbol{\mu}(s)\end{aligned}$$

Taking square-roots, then followed by expectations, and then employing the Cauchy-Schwarz inequality gives:

$$\mathbb{E}\left[\|\mathbf{x}_k^C - \mathbf{x}_k^D\|\right] \leq \sqrt{2}\mathbb{E}\left[\sqrt{\sup_{s\in[0,k]} \gamma(\mathbf{y}_s^C - \mathbf{y}_{\lfloor s\rfloor}^C|\mathcal{K})}\sqrt{\int_0^k \delta^\star(\mathbf{v}_s|\mathcal{K})d\boldsymbol{\mu}(s)}\right]$$
$$\leq \sqrt{2\mathbb{E}\left[\sup_{s\in[0,k]} \gamma(\mathbf{y}_s^C - \mathbf{y}_{\lfloor s\rfloor}^C|\mathcal{K})\right]\mathbb{E}\left[\int_0^k \delta^\star(\mathbf{v}_s|\mathcal{K})d\boldsymbol{\mu}(s)\right]}. \tag{26}$$

So, now it suffices to bound both terms on the right of (26).

We first bound the γ term. The methodology deviates from that of Bubeck et al. (2018), as this term is now a bit more complicated. First note that $\gamma(x|\mathcal{K}) \leq r^{-1}\|x\|$ because $\mathcal{K}$ contains a ball of radius r around the origin. Thus, plugging in the defition for $\mathbf{y}_t^C$ and using the triangle inequality gives:

$$\gamma(\mathbf{y}_s^C - \mathbf{y}_{\lfloor s\rfloor}^C|\mathcal{K}) \leq r^{-1}\|\mathbf{y}_s^C - \mathbf{y}_{\lfloor s\rfloor}^C\|$$
$$\leq r^{-1}\eta\int_{\lfloor s\rfloor}^s \|\nabla_x f(\mathbf{x}_\tau^C, \mathbf{z}_{\lfloor s\rfloor})\|d\tau + r^{-1}\sqrt{\frac{2\eta}{\beta}}\|\mathbf{w}_s - \mathbf{w}_{\lfloor s\rfloor}\|$$

For compact notation, set $i = \lfloor s\rfloor$ and $\mathbf{g}_\tau = \nabla_x f(\mathbf{x}_\tau^C, \mathbf{z}_i) - \nabla_x \bar{f}(\mathbf{x}_\tau^C)$. To bound the integral term, note that

$$\|\nabla_x f(\mathbf{x}_\tau^C, \mathbf{z}_{\lfloor s\rfloor})\| = \|\nabla_x \bar{f}(\mathbf{x}_\tau^C) + (\mathbf{g}_\tau - \mathbf{g}_i) + \mathbf{g}_i\|$$
$$\leq u + \ell D + \|\mathbf{g}_i\|. \tag{27}$$

The inequality arises because of the bound on $\|\nabla_x \bar{f}\|$ and the Lipschitz property of $\nabla_x f$.

It follows that

$$\gamma(\mathbf{y}_s^C - \mathbf{y}_{\lfloor s\rfloor}^C|\mathcal{K}) \leq \frac{\eta(u+\ell D)}{2r} + \frac{\eta}{2r}\|\mathbf{g}_i\| + r^{-1}\sqrt{\frac{2\eta}{\beta}}\|\mathbf{w}_s - \mathbf{w}_{\lfloor s\rfloor}\|. \tag{28}$$

Thus, to bound the first term on the right of (26), it suffices to bound $\max_{i=0,\ldots,k-1}\|\mathbf{g}_i\|$ and $\sup_{s\in[0,k]}\|\mathbf{w}_s - \mathbf{w}_{\lfloor s\rfloor}\|$. The $\|\mathbf{g}_i\|$ terms can be bounded using a modification of a standard sub-Gaussian bounding method from exercise 2.21 of Wainwright (2019). We show it explicitly, as the method will be generalized when bounding the $\|\mathbf{w}_s - \mathbf{w}_{\lfloor s\rfloor}\|$ terms.

Let e_j be the standard basis vectors of $\mathbb{R}^n$. Then the following bound follows from the triangle inequality:

$$\|\mathbf{g}_i\| \leq \sum_{j=1}^n |e_j^\top \mathbf{g}_i|$$
$$= \sum_{j=1}^n \max_{\varepsilon\in\{-1,1\}} \varepsilon e_j^\top \mathbf{g}_i.$$

Then for any $\lambda > 0$ we have that

$$\begin{aligned}\max_{i=0,\dots,k-1} \|\mathbf{g}_i\| &\leq \max_{i=0,\dots,k-1} \sum_{j=1}^{n} \max_{\varepsilon\in\{-1,1\}} \varepsilon e_j^\top \mathbf{g}_i \\ &\leq \sum_{j=1}^{n} \max_{i\in\{0,\dots,k-1\},\varepsilon\in\{-1,1\}} \varepsilon e_j^\top \mathbf{g}_i \\ &\leq \sum_{j=1}^{n} \lambda^{-1} \log\left(\sum_{i=0}^{k-1}\sum_{\varepsilon\in\{-1,1\}} \exp\left(\lambda\varepsilon e_j^\top \mathbf{g}_i\right)\right).\end{aligned}$$

Taking expectations and using Jensen's inequality, followed by the sub-Gaussian property of $\mathbf{g}_i$ gives

$$\begin{aligned}\mathbb{E}\left[\max_{i=0,\dots,k-1} \|\mathbf{g}_i\|\right] &\leq \sum_{j=1}^{n} \lambda^{-1} \log\left(\sum_{i=0}^{k-1}\sum_{\varepsilon\in\{-1,1\}} \mathbb{E}\left[\exp\left(\lambda\varepsilon e_j^\top \mathbf{g}_i\right)\right]\right) \\ &\leq \frac{n}{\lambda}\log\left(2ke^{\lambda^2\sigma^2/2}\right) \qquad (29) \\ &= \frac{n\log(2k)}{\lambda} + \frac{n\lambda\sigma^2}{2}. \qquad (30)\end{aligned}$$

Optimizing over λ gives:

$$\mathbb{E}\left[\max_{i=0,\dots,k-1} \|\mathbf{g}_i\|\right] \leq n\sigma\sqrt{2\log(2k)}. \qquad (31)$$

Now we will bound $\mathbb{E}\left[\sup_{s\in[0,k]} \|\mathbf{w}_s - \mathbf{w}_{\lfloor s\rfloor}\|\right]$ using an extension of the argument just used. Note that

$$\mathbb{E}\left[\sup_{s\in[0,k]} \|\mathbf{w}_s - \mathbf{w}_{\lfloor s\rfloor}\|\right] = \mathbb{E}\left[\max_{i=0,\dots,k-1}\sup_{s\in[i,i+1]} \|\mathbf{w}_s - \mathbf{w}_i\|\right]$$

Then the triangle inequality implies that

$$\|\mathbf{w}_s - \mathbf{w}_i\| \leq \sum_{j=1}^{n} |e_j^\top(\mathbf{w}_s - \mathbf{w}_i)| = \sum_{j=1}^{n} \max_{\varepsilon\in\{-1,1\}} \varepsilon e_j^\top(\mathbf{w}_s - \mathbf{w}_i).$$

It follows that for all $\lambda > 0$, we get:

$$\begin{aligned}\max_{i=0,\dots,k-1}\sup_{s\in[i,i+1]} \|\mathbf{w}_s - \mathbf{w}_i\| &\leq \sum_{j=1}^{n} \max_{i\in\{0,\dots,k\},\varepsilon\in\{-1,1\}}\sup_{s\in[i,i+1]} \varepsilon e_j^\top(\mathbf{w}_s - \mathbf{w}_i) \\ &\leq \sum_{j=1}^{n} \lambda^{-1}\log\left(\sum_{i=0}^{k-1}\sum_{\varepsilon\in\{-1,1\}}\sup_{s\in[i,i+1]} e^{\lambda\varepsilon e_j^\top(\mathbf{w}_s - \mathbf{w}_i)}\right)\end{aligned}$$

So, taking expectations and using Jensen's inequality gives:

$$\begin{aligned}
&\mathbb{E}\left[\max_{i=0,\ldots,k-1}\sup_{s\in[i,i+1]}\|\mathbf{w}_s-\mathbf{w}_i\|\right] \\
&\quad\le\sum_{j=1}^{n}\lambda^{-1}\mathbb{E}\left[\log\left(\sum_{i=0}^{k-1}\sum_{\varepsilon\in\{-1,1\}}\sup_{s\in[i,i+1]}e^{\lambda\varepsilon e_j^\top(\mathbf{w}_s-\mathbf{w}_i)}\right)\right] \\
&\quad\le\sum_{j=1}^{n}\lambda^{-1}\log\left(\sum_{i=0}^{k-1}\sum_{\varepsilon\in\{-1,1\}}\mathbb{E}\left[\sup_{s\in[i,i+1]}e^{\lambda\varepsilon e_j^\top(\mathbf{w}_s-\mathbf{w}_i)}\right]\right) \qquad (32)
\end{aligned}$$

Now we bound the expectation of each term on the right of (32). For simple notation, let $\alpha=\varepsilon e_j$ correspond to one of the terms in the sum. Note that α is a unit vector and that $e^{\lambda\alpha^\top(\mathbf{w}_s-\mathbf{w}_i)}$ is convex with respect to $\mathbf{w}_s$. Now since, $\mathbf{w}_s$ is martingale, it follows that $e^{\lambda\alpha^\top(\mathbf{w}_s-\mathbf{w}_i)}$ is a submartingale for $s\in[i,i+1]$. So, a Cauchy-Schwarz bound followed by Doob's maximal inequality, and then direct computation gives:

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\in[i,i+1]}e^{\lambda\alpha^\top(\mathbf{w}_s-\mathbf{w}_i)}\right] &\le\sqrt{\mathbb{E}\left[\sup_{s\in[i,i+1]}e^{2\lambda\alpha^\top(\mathbf{w}_s-\mathbf{w}_i)}\right]} \\
&\le 2\sqrt{\mathbb{E}\left[e^{2\lambda\alpha^\top(\mathbf{w}_{i+1}-\mathbf{w}_i)}\right]} \\
&= 2e^{\lambda^2}
\end{aligned}$$

Combining this result with (32) shows that

$$\mathbb{E}\left[\sup_{s\in[0,k]}\|\mathbf{w}_s-\mathbf{w}_{\lfloor s\rfloor}\|\right]\le\frac{n}{\lambda}\log(4ke^{\lambda^2})=\frac{n\log(4k)}{\lambda}+n\lambda$$

for all $\lambda>0$.

Optimizing over λ shows that

$$\mathbb{E}\left[\sup_{s\in[0,k]}\|\mathbf{w}_s-\mathbf{w}_{\lfloor s\rfloor}\|\right]\le 2n\sqrt{\log(4k)}$$

Combining this result with (28) and (31) shows that

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\in[0,k]}\gamma(\mathbf{y}_s^C-\mathbf{y}_{\lfloor s\rfloor}^C|\mathcal{K})\right] &\le\frac{\eta(u+\ell D+n\sigma\sqrt{2\log(2k)})}{2r}+\frac{2n}{r}\sqrt{\frac{2\eta\log(4k)}{\beta}} \\
&\le\sqrt{\eta\log(4k)}\left(\frac{u+\ell D}{2r}+\frac{n\sigma}{\sqrt{2}r}+\frac{2n\sqrt{2}}{r\sqrt{\beta}}\right). \qquad (33)
\end{aligned}$$

The second inequality used the assumption that $\eta\le 1$ so that $\eta\le\sqrt{\eta}$, and the fact that $\log(4k)\ge 1$ for $k\ge 1$.

Now we bound the second term on the right of (26). Note that $\mathbf{x}_t^C$ is a continuous semimartingale and the process $\int_0^t \mathbf{v}_s d\boldsymbol{\mu}(s)$ has bounded variation. Thus, from Itô's formula, Kallenberg (2002), we have that

$$d\|\mathbf{x}_t^C\|^2 = 2(\mathbf{x}_t^C)^\top \left(-\eta \nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor}) + \sqrt{\frac{2\eta}{\beta}} d\mathbf{w}_t - \mathbf{v}_t d\boldsymbol{\mu}(t)\right) + \frac{2\eta n}{\beta} dt \tag{34}$$

Reasoning as in (24) shows that

$$\mathbb{E}\left[\left|(\mathbf{x}_t^C)^\top \nabla f_x(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor})\right|\right] \leq Du + D\mathbb{E}\left[\|\mathbf{g}_t\|\right] \leq Du + 2Dn\sigma.$$

By construction, $(\mathbf{x}_t^C)^\top \mathbf{v}_t = \sup\{x^\top \mathbf{v}_t | x \in \mathcal{K}\} = \delta^\star(\mathbf{v}_t|\mathcal{K})$. Thus, re-arranging, integrating, and taking expectations gives the bound:

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t \delta^\star(\mathbf{v}_s|\mathcal{K}) d\boldsymbol{\mu}(s)\right] \\
&\quad = \frac{\eta n t}{\beta} - \eta \mathbb{E}\left[\int_0^t (\mathbf{x}_s^C)^\top \nabla_x f(\mathbf{x}_s^C, \mathbf{z}_{\lfloor s \rfloor}) ds\right] + \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_0^C\|^2 - \|\mathbf{x}_t^C\|^2\right] \\
&\quad \leq \eta t \left(\frac{n}{\beta} + Du + 2Dn\sigma\right) + \frac{D^2}{2}
\end{aligned} \tag{35}$$

Combining (26), (33), and (35) shows that

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_k^C - \mathbf{x}_k^D\|\right] \leq \sqrt{2} \sqrt{\sqrt{\eta \log(4k)} \left(\frac{u + \ell D}{2r} + \frac{n\sigma}{\sqrt{2}r} + \frac{2n\sqrt{2}}{r\sqrt{\beta}}\right)} \\
\cdot \sqrt{\eta k \left(\frac{n}{\beta} + Du + 2Dn\sigma\right) + \frac{D^2}{2}}
\end{aligned}$$

Combining this result with (21) and (25) finishes the proof. ■

Proof of Lemma 10 Recall that $\mathbf{x}_t^A$ and $\mathbf{x}_t^D$ are discretized processes: $\mathbf{x}_t^A = \mathbf{x}_{\lfloor t \rfloor}^A$ and $\mathbf{x}_t^D = \mathbf{x}_{\lfloor t \rfloor}^D$. Furthermore, if we set $\mathbf{y}_t^A$ as

$$\mathbf{y}_t^A = \mathbf{x}_0^A + \eta \int_0^t \nabla_x f(\mathbf{x}_{\lfloor s \rfloor}^A, \mathbf{z}_{\lfloor s \rfloor}) ds + \sqrt{\frac{2\eta}{\beta}} \mathbf{w}_t,$$

then we have $\mathbf{x}^A = \mathcal{S}(\mathcal{D}(\mathbf{y}^A))$ and $\mathbf{x}^D = \mathcal{S}(\mathcal{D}(\mathbf{y}^C))$, where $\mathcal{D}$ is the discretization operator and $\mathcal{S}$ is the Skorokhod solution operator. In particular

$$\mathbf{x}_{k+1}^A = \Pi_{\mathcal{K}}(\mathbf{x}_k^A + \mathbf{y}_{k+1}^A - \mathbf{y}_k^A).$$

Define a difference process, $\boldsymbol{\rho}_t$, by:

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

Note that for integers k, $\boldsymbol{\rho}_k = \mathbf{x}_k^A - \mathbf{x}_k^D$. While $\boldsymbol{\rho}_t$ can jump at the integers, non-expansiveness of convex projections implies that

$$\|\boldsymbol{\rho}_k\| = \|\mathbf{x}_k^A - \mathbf{x}_k^D\| \leq \lim_{t \uparrow k} \|\boldsymbol{\rho}_t\| \tag{36}$$

Let $k \geq 0$ be an integer. For $t \in [k, k+1)$ we have that

$$d\boldsymbol{\rho}_t = d(\mathbf{y}_t^A - \mathbf{y}_t^C) = \eta(\nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_{\lfloor t \rfloor}))dt$$

It follows that $\boldsymbol{\rho}_t$ is a continuous bounded variation process on the interval $[k, k+1)$. When $\boldsymbol{\rho}_t \neq 0$, we can bound the growth of $\|\boldsymbol{\rho}_t\|$ using the chain rule, followed by the Cauchy-Schwarz inequality, the Lipshitz property of $\nabla_x f$, and the triangle inequality:

$$\begin{aligned} d\|\boldsymbol{\rho}_t\| &= \left(\frac{\boldsymbol{\rho}_t}{\|\boldsymbol{\rho}_t\|}\right)^\top \eta(\nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x f(\mathbf{x}_t^A, \mathbf{z}_{\lfloor t \rfloor}))dt \\ &\leq \eta\ell\|\mathbf{x}_t^C - \mathbf{x}_t^A\|dt \qquad (37) \\ &\leq \eta\ell(\|\mathbf{x}_t^C - \mathbf{x}_t^D\| + \|\mathbf{x}_t^D - \mathbf{x}_t^A\|)dt. \qquad (38) \end{aligned}$$

While we have not characterized the behavior when $\boldsymbol{\rho}_t = 0$, the Lemma 19 from Appendix G can be used to show that this behavior does not cause problems. Specifically for $t \in [k, k+1)$

$$\begin{aligned} \|\boldsymbol{\rho}_t\| &= \|\boldsymbol{\rho}_k\| + \int_k^t d\|\boldsymbol{\rho}_s\| \\ &\overset{Lem.\ 19}{=} \|\boldsymbol{\rho}_k\| + \lim_{\epsilon \downarrow 0} \int_k^t \mathbb{1}(\|\boldsymbol{\rho}_s\| \geq \epsilon) d\|\boldsymbol{\rho}_s\| \\ &\overset{(38)}{\leq} \|\boldsymbol{\rho}_k\| + \lim_{\epsilon \downarrow 0} \int_k^t \mathbb{1}(\|\boldsymbol{\rho}_s\| \geq \epsilon)\eta\ell(\|\mathbf{x}_s^C - \mathbf{x}_s^D\| + \|\mathbf{x}_s^D - \mathbf{x}_s^A\|)ds \\ &\leq (1+\eta\ell)\|\boldsymbol{\rho}_k\| + \eta\ell \int_k^t \|\mathbf{x}_s^C - \mathbf{x}_s^D\|ds \end{aligned}$$

The final inequality used the fact that $\boldsymbol{\rho}_k = \mathbf{x}_s^A - \mathbf{x}_s^D$ for all $s \in [k, k+1)$.

Now using (36) we see that

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

Then using the assumption that $\boldsymbol{\rho}_0 = \mathbf{x}_0^A - \mathbf{x}_0^D = 0$, we have that

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

Taking expectations and using Lemma 9 gives that

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{\rho}_k\|] &\leq \eta\ell \sum_{i=1}^{k-1}(1+\eta\ell)^{k-i-1} \int_i^{i+1} (\eta \log(4\max\{1,s\}))^{1/4} \left(c_9\sqrt{\eta s} + c_{10}\right) ds \\ &\leq \eta\ell \left(\eta \log(4\max\{1,k\})\right)^{1/4} \left(c_9\sqrt{\eta k} + c_{10}\right) \sum_{i=1}^{k-1}(1+\eta\ell)^{k-i-1} \\ &\leq \left(\eta \log(4\max\{1,k\})\right)^{1/4} \left(c_9\sqrt{\eta k} + c_{10}\right) \left((1+\eta\ell)^k - 1\right) \end{aligned}$$

The result now follows because $\mathbf{x}_t^A$ and $\mathbf{x}_t^D$ are constant for $t \in [k, k+1)$ and the bound above is monotinically increasing in k. ■

Proof of Lemma 11 Let $-\int_0^t \mathbf{v}_s^B d\boldsymbol{\mu}^B(s)$ be the unique finite-variation process that enforces that $\mathbf{x}_t^B \in \mathcal{K}$ in the Skorokhod solution. Lemma 2.2 of Tanaka et al. (1979) implies that

$$\|\mathbf{x}_t^B - \mathbf{x}_t^M\|^2 \leq \|\mathbf{y}_t^B - \mathbf{y}_t^M\|^2 + 2\int_0^t \left(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M\right)^\top \left(\mathbf{v}_s^M d\boldsymbol{\mu}^M(s) - \mathbf{v}_s^B d\boldsymbol{\mu}^B(s)\right)$$

Thus, taking square roots and using the triangle inequality gives

$$\begin{aligned}\|\mathbf{x}_t^B - \mathbf{x}_t^M\| \leq \|\mathbf{y}_t^B - \mathbf{y}_t^M\| &+ \sqrt{2\left|\int_0^t \left(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M\right)^\top \mathbf{v}_s^M d\boldsymbol{\mu}^M(s)\right|} \\ &+ \sqrt{2\left|\int_0^t \left(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M\right)^\top \mathbf{v}_s^B d\boldsymbol{\mu}^B(s)\right|} \quad (39)\end{aligned}$$

Now we analyze the various terms of this equation.

First we will bound $\mathbb{E}[\|\mathbf{y}_t^B - \mathbf{y}_t^M\|] \leq \sqrt{\mathbb{E}[\|\mathbf{y}_t^B - \mathbf{y}_t^M\|^2]}$.

Let $\mathcal{F}_\infty$ be the σ-algebra generated by the Brownian motion. In the following discussion, we will assume that the realization of the Brownian motion is fixed and examine the effects of $\mathbf{z}_t$. Note that the initial condition assumption and the definition of $\mathbf{y}_t^B$ from (18) imply that $\mathbf{y}_t^M = \mathbb{E}[\mathbf{y}_t^B|\mathcal{F}_\infty]$. In other words, $\mathbf{y}_t^B - \mathbf{y}_t^M$ is a zero-mean function of the random variables $\mathbf{z}_0, \mathbf{z}_1, \ldots$.

To bound $\mathbb{E}\left[\|\mathbf{y}_t^B - \mathbf{y}_t^M\|^2|\mathcal{F}_\infty\right]$, it suffices to bound the individual coordinates. Each coordinate can be represented as $e_j^\top(\mathbf{y}_t^B - \mathbf{y}_t^M)$, where e_j is a corresponding unit basis vector. Thus, it suffices to bound $\mathbb{E}\left[\left(\alpha^\top(\mathbf{y}_t^B - \mathbf{y}_t^M)\right)^2|\mathcal{F}_\infty\right]$ for an arbitrary unit vector, α.

With the realization of the Brownian motion fixed, $\mathbf{v}_t := \alpha^\top(\mathbf{y}_t^B - \mathbf{y}_t^M)$ can be decomposed as a sum of independent, sub-Gaussian random variables:

$$\begin{aligned}\mathbf{v}_t = \alpha^\top(\mathbf{y}_t^B - \mathbf{y}_t^M) &= \eta\sum_{i=0}^{\lfloor t\rfloor - 1}\int_i^{i+1} \alpha^\top\left(\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)\right) ds \\ &\quad + \eta\int_{\lfloor t\rfloor}^t \alpha^\top\left(\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)\right) ds \\ &=: \sum_{i=0}^{\lfloor t\rfloor} \boldsymbol{\rho}_i\end{aligned}$$

Recall that $\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)$ is sub-Gaussian for all $\mathbf{x}_s^M$. In particular, for $i < \lfloor t \rfloor$, we have for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda \boldsymbol{\rho}_i)|\mathcal{F}_\infty] \\
&\quad = \mathbb{E}\left[\exp\left(\int_i^{i+1} \lambda\eta\alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)\right) ds\right)\middle|\mathcal{F}_\infty\right] \\
&\quad \overset{Jensen+Fubini}{=} \int_i^{i+1} \mathbb{E}\left[\exp\left(\lambda\eta\alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)\right)\right)\middle|\mathcal{F}_\infty\right] ds \\
&\quad \overset{sub-Gaussian}{\leq} \int_i^{i+1} \exp\left(\frac{1}{2}\lambda^2\eta^2\sigma^2\right) ds \\
&\quad = \exp\left(\frac{1}{2}\lambda^2\eta^2\sigma^2\right).
\end{aligned}$$

Now consider the case that $i = \lfloor t \rfloor$. When $t = \lfloor t \rfloor = i$, we have that $\boldsymbol{\rho}_i = 0$ and so $\mathbb{E}[\exp(\lambda\boldsymbol{\rho}_i)|\mathcal{F}_\infty] = 1$. When $t > \lfloor t \rfloor$, a similar argument as above gives:

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda \boldsymbol{\rho}_i)|\mathcal{F}_\infty] \\
&= \mathbb{E}\left[\exp\left(\frac{1}{t-\lfloor t \rfloor}\int_{\lfloor t \rfloor}^t \lambda\eta(t-\lfloor t \rfloor)\alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)\right) ds\right)\middle|\mathcal{F}_\infty\right] \\
&\overset{J+F}{=} \frac{1}{t-\lfloor t \rfloor}\int_{\lfloor t \rfloor}^t \mathbb{E}\left[\exp\left(\lambda\eta(t-\lfloor t \rfloor)\alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_s^M) - \nabla_x f(\mathbf{x}_s^M, \mathbf{z}_i)\right)\right)\middle|\mathcal{F}_\infty\right] ds \\
&\overset{sub-Gaussian}{\leq} \frac{1}{t-\lfloor t \rfloor}\int_{\lfloor t \rfloor}^t \exp\left(\frac{1}{2}\lambda^2\eta^2(t-\lfloor t \rfloor)^2\sigma^2\right) ds \\
&\leq \exp\left(\frac{1}{2}\lambda^2\sigma^2\eta^2(t-\lfloor t \rfloor)\right).
\end{aligned}$$

The final inequality used the fact that $(t-\lfloor t \rfloor)^2 \leq t - \lfloor t \rfloor$.

Now using the fact that $\boldsymbol{\rho}_i$ are independent, conditioned on $\mathcal{F}_\infty$, we have that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \mathbf{v}_t\right)|\mathcal{F}_\infty\right] &= \prod_{i=0}^{\lfloor t \rfloor} \mathbb{E}\left[e^{\lambda\boldsymbol{\rho}_i}|\mathcal{F}_\infty\right] \\
&\leq e^{(\lambda\eta\sigma)^2(t-\lfloor t \rfloor))/2} \prod_{i=0}^{\lfloor t \rfloor - 1} e^{(\lambda\eta\sigma)^2/2} \\
&\leq e^{\lambda^2\eta^2\sigma^2 t/2}.
\end{aligned}$$

Thus we have shown that $\mathbf{v}_t$ is sub-Gaussian with parameter $\hat{\sigma}^2 = \eta^2\sigma^2 t$. Then a standard Chernoff bound argument shows that $\mathbb{P}(|\mathbf{v}_t|^2 > \epsilon|\mathcal{F}_\infty) \leq 2e^{-\epsilon/(2\hat{\sigma}^2)}$. Then we can bound the

variance by:

$$\begin{aligned}\mathbb{E}\left[\left(\alpha^\top(\mathbf{y}_t^B - \mathbf{y}_t^M)\right)^2 |\mathcal{F}_\infty\right] &= \int_0^\infty \mathbb{P}(|\mathbf{v}_t|^2 > \epsilon|\mathcal{F}_\infty)d\epsilon \\ &\leq 2\int_0^\infty e^{-\epsilon/(2\hat{\sigma}^2)}d\epsilon \\ &= 4\hat{\sigma}^2 \\ &= 4\eta^2\sigma^2 t. \end{aligned} \tag{40}$$

Applying (40) to $\alpha = e_j$ for all of the standard basis vectors, then summing and using the tower property gives:

$$\mathbb{E}\left[\|\mathbf{y}_t^B - \mathbf{y}_t^M\|^2\right] \leq 4n\eta^2\sigma^2 t.$$

Taking square roots gives

$$\mathbb{E}\left[\|\mathbf{y}_t^B - \mathbf{y}_t^M\|\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{y}_t^B - \mathbf{y}_t^M\|^2\right]} \leq 2\sigma\eta\sqrt{nt}. \tag{41}$$

Bounding the integral terms from (39) is more complex. First we consider the integral with rerspect to $\boldsymbol{\mu}^M$. The integral with respect to $\boldsymbol{\mu}^B$ is similar. As in the proof of Lemma 9 we will use a Hölder inequality bound, followed by a Cauchy-Schwarz bound:

$$\begin{aligned}&\mathbb{E}\left[\sqrt{\left|\int_0^t \left(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M\right)^\top \mathbf{v}_s^M d\boldsymbol{\mu}^M(s)\right|}\right] \\ &\quad\leq \mathbb{E}\left[\sqrt{\int_0^t \gamma(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M|\mathcal{K})\delta^\star(\mathbf{v}_s^M|\mathcal{K})d\boldsymbol{\mu}^M(s)}\right] \\ &\quad\leq \mathbb{E}\left[\sqrt{\sup_{s\in[0,t]}\gamma(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M|\mathcal{K})}\sqrt{\int_0^t \delta^\star(\mathbf{v}_s^M|\mathcal{K})d\boldsymbol{\mu}^M(s)}\right] \qquad (42) \\ &\quad\leq \sqrt{\mathbb{E}\left[\sup_{s\in[0,t]}\gamma(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M|\mathcal{K})\right]}\sqrt{\mathbb{E}\left[\int_0^t \delta^\star(\mathbf{v}_s^M|\mathcal{K})d\boldsymbol{\mu}^M(s)\right]} \qquad (43)\end{aligned}$$

The integral bound follows from (35) applied to $\bar{f}$ in place of f:

$$\mathbb{E}\left[\int_0^t \delta^\star(\mathbf{v}_s^M|\mathcal{K})d\boldsymbol{\mu}^M(s)\right] \leq \eta t\left(\frac{n}{\beta} + Du + 2Dn\sigma\right) + \frac{D^2}{2} \tag{44}$$

Bounding the supremum will take more work. The eventual plan is to bound the individual components using the Dudley entropy integral. To this end, we first note that for any vector in $x \in \mathbb{R}^n$, we have that

$$\gamma(x|\mathcal{K}) \leq r^{-1}\|x\| \leq r^{-1}\sum_{i=1}^n |x_i|. \tag{45}$$

Thus, it suffices to bound $\mathbb{E}[\sup_{s\in[0,t]} |\mathbf{v}_t - \mathbf{v}_s||\mathcal{F}_\infty]$, where

$$\mathbf{v}_s = \alpha^\top \left(\mathbf{y}_s^B - \mathbf{y}_s^M\right).$$

and α is an arbitrary unit vector.

Also note that

$$\sup_{s\in[0,t]} |\mathbf{v}_t - \mathbf{v}_s| \leq \sup_{s,\hat{s}\in[0,t]} (\mathbf{v}_s - \mathbf{v}_{\hat{s}}).$$

The expectation of the expression on the right will now be bounded via the Dudley entropy integral. To derive the bound, we must show that $\mathbf{v}_s$ has sub-Gaussian increments. A mild extension of the argument that $\mathbf{v}_t$ is sub-Guassian will suffice.

Without loss of generality, assume that $\hat{s} \geq s$. Then

$$\begin{aligned}
\mathbf{v}_{\hat{s}} - \mathbf{v}_s &= \eta \int_s^{\hat{s}} \alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_\tau^M) - \nabla_x \bar{f}(\mathbf{x}_\tau^M, \mathbf{z}_{\lfloor\tau\rfloor})\right) d\tau \\
&= \eta \int_s^{\lceil s\rceil} \alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_\tau^M) - \nabla_x \bar{f}(\mathbf{x}_\tau^M, \mathbf{z}_{\lfloor\tau\rfloor})\right) d\tau + \\
&\quad \eta \sum_{i=\lceil s\rceil}^{\lfloor\hat{s}\rfloor - 1} \int_i^{i+1} \alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_\tau^M) - \nabla_x \bar{f}(\mathbf{x}_\tau^M, \mathbf{z}_{\lfloor\tau\rfloor})\right) d\tau + \\
&= \int_{\lfloor s\rfloor}^{s} \alpha^\top \left(\nabla_x \bar{f}(\mathbf{x}_\tau^M) - \nabla_x \bar{f}(\mathbf{x}_\tau^M, \mathbf{z}_{\lfloor\tau\rfloor})\right) d\tau \\
&=: \sum_{i=\lceil s\rceil - 1}^{\lfloor\hat{s}\rfloor} \boldsymbol{\rho}_i
\end{aligned}$$

Then, similar to the case above, $\boldsymbol{\rho}_i$ are independent sub-Gaussian random variables with the following bounds for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}[e^{\lambda \boldsymbol{\rho}_i}|\mathcal{F}_\infty] \leq \begin{cases} e^{(\eta\sigma\lambda)^2(\lceil s\rceil - s)/2} & \text{if } i = \lceil s\rceil - 1 \\ e^{(\eta\sigma\lambda^2)/2} & \text{if } i = \lceil s\rceil, \ldots, \lfloor\hat{s}\rfloor - 1 \\ e^{(\eta\sigma\lambda)^2(\hat{s} - \lfloor s\rfloor)/2} & \text{if } i = \lfloor\hat{s}\rfloor \end{cases}$$

Then independence implies that for all $\lambda \in \mathbb{R}$, the following bound holds:

$$\begin{aligned}
\mathbb{E}[e^{\lambda(\mathbf{v}_s - \mathbf{v}_{\hat{s}})}|\mathcal{F}_\infty] &\leq \exp\left(\frac{\lambda^2\eta^2\sigma^2}{2}\left((\lfloor s\rfloor - s)^2 + \sum_{i=\lceil s\rceil}^{\lfloor\hat{s}\rfloor - 1} 1 + (\hat{s} - \lfloor\hat{s}\rfloor)^2\right)\right) \\
&\leq \exp\left(\frac{\lambda^2\eta^2\sigma^2|s - \hat{s}|}{2}\right).
\end{aligned}$$

It follows that $\mathbf{v}_s$ is sub-Guassian with respect to the metric defined by $d(s,\hat{s}) = \eta\sigma\sqrt{|s - \hat{s}|}$. See Definition 5.16 of Wainwright (2019).

Let $N([0,t], d, \epsilon)$ be the covering number of the interval $[0,t]$ via closed balls of radius ϵ under the metric d, and similarly $N([0,t], |\cdot|, \epsilon)$ is the corresponding covering number with respect to the

absolute value metric. A standard argument shows that $N([0,t],|\cdot|,\epsilon) = 1$ when $\epsilon \geq t/2$ and when $\epsilon \leq t/2$, we have that

$$N([0,t],|\cdot|,\epsilon) \leq \frac{t}{2\epsilon} + 1 \leq t/\epsilon.$$

See Example 5.2 of Wainwright (2019).

By the definition of d, a ball of radius ϵ in the d metric corresponds to a ball of radius $\left(\frac{\epsilon}{\eta\sigma}\right)^2$ in the absolute value metric. And so, when $\epsilon \geq \eta\sigma\sqrt{\frac{t}{2}}$, we have that $N([0,t],d,\epsilon) = 1$ and when $\epsilon \leq \eta\sigma\sqrt{\frac{t}{2}}$ the following bound holds:

$$N([0,t],d,\epsilon) \leq \frac{t\eta^2\sigma^2}{\epsilon^2}$$

Thus, the Dudley entropy integral bound (see Theorem 5.22 of Wainwright (2019)) implies that

$$\begin{aligned}
&\mathbb{E}[\sup_{s,\hat{s}\in[0,t]} (\mathbf{v}_s - \mathbf{v}_{\hat{s}})|\mathcal{F}_\infty] \\
&\quad\leq 32\int_0^{\eta\sigma\sqrt{\frac{t}{2}}} \sqrt{\log N([0,t],d,\epsilon)}d\epsilon \\
&\quad\leq 32\int_0^{\eta\sigma\sqrt{t}} \sqrt{\log\left(\frac{t\eta^2\sigma^2}{\epsilon^2}\right)}d\epsilon \\
&\quad= 32\eta\sigma\sqrt{2t}\int_0^\infty x^{1/2}e^{-x}dx \quad \left(\text{using } 2x = \log\left(\frac{t\eta^2\sigma^2}{\epsilon^2}\right)\right) \\
&\quad= 16\eta\sigma\sqrt{2\pi t} \qquad (46)
\end{aligned}$$

Thus, applying (46) for all of the standard basis vectors and plugging the bound into (45) and using the tower property gives

$$\mathbb{E}\left[\sup_{s\in[0,t]} \gamma(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M|\mathcal{K})\right] \leq \frac{16n\eta\sigma\sqrt{2\pi t}}{r} \qquad (47)$$

Combining (43), (44), and (47) shows that

$$\mathbb{E}\left[\sqrt{\left|\int_0^t \left(\mathbf{y}_t^B - \mathbf{y}_t^M - \mathbf{y}_s^B + \mathbf{y}_s^M\right)^\top \mathbf{v}_s^M d\boldsymbol{\mu}^M(s)\right|}\right] \leq \sqrt{\frac{16n\eta u\sqrt{2\pi t}}{r}\left(\eta t\left(\frac{n}{\beta} + Du + 2Dn\sigma\right) + \frac{1}{2}D^2.\right)} \qquad (48)$$

An identical argument holds for the integral with respect to $\mathbf{v}_s^B d\boldsymbol{\mu}^M(s)$. So, multiplying the bound from (48) by $2\sqrt{2} = \sqrt{8}$ and adding it to (41) shows that

$$\begin{aligned}
&W_1(\mathcal{L}(\mathbf{x}_t^B), \mathcal{L}(\mathbf{x}_t^M)) \\
&\quad\leq \mathbb{E}\left[\|\mathbf{x}_t^B - \mathbf{x}_t^M\|\right] \\
&\quad\leq 2\sigma\eta\sqrt{nt} + \sqrt{\frac{128n\eta\sigma\sqrt{2\pi t}}{r}\left(\eta t\left(\frac{n}{\beta} + Du + 2Dn\sigma\right) + \frac{1}{2}D^2\right)}.
\end{aligned}$$

The result follows by factoring out the terms depending on η and t. ■

Proof of Lemma 12 Note that

$$d(\mathbf{x}_t^C - \mathbf{x}_t^B) = \eta\left(\nabla_x f(\mathbf{x}_t^M, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor})\right) dt + \mathbf{v}_t^B d\boldsymbol{\mu}^B(t) - \mathbf{v}_t^C d\boldsymbol{\mu}^C(t)$$

so that $\mathbf{x}_t^C - \mathbf{x}_t^B$ is a continuous bounded-variation process. Thus, whenever $\mathbf{x}_t^C \neq \mathbf{x}_t^B$, we have that:

$$\begin{aligned} d\|\mathbf{x}_t^C - \mathbf{x}_t^B\| &= \left(\frac{\mathbf{x}_t^C - \mathbf{x}_t^B}{\|\mathbf{x}_t^C - \mathbf{x}_t^B\|}\right)^\top \eta\left(\nabla_x f(\mathbf{x}_t^M, \mathbf{z}_{\lfloor t \rfloor}) - \nabla_x f(\mathbf{x}_t^C, \mathbf{z}_{\lfloor t \rfloor})\right) dt \\ &\quad + \left(\frac{\mathbf{x}_t^C - \mathbf{x}_t^B}{\|\mathbf{x}_t^C - \mathbf{x}_t^B\|}\right)^\top \left(\mathbf{v}_t^B d\boldsymbol{\mu}^B(t) - \mathbf{v}_t^C d\boldsymbol{\mu}^C(t)\right) \\ &\leq \eta\ell\|\mathbf{x}_t^M - \mathbf{x}_t^C\| dt \\ &\leq \eta\ell(\|\mathbf{x}_t^M - \mathbf{x}_t^B\| + \|\mathbf{x}_t^C - \mathbf{x}_t^B\|) dt. \end{aligned}$$

The first inequality uses the definitions of $\mathbf{v}_t^B$ and $\mathbf{v}_t^C$ to imply that the corresponding terms are non-positive. It also simplifies the inner product with the gradients via the Lipschitz property and the Cauchy-Schwarz inequality. The second inequality uses the triangle inequality.

Now we will use an argument to rule out any expected behavior from the dynamics when $\mathbf{x}_t^C = \mathbf{x}_t^B$. Indeed, using Lemma 19 from Appendix G shows that

$$\begin{aligned} \|\mathbf{x}_t^C - \mathbf{x}_t^B\| &= \int_0^t d\|\mathbf{x}_s^C - \mathbf{x}_s^B\| \\ &= \lim_{\epsilon \downarrow 0} \int_0^t \mathbb{1}(\|\mathbf{x}_s^C - \mathbf{x}_s^B\| \geq \epsilon) d\|\mathbf{x}_s^C - \mathbf{x}_s^B\| \\ &\leq \lim_{\epsilon \downarrow 0} \int_0^t \eta\ell \mathbb{1}(\|\mathbf{x}_s^C - \mathbf{x}_s^B\| \geq \epsilon)(\|\mathbf{x}_s^M - \mathbf{x}_s^B\| + \|\mathbf{x}_s^C - \mathbf{x}_s^B\|) ds \\ &\leq \int_0^t \eta\ell(\|\mathbf{x}_s^M - \mathbf{x}_s^B\| + \|\mathbf{x}_s^C - \mathbf{x}_s^B\|) ds. \end{aligned}$$

Thus Gronwall's inequality implies that

$$\|\mathbf{x}_t^C - \mathbf{x}_t^B\| \leq \eta\ell \int_0^t e^{\eta\ell(t-s)} \|\mathbf{x}_s^M - \mathbf{x}_s^B\| ds$$

Taking expectations and using Lemma 11 gives the desired bound:

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}_t^C - \mathbf{x}_t^B\|] &\leq \eta\ell \int_0^t e^{\eta\ell(t-s)} \left(c_{11}\eta s^{1/2} + c_{12}\eta^{1/2} s^{1/4} + c_{13}\eta s^{3/4}\right) ds \\ &\leq \eta\ell \left(c_{11}\eta t^{1/2} + c_{12}\eta^{1/2} t^{1/4} + c_{13}\eta t^{3/4}\right) \int_0^t e^{\eta\ell(t-s)} ds \\ &= \left(c_{11}\eta t^{1/2} + c_{12}\eta^{1/2} t^{1/4} + c_{13}\eta t^{3/4}\right)(e^{\eta\ell t} - 1). \end{aligned}$$

■

Appendix C. Bounding the Constants

In this expression, we bound the size of the constants based on the problem data. First we get simplified bounds on the constants from Proposition 7. Then we use this result prove Proposition 2, which bounds the overall constants for the algorithm.

Lemma 14 *If $D^2\ell\beta < 8$ and $a = \frac{4}{D^2\beta}$, then*

$$a \geq \frac{\ell}{2} \quad \textit{and} \quad c_8 \leq \frac{2e}{\left(1 - \frac{D^2\ell\beta}{8}\right)^2}.$$

Otherwise, if $a = \frac{D^2\ell^2\beta}{16}\left(1 - \tanh^2\left(\frac{D^2\ell\beta}{8}\right)\right)$, then

$$a \geq \frac{D^2\ell^2\beta}{16}\exp\left(-\frac{D^2\ell\beta}{4}\right) \quad \textit{and} \quad c_8 \leq \frac{4}{D^2\ell\beta}\exp\left(\frac{D^2\ell\beta}{2}\right).$$

Proof For consider the case that $D^2\ell\beta < 8$ and $a = \frac{4}{D^2\beta}$. The lower bound on a is derived by combining the inequality $D^2\ell\beta < 8$ with the expression for a. To derive the upper bound on c_8, first note that

$$D\omega_N\xi = \frac{D^2\ell\beta}{8} < 1$$

and so the numerator is bounded by $e^{D\omega_N\xi} \leq e$.

Now we compute a lower bound on the denominator. By the choice of a, we have that $D\omega_N = 1$ and $\xi = \frac{D^2\ell\beta}{8} < 1$. We use the fact that $\sin\left(\sqrt{1-\xi^2}\right) < \sqrt{1-\xi^2}$ to give

$$\cos\left(\sqrt{1-\xi^2}\right) - \frac{\xi}{\sqrt{1-\xi^2}}\sin\left(\sqrt{1-\xi^2}\right) \geq \cos\left(\sqrt{1-\xi^2}\right) - \xi$$

Then we use the elementary bound

$$\cos(\theta) = \cos(0) - \int_0^\theta \sin(t)dt \geq 1 - \int_0^\theta tdt = 1 - \frac{\theta^2}{2}$$

to give

$$\cos\left(\sqrt{1-\xi^2}\right) - \frac{\xi}{\sqrt{1-\xi^2}}\sin\left(\sqrt{1-\xi^2}\right) \geq 1 - \frac{1}{2}\left(1-\xi^2\right) - \xi = \frac{1}{2}(1-\xi)^2.$$

Combining the bounds for the numerator and the denominator, along with the fact that $\xi = \frac{D^2\ell\beta}{8}$ gives the desired bound on c_8.

Now consider the case that $a = \frac{D^2\ell^2\beta}{16}\left(1 - \tanh^2\left(\frac{D^2\ell\beta}{8}\right)\right)$. For simpler notation, set $x = \frac{D^2\ell\beta}{8}$ so that $a = \frac{\ell}{2}x\left(1 - \tanh^2(x)\right)$. Then the exponential decay factor can be bounded using the fact that:

$$x\left(1 - \tanh^2(x)\right) = \frac{4x}{\left(e^x + e^{-x}\right)^2} \geq xe^{-2x}.$$

Now we bound c_8. Using the definition of x, the numerator is given by e^x.

Now we derive a lower bound on the denominator of c_8. Let $y = D\omega_N\sqrt{\xi^2-1}$. Plugging in the expressions for ξ, a, and x shows that

$$y = D\omega_N\sqrt{\xi^2-1} = x\tanh(x) \quad \text{and} \quad \frac{\sqrt{\xi^2-1}}{\xi} = \tanh(x).$$

Then the denominator of c_8 can be expressed as

$$\cosh(y) - \frac{\sinh(y)}{\tanh(x)} = \frac{\cosh(y)}{\tanh(x)}\left(\tanh(x) - \tanh(y)\right) \geq \frac{\tanh(x)-\tanh(y)}{\tanh(x)}, \tag{49}$$

where the inequality follows from the fact that $\cosh(y) \geq 1$.

Using the fact that $\frac{d}{dx}\tanh(x) = 1 - \tanh^2(x)$ and that $\tanh$ is monotonically increasing gives the bound:

$$\begin{aligned}
\tanh(x) - \tanh(y) &= \int_y^x \left(1 - \tanh^2(z)\right) dz \\
&\geq (x-y)(1-\tanh^2(x)) \\
&= x(1-\tanh(x))^2(1+\tanh(x)) \\
&= \frac{8xe^{-2x}}{\left(e^x + e^{-x}\right)^3}.
\end{aligned} \tag{50}$$

The second-to-last line uses the fact that $y = x\tanh(x)$, while the last line uses that $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Combining (49) and (50) shows that

$$\cosh(y) - \frac{\sinh(y)}{\tanh(x)} \geq \frac{8xe^{-2x}}{\left(e^x + e^{-x}\right)^2\left(e^x - e^{-x}\right)} \geq 2xe^{-5x}.$$

Combining the numerator and denominator bounds shows that

$$c_8 \leq \frac{e^{4x}}{2x}.$$

Plugging in the expression for x gives the result. ■

Proof of Proposition 2 Collecting the constant definitions from the proofs above gives the following relations, in addition to the definitions of c_8 and a:

$$
\begin{aligned}
c_1 &= c_8 D \\
c_2 &= c_6 + c_7 \\
c_6 &= 2^{1/4}\,(c_9 + c_{10})\, e^{\ell} \left(1 + \frac{c_8}{1 - e^{-a/2}}\right) \\
c_7 &= (c_{11} + c_{12} + c_{13}) e^{\ell} \left(1 + \frac{c_8}{1 - e^{-a/2}}\right) \\
c_9 &= \sqrt{2\left(\frac{u + \ell D}{2r} + \frac{n\sigma}{\sqrt{2}r} + \frac{2n\sqrt{2}}{r\sqrt{\beta}}\right)\left(\frac{n}{\beta} + Du + 2Dn\sigma\right)} \\
c_{10} &= \sqrt{2\left(Du + 2n\sigma + \frac{n}{\beta}\right)} + D\sqrt{\frac{u + \ell D}{2r} + \frac{n\sigma}{\sqrt{2}r} + \frac{2n\sqrt{2}}{r\sqrt{\beta}}} \\
c_{11} &= 2\sigma\sqrt{n} \\
c_{12} &= \sqrt{\frac{64 n\sigma D\sqrt{2\pi}}{r}} \\
c_{13} &= \sqrt{\frac{128 n\sigma\sqrt{2\pi}}{r}\left(\frac{n}{\beta} + Du + 2Dn\sigma\right)}
\end{aligned}
$$

Since neither c_8 nor a depend on the state dimension, n, we can see that the constants grow linearly with n.

In the case of $D^2\ell\beta < 8$ and $a = \frac{4}{D^2\beta}$, Lemma 14 implies that $(1 - e^{-a/2})^{-1} \le (1 - e^{-\ell/4})^{-1}$. So, the only way for the constants to become large as β varies is for β^{-1} or $\left(1 - \frac{D^2\ell\beta}{8}\right)^{-1}$ to approach ∞. In particular the terms that can go to ∞ are $\beta^{-1/4}$ and $\left(1 - \frac{D^2\ell\beta}{8}\right)^{-2}$ in this case.

Now consider the case that

$$
\begin{aligned}
a &= \frac{D^2\ell^2\beta}{16}\left(1 - \tanh^2\left(\frac{D^2\ell\beta}{8}\right)\right) \\
c_8 &= \frac{e^{D\omega_N\xi}}{\cosh(D\omega_N\sqrt{\xi^2 - 1}) - \frac{\xi}{\sqrt{\xi^2 - 1}}\sinh(D\omega_N\sqrt{1 - \xi^2})}.
\end{aligned}
$$

The general lower bound on a is taken directly from Lemma 14.

The main term that remains to be bounded is $\frac{1}{1 - e^{-a/2}}$. To perform this bound, we first note that for all $y > 0$,

$$
\frac{1}{1 - e^{-y}} \le \max\left\{\frac{2}{y}, \frac{1}{1 - e^{-1}}\right\}. \tag{51}
$$

Indeed, the left side is monotonically decreasing, and so for all $y \ge 1$, the bound

$$
\frac{1}{1 - e^{-y}} \le \frac{1}{1 - e^{-1}}
$$

holds.

Now, we use the elementary bound that for all $y \geq 0$, $e^{-y} \leq 1 - y + \frac{1}{2}y^2$. This inequality appears in Lattimore and Szepesvári (2019) without proof, but can be proved by showing that $e^y \left(1 - y + \frac{1}{2}y^2\right)$ is monotonically increasing. In particular, when $0 < y \leq 1$, we have that

$$\frac{1}{1 - e^{-y}} \leq \frac{1}{y\left(1 - \frac{1}{2}y\right)} \leq \frac{2}{y}.$$

Combining the bounds for $y \geq 1$ an $0 < y \leq 1$ gives (51).

So, now combining (51) with the results of Lemma 14 shows that

$$\frac{c_8}{1 - e^{-a/2}} \leq \frac{4}{D^2\ell\beta} \exp\left(\frac{D^2\ell\beta}{2}\right) \max\left\{\frac{32}{D^2\ell^2\beta} \exp\left(\frac{D^2\ell\beta}{4}\right), \frac{1}{1 - e^{-1}}\right\}$$

Then combining this above bound with the various expressions for the constants shows that there is a polynomial p such c_1 and c_2 can be bounded by

$$c_i \leq p(\beta^{-1/4}) \exp\left(\frac{3D^2\ell\beta}{4}\right)$$

■

Appendix D. Near-Optimality of Gibbs Distributions

In this appendix, we prove Proposition 5, which states that the algorithm can produce near-optimal samples, provided that β is sufficiently large. This proposition depends on an elementary result on the properties of Gibbs distributions constrained to $\mathcal{K}$, shown next.

Lemma 15 *For any function $g : \mathcal{K} \to \mathbb{R}$, let π_g be the probability measure defined by $\pi_g(A) = \frac{\int_A e^{-g(x)}dx}{\int_{\mathcal{K}} e^{-g(y)}dy}$. In particular, π_0 corresponds to the uniform measure. If g is ℓ-lipschitz, then the KL divergence of π_g from the uniform measure is bounded by:*

$$0 \leq \mathrm{KL}(\pi_g, \pi_0) \leq \min_{x \in \mathcal{K}} g(x) - \mathbb{E}_{\pi_g}[g(\mathbf{x})] + n \log\left(\max\left\{\frac{2}{r}, \frac{(r + \sqrt{r^2 + D^2})\ell}{r \log 2}\right\}\right) + \log(2D^n).$$

Proof The lower-bound on the KL divergence is standard Cover and Thomas (2012); Gray (2011). Now we prove the upper bound.

Say $x^\star$ minimizes $g(x)$ over $\mathcal{K}$. A minimizer exists because g is Lipschitz and $\mathcal{K}$ is compact. Multiplying the numerator and denominator of the definition of π_g by $e^{g(x^\star)}$ gives

$$\pi_g(A) = \frac{\int_A e^{g(x^\star) - g(x)}dx}{\int_{\mathcal{K}} e^{g(x^\star) - g(y)}dy}.$$

Note that $\pi_0(dx) = \frac{dx}{\mathrm{vol}(\mathcal{K})}$. Thus, the definition of KL divergence gives:

$$\mathrm{KL}(\pi_g, \pi_0) = \mathbb{E}_{\pi_g}[g(x^\star) - g(\mathbf{x})] + \log\left(\mathrm{vol}(\mathcal{K})\right) - \log\left(\int_{\mathcal{K}} e^{g(x^\star) - g(x)}dx\right)$$

Note that $\mathcal{K}$ is contained in a ball of radius D, so that $\mathrm{vol}(\mathcal{K}) \leq D^n \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$.

So, the desired upper bound is obtained by providing a lower bound on $\int_{\mathcal{K}} e^{g(x^\star)-g(x)}dx$. Note that

$$0 \geq g(x^\star) - g(x) \geq -\ell\|x - x^\star\|$$

Also, note that $e^{-\ell\|x-x^\star\|} \geq \frac{1}{2}$ if and only if $\|x - x^\star\| \leq \frac{\log 2}{\ell}$.

Set $\epsilon = \frac{\log 2}{\ell}$ and let $\mathcal{B}_{x^\star}(\epsilon)$ be the ball of radius ϵ centered at $x^\star$. Then for any $\mathcal{S} \subset \mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$ we have that

$$\int_{\mathcal{K}} e^{g(x^\star)-g(x)}dx \geq \frac{1}{2}\mathrm{vol}\left(\mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)\right) \geq \frac{1}{2}\mathrm{vol}(\mathcal{S})$$

We will show that $\mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$ always contains a ball of radius $\min\{\frac{r}{2}, \frac{r\epsilon}{r+\sqrt{r^2+D^2}}\}$. The lemma then follows by using the fact that a ball of radius ρ has volume given by $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}\rho^n$. Note that the constant factors of $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$ cancel in the bound.

To find the desired ball, we consider three cases: 1) $0 \notin \mathcal{B}_{x^\star}(\epsilon)$, 2) $0 \in \mathcal{B}_{x^\star}(\epsilon)$ and $\epsilon \leq r$, and 3) $0 \in \mathcal{B}_{x^\star}(\epsilon)$ and $\epsilon > r$.

When $0 \notin \mathcal{B}_{x^\star}(\epsilon)$, we construct the desired ball from the geometry of Fig. 1. Without loss of generality, we can assume that $x^\star = -\|x^\star\|e_1$, where e_1 is the first standard unit vector. Also, since $0 \notin \mathcal{B}_{x^\star}(\epsilon)$, we must have that $\|x^\star\| > 0$. Consider the convex set defined by:

$$\|x - x^\star\| \leq \epsilon \tag{52a}$$

$$-\|x^\star\| \leq x_1 \leq 0 \tag{52b}$$

$$\sqrt{\sum_{i=2}^{n} x_i^2} \leq r + \frac{r}{\|x^\star\|}x_1 \tag{52c}$$

The set defined by (52) is a subset of $\mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$. The angle between the $x^\star$ and the conic constraint boundary, from (52c), is given by $\theta = \tan^{-1}\frac{r}{\|x^\star\|}$. For any $d > 0$, the largest ball centered at $(-\|x^\star\| + d)e_1$ which fits into the conic set from (52c) has radius $d\sin\theta$. The largest such ball that is also contained in $\mathcal{B}_{x^\star}(\epsilon)$ is found by setting $d + d\sin\theta = \epsilon$. Plugging the definitions of d and θ shows that the corresponding ball has radius ρ, which satifies

$$\frac{r\epsilon}{r+\sqrt{r^2+D^2}} \leq \rho = \frac{r\epsilon}{r+\sqrt{r+\|x^\star\|^2}} < \frac{\epsilon}{2}.$$

Now consider the case that $0 \in \mathcal{B}_{x^\star}(\epsilon)$ and $\epsilon \leq r$. It follows that $x^\star \in \mathcal{B}_0(r)$. Then applications of the triangle inequality show that $\mathcal{B}_{x^\star/2}(\epsilon/2) \subset \mathcal{B}_0(r) \cap \mathcal{B}_{x^\star}(\epsilon) \subset \mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$. Thus, a ball of radius $\epsilon/2 > \frac{r\epsilon}{r+\sqrt{r+D^2}}$ has been constructed in $\mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$.

Finally, consider the case that $0 \in \mathcal{B}_{x^\star}(\epsilon)$ and $\epsilon > r$. If $\|x^\star\| \geq r/2$, then $\mathcal{B}_{\frac{rx^\star}{2\|x^\star\|}}(r/2) \subset \mathcal{B}_0(r) \cap \mathcal{B}_{x^\star}(\epsilon) \subset \mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$. Otherwise, if $\|x^\star\| < r/2$, then $\mathcal{B}_0(r/2) \subset \mathcal{B}_0(r) \cap \mathcal{B}_{x^\star}(\epsilon) \subset \mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$. In either case, a ball of radius $r/2$ has been constructed in $\mathcal{K} \cap \mathcal{B}_{x^\star}(\epsilon)$. ■

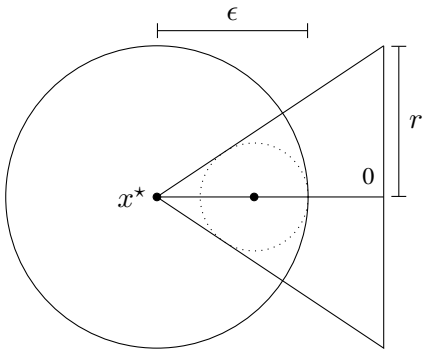

Figure 1: **Case of $0 \notin \mathcal{B}_{x^\star}(\epsilon)$.** Using elementary trigonometry the largest inscribed circle can be calculated.

Proof of Proposition 5 Recall that $\bar{f}$ is u-Lipschitz, so that $\beta\bar{f}$ is βu-Lipschitz. Assume that $\tilde{\mathbf{x}}$ is drawn according to $\pi_{\beta\bar{f}}$. So, applying Lemma 15 to $\beta\bar{f}$ and dividing by β implies that

$$\mathbb{E}[\bar{f}(\tilde{\mathbf{x}})] \leq \min_{x\in\mathcal{K}} \bar{f}(x) + \frac{n}{\beta}\log\left(2D\max\left\{\frac{2}{r}, \frac{(r+\sqrt{r^2+D^2})u\beta}{r\log 2}\right\}\right). \tag{53}$$

Let $\mathbf{x}_k$ be the k-th iterate of the algorithm.

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_k)] \\
&\overset{\text{Kantorovich Duality}}{\leq} \mathbb{E}_{\pi_{\beta\bar{f}}}[\bar{f}(\mathbf{x})] + uW_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \\
&\overset{(53)}{\leq} \min_{x\in\mathcal{K}} \bar{f}(x) + uW_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) + \frac{n\log(c_5\max\{1,\beta\})}{\beta},
\end{aligned}$$

where $c_5 = 2D\max\left\{\frac{2}{r}, \frac{(r+\sqrt{r^2+D^2})u}{r\log 2}\right\}$.

Now we will show how to tune the parameters to achieve an average suboptimality of ϵ.

First, we choose β so that $\frac{n\log(c_5\max\{1,\beta\})}{\beta} \leq \frac{\epsilon}{2}$. Without loss of generality, assume that $\beta \geq 1$. Set $x = \log(c_5\beta)$, so that $\beta = c_5^{-1}e^x$ and the required bound becomes:

$$xe^{-x} \leq \frac{c_5\epsilon}{2n}$$

Fix any $\lambda \in (0,1)$. Then the maximum value of $xe^{-(1-\lambda)x}$ occurs at $x = (1-\lambda)^{-1}$, so that for all $x \in \mathbb{R}$:

$$xe^{-x} \leq \frac{1}{(1-\lambda)e}e^{-\lambda x}. \tag{54}$$

So, it suffices to set $e^{-\lambda x} \leq \frac{c_5\epsilon(1-\lambda)e}{2n}$. Plugging in the definition of x and re-arranging shows that a sufficient condition for $\frac{n\log(c_5\beta)}{\beta} \leq \frac{\epsilon}{2}$ is given by:

$$\beta \geq c_5^{-1}\left(\frac{2n}{c_5(1-\lambda)\epsilon e}\right)^{1/\lambda}. \tag{55}$$

Now, for fixed $\beta \geq 1$, the bounds from Theorem 1 and Proposition 2 to give that

$$\begin{aligned} W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \left(c_1 + \frac{c_2}{(4a)^{1/4}} \right) T^{-1/4} (\log T)^{1/2} \\ \leq p(1) e^{\frac{3D^2\ell\beta}{4}} \left(1 + \frac{e^{\frac{D^2\ell\beta}{16}}}{4^{1/4} c_3} \right) T^{-1/4} (\log T)^{1/2} \\ \leq p(1) \left(1 + \frac{1}{4^{1/4} c_3} \right) \exp\left(\frac{13 D^2 \ell \beta}{16} \right) T^{-1/4} (\log T)^{1/2} \end{aligned}$$

Similar to the derivation of (54) we have for all $\delta \in (0, 1/2)$ and all $T > 0$:

$$T^{-1/4} (\log T)^{1/2} \leq \sqrt{\frac{T^{-\frac{1}{2}+\delta}}{e\delta}}$$

Thus, to have $uW_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \leq \frac{\epsilon}{2}$, it suffices to have

$$T^{-\frac{1}{2}+\delta} \leq e\delta \left(\frac{\epsilon}{2} \right)^2 \left(p(1) \left(1 + \frac{1}{4^{1/4} c_3} \right) \exp\left(\frac{13 D^2 \ell \beta}{16} \right) \right)^{-2} =: \hat{\epsilon},$$

which occurs whenever

$$T \geq \frac{1}{\hat{\epsilon}^{\frac{2}{1-2\delta}}}$$

In particular, there is a constant, c_{14}, independent of η, β, ϵ, λ, and δ, such that the bound above holds whenever

$$T \geq \frac{c_{14}^{\frac{2}{1-2\delta}}}{\epsilon^{\frac{4}{1-2\delta}}} \exp\left(\frac{13 D^2 \ell \beta}{4(1-2\delta)} \right). \tag{56}$$

The result now follows by combining (55) and (56), noting that $\frac{4}{1-2\delta}$ can take any value $\rho > 4$ and $1/\lambda$ can take any value $\zeta > 1$. ■

Appendix E. The Skorokhod Problem

This appendix describes basic results on the Skorokhod problem, which is used to construct solutions to reflected SDEs. First we describe some existing theory. Then we present limiting argument that is used to translate results on compact convex sets with smooth boundaries general compact convex sets.

E.1. Background

A classical construction for constraining stochastic processes to remain in a set is based on the Skorokhod problem, which we describe below. This will be useful, in particular, for analyzing projected gradient algorithms in continuous time.

Let $\mathcal{K}$ be a convex subset of $\mathbb{R}^n$ with non-empty interior. Let $w : [0, \infty) \to \mathbb{R}^n$ be a piecewise-continuous function with $w_0 \in \mathcal{K}$. 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 w_t if the following conditions hold:

- $x_t = w_t + \phi_t \in \mathcal{K}$ 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 $T > 0$.

For each w, the corresponding functions x_t and ϕ_t exist and are unique Tanaka et al. (1979). Note that if $x_t \in \text{int}(\mathcal{K})$, then $N_{\mathcal{K}}(x_t) = \{0\}$, and so $v_t = 0$. Thus, without loss of generality, we can assume that μ is supported entirely on the times in which $x_t \in \partial\mathcal{K}$. In many cases, we are primarily interested in x_t and so we will often refer to x_t as the solution of the Skorokhod problem corresponding to w_t. By existence and uniqueness, we can view the Skorokhod problem solution as a mapping: $x = \mathcal{S}(w)$.

The connection between Skorokhod problems and projection algorithms becomes more concrete when w_t is piecewise constant. Specifically, assume that $0 = t_0 < t_1 < \cdots < t_{M-1} \leq T$ are the jump points of w_t, and let $S_k = [t_k, t_{k+1})$ for $k < M - 1$ and $S_{M-1} = [t_{M-1}, T]$. Then w_t can be represented as

$$w_t = \sum_{k=0}^{M-1} w_{t_k} \mathbb{1}_{S_k}(t).$$

Then the solution of the Skorokhod problem has the form

$$x_t = \sum_{k=0}^{M-1} x_{t_k} \mathbb{1}_{S_k}(t), \quad \phi_t = -\int_0^t \sum_{k=0}^{M-1} v_{t_k} d_{k+1} \delta(s - t_k) ds,$$

where $x_0 = w_0$, $v_0 = 0$, and

$$\begin{aligned} x_{t_{k+1}} &= \Pi_{\mathcal{K}}(x_{t_k} + w_{t_{k+1}} - w_{t_k}) \\ d_{k+1} &= \|(x_{t_k} + w_{t_{k+1}} - w_{t_k}) - x_{t_{k+1}}\| \\ v_{t_{k+1}} &= \begin{cases} 0 & x_{t_k} + w_{t_{k+1}} - w_{t_k} \in \mathcal{K} \\ \frac{(x_{t_k} + w_{t_{k+1}} - w_{t_k}) - x_{t_{k+1}}}{d_{k+1}} & x_{t_k} + w_{t_{k+1}} - w_{t_k} \notin \mathcal{K}. \end{cases} \end{aligned}$$

In Tanaka et al. (1979), a construction for the Skorokhod solution for a continuous trajectory, w, proceeds as follows. The continuous trajectory is approximated by piecewise constant trajectories of the form $w_{\lfloor ti \rfloor / i}$ for positive integers i. Then the Skorokhod problems are solved for these discretized trajectories and shown to converge to a unique solution for the original Skorokhod problem for w.

The existence of a solution to the Skorokhod problem for arbitrary continuous trajectories can be used to construct unique solutions to reflected stochastic differential equations. In particular, the integrated form of a reflected SDE can be expressed as:

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t f(s, \mathbf{x}_s) ds + \int_0^t \sigma(s, \mathbf{x}_s) d\mathbf{w}_s - \int_0^t \mathbf{v}_s d\boldsymbol{\mu}(s), \tag{57}$$

where $\mathbf{x}_0 \in \mathcal{K}$, $-\int_0^t \mathbf{v}_s d\boldsymbol{\mu}(s)$ is the reflection process that ensures that $\mathbf{x}(t) \in \mathcal{K}$ for all $t \geq 0$. Note that $\mathbf{x}$ is the Skorokhod solution to the process:

$$\mathbf{y}_t = \mathbf{x}_0 + \int_0^t f(\mathbf{x}_s) ds + \int_0^t \sigma(\mathbf{x}_s) d\mathbf{w}_s.$$

A construction for $\mathbf{x}_t$ based on Picard iteration was given in Tanaka et al. (1979). The paper Słomiński (2001), examines the Euler scheme defined by: $\bar{\mathbf{x}}_0^m = \mathbf{x}_0$ and for integers $k \geq 0$:

$$\bar{\mathbf{x}}^m_{(k+1)/m} = \Pi_{\mathcal{K}}\left(\bar{\mathbf{x}}^m_{k/m} + \frac{1}{m} f(\bar{\mathbf{x}}^m_{k/m}) + \sigma(\bar{\mathbf{x}}^m_{k/m})(\mathbf{w}_{(k+1)/m} - \mathbf{w}_{k/m})\right). \tag{58}$$

Then for $t \in [k/m, (k+1)/m)$, we set $\bar{\mathbf{x}}^m_t = \bar{\mathbf{x}}^m_{k/m}$. Corollary 3.3 of Słomiński (2001) shows that $\bar{\mathbf{x}}^m$ converges uniformly to $\mathbf{x}$ on compact subsets of $[0, \infty)$.

E.2. Approximating the Domain

In Appendix F we show that the distribution $\pi_{\beta \tilde{f}}$ from (3) is invariant for the process $\mathbf{x}^M$. However, many of the arguments are easier when $\mathcal{K}$ has a smooth boundary. To handle the general case, we examine Skorokhod solutions on smooth approximations of $\mathcal{K}$ and then use a limiting argument. Here we build the approximation results needed for this argument. The basic idea for this approximation is discussed in Section 2 of Lions and Sznitman (1984), but not proved explicitly.

Let $\mathcal{K}_1 \subset \mathcal{K}_2 \subset \cdots \subset \mathcal{K}$ be an increasing family of convex compact sets such that if S is any compact subset of $\text{int}(\mathcal{K})$, then $S \subset \mathcal{K}_i$ for all sufficiently large i.

The approximation results are proved using the following fact about projections.

Lemma 16 *The functions $\Pi_{\mathcal{K}_i}$ converge uniformly to $\Pi_{\mathcal{K}}$ on compact subsets of $\mathbb{R}^n$.*

Proof We will show that for any $\epsilon > 0$, there is an i such that $\|\Pi_{\mathcal{K}_i}(x) - \Pi_{\mathcal{K}}(x)\| \leq \epsilon$ for all $x \in \mathbb{R}^n$ with $\|x\| \leq R$. By the monotonicity of $\mathcal{K}_i$, the result then holds for all $j \geq i$. Note that if $x \in \mathcal{K}_i$, then $\Pi_{\mathcal{K}_i}(x) = \Pi_{\mathcal{K}}(x) = x$, so we only need to analyze the case that $x \notin \mathcal{K}_i$.

Let $\delta > 0$ and let S be a compact subset of $\text{int}(\mathcal{K})$ such that $\text{dist}(x, S) \leq \delta$ for all $x \in \mathcal{K}$. Here $\text{dist}(x, S)$ is the distance function. Such an S can be chosen as $S = \lambda\mathcal{K}$ for $\lambda \in (0, 1)$ sufficiently close to 1. Take $\mathcal{K}_i$ such that $S \subset \mathcal{K}_i$. This implies, in particular that for any $x \in \mathcal{K}$, that $\text{dist}(x, \mathcal{K}_i) \leq \delta$.

Consider any $x \notin \mathcal{K}_i$. By the distance assumption, there is a point $y \in \mathcal{K}_i$ such that $\|y - \Pi_{\mathcal{K}}(x)\| \leq \delta$. The generalized Pythagorean inequality applied to $\|\cdot\|^2$ and $\mathcal{K}_i$ implies that

$$\|y - x\|^2 \geq \|y - \Pi_{\mathcal{K}_i}(x)\|^2 + \|x - \Pi_{\mathcal{K}_i}(x)\|^2. \tag{59}$$

(See Herbster and Warmuth (2001) for more on the generalized Pythagorean inequality.)

Note that $\|x - \Pi_{\mathcal{K}_i}(x)\| \geq \|x - \Pi_{\mathcal{K}}(x)\|$, since $\Pi_{\mathcal{K}_i}(x) \in \mathcal{K}$. Furthermore, the triangle inequality, followed by the distance assumption on y imply that:

$$\|y - x\| \leq \|y - \Pi_{\mathcal{K}}(x)\| + \|x - \Pi_{\mathcal{K}}(x)\| \leq \delta + \|x - \Pi_{\mathcal{K}}(x)\|$$

Plugging the upper and lower bounds into (59) shows that

$$\delta^2 + 2\delta\|x - \Pi_{\mathcal{K}}(x)\| + \|x - \Pi_{\mathcal{K}}(x)\|^2 \geq \|y - \Pi_{\mathcal{K}_i}(x)\|^2 + \|x - \Pi_{\mathcal{K}}(x)\|^2.$$

Using the fact that $0 \in \mathcal{K}$ and $\|x\| \leq R$ shows that $\|x - \Pi_{\mathcal{K}}(x)\| \leq R$. So, rearranging the inequality above and plugging in this bound shows that:

$$\|y - \Pi_{\mathcal{K}_i}(x)\| \leq \sqrt{\delta^2 + 2\delta R}.$$

Also, note that by the triangle inequality, the assumption on y, and the inequality above:

$$\|\Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}_i}(x)\| \leq \|\Pi_{\mathcal{K}}(x) - y\| + \|y - \Pi_{\mathcal{K}_i}(x)\| \leq \delta + \sqrt{\delta^2 + 2\delta R}.$$

So the result holds by choosing δ such that $\delta + \sqrt{\delta^2 + 2\delta R} \leq \epsilon$. ■

Lemma 17 *Let* $\mathbf{x}$ *and* $\mathbf{x}_i$ *be solutions of the reflected SDE from (57) over domains* $\mathcal{K}$ *and* $\mathcal{K}_i$ *respectively, with* $f(x)$ *and* $\sigma(x)$ *both Lipschitz in* x*. For almost all realizations of the Brownian motion,* $\mathbf{x}_i$ *converges uniformly to* $\mathbf{x}$ *on compact subsets of* $[0, \infty)$*.*

Proof In the proof we denote the trajectories like $\mathbf{x}(t)$ and $\mathbf{x}_i(t)$ to reduce the complexity of the subscripts and superscripts.

Consider the Euler approximation from (58). For almost all realizations of the Brownian motion, the resulting solution converges uniformly on compacts to $\mathbf{x}$. Similarly, for each $\mathcal{K}_i$, the corresponding Euler scheme converges uniformly on compacts to $\mathbf{x}_i$ for almost all Brownian motion realizations. Now since the intersection of a countable collection of almost sure events is again an almost sure event, we have for almost all Brownian motion realizations, all of the corresponding Euler schemes converge uniformly on compacts.

Let $\mathbf{w}$ be a realization for which all of the corresponding Euler schemes converge uniformly on compacts. Fix $T > 0$ and let $\bar{\mathbf{x}}_i^m$ be the Euler approximations of $\mathbf{x}_i$. Corollary 3.3 of Słomiński (2001) gives a convergence rate for the Euler scheme which implies that there is a constant $c > 0$ such that

$$\forall i, \sup_{t\in[0,T]} \|\bar{\mathbf{x}}_i^m(t) - \mathbf{x}_i(t)\| \leq cm^{-1/5} \quad \text{and} \quad \sup_{t\in[0,T]} \|\bar{\mathbf{x}}^m(t) - \mathbf{x}(t)\| \leq cm^{-1/5}.$$

So fix $\epsilon > 0$ and choose m sufficiently large so that

$$\forall i, \sup_{t\in[0,T]} \|\bar{\mathbf{x}}_i^m(t) - \mathbf{x}_i(t)\| \leq \epsilon \quad \text{and} \quad \sup_{t\in[0,T]} \|\bar{\mathbf{x}}^m(t) - \mathbf{x}(t)\| \leq \epsilon.$$

Now we will show that i can be chosen so that $\sup_{t\in[0,T]} \|\bar{\mathbf{x}}_i^m(t) - \mathbf{x}^m(t)\| \leq \epsilon$. If we can show this, the result will follow by the triangle inequality.

Let $d = \sup_{s,t\in[0,T]} \|\mathbf{w}_s - \mathbf{w}_t\|$. Since f and σ are continuous, they are bounded on $\mathcal{K}$. It follows that all of the arguments of the projection used in the Euler schemes are bounded in norm by:

$$D + \sup_{x\in\mathcal{K}} \|f(x)\| + \sup_{x\in\mathcal{K}} \|\sigma(x)\|_2 d.$$

Here $\|\cdot\|_2$ is the matrix 2-norm.

Thus, for any fixed m, Lemma 16 implies that all of the projections converge uniformly as $i \to \infty$, which in turn implies that $\bar{\mathbf{x}}_i^m$ converges to $\bar{\mathbf{x}}^m$ uniformly on $[0, T]$. In particular, we can choose m such that $\sup_{t\in[0,T]} \|\mathbf{x}(t) - \bar{\mathbf{x}}^m(t)\| \leq \epsilon$, and the proof is complete. ■

Appendix F. Invariance of the Gibbs Distribution

Here we prove a basic result that $\pi_{\beta\bar{f}}$ is invariant for $\mathbf{x}^M$. Our proof extends the methodology from Lemma 2.1 of Harrison and Williams (1987), which examines the case that $\bar{f}$ is affine and the boundary is smooth. Bubeck et al. (2018) gives a brief outline of the analysis when $\bar{f}$ is convex and $\mathcal{K}$ is a general compact convex set. The basic idea follows through in the more general case in which $\bar{f}$ is only assumed to be differentiable.

Lemma 18 *The measure $\pi_{\beta\bar{f}}$ is a stationary distribution for (5).*

Proof We first assume that $\mathcal{K}$ has a smooth boundary. Later, we will use a limiting argument to show that the result still holds for general compact convex $\mathcal{K}$.

The generator associated with $\mathbf{x}^M$ on the interior of $\mathcal{K}$ is given by:

$$Lg(x) = -\eta\nabla\bar{f}(x)^\top\nabla g(x) + \frac{\eta}{\beta}(\Delta g)(x),$$

where Δ is the Laplacian operator. (In this proof we will drop the subscript of x from the gradient operators, since $\mathbf{z}_k$ does not influence $\mathbf{x}^M$.)

Define the diffusion operator P_t by:

$$(P_t g)(x) = \mathbb{E}[g(\mathbf{x}_t^M)|\mathbf{x}_0 = x].$$

To show invariance, it suffices to show that for all $g \in L_2(\pi_{\beta\bar{f}})$ and all $t > 0$ the follow holds:

$$\int_{\mathcal{K}} g(x)d\pi_{\beta\bar{f}}(x) = \int_{\mathcal{K}} (P_t g)(x)d\pi_{\beta\bar{f}}(x) \tag{60}$$

Now, since the set of differentiable functions is dense in $L_2(\pi_{\beta\bar{f}})$, we can assume without loss of generality that g is differentiable. In the case that g is differentiable, Theorem 6.31 of Gilbarg and Trudinger (1998) shows that there is a unique twice-differentiable h such that

$$(Lh)(x) - \lambda h(x) = -g(x) \quad \forall x \in \text{int}(\mathcal{K}) \tag{61a}$$

$$\nabla h(x)^\top v = 0 \quad \forall x \in \partial\mathcal{K} \text{ and } \forall v \in N_{\mathcal{K}}(x). \tag{61b}$$

Note that since $\partial\mathcal{K}$ is smooth, $N_{\mathcal{K}}(x)$ is a half-line for all $x \in \partial\mathcal{K}$.

Let

$$\mathbf{q}_t = e^{-\lambda t}h(\mathbf{x}_t^M) + \int_0^t e^{-\lambda s}g(\mathbf{x}_s^M)ds.$$

Then Itô's formula combined with (61b) followed by (61a) gives:

$$\begin{aligned}
&d\mathbf{q}_t \\
&= e^{-\lambda t}(-\lambda h(\mathbf{x}_t^M) - \eta\nabla\bar{f}(\mathbf{x}_t^M)^\top\nabla h(\mathbf{x}_t^M) + \frac{\eta}{\beta}\Delta h(\mathbf{x}_t^M) + g(\mathbf{x}_t^M))dt + \sqrt{\frac{2\eta}{\beta}}\nabla g(\mathbf{x}_t^M)^\top d\mathbf{w}_t \\
&= \sqrt{\frac{2\eta}{\beta}}\nabla g(\mathbf{x}_t^M)^\top d\mathbf{w}_t.
\end{aligned}$$

So, in particular, $\mathbf{q}_t$ is a martingale.

If $x = \mathbf{x}_0^M$, then

$$\begin{aligned} h(x) &= \mathbf{q}_0 \\ &= \lim_{t\to\infty} \mathbb{E}[\mathbf{q}_t | \mathbf{x}_0 = x] \\ &= \mathbb{E}\left[\int_0^\infty e^{-\lambda t} g(\mathbf{x}_t^M) dt \middle| \mathbf{x}_0^M = x\right] \\ &= \int_0^\infty e^{-\lambda t} (P_t g)(x) dt. \end{aligned} \tag{62}$$

The last equality follows from Fubini's theorem, which is justified by the fact that g is differentiable, and so the integrand is bounded on $\mathcal{K}$.

Taking the Laplace transform of both sides of (60) gives the condition:

$$\lambda^{-1} \int_{\mathcal{K}} g(x) d\pi_{\beta\bar{f}}(x) = \int_{\mathcal{K}} \int_0^\infty e^{-\lambda t} (P_t g)(x) dt d\pi_{\beta\bar{f}}(x). \tag{63}$$

Note that Fubini's theorem was again used to switch the order of integrals on the right. Now, uniqueness of Laplace transforms implies that (60) holds for all $t > 0$ if and only if (63) holds for all $\lambda > 0$.

Using (61a), the left side of (63) becomes:

$$\lambda^{-1} \int_{\mathcal{K}} g(x) d\pi_{\beta\bar{f}}(x) = \lambda^{-1} \int_{\mathcal{K}} (\lambda h(x) - Lh(x)) d\pi_{\beta\bar{f}}(x)$$

Using (62), the right side of (63) becomes:

$$\int_{\mathcal{K}} \int_0^\infty e^{-\lambda t} (P_t g)(x) dt d\pi_{\beta\bar{f}}(x) = \int_{\mathcal{K}} h(x) d\pi_{\beta\bar{f}}(x).$$

Thus, we see that (63) holds if and only if

$$\int_{\mathcal{K}} Lh(x) d\pi_{\beta\bar{f}}(x) = 0. \tag{64}$$

Using the specific form of L and $\pi_{\beta\bar{f}}$, we see that (64) holds if and only if

$$\int_{\mathcal{K}} \left(-\eta \nabla \bar{f}(x)^\top \nabla h(x) + \frac{\eta}{\beta} \Delta h(x)\right) e^{-\beta \bar{f}(x)} dx = 0 \tag{65}$$

We will prove (65) via Stokes theorem, which states that $\int_{\mathcal{K}} d\omega(x) = \int_{\partial\mathcal{K}} \omega(x)$ for a differential $n-1$ form. Consider the $(n-1)$-form defined by:

$$\omega(x) = \sum_{i=1}^n (-1)^{i+1} \left(\frac{\partial h(x)}{\partial x_i} e^{-\beta \bar{f}(x)}\right) \bigwedge_{j \neq i} dx_j$$

Here the wedge product follows the standard ordering over the integers.

By construction, we have

$$d\omega(x) = \left(-\eta \nabla \bar{f}(x)^\top \nabla h(x) + \frac{\eta}{\beta} \Delta h(x)\right) e^{-\beta \bar{f}(x)} dx_1 \wedge \cdots \wedge dx_n.$$

Thus, Stokes theorem implies (65) if and only if

$$\int_{\partial\mathcal{K}} \omega(x) = 0$$

To evaluate this integral, we follow a typical construction from the integration of differential forms from Lee (2013). Choose a finite open cover of $\mathcal{K}$ in the subspace topolgy, $U^1, \ldots, U^m$, with corresponding smooth charts, $\phi^1, \ldots, \phi^m$, and a corresponding partition of unity $\psi^1, \ldots, \psi^m$. The partition of unity has the property that ψ^i is supported in U^i. Without loss of generality, we assume that ϕ^i preserve the orientation of $\mathcal{K}$. Furthermore, since $\partial\mathcal{K}$ is smooth, the neighborhoods and charts can be chosen such that if $U^i \cap \partial\mathcal{K} \neq \emptyset$, then ϕ^i maps U^i to a half-space, $\mathbb{H}$:

$$\begin{aligned}\phi^i(U^i) &\subset \{y \in \mathbb{R}^n | y_n \geq 0\} \\ \phi^i(U^i \cap \partial\mathcal{K}) &\subset \{y \in \mathbb{R}^n | y_n = 0\}.\end{aligned}$$

As in Lee (2013), the desired integral can be evaluated as

$$\begin{aligned}\int_{\partial\mathcal{K}} \omega(x) &= \sum_{i=1}^{M} \int_{\partial\mathcal{K}} \psi^i(x)\omega(x) \\ &= \sum_{i=1}^{M} \int_{\partial\mathbb{H}^n} (((\phi^i)^{-1})^\star(\psi^i\omega))(y),\end{aligned}$$

where $((\phi^i)^{-1})^\star$ denotes the pullback operation.

To evaluate $\int_{\mathcal{K}} \omega(x)$, it suffices to evaluate this integral over elements of the cover that intersect $\partial\mathcal{K}$. We will show that each of these elements integrates to zero. To this end, let U be a set in the cover with $U \cap \partial\mathcal{K}$ with associated chart ϕ and partition of unity element ψ. For compact notation, set

$$\alpha_i(x) = \psi(x)\frac{\partial h(x)}{\partial x_i} e^{-\beta \bar{f}(x)}$$

so that

$$\psi(x)\omega(x) = \sum_{i=1}^{n} (-1)^{i+1} \alpha_i(x) \bigwedge_{j \neq i} dx_j.$$

Let $J(y)$ be the Jacobian matrix of $\phi^{-1}(y)$ and let $M_{ij}(y)$ be the associated minors. Then the definition of the pullback followed by Proposition 14.11 of Lee (2013) shows that:

$$\begin{aligned}(\phi^\star(\psi\omega))(y) &= \sum_{i=1}^{n} (-1)^{i+1} \alpha_i(\phi^{-1}(y)) \bigwedge_{j \neq i} \left(\sum_{k=1}^{n} J_{jk}(y) dy_k \right) \\ &= \sum_{i=1}^{n} (-1)^{i+1} \alpha_i(\phi^{-1}(y)) \sum_{k=1}^{n} M_{ik}(y) \bigwedge_{\ell \neq k} dy_\ell.\end{aligned}$$

As in the proof of Stokes theorem from Lee (2013), all of the terms of the pullback that include dy_n integrate to zero. This is because the y_n is fixed at 0 on the boundary, so the integrals over y_n must be zero. Thus, the integral simplifies to:

$$\int_{\partial\mathbb{H}^n} ((\phi^{-1})^\star(\psi\omega))(y) = \int_{\partial\mathbb{H}^n} \sum_{i=1}^{n} (-1)^{i+1} \alpha_i(\psi^{-1}(y)) M_{in}(y) \bigwedge_{j \neq n} dy_j \tag{66}$$

To show that the right side is zero, it suffices to show that the integrand on the right is zero. The inverse function theorem, followed by Cramer's rule shows that

$$\left.\frac{\partial \phi_n}{\partial x_i}\right|_{x=\phi^{-1}(y)} = (J(y)^{-1})_{ni} = \frac{1}{\det(J(y))}(-1)^{i+n} M_{in}(y).$$

Letting $x = \phi^{-1}(y)$, it follows that the integrand on the right of (66) is given by

$$\sum_{i=1}^{n} (-1)^{i+1} \alpha_i(\phi^{-1}(y)) M_{in}(y) = \det(J(y))(-1)^{1-n} \psi(x) e^{-\beta \bar{f}(x)} \nabla h(x)^\top \nabla \phi_n(x).$$

Note here that if $y \in \partial \mathbb{H}^n$, then $x \in \partial \mathcal{K}$. So, (61b) implies that the integrand is zero if $-\nabla \phi_n(x) \in N_{\mathcal{K}}(x)$. This follows because for all $z \in \mathcal{K}$ and all $t > 0$ sufficiently small, we have that $x + t(z - x) \in \mathcal{K}$ and

$$0 \leq \phi_n(x + t(z-x)) = \phi_n(x) + t \nabla \phi_n(x)^\top (z - x) + o(t) = t \nabla \phi_n(x)^\top (z - x) + o(t).$$

Thus $\nabla \phi_n(x)^\top x \leq \nabla \phi_n(x)^\top z$ and so $-\nabla \phi_n(x) \in N_{\mathcal{K}}(x)$. Thus, (65) has been proved and so the lemma has been proved for $\mathcal{K}$ with smooth boundaries.

Now we cover the general case. Let b be a self-concordant barrier function for $\mathcal{K}$ such that for any sequence $x_i \in \text{int}$ with $x_i \to \partial \mathcal{K}$, we have $b(x_i) \to \infty$. By Theorem 2.5.1 of Nesterov and Nemirovskii (1994), such a barrier function exists. Let $\mathcal{K}_i = \{x | b(x) \leq i\}$. For all sufficiently large i, we have that $\mathcal{K}_i$ is non-empty. Whenever $\mathcal{K}_i$ is nonempty, it has a smooth boundary. Furthermore, if S is a compact subset of $\text{int}(\mathcal{K})$, then $S \subset \mathcal{K}_i$ for all sufficiently large i. Let $\mathbf{x}^{M,i}$ be the Skorokhod solutions corresponding to the sets $\mathcal{K}_i$.

Let $Z_i = \int_{\mathcal{K}_i} e^{-\beta \bar{f}(x)} dx$. Then $Z_i \uparrow Z =: \int_{\mathcal{K}} e^{-\beta \bar{f}(x)} dx$ by monotone convergence. Let $\mathbf{x}^{M,i}$ be the solution from (5) in which the Skorokhod problem is solved over $\mathcal{K}_i$ in place of $\mathcal{K}$. Let P_t^i be the diffusion operator corresponding to $\mathbf{x}_t^{M,i}$. Then, for each non-empty $\mathcal{K}_i$, the corresponding version of (60) can be written as

$$\frac{1}{Z_i} \int_{\mathcal{K}_i} g(x) e^{-\beta \bar{f}(x)} dx = \frac{1}{Z_i} \int_{\mathcal{K}_i} (P_t^i g)(x) e^{-\beta \bar{f}(x)} dx \tag{67}$$

Using the fact that g is bounded on $\mathcal{K}$, dominated convergence implies that

$$\lim_{i \to \infty} \frac{1}{Z_i} \int_{\mathcal{K}_i} g(x) e^{-\beta \bar{f}(x)} dx = \frac{1}{Z} \int_{\mathcal{K}} g(x) e^{-\beta \bar{f}(x)} dx = \int_{\mathcal{K}} g(x) d\pi_{\beta \bar{f}}(x).$$

The proof will be completed if we can show that the right side of (67) converges to the right side of (60).

Note that the right side of (67) can be expressed as

$$\frac{1}{Z_i} \int_{\mathcal{K}_i} (P_t^i g)(x) e^{-\beta \bar{f}(x)} dx = \frac{1}{Z_i} \int_{\mathcal{K}_i} \mathbb{E}[g(\mathbf{x}_t^{M,i}) | \mathbf{x}_0^{M,i} = x] e^{-\beta \bar{f}(x)} dx$$

Now, Lemma 17 from Appendix E shows that for almost all realizations of the Brownian motion, $\mathbf{w}$, the Skorokhod solution converges pointwise $\lim_{i \to \infty} \mathbf{x}_t^{M,i} = \mathbf{x}_t^M$ for all $t \geq 0$. (In fact it converges uniformly on compacts.) As a result the integrand on the right converges pointwise almost surely to the integrand on the right side of (60). Thus, the integrals on the right of (67) converge to the integral on the right of (60) via dominated convergence. ■

Appendix G. An Elementary Result on Stieltjes Integration

The following basic result is used a few times to examine bounded variation functions, $x(t)$ whose differentials $dx(t)$ are only known when $x(t) \neq 0$.

Lemma 19 *Let $x(t)$ be a continuous non-negative function with bounded variation. Then*

$$x(t) - x(0) = \lim_{\epsilon \downarrow 0} \int_0^t \mathbb{1}(x(s) \geq \epsilon) dx(s) \tag{68}$$

Proof Fix any $\epsilon > 0$. Then we have the Stieltjes integral representation:

$$\begin{aligned} x(t) - x(0) &= \int_0^t dx(s) \\ &= \int_0^t \mathbb{1}(x(s) \geq \epsilon) dx(s) + \int_0^t \mathbb{1}(x(s) < \epsilon) dx(s). \end{aligned}$$

We will show that the second integral on the right goes to 0 as $\epsilon \to 0$. This would imply the desired result by re-arranging and taking limits.

Fix any $\delta > 0$. Then since $x(t)$ has bounded variation, there are numbers $0 = s_0 < s_1 < \cdots < s_N = t$ such that

$$\left| \int_0^t \mathbb{1}(x(s) < \epsilon) dx(s) - \sum_{i=0}^{N-1} \mathbb{1}(x(\bar{s}_i) < \epsilon)(x(s_{i+1}) - x(s_i)) \right| \leq \delta,$$

where $\bar{s}_i = \frac{1}{2}(s_{i+1} + s_i)$.

Continuity of $x(t)$ implies that the s_i can be chosen such that if $x(\bar{s}_i) < \epsilon$ then $x(s_i) \leq \epsilon$ and $x(s_{i+1}) \leq \epsilon$.

If $x(\bar{s}_{\hat{i}}) < \epsilon$ for some $\hat{i}$, then let $\mathcal{I} = \{j, j+1, \ldots, k\}$ be the largest sequence of integers such that $0 \leq j \leq \hat{i} \leq k \leq N-1$ and $x(\bar{s}_i) < \epsilon$ for $i = j, \ldots, k$. Then

$$\sum_{i=j}^{k} \mathbb{1}(x(\bar{s}_i) < \epsilon)(x(s_{i+1}) - x(s_i)) = x(s_{k+1}) - x(s_j) \tag{69}$$

Maximality of the interval and our choice of s_i imply that either $j = 0$ or $x(s_j) = \epsilon$ and either $k+1 = N$ or $x(s_{k+1}) = \epsilon$.

Note that in all cases $|x(s_{k+1}) - x(s_j)| \leq \epsilon$. If $x(s_j) = x(s_{k+1}) = \epsilon$ then the sum from (69) is zero. So the sum can only be non-zero if $j = 0$ or $k+1 = N$ (or both).

Since every term such that $x(\bar{s}_i) < \epsilon$ can be included in one of the intervals constructed above, and a most two of them can give rise to a non-zero sum, we see that the Riemann sum is bounded as:

$$\left| \sum_{i=0}^{N-1} \mathbb{1}(x(\bar{s}_i) < \epsilon)(x(s_{i+1}) - x(s_i)) \right| \leq 2\epsilon$$

Using the fact that δ is arbtrary and using the triangle inequality shows that

$$\left| \int_0^t \mathbb{1}(x(s) < \epsilon) dx(s) \right| \leq 2\epsilon$$

■