Proceedings of Machine Learning Research vol 313:1–39, 2026 37th International Conference on Algorithmic Learning Theory

Quantitative Convergence Analysis of Projected Stochastic Gradient Descent for Non-Convex Losses via the Goldstein Subdifferential

Yuping Zheng ZHEN0348@UMN.EDU and **Andrew Lamperski** ALAMPERS@UMN.EDU
200 Union St. Se, Keller Hall 4-174, Minneapolis, MN 55455, USA

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Stochastic gradient descent (SGD) is the main algorithm behind a large body of work in machine learning. In many cases, constraints are enforced via projections, leading to projected stochastic gradient algorithms. In recent years, a large body of work has examined the convergence properties of projected SGD for non-convex losses in asymptotic and non-asymptotic settings. Strong quantitative guarantees are available for convergence measured via Moreau envelopes. However, these results cannot be compared directly with work on unconstrained SGD, since the Moreau envelope construction changes the gradient. Other common measures based on gradient mappings have the limitation that convergence can only be guaranteed if variance reduction methods, such as mini-batching, are employed. This paper presents an analysis of projected SGD for non-convex losses over compact convex sets. Convergence is measured via the distance of the gradient to the Goldstein subdifferential generated by the constraints. Our proposed convergence criterion directly reduces to commonly used criteria in the unconstrained case, and we obtain convergence without requiring variance reduction. We obtain results for data that are independent, identically distributed (IID) or satisfy mixing conditions (L-mixing). In these cases, we derive asymptotic convergence and $O(N^{-1/3})$ non-asymptotic bounds in expectation, where N is the number of steps. In the case of IID sub-Gaussian data, we obtain almost-sure asymptotic convergence and high-probability $\tilde{O}(N^{-1/5})$ non-asymptotic bounds. In particular, these are the first non-asymptotic high-probability bounds for projected SGD with non-convex losses.

Keywords: Stochastic Optimization, Projected Stochastic Gradient Descent, Non-convex Learning, Non-asymptotic Analysis

1. Introduction

This paper focuses on the analysis of projected stochastic gradient descent (SGD) for solving optimization problems of the form:

$$\min_{x\in\mathcal{X}} \mathbb{E}[f(x,\mathbf{z})] = \min_{x\in\mathcal{X}} \bar{f}(x),$$

where $\mathcal{X}$ is a compact convex constraint set, $\mathbb{E}$ denotes the expected value over the random variable $\mathbf{z}$, and $\bar{f}$ is a smooth, but possibly non-convex loss.

Stochastic gradient descent and its variants have a plethora of applications in machine learning. See e.g. (Bottou et al., 2018; McMahan et al., 2013; Koren et al., 2009, 2021; Zinkevich et al., 2010; Zinkevich, 2003; Goodfellow et al., 2016). Projected SGD is commonly employed for stabilization and regularization in machine learning and neural networks, (Bottou et al., 2018), though often under different names. For example, the projection scheme is called "reprojection" in (Goodfellow et al., 2016) and a specific variant is called "max-norm regularization" in (Srivastava et al., 2014).

Related Work. Due to its practical significance, a large body of literature has examined projected SGD and generalized families of algorithms that include projected SGD. We review work on asymptotic convergence and non-asymptotic bounds for non-convex problems next.

Asymptotic convergence for projected SGD with non-convex objectives has a long history, with proofs dating back to at least (Ermol'ev and Norkin, 1998; Ermoliev and Norkin, 2003). More recent work on asymptotic properties of projected SGD and its generalizations, such as proximal gradients, includes (Davis et al., 2020; Bianchi et al., 2022; Majewski et al., 2018; Nguyen and Yin, 2023; Josz et al., 2024; Duchi and Ruan, 2018; Asi and Duchi, 2019b,a; Li and Milzarek, 2022). These works, and the work of the present paper, are largely based on continuous-time approximation methods described in (Kushner and Yin, 2003; Borkar, 2023; Benaïm, 2006).

Non-asymptotic bounds in expectation, measured with respect to Moreau envelopes and related measures, are given for IID data, $\mathbf{z}_k$, in (Davis and Drusvyatskiy, 2019; Deng and Gao, 2021; Zhu et al., 2023; Gao and Deng, 2024; Alacaoglu et al., 2020; Davis et al., 2025; Fatkhullin et al., 2025) and dependent data under mixing conditions in (Alacaoglu and Lyu, 2023). Non-asymptotic bounds in expectation, measured special variants of the proximal gradient mapping are given in (Ghadimi et al., 2016; Lan et al., 2024) with similar measures used in (He et al., 2025; Xie et al., 2025).

We will show in Section 3 that the Moreau envelope measure from (Davis and Drusvyatskiy, 2019) and subsequent works do not reduce to the gradient norm, $\|\nabla \bar{f}(x)\|$, in the unconstrained case, which is arguably the most common measure for non-convex unconstrained problems. In contrast, we will show that measures from (Ghadimi et al., 2016) and related works do reduce to $\|\nabla \bar{f}(x)\|$, but result in a non-shrinking term that can only be mitigated by variance reduction methods, such as mini-batching.

For convex losses, the convergence theory for projected SGD is more mature, with overviews given in (Hazan et al., 2016; Shalev-Shwartz and Ben-David, 2014).

Beyond projected SGD and generalizations, a variety of alternative methods for enforcing constraints in stochastic optimization have been proposed. These include penalty methods (Lin et al., 2022; Alacaoglu and Wright, 2024), Frank-Wolfe methods (Reddi et al., 2016; Lacoste-Julien, 2016), and Lagrangian methods (Papadimitriou and Vu, 2025).

Contributions. We present an analysis of projected SGD with performance measured by distance of $-\nabla \bar{f}(x)$ to the Goldstein subdifferential, (Goldstein, 1977), associated with the constraints. Unlike Moreau envelope measures, our measure reduces to $\|\nabla \bar{f}(x)\|$ in the unconstrained case, and unlike the proximal gradient mapping measures from (Ghadimi et al., 2016), we can show convergence without variance reduction / mini-batching.

For IID and L-mixing data, $\mathbf{z}_k$, we show that our proposed measure converges asymptotically to 0 in expectation under stochastic approximation step size conditions. For fixed step sizes, we give a non-asymptotic bound in expectation of $O(N^{-1/3})$, where N is the number of steps. Currently, our bound is weaker than the $O(N^{-1/2})$ bound obtained with respect to the Moreau envelope in (Davis and Drusvyatskiy, 2019). More work is required to determine if this is due to a fundamental difference in the measures, or a limitation of the current analysis.

For IID sub-Gaussian data, we show that our measure converges asymptotically to 0 with probability 1 under stochastic approximation step size conditions. For fixed step sizes, we give a non-asymptotic bound of $\tilde{O}(N^{-1/5})$, which holds with high probability. In particular, these are the first non-asymptotic high probability bounds for projected SGD with non-convex losses.

2. Problem Setup

2.1. Notation and terminology

$\mathbb{N}$ denotes non-negative integers and $\mathbb{R}$ denotes the real numbers. Random variables are denoted in bold. If $\mathbf{x}$ is random variable, then $\mathbb{E}[\mathbf{x}]$ denotes its expected value. $\|x\|$ denotes the Euclidean norm over $\mathbb{R}^n$. The probabilistic indicator function is denoted by $\mathbb{1}$. (The indicator function from variational / convex analysis will be denoted by $\mathcal{I}_\mathcal{X}$ below.) $\mathbb{P}$ denotes probability measure. If $\mathcal{F}$ and $\mathcal{G}$ are σ-algebras, then $\mathcal{F} \vee \mathcal{G}$ denotes the σ-algebra generated by the union of $\mathcal{F}$ and $\mathcal{G}$.

$\Pi_\mathcal{X}(y)$ denotes the projection of y onto a convex set $\mathcal{X}$, i.e. $\Pi_\mathcal{X}(y) = \arg\min_{x\in\mathcal{X}} \|y-x\|$. The Euclidean distance of y to the set $\mathcal{X}$ is denoted by $\mathrm{dist}(y, \mathcal{X})$.

The boundary of $\mathcal{X}$ is denoted as $\partial\mathcal{X}$, the normal cone of $\mathcal{X}$ at a point x is denoted by $\mathcal{N}_\mathcal{X}(x)$, the tangent cone of $\mathcal{X}$ at a point x is denoted by $T_\mathcal{X}(x)$. $\mathcal{N}_\mathcal{X}(x) = \{\phi | \phi^\top x \geq \phi^\top z,\ \forall z \in \mathcal{X}\}$. $T_\mathcal{X}(x) = \{t(y-x) | y \in \mathcal{X},\ t \geq 0\}$.

Let $osc(\bar{f})$ denote the oscillation of a bounded function $\bar{f}$, which is defined by $osc(\bar{f}) = \sup_{x,x'\in\mathcal{X}} |\bar{f}(x) - \bar{f}(x')|$.

2.2. Projected SGD

Assume that the initial value of $\mathbf{x}_0 \in \mathcal{X}$ is independent of $\mathbf{z}_i$ for all $i \in \mathbb{N}$. Projected SGD is the algorithm:

$$\mathbf{x}_{k+1} = \Pi_\mathcal{X}\left(\mathbf{x}_k - \alpha_k \nabla_x f(\mathbf{x}_k, \mathbf{z}_k)\right) \tag{1}$$

where α_k is the step size. Our main result holds for any deterministic step size sequence with $0 < \alpha_k \leq \frac{1}{2}$. We also describe special cases of constant step size, $\alpha_k = \alpha$, and standard stochastic approximation conditions:

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{2}$$

2.3. Approximate Stationarity via the Goldstein Subdifferential

The Goldstein subdifferential is a relaxed version of the Clarke subdifferential and is widely used in nonsmooth optimization. It was first introduced in (Goldstein, 1977) and has been used for measuring the stationarity for optimization algorithms, e.g. (Davis et al., 2022; Zhang et al., 2020a).

Let $\mathcal{X}$ denote a closed convex set. If $\mathcal{I}_\mathcal{X}$ is the corresponding convex indicator function:

$$\mathcal{I}_\mathcal{X}(x) = \begin{cases} 0 & x \in \mathcal{X} \\ +\infty & x \notin \mathcal{X}. \end{cases}$$

then the Clarke subdifferential reduces to the standard convex subdifferential, and corresponds to the normal cone:

$$\overline{\partial}\mathcal{I}_\mathcal{X}(x) = \partial\mathcal{I}_\mathcal{X}(x) = \mathcal{N}_\mathcal{X}(x).$$

See Rockafellar and Wets (2009) for details on these definitions.

For $\epsilon > 0$, the Goldstein subdifferential is defined in terms of the Clarke subdifferential by:

$$\overline{\partial}_\epsilon g(x) = \mathrm{conv}\left(\bigcup_{\|y-x\|\leq\epsilon} \overline{\partial} g(y)\right).$$

Thus, in the simple case that $g = \mathcal{I}_\mathcal{X}$, we have

$$\overline{\partial}_\epsilon \mathcal{I}_\mathcal{X}(x) = \mathrm{conv}\left(\bigcup_{\|y-x\|\le\epsilon} \mathcal{N}_\mathcal{X}(y)\right).$$

The standard first-order necessary optimality conditions give that if x is a local minimizer of $\bar{f}$ then $-\nabla \bar{f}(x) \in \mathcal{N}_\mathcal{X}(x)$. This occurs if and only if $\mathrm{dist}(-\nabla \bar{f}(x), \overline{\partial}\mathcal{I}_\mathcal{X}(x)) = 0$. In this work, we will bound the relaxed stationarity measure, $\mathrm{dist}(-\nabla \bar{f}(x), \overline{\partial}_\epsilon \mathcal{I}_\mathcal{X}(x))$.

2.4. L-mixing processes

In this paper, we consider the case that the external data variables, $\mathbf{z}_k$, can have dependencies over time, but these dependencies satisfy a property known as L-mixing. The class of L-mixing processes was introduced in (Gerencsér, 1989) and has been used to quantify the time-correlation in stochastic optimization in recent years (see Barkhagen et al., 2021; Chau et al., 2019, 2021; Zheng and Lamperski, 2022, 2025a,b). It contains a wide variety of processes including measurements of geometrically ergodic Markov chain (Gerencsér et al., 2002), which is suitable to model various of stable nonlinear stochastic systems. Furthermore, the class of L-mixing processes is closed under a variety of operations. In particular, L-mixing random variables results in another L-mixing sequence after passing through a stable, causal linear filter (Zheng and Lamperski, 2025a). Therefore, the class of L-mixing processes contains a wide variety of data streams from system identification and time-series analysis.

Now we introduce the definition of the discrete-time L-mixing processes. Let $\mathcal{F} = (\mathcal{F}_k)_{k\ge0}$ be an increasing family of σ-algebras and let $\mathcal{F}^+ = (\mathcal{F}_k^+)_{k\ge0}$ be a decreasing family of σ-algebras such that $\mathcal{F}_k$ and $\mathcal{F}_k^+$ are independent for all $k \ge 0$. A discrete-time stochastic process $\mathbf{z}_k$ is called L-mixing with respect to $(\mathcal{F}, \mathcal{F}^+)$ if

- $\mathbf{z}_k$ is $\mathcal{F}_k$-measurable for all integers $k \ge 0$
- $\mathcal{M}_m(\mathbf{z}) := \sup_{k\ge0} \mathbb{E}^{1/m}\left[\|\mathbf{z}_k\|^m\right] < \infty$ for all $m \ge 1$
- $\Psi_m(\mathbf{z}) := \sum_{\tau=0}^{\infty} \psi_m(\tau, \mathbf{z}) < \infty$ for all integers $k \ge 1$ and all $m \ge 1$, where $\psi_m(\tau, \mathbf{z}) = \sup_{k\ge\tau} \mathbb{E}^{1/m}\left[\left\|\mathbf{z}_k - \mathbb{E}[\mathbf{z}_k|\mathcal{F}_{k-\tau}^+]\right\|^m\right]$.

The value of $\Psi_m(\mathbf{z})$ measures how fast the time-dependence between data decays.

2.5. Assumptions

Assumptions on the domain. For the rest of the paper, $\mathcal{X}$ denotes a compact convex subset of $\mathbb{R}^n$ of diameter D which contains a ball of radius $r > 0$ around the origin.

Assumptions on the objective function. Assume that for each z, $\nabla_x f(x, z)$ is ℓ-Lipschitz in both x and z, i.e. $\|\nabla_x f(x_1, z) - \nabla_x f(x_2, z)\| \le \ell\|x_1 - x_2\|$ and $\|\nabla_x f(x, z_1) - \nabla_x f(x, z_2)\| \le \ell\|z_1 - z_2\|$. This implies that $\|\nabla \bar{f}(x_1) - \nabla \bar{f}(x_2)\| \le \ell\|x_1 - x_2\|$, $\|\nabla \bar{f}(x)\| \le u$ where $u \le \|\nabla \bar{f}(0)\| + \ell D$ as well as $osc(\bar{f}) \le Du$.

Note that without further specification in the paper, we simply use $\nabla f(x, z)$ to indicate $\nabla_x f(x, z)$.

Remark 1 *Our assumptions on the domain and the objective function are strong. Convexity and compactness of the domain are used for analysis of continuous-time approximations of projected SGD via Skorokhod problems. Relaxing convexity and compactness via more general approaches to Skorokhod problems, as in Lions and Sznitman (1984), would be an interesting future direction. The assumption that $\nabla f(x, z)$ is ℓ-Lipschitz in both x and z is stronger than the assumption that $\nabla \bar{f}(x)$ is Lipschitz. However, this stronger assumption also appears in other work on SGD-based learning theory, see e.g. (Chau et al., 2021), (Barkhagen et al., 2021), (Durmus et al., 2025). A more general smoothness assumption is made in (Zhang et al., 2023) and can be reduced to our smoothness assumption when the parameter and data domains are bounded.*

Assumptions on the external random variables $\mathbf{z}_k$. In this work, we present the convergence bound under different assumptions on the external random variables $\mathbf{z}_k \in \mathcal{Z}$:

A1) $\nabla f(x, \mathbf{z}_k) = \nabla \bar{f}(x) + \mathbf{z}_k$, where $\mathbf{z}_k$ are IID zero mean sub-Gaussian random vectors, independent of the initial state, $\mathbf{x}_0$. Specifically, there exists a number $\hat{\sigma} > 0$ such that for all $v \in \mathbb{R}^n$, the following bound holds:

$$\mathbb{E}\left[e^{v^\top \mathbf{z}}\right] \le e^{\frac{1}{2}\hat{\sigma}^2 \|v\|^2}. \tag{3}$$

A2) $\mathbb{E}\left[\|\nabla f(x, \mathbf{z}_k) - \nabla \bar{f}(x)\|^2\right] \le \sigma^2$ and $\mathbf{z}_k$ are independent for all $k \in \mathbb{N}$.

A3) $\mathbf{z}_k$ is L-mixing processes with respect to $(\mathcal{F}, \mathcal{F}^+)$, independent of the initial state, $\mathbf{x}_0$.

Note that A1 is a special case of both A2 and A3. Indeed, using that $\mathbb{E}[(e_i^\top \mathbf{z})^2] \le \hat{\sigma}^2$ for each standard basis vector, e_i, gives that $\mathbb{E}[\|\mathbf{z}\|^2] \le n\hat{\sigma}^2$. To see that A3 holds, we can set $\mathcal{F}_k = \sigma(\{\mathbf{z}_0, \dots, \mathbf{z}_k\})$ and $\mathcal{F}_k^+ = \sigma(\{\mathbf{z}_{k+1}, \mathbf{z}_{k+2}, \dots\})$. Then we can bound the moments via bounds on the moment generating function, noting that for all $m \ge 1$: $\Psi_m(\mathbf{z}) = \psi_m(0, \mathbf{z})$. In particular, $\Psi_2(\mathbf{z}) \le \sqrt{n}\hat{\sigma}$.

3. Discussion on Convergence Criteria

In this section, we review various convergence criteria used for analyzing gradient-descent algorithms under different hypotheses.

In preparation for the following discussion, we first present a lemma which is the key to the application of ODE method to approximate the discrete-time processes with continuous-time processes. More details on the approximation method are shown in Section 4.

Lemma 2 *For all $x \in \mathcal{X}$, $g \in \mathbb{R}^n$, the following holds:*

$$\lim_{\alpha \downarrow 0} \frac{\Pi_{\mathcal{X}}(x + \alpha g) - x}{\alpha} = \Pi_{T_{\mathcal{X}}(x)}(g)$$

This result appears in (Calamai and Moré, 1987; McCormick and Tapia, 1972) and the corresponding proof can be found in Proposition 2 of (McCormick and Tapia, 1972).

Lemma 2 implies that projected SGD can be viewed as a constrained stochastic Euler approximation to the following ODE:

$$\frac{d}{dt}\mathbf{x}_t^C = \Pi_{T_{\mathcal{X}}(\mathbf{x}_t^C)}(-\nabla \bar{f}(\mathbf{x}_t^C)). \tag{4}$$

Note $\mathbf{x}^C$ is called the *continuous* process in the rest of the paper.

For GD and projected GD algorithms with convex objectives, we can use $\bar{f}(x_k) - \bar{f}(x^*)$ to measure the convergence rate since all critical points, x^*, are actually global minima (Boyd and Vandenberghe, 2004; Bubeck et al., 2015; Nesterov et al., 2018). In the strongly convex case, $\|x_k - x^*\|^2$ is often used to measure the convergence (Nesterov et al., 2018), since minimizers are unique. A stochastic variation $\mathbb{E}[\bar{f}(\mathbf{x}_k) - \bar{f}(x^*)]$ is used under the conditions that $\bar{f}$ is non-strongly convex and global minimum exists (not necessarily unique) (Moulines and Bach, 2011) or if $\bar{f}$ satisfies the Polyak-Lojasiewicz condition (Khaled and Richtárik, 2020; Gower et al., 2021). The stochastic version $\mathbb{E}[\|\mathbf{x}_k - x^*\|^2]$ is used when $\bar{f}$ is strongly convex for both unconstrained and projected SGD (Moulines and Bach, 2011).

For non-convex problems, algorithms may converge to critical points which are not necessarily global minima. In general, there could be multiple critical points. So, measures based on $\bar{f}(x) - \bar{f}(x^*)$ or $\|x - x^*\|$ with fixed critical points, x^*, will not be suitable. In asymptotic analysis, it is common to measure convergence of the algorithms to the set of critical points (Bianchi et al., 2022; Ermol'ev and Norkin, 1998; Ermoliev and Norkin, 2003).

For non-asymptotic bounds for unconstrained non-convex problems, most analyses utilize variations on the size $\|\nabla \bar{f}(x)\|$ to measure stationarity. For example, in unconstrained deterministic problems, (Nesterov et al., 2018) uses $\min_{0\le k<N} \|\nabla \bar{f}(x_k)\|$. For stochastic problems, $\min_{0\le k<N} \mathbb{E}[\|\nabla \bar{f}(\mathbf{x}_k)\|]$ is used in (Khaled and Richtárik, 2020; Yuan et al., 2022; Lei et al., 2019; Wu et al., 2020), $\frac{1}{N}\mathbb{E}[\sum_{k=0}^{N-1} \|\nabla \bar{f}(\mathbf{x}_k)\|^2]$ is used for constant step sizes in (Bottou et al., 2018; Zhang et al., 2020b; Chen and Zhao, 2023), and $\mathbb{E}[\frac{1}{\tau_N}\sum_{k=0}^{N-1} \alpha_k \|\nabla \bar{f}(\mathbf{x}_k)\|^2]$ is used for diminishing step sizes in (Bottou et al., 2018). For a more thorough review of unconstrained SGD, see (Garrigos and Gower, 2023).

The most common measure for non-asymptotic analysis of projected SGD and its generalizations is based on Moreau envelopes. See (Davis and Drusvyatskiy, 2019; Deng and Gao, 2021; Zhu et al., 2023; Gao and Deng, 2024; Alacaoglu et al., 2020; Davis et al., 2025; Fatkhullin et al., 2025). For $\lambda > 0$, the Moreau envelope and proximal map of a function $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined respectively by:

$$\psi_\lambda(x) = \min_{y\in\mathbb{R}^n} \left(\psi(y) + \frac{1}{2\lambda}\|x-y\|^2\right) \quad \text{and} \quad \operatorname{prox}_{\lambda\psi}(x) = \arg\min_{y\in\mathbb{R}^n} \left(\psi(y) + \frac{1}{2\lambda}\|x-y\|^2\right).$$

For projected SGD, the function $\psi = \bar{f} + \mathcal{I}_{\mathcal{X}}$ is used in Moreau envelope analysis.

The gradient of the Moreau envelope is given by $\nabla\psi_\lambda(x) = \frac{1}{\lambda}\left(x - \operatorname{prox}_{\lambda\psi}(x)\right)$. In (Davis and Drusvyatskiy, 2019) and subsequent work, convergence of projected SGD is measured via

$$\frac{1}{\tau_N}\sum_{k=0}^{N-1} \alpha_k \mathbb{E}\left[\|\nabla\psi_\lambda(\mathbf{x}_k)\|^2\right],$$

where $\lambda > 0$ is a fixed number with bounds scaling with λ^{-1}.

While the Moreau envelope measure resembles the common sum-of-squared norms measure from unconstrained SGD, it does not reduce to the value in the unconstrained case. Indeed, if

$$\psi(x) = \bar{f}(x) = \frac{1}{2}x^\top P x + q^\top x,$$

with positive definite P, then $\nabla \bar{f}(x) = Px + q$ and $\nabla \psi_\lambda(x) = (\lambda P + I)^{-1}(Px + q)$. For more complex objectives, the relationship between $\nabla \bar{f}$ and $\nabla \psi_\lambda$ will be more complex. These differences make direct comparison of Moreau envelope results with work on unconstrained SGD challenging.

An alternative measure, proposed in (Ghadimi et al., 2016) and used later in (Lan et al., 2024) is

$$\mathbb{E}\left[\frac{1}{\alpha_{\mathbf{r}}^2}\left\|\mathbf{x}_{\mathbf{r}} - \Pi_{\mathcal{X}}\left(\mathbf{x}_{\mathbf{r}} - \alpha_k \frac{1}{m_{\mathbf{r}}}\sum_{i=1}^{m_{\mathbf{r}}} \nabla f(\mathbf{x}_{\mathbf{r}}, \mathbf{z}_{\mathbf{r},i})\right)\right\|^2\right]$$

where $\mathbf{r}$ is a randomly drawn iteration and $\{\mathbf{z}_{\mathbf{r},1}, \ldots, \mathbf{z}_{\mathbf{r},m_{\mathbf{r}}}\}$ is a minibatch of noise variables. (Note that the convex projection here is a special case covered by their theory.)

The convergence measures in (Ghadimi et al., 2016; Lan et al., 2024) are modifications of the reduced gradient described in (Nesterov et al., 2018). Related measures are commonly used in deterministic settings. In projected GD, the measure $\frac{1}{\alpha_k}\|x_k - \Pi_{\mathcal{X}}(x_k - \alpha_k \nabla \bar{f}(x_k))\|$ ($\alpha_k = 1$) is used in (Royer et al., 2024), the measure $\text{dist}\left(-\nabla \bar{f}(x_k), \mathcal{N}_{\mathcal{X}}(x_k)\right)$ is used in (Olikier and Waldspurger, 2025), while $\|T_{\mathcal{X}(x_k)}(-\nabla \bar{f}(x_k))\|$ is used in (di Serafino et al., 2024; Calamai and Moré, 1987; Balashov and Tremba, 2022). These measures all reduce to $\|\nabla \bar{f}(x)\|$ in the unconstrained case.

The example below shows that it is impossible to achieve low error with respect to the measure from (Ghadimi et al., 2016) and related measures, unless the variance of the randomness is reduced. As a result, to achieve low error, (Ghadimi et al., 2016; Lan et al., 2024) propose large mini-batches. Similar limitations appear in the work of (He et al., 2025; Xie et al., 2025).

Example 1 *Let $f(x,z) = -x + xz$ so that $\nabla_x f(x,z) = -1 + z$. Set the constraint to be $\mathcal{X} = [-1,1]$. Let $\mathbf{z}_k$ follows the scaled binary Rademacher distribution such that $\mathbb{P}(\mathbf{z}_k = 2) = 0.5$ and $\mathbb{P}(\mathbf{z}_k = -2) = 0.5$.*

The normal cone of $\mathcal{X}$ is given by:

$$\mathcal{N}_{\mathcal{X}}(x) = \begin{cases} 0 & x \in (-1,1) \\ (-\infty, 0] & x = -1 \\ [0, \infty) & x = 1. \end{cases}$$

Projected SGD becomes

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_k + \alpha_k(1 - \mathbf{z}_k)\right).$$

Note that $\nabla \bar{f}(x) = -1$ for all x. Furthermore, for all $y \in \mathcal{X}$,

$$\text{dist}(-\nabla \bar{f}(x), \mathcal{N}_{\mathcal{X}}(y)) = \begin{cases} 1 & y \in [-1,1) \\ 0 & y = 1. \end{cases} \tag{5}$$

Say that $0 < \alpha_k \le \frac{3}{8}$ and $\alpha_{k+1} \le \alpha_k$. Then for any $\mathbf{x}_k \in \mathcal{X}$, we have $\mathbf{x}_{k+1} \in (-1, 1 - \alpha_{k+1}]$ with probability at least $\frac{1}{2}$. Thus, for all $k \ge 0$, with probability at least $\frac{1}{2}$, we have

$$\frac{1}{\alpha_k}\|\mathbf{x}_k - \Pi_{\mathcal{X}}\left(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k, \mathbf{z}_k)\right)\| = |1 - \mathbf{z}_k| \ge 1$$

and

$$\frac{1}{\alpha_{k+1}}\left\|\mathbf{x}_{k+1}-\Pi_{\mathcal{X}}(\mathbf{x}_{k+1}-\alpha_{k+1}\nabla\bar{f}(\mathbf{x}_{k+1}))\right\| =$$
$$\mathrm{dist}(-\nabla\bar{f}(\mathbf{x}_{k+1}),\mathcal{N}_{\mathcal{X}}(\mathbf{x}_{k+1}))=\|\Pi_{T_{\mathcal{X}}(\mathbf{x}_{k+1})}(-\nabla\bar{f}(\mathbf{x}_{k+1}))\|=1.$$

So, the average of any of these criteria will be at least $\frac{1}{2}$.

While these common convergence metrics remain bounded away from zero, on average, Fig. 1 (along with the theory in this paper) shows that the projected SGD solutions closely follow the continuous-time trajectory, $\mathbf{x}_t^C$, when the step size is small. The issue is that these measures amplify small random fluctuations near the boundary.

4. Approximation and Main Results

In this section, we present the continuous-time approximation of the algorithm via ordinary differential equations (ODEs), which is briefly presented in Section 3. Then, we present the main results under our proposed convergence criterion.

4.1. Continuous-Time Approximation

Let $\tau_k=\sum_{j=0}^{k-1}\alpha_j$, which measures the total amount of continuous time that has been simulated prior to the computation of $\mathbf{x}_k$. To analyze projected SGD in terms of continuous-time processes, we let $\mathbf{x}_t^A$ denote the iterates of (1) embedded into continuous-time as:

$$\mathbf{x}_t^A=\mathbf{x}_k \quad \text{if } t\in[\tau_k,\tau_{k+1}).$$

As seen in Fig. 1, projected SGD and its continuous-time approximation can drift apart due to instabilities. So, for our convergence analysis, we will construct a sequence of restarted continuous-time processes, defined as follows.

For a fixed number of iterates, N, define break points by:

$$\begin{aligned}
&s_0=0\\
&s_{i+1}=\max\{\tau_j|\tau_j-s_i\le 1, 0\le\tau_j\le\tau_N\} \text{ if } s_i<\tau_N.
\end{aligned}$$

Then, for $t\in[s_i,s_{i+1}]$, set:

$$\begin{aligned}
\frac{d}{dt}\mathbf{x}_t^{C_i}&=\Pi_{T_{\mathcal{X}}(\mathbf{x}_t^{C_i})}(-\nabla\bar{f}(\mathbf{x}_t^{C_i}))\\
\mathbf{x}_{s_i}^{C_i}&=\mathbf{x}_{s_i}^A.
\end{aligned}$$

For compact notation, define $\mathbf{x}_t^J$ to be the process that jumps between the continuous processes: $\mathbf{x}_t^J=\mathbf{x}_t^{C_i}$ when $t\in[s_i,s_{i+1})$.

For $k\ge 0$, let

$$\mathbf{b}_k=\sup_{t\in[\tau_k,\tau_{k+1})}\|\mathbf{x}_t^J-\mathbf{x}_{\tau_k}^A\|.$$

Denote $\chi(N)=\max\{i|s_i<\tau_N\}$ so that $s_{\chi(N)+1}=\tau_N$. Then the total number of subintervals partitioning the interval $[0,\tau_N]$ is $\chi(N)+1$. Let $\mathcal{K}(i)$ denote the value of j such that $\tau_j=s_i$, and let $\zeta(j)$ denote the value of i such that $\mathcal{K}(i)\le j<\mathcal{K}(i+1)$.

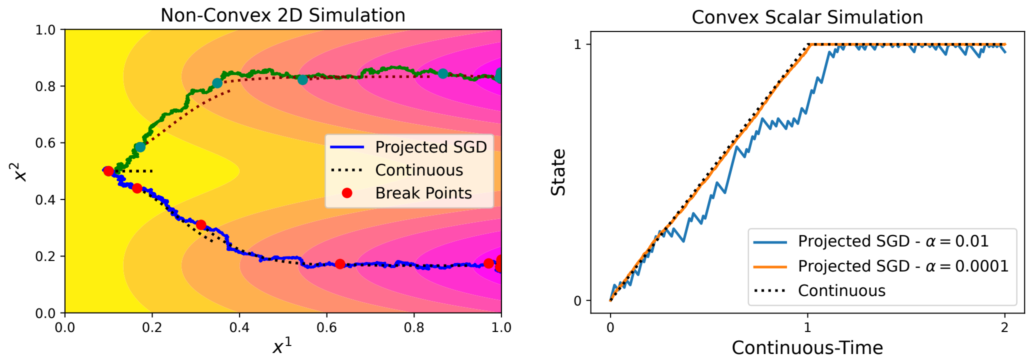
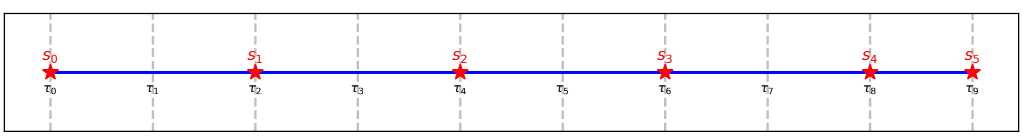


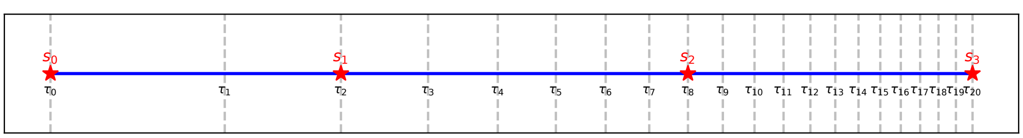
Figure 1: **Simulations.** The left shows two runs of projected SGD for a non-convex system from the same starting point. The combination of stochasticity and non-convexity implies that two trajectories with the same starting point can diverge over time. Here, the solid lines show the result of projected SGD, the dotted lines show the continuous-time approximations, and the filled circles indicate the break points. The right shows two runs of projected SGD on a convex scalar problem. With small step size, $\alpha = 0.0001$, the trajectory converges to a small region near the optimal solution. However, the existing convergence measures for constrained problems amplify the small fluctuations.

Assume that $\alpha_k \leq \frac{1}{2}$ for all $k \in [0, N-1]$. Then for any $i \in [0, \chi(N) - 1]$, there exists $j \in [0, N-1]$ s.t. $s_{i+1} = \tau_j \leq s_i + 1 \leq \tau_{j+1} = \tau_j + \alpha_j \leq \tau_j + \frac{1}{2}$, which implies that $s_i + \frac{1}{2} \leq s_{i+1} \leq s_i + 1$, i.e. $\frac{1}{2} \leq s_{i+1} - s_i \leq 1$. The last interval is $[s_{\chi(N)}, s_{\chi(N)+1}]$ whose length is at most 1, but is not necessarily greater than $\frac{1}{2}$.

Figure 2 shows the partitions of the interval $[0, \tau_N]$ for constant step size and diminishing step size according to the construction rules above. For constant step size, set $\alpha = \frac{3}{8}$ and $N = 9$, the interval $[0, \tau_N]$ is partitioned into 5 subintervals. For diminishing step size, set $\alpha_k = \frac{1}{k+2}$ and $N = 20$, the interval $[0, \tau_N]$ is partitioned into 3 subintervals.



(a) Constant Step Size



(b) Diminishing Step Size

Figure 2: Demonstration of the construction of subintervals $[s_i, s_{i+1}]$.

In the results below, we will use the following constants:

$$c_1 = \begin{cases} e^{\ell}\sqrt{n}\hat{\sigma} & \text{Under Assumption } A1 \\ e^{\ell}\sigma & \text{Under Assumption } A2 \\ 2\ell e^{\ell}\Psi_2(\mathbf{z}) & \text{Under Assumption } A3 \end{cases} \tag{6a}$$

$$c_2 = \left(u + \sqrt{2r^{-1}u(Du + D^2)}\right) e^{\ell} \tag{6b}$$

$$c_3 = 2\sqrt{2}e^{2\ell}\hat{\sigma}D \tag{6c}$$

$$c_4 = 4e^{2\ell}\hat{\sigma}^2 \tag{6d}$$

$$c_5 = e^{2\ell}(n+1)\hat{\sigma}^2 \tag{6e}$$

Remark 3 *The constants above have exponential dependence on ℓ. Under the assumption that $0 < \alpha_k \leq \frac{1}{2\ell}$, the exponential dependence can be removed by redefining the breakpoints by $s_0 = 0$, $s_{i+1} = \max\{\tau_j | \tau_j - s_i \leq 1/\ell, 0 \leq \tau_j \leq \tau_N\}$.*

The following lemma gives bounds in expectation and with high probability on the deviations of the algorithm from the jumping continuous process $\|\mathbf{x}_t^A - \mathbf{x}_t^J\|$. It is proved in Appendix C.

Lemma 4 *Assume that $0 < \alpha_k \leq \frac{1}{2}$ for all $k \in \mathbb{N}$. Let $\mathcal{K}(i)$ be the sequence of integers defined in Section 4.1. The following hold:*

(i) If assumption A1, A2, or A3 holds, then for all integers $k \in [\mathcal{K}(i), \mathcal{K}(i+1))$:

$$\mathbb{E}\left[\mathbf{b}_k\right] \leq c_1 \sqrt{\sum_{j=\mathcal{K}(i)}^{k-1} \alpha_j^2} + c_2 \max_{j \in [\mathcal{K}(i), k]} \sqrt{\alpha_j}.$$

(ii) If Assumption A1 holds and $\delta \in (0,1)$, then with probability at least $1-\delta$,

$$\max_{k \in [\mathcal{K}(i), \mathcal{K}(i+1))} \mathbf{b}_k \leq h_i(\delta),$$

where

$$h_i(\delta) = \left(c_3 \sqrt{\log(2\delta^{-1})} \sqrt{\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2} + \left(c_4 \log(2\delta^{-1}) + c_5\right) \sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2 \right)^{1/2} + c_2 \max_{j \in [\mathcal{K}(i), \mathcal{K}(i+1))} \sqrt{\alpha_j}.$$

The next result shows that in the decaying step size case, the algorithm, $\mathbf{x}_t^A$ converges to the jumping continuous process, $\mathbf{x}_t^J$, asymptotically. Note that $\max_{k \in [\mathcal{K}(i), \mathcal{K}(i+1))} \mathbf{b}_k = \sup_{t \in [s_i, s_{i+1})} \|\mathbf{x}_t^A - \mathbf{x}_t^J\|$.

Proposition 5 *Assume that* $0 < \alpha_k \leq \frac{1}{2}$*,* $\sum_{k=0}^{\infty} \alpha_k = \infty$*, and* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$*. Let* h_i *be the bounding function defined in Lemma 4. Set* $\delta_i = \frac{\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2}{\sum_{k=0}^{\infty} \alpha_k^2}$*. If Assumption A1 holds, then* $\lim_{i\to\infty} h_i(\delta_i) = 0$*, and with probability* 1*, the event*

$$\sup_{t\in[s_i,s_{i+1})} \|\mathbf{x}_t^A - \mathbf{x}_t^J\| > h_i(\delta_i)$$

occurs at most finitely many times. In particular, $\lim_{t\to\infty} \|\mathbf{x}_t^A - \mathbf{x}_t^J\| = 0$ *with probability* 1*.*

4.2. Main Results

Here we present the main results of the paper. All of the results in this section are proved in Appendix B.

Theorem 6 below presents general error bounds under our proposed measure which hold for any sequence of step sizes satisfying $0 < \alpha_k \leq \frac{1}{2}$. More explicit versions for step size sequences following stochastic approximation conditions and constant step sizes are shown in Corollaries 8 and 9, respectively.

Theorem 6 *Assume that* $0 < \alpha_k \leq \frac{1}{2}$ *for all integers* $k \in [0, N-1]$*. Let* $\chi(N)$ *and* $\mathcal{K}(i)$ *be the integers defined in Section 4.1.*

- *If Assumption A1, A2, or A3 holds, then*

$$\frac{1}{\tau_N}\sum_{k=0}^{N-1} \alpha_k \mathbb{E}\left[\mathrm{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2\right] \leq \frac{1}{\tau_N}\sum_{i=0}^{\chi(N)} \left(c_6\sqrt{\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2} + c_7 \max_{j\in[\mathcal{K}(i),\mathcal{K}(i+1))} \sqrt{\alpha_j}\right) + \frac{Du}{\tau_N},$$

  where

$$c_6 = (u + 2u\ell)c_1 \quad \text{and} \quad c_7 = (u + 2u\ell)c_2.$$

- *If Assumption A1 holds, then for any collection of numbers* $\delta_0, \ldots, \delta_{\chi(N)}$ *such that* $0 < \delta_i$ *and* $\sum_{i=0}^{\chi(N)} \delta_i < 1$*, with probability at least* $1 - \sum_{i=0}^{\chi(N)} \delta_i$*, the following bound holds:*

$$\frac{1}{\tau_N}\sum_{k=0}^{N-1} \alpha_k \mathrm{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{h_{\zeta(k)}(\delta_{\zeta(k)})}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2 \leq \frac{u + 2\ell u}{\tau_N}\sum_{i=0}^{\chi(N)} h_i(\delta_i) + \frac{Du}{\tau_N}.$$

Remark 7 *The convergence criterion in Theorem 6 generalizes the common sum of norm square convergence criterion for unconstrained SGD. In the unconstrained case,* $\bar{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k) = \{0\}$*, which gives*

$$\frac{1}{\tau_N}\sum_{k=0}^{N-1} \alpha_k \mathbb{E}\left[\mathrm{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2\right] = \frac{1}{\tau_N}\sum_{k=0}^{N-1} \alpha_k \mathbb{E}\left[\|\nabla \bar{f}(\mathbf{x}_{\tau_k}^A)\|^2\right]. \tag{7}$$

In particular, when $\alpha_k \equiv \alpha$*, then* (7) $= \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\|\nabla \bar{f}(\mathbf{x}_{\tau_k}^A)\|^2\right]$*. In both the variable or constant step size cases,* (7) *matches convergence criteria for non-convex functions in (Bottou et al., 2018).*

Corollary 8 *Assume that* $0 < \alpha_k \le \frac{1}{2}$ *for all integers* $k \ge 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

- *If Assumption A1, A2, or A3 holds, then*

$$\lim_{N\to\infty} \frac{1}{\tau_N} \sum_{k=0}^{N-1} \alpha_k \mathbb{E}\left[\operatorname{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{\mathbf{b}_k} \mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2\right] = 0 \quad \textit{and} \quad \lim_{k\to\infty} \mathbb{E}[\mathbf{b}_k] = 0.$$

- *If Assumption A1 holds, then with probability* 1

$$\lim_{N\to\infty} \frac{1}{\tau_N} \sum_{k=0}^{N-1} \alpha_k \operatorname{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{\mathbf{b}_k} \mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2 = 0 \quad \textit{and} \quad \lim_{k\to\infty} \mathbf{b}_k = 0.$$

Corollary 9 *Assume that* $0 < \alpha_k = \alpha \le \frac{1}{2}$ *for all integers* $k \in [0, N-1]$.

- *If Assumption A1, A2, or A3 holds, then*

$$\frac{1}{\tau_N} \sum_{k=0}^{N-1} \alpha_k \mathbb{E}\left[\operatorname{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{\mathbf{b}_k} \mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2\right] \le c_8\sqrt{\alpha} + \frac{c_9}{N}\alpha^{-1} \;\; \textit{and} \;\; \mathbb{E}[\mathbf{b}_k] \le (c_1 + c_2)\sqrt{\alpha}, \tag{8}$$

where the constants are given by

$$c_8 = 2(c_6 + c_7) \quad \textit{and} \quad c_9 = Du + c_6 + c_7.$$

In particular, if $\alpha = O(N^{-2/3})$ *then both bounds in* (8) *are of* $O(N^{-1/3})$.

- *If Assumption A1 holds, then for any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$*:*

$$\frac{1}{N} \sum_{k=0}^{N-1} \operatorname{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{q\left(\frac{\delta}{2\alpha N+1}\right)\alpha^{1/4}} \mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2 \le 2q\left(\frac{\delta}{2\alpha N + 1}\right)\alpha^{1/4} + \frac{q\left(\frac{\delta}{2\alpha N+1}\right) + Du}{\alpha N}, \tag{9}$$

where

$$q(\hat{\delta}) = (u + 2u\ell)\left(c_3\sqrt{\log(2\hat{\delta}^{-1})} + \left(c_4 \log(2\hat{\delta}^{-1}) + c_5\right)\right)^{1/2} + (u + 2u\ell)c_7.$$

In particular if $\alpha = O(N^{-4/5})$, *then the bound in* (9) *is of* $O\left(N^{-1/5}\sqrt{\log\left(N^{1/5}\delta^{-1}\right)}\right)$.

5. Conclusion and Future Work

In this work, we gave a new convergence analysis of projected SGD where stationarity is measured by the distance of the gradient from the Goldstein subdifferential generated by the constraints. This proposed convergence measure allows direct comparison with results on unconstrained problems and does not require variance reduction techniques to achieve convergence. Our results hold in expectation for both IID and mixing data sequences, giving both asymptotic convergence and non-asymptotic bounds. In the special case of IID data sequences, we obtain asymptotic convergence almost surely and give the first non-asymptotic high probability bounds.

Future work is needed to clarify the relation of our results and prior work. In particular, faster bounds are achieved with respect to the Moreau envelope in (Davis and Drusvyatskiy, 2019), and it would be useful to understand if this is due to fundamental differences in the measure or limitations of our analytic technique. Extensions of the work include the analysis of adaptive step size rules, as commonly arise in applications, or incorporation into more complex algorithmic schemes, such as policy gradient algorithms within actor-critic reinforcement learning algorithms.

Acknowledgments

This work was supported in part by NSF ECCS 2412435 and NIH 1R01NS147767-0. The authors thank the reviewers for helpful suggestions for improving the paper.

References

Ahmet Alacaoglu and Hanbaek Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *International Conference on Machine Learning*, pages 458–489. PMLR, 2023.

Ahmet Alacaoglu and Stephen J Wright. Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 4627–4635. PMLR, 2024.

Ahmet Alacaoglu, Yura Malitsky, and Volkan Cevher. Convergence of adaptive algorithms for weakly convex constrained optimization. *arXiv preprint arXiv:2006.06650*, 2020.

Hilal Asi and John C Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019a.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019b.

MV Balashov and AA Tremba. Error bound conditions and convergence of optimization methods on smooth and proximally smooth manifolds. *Optimization*, 71(3):711–735, 2022.

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

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 2006.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30 (3):1117–1147, 2022.

Vivek S Borkar. Stochastic approximation: A dynamical systems viewpoint, 2023.

Michał Borowski and Błażej Miasojedow. Convergence of projected stochastic approximation algorithm. *arXiv preprint arXiv:2501.08256*, 2025.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

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

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Paul H Calamai and Jorge J Moré. Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116, 1987.

Huy N Chau, Chaman Kumar, Miklós Rásonyi, and Sotirios Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23:217–244, 2019.

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36:7017–7049, 2023.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in neural information processing systems*, 35:6692–6703, 2022.

Damek Davis, Dmitriy Drusvyatskiy, and Zhan Shi. Stochastic optimization over proximally smooth sets. *SIAM Journal on Optimization*, 35(1):157–179, 2025.

Qi Deng and Wenzhi Gao. Minibatch and momentum model-based methods for stochastic weakly convex optimization. *Advances in Neural Information Processing Systems*, 34:23115–23127, 2021.

Daniela di Serafino, William W Hager, Gerardo Toraldo, and Marco Viola. On the stationarity for nonlinear optimization problems with polyhedral constraints. *Mathematical Programming*, 205 (1):107–134, 2024.

John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

Alain Durmus, Aurélien Enfroy, Éric Moulines, and Gabriel Stoltz. Uniform minorization condition and convergence bounds for discretizations of kinetic langevin dynamics. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 61, pages 629–664. Institut Henri Poincaré, 2025.

Yu M Ermol'ev and VI Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2):196–215, 1998.

Yu M Ermoliev and VI Norkin. Solution of nonconvex nonsmooth stochastic optimization problems. *Cybernetics and Systems Analysis*, 39(5):701–715, 2003.

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can sgd handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025.

Wenzhi Gao and Qi Deng. Stochastic weakly convex optimization beyond lipschitz continuity. *arXiv preprint arXiv:2401.13971*, 2024.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

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

László Gerencsér, Gábor Molnár-Sáska, György Michaletzky, Gábor Tusnády, and Zsuzsanna Vágó. New methods for the statistical analysis of hidden markov models. In *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, volume 2, pages 2272–2277. IEEE, 2002.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Yue-Hong He, Gao-Xi Li, and Xian-Jun Long. Non-asymptotic analysis of hybrid spg for nonconvex stochastic composite optimization. *Journal of Optimization Theory and Applications*, 207(1):19, 2025.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Cedric Josz, Lexiao Lai, and Xiaopeng Li. Proximal random reshuffling under local lipschitz continuity. *arXiv preprint arXiv:2408.07182*, 2024.

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2021.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

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

Guanghui Lan, Tianjiao Li, and Yangyang Xu. Projected gradient methods for nonconvex and stochastic optimization: new complexities and auto-conditioned stepsizes. *arXiv preprint arXiv:2412.14291*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10):4394–4400, 2019.

Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. *Advances in Neural Information Processing Systems*, 35:33107–33119, 2022.

Qihang Lin, Runchao Ma, and Yangyang Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational optimization and applications*, 82(1):175–224, 2022.

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.

Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

GP McCormick and RA Tapia. The gradient projection method under mild differentiability conditions. *SIAM Journal on Control*, 10(1):93–98, 1972.

H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230, 2013.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nhu Nguyen and George Yin. Stochastic approximation with discontinuous dynamics, differential inclusions, and applications. *The Annals of Applied Probability*, 33(1):780–823, 2023.

Guillaume Olikier and Irène Waldspurger. Projected gradient descent accumulates at bouligand stationary points. *SIAM Journal on Optimization*, 35(2):1004–1029, 2025.

Dimitri Papadimitriou and Bang Vu. A stochastic lagrangian-based method for nonconvex optimization with nonlinear constraints. 2025.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1244–1251. IEEE, 2016.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

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

Clément W Royer, Oumaima Sohab, and Luis Nunes Vicente. Full-low evaluation methods for bound and linearly constrained derivative-free optimization. *Computational Optimization and Applications*, 89(2):279–315, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

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

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Yue Xie, Jiawen Bi, and Hongcheng Liu. On tackling high-dimensional nonconvex stochastic optimization via stochastic first-order methods with non-smooth proximal terms and variance reduction. *arXiv preprint arXiv:2509.13992*, 2025.

Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020a.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020b.

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25, 2023.

Yuping Zheng and Andrew Lamperski. Constrained langevin algorithms with l-mixing external random variables. *Advances in Neural Information Processing Systems*, 35:20511–20521, 2022.

Yuping Zheng and Andrew Lamperski. Non-asymptotic analysis of classical spectrum estimators with l-mixing time-series data. In *2025 American Control Conference (ACC)*, pages 1896–1901. IEEE, 2025a.

Yuping Zheng and Andrew Lamperski. Non-asymptotic analysis of classical spectrum estimators for l-mixing time-series data with unknown means. *arXiv preprint arXiv:2504.00217*, 2025b.

Daoli Zhu, Lei Zhao, and Shuzhong Zhang. A unified analysis on the subgradient upper bounds for the subgradient methods minimizing composite nonconvex, nonsmooth and non-lipschitz functions. *arXiv preprint arXiv:2308.16362*, 2023.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Appendix A. Convergence Analysis via Intermediate Processes

In this section, we introduce some intermediate processes to bound $\mathbb{E}\left[\left\|\mathbf{x}_t^A - \mathbf{x}_t^C\right\|\right]$. The bound of $\mathbb{E}\left[\left\|\mathbf{x}_t^A - \mathbf{x}_t^C\right\|\right]$ and all the supporting lemmas are shown in Appendix C. The ideas of using the intermediate processes and the proofs on the quantitative bounds are similar to those in (Lamperski, 2021; Zheng and Lamperski, 2022).

Other than the *continuous* time process $\mathbf{x}_t^C$ defined in (4), we further introduce the *mean* process, $\mathbf{x}_{\tau_k}^M$, and the *discretized* process, $\mathbf{x}_{\tau_k}^D$ with $\mathbf{x}_{\tau_{k_0}}^A = \mathbf{x}_{\tau_{k_0}}^M = \mathbf{x}_{\tau_{k_0}}^C = \mathbf{x}_{\tau_{k_0}}^D$ where integers $k_0 \le k$:

$$\mathbf{x}_{\tau_{k+1}}^M = \Pi_{\mathcal{X}}\left(\mathbf{x}_{\tau_k}^M - \alpha_k \nabla \bar{f}(\mathbf{x}_{\tau_k}^M)\right). \tag{10}$$

The discretized processes $\mathbf{x}^D_{\tau_k}$ uses the form of Skorokhod solution introduced in Appendix E.

Here is some preliminary to define $\mathbf{x}^D_{\tau_k}$. To enable the construction of Skorokhod problem, which is key to prove Lemma 14 relying on Lemma 2.2 (i) in (Tanaka, 1979), we first show that the alternative representation of projected ODE (4):

$$\frac{d}{dt}\mathbf{x}^C_t = -\nabla \bar{f}(\mathbf{x}^C_t) - \mathbf{v}^C_t \tag{11}$$

where $\mathbf{v}^C_t \in \mathcal{N}_{\mathcal{X}}(\mathbf{x}^C_t)$.

The projected ODE (11) in the context of constrained stochastic approximation can be found in (Kushner and Yin, 2003) and these two equivalent forms of projected ODE are also mentioned in (Borowski and Miasojedow, 2025) but there was no proof. The equivalence of (4) and (11) follows from the Moreau decomposition (e.g. Hiriart-Urruty and Lemaréchal (2004)), which implies that that for any vector $g \in \mathbb{R}^n$, $\Pi_{T_{\mathcal{X}}(x)}(g) = g - \Pi_{\mathcal{N}_{\mathcal{X}}(x)}(g)$. In particular, $\mathbf{v}^C_t = \frac{\Pi_{\mathcal{N}(x)}(-\nabla \bar{f}(\mathbf{x}^C_t))}{\|\Pi_{\mathcal{N}(x)}(-\nabla \bar{f}(\mathbf{x}^C_t))\|}$ when $\mathbf{x}^C_t \in \partial\mathcal{X}$.

Then, the projected ODE (11) can be written as

$$d\mathbf{x}^C_t = -\nabla \bar{f}(\mathbf{x}^C_t)dt - \mathbf{v}^C_t d\boldsymbol{\mu}^C(t). \tag{12}$$

Here, $-\int_0^t \mathbf{v}^C_t d\boldsymbol{\mu}^C(t)$ is a bounded variation reflection process that keeps $\mathbf{x}^C_t \in \mathcal{X}$ for all $t \in [0, \tau_N]$, as long as $\mathbf{x}^C_0 \in \mathcal{X}$. The measure, $\boldsymbol{\mu}^C$, is non-negative and supported on $\{s|\mathbf{x}^C_s \in \partial\mathcal{X}\}$, while $\mathbf{v}^C_s \in \mathcal{N}_{\mathcal{X}}(\mathbf{x}^C_s)$. With these conditions on (12), $\mathbf{v}^C_t d\boldsymbol{\mu}^C(t)$ is uniquely defined and $\mathbf{x}^C$ is the unique solution to the Skorokhod problem for a process defined below:

$$\mathbf{y}^C_t = \mathbf{x}^C_0 - \int_0^t \nabla \bar{f}(\mathbf{x}^C_s)ds. \tag{13}$$

More details on Skorokhod problems are given in Appendix E.

In the following, we denote the Skorokhod solution for given trajectory, $\mathbf{y}$, by $\mathcal{S}(\mathbf{y})$.

Let $\mathbf{y}^D_t = \mathbf{y}^C_{\tau_k}$ for all $t \in [\tau_k, \tau_{k+1})$. Such discretization operator is denoted by $\mathcal{D}(\cdot)$. Then, we define $\mathbf{x}^D = \mathcal{S}(\mathcal{D}(\mathbf{y}^C))$, i.e. $\mathbf{x}^D_t = \mathbf{y}^C_{\tau_k} + \boldsymbol{\phi}^D_t$ for all $t \in [\tau_k, \tau_{k+1})$, where $\boldsymbol{\phi}^D_t = -\int_0^t \mathbf{v}^C_t d\boldsymbol{\mu}^C(t)$.

Therefore, we have

$$\begin{aligned}\mathbf{x}^D_{\tau_{k+1}} &= \Pi_{\mathcal{X}}\left(\mathbf{x}^D_{\tau_k} + \mathbf{y}^C_{\tau_{k+1}} - \mathbf{y}^C_{\tau_k}\right) \\ &= \Pi_{\mathcal{X}}\left(\mathbf{x}^D_{\tau_k} - \int_{\tau_k}^{\tau_{k+1}} \nabla \bar{f}(\mathbf{x}^C_t)dt\right).\end{aligned} \tag{14}$$

The intermediate processes are used to bound the individual terms from the following triangle inequality:

$$\|\mathbf{x}^A_t - \mathbf{x}^C_t\| \le \|\mathbf{x}^A_{\tau_k} - \mathbf{x}^M_{\tau_k}\| + \|\mathbf{x}^M_{\tau_k} - \mathbf{x}^D_{\tau_k}\| + \|\mathbf{x}^D_{\tau_k} - \mathbf{x}^C_{\tau_k}\| + \|\mathbf{x}^C_t - \mathbf{x}^C_{\tau_k}\|.$$

Appendix B. Proof of Main Results

The section presents the proofs of the main results.

Proof of Theorem 6

Firstly, we have

$$\begin{aligned}\bar{f}(\mathbf{x}_{s_{i+1}}^C) - \bar{f}(\mathbf{x}_{s_i}^C) &= \int_0^{\tau_N} \frac{d}{dt}\bar{f}(\mathbf{x}_t^C)dt \\ &= \int_0^{\tau_N} \nabla \bar{f}(\mathbf{x}_t^C)^\top \Pi_{T_\mathcal{X}(\mathbf{x}_t^C)}\left(-\nabla \bar{f}(\mathbf{x}_t^C)\right) dt \\ &= -\int_0^{\tau_N} \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^C)}\left(-\nabla \bar{f}(\mathbf{x}_t^C)\right)\right\|^2 dt \end{aligned} \tag{15}$$

where the first and second equalities use the fundamental theorem of calculus and the chain rule respectively and the last equality uses Lemma 19 in Appendix D.

For one time interval, $[s_i, s_{i+1}]$, we have the following decomposition:

$$\begin{aligned}\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A) &= \bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A) - \left(\bar{f}(\mathbf{x}_{s_{i+1}}^{C_i}) - \bar{f}(\mathbf{x}_{s_i}^{C_i})\right) + \left(\bar{f}(\mathbf{x}_{s_{i+1}}^{C_i}) - \bar{f}(\mathbf{x}_{s_i}^{C_i})\right) \\ &= \bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_{i+1}}^{C_i}) - \int_{s_i}^{s_{i+1}} \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^{C_i})\right)\right\|^2 dt\end{aligned}$$

where the second equality uses $\mathbf{x}_{s_i}^A = \mathbf{x}_{s_i}^{C_i}$ and (15).

Adding and subtracting $\Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)$ inside the norm of the second term on the RHS and rearranging gives

$$\begin{aligned}&\int_{s_i}^{s_{i+1}} \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right) + \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^{C_i})\right) - \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\|^2 dt \\ &= \bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_{i+1}}^{C_i}) - \left(\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A)\right) \\ \Rightarrow &\int_{s_i}^{s_{i+1}} \left(\left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\| - \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^{C_i})\right) - \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\|\right)^2 dt \\ &\leq \bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_{i+1}}^{C_i}) - \left(\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A)\right) \\ \Rightarrow &\int_{s_i}^{s_{i+1}} \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\|^2 dt \\ &\leq \left(\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_{i+1}}^{C_i}) - \left(\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A)\right)\right) \\ &+ 2\int_{s_i}^{s_{i+1}} \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\| \left\| \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^{C_i})\right) - \Pi_{T_\mathcal{X}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\| dt \\ &\leq u\left\|\mathbf{x}_{s_{i+1}}^A - \mathbf{x}_{s_{i+1}}^{C_i}\right\| - \left(\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A)\right) + 2u\ell \int_{s_i}^{s_{i+1}} \left\|\mathbf{x}_t^A - \mathbf{x}_t^{C_i}\right\| dt\end{aligned}$$

where the right arrow uses the fact that $(\|a\| - \|b\|)^2 \leq \|a+b\|^2$ for all $a, b \in \mathbb{R}^n$. The last inequality uses the fact that $\bar{f}$ is u-Lipschitz and $\nabla \bar{f}$ is ℓ-Lipschitz as well as the non-expansiveness of the convex projection.

Summing over $\chi(N)+1$ terms gives

$$\begin{aligned}
&\sum_{i=0}^{\chi(N)} \int_{s_i}^{s_{i+1}} \left\|\Pi_{T_{\mathcal{X}}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\|^2 dt \\
&\le u\sum_{i=0}^{\chi(N)} \left\|\mathbf{x}_{s_{i+1}}^A - \mathbf{x}_{s_{i+1}}^{C_i}\right\| - \sum_{i=0}^{\chi(N)} \left(\bar{f}(\mathbf{x}_{s_{i+1}}^A) - \bar{f}(\mathbf{x}_{s_i}^A)\right) + 2u\ell \sum_{i=0}^{\chi(N)} \int_{s_i}^{s_{i+1}} \left\|\mathbf{x}_t^A - \mathbf{x}_t^{C_i}\right\| dt \\
&= u\sum_{i=0}^{\chi(N)} \left\|\mathbf{x}_{s_{i+1}}^A - \mathbf{x}_{s_{i+1}}^{C_i}\right\| + \left(\bar{f}(\mathbf{x}_0^A) - \bar{f}(\mathbf{x}_{\tau_N}^A)\right) + 2u\ell \sum_{i=0}^{\chi(N)} \int_{s_i}^{s_{i+1}} \left\|\mathbf{x}_t^A - \mathbf{x}_t^{C_i}\right\| dt \\
&\le u\sum_{i=0}^{\chi(N)} \left\|\mathbf{x}_{s_{i+1}}^A - \mathbf{x}_{s_{i+1}}^{C_i}\right\| + 2u\ell \sum_{i=0}^{\chi(N)} \int_{s_i}^{s_{i+1}} \left\|\mathbf{x}_t^A - \mathbf{x}_t^{C_i}\right\| dt + osc(\bar{f})
\end{aligned} \tag{16}$$

where the equality uses a telescoping sum.

Now, we examine the expected value case, which holds for all of the assumptions.

Lemma 10 gives the bound below

$$\mathbb{E}\left[\left\|\mathbf{x}_{s_{i+1}}^A - \mathbf{x}_{s_{i+1}}^{C_i}\right\|\right] \le g_1(s_{i+1}-s_i)\sqrt{\sum_{\{j|s_i\le\tau_j<s_{i+1}\}} \alpha_j^2} + g_2(s_{i+1}-s_i) \max_{\{j|s_i\le\tau_j\le s_{i+1}\}} \sqrt{\alpha_j}$$

where $g_1(q)=\sigma e^{\ell q}$ under Assumption A2, $g_1(q)=2\ell\Psi_2(\mathbf{z})e^{\ell q}$ under Assumption A3 and $g_2(q) = e^{\ell q}\left(u+\sqrt{2r^{-1}u\left(qDu+D^2\right)}\right)$.

Therefore, we have

$$\begin{aligned}
&\frac{1}{\tau_N}\sum_{i=0}^{\chi(N)} \mathbb{E}\left[\int_{s_i}^{s_{i+1}} \left\|\Pi_{T_{\mathcal{X}}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\|^2 dt\right] \\
&\le \frac{1}{\tau_N}\left(\sum_{i=0}^{\chi(N)} \left(u\mathbb{E}\left[\left\|\mathbf{x}_{s_{i+1}}^A - \mathbf{x}_{s_{i+1}}^{C_i}\right\|\right] + 2u\ell(s_{i+1}-s_i)\max_{j\in[s_i,s_{i+1}]}\mathbb{E}\left[\left\|\mathbf{x}_j^A - \mathbf{x}_j^{C_i}\right\|\right]\right) + osc(\bar{f})\right) \\
&\le \frac{1}{\tau_N}\left(\sum_{i=0}^{\chi(N)} \left(u+2u\ell(s_{i+1}-s_i)\right)\left(g_1(s_{i+1}-s_i)\sqrt{\sum_{\{j|s_i\le\tau_j<s_{i+1}\}}\alpha_j^2}\right.\right. \\
&\qquad\qquad \left.\left. + g_2(s_{i+1}-s_i)\max_{\{j|s_i\le\tau_j\le s_{i+1}\}}\sqrt{\alpha_j}\right) + osc(\bar{f})\right).
\end{aligned} \tag{17}$$

Note that if $t\in[s_i,s_{i+1})$, we must have that $\mathbf{x}_t^A = \mathbf{x}_{\tau_k}^A = \mathbf{x}_k$ for some integer $\mathcal{K}(i)\le k<\mathcal{K}(i+1)$. In this case, $\|\mathbf{x}_t^A - \mathbf{x}_t^{C_i}\| \le \mathbf{b}_k$, by the definition of $\mathbf{b}_k$.

Lemma 20, followed by the definitions of the convex projection and the expression for the Goldstein subdifferential give

$$\begin{aligned}\left\|\Pi_{T_{\mathcal{X}}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\| &= \left\|-\nabla \bar{f}(\mathbf{x}_t^A) - \Pi_{\mathcal{N}_{\mathcal{X}}(\mathbf{x}_t^{C_i})}(-\nabla \bar{f}(\mathbf{x}_t^A)\right\| \\ &= \mathrm{dist}\left(-\nabla \bar{f}(\mathbf{x}_t^A), \mathcal{N}_{\mathcal{X}}(\mathbf{x}_t^{C_i})\right) \\ &\geq \mathrm{dist}\left(-\nabla \bar{f}(\mathbf{x}_t^A), \overline{\partial}_{\|\mathbf{x}_t^A - \mathbf{x}_t^{C_i}\|}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_t^A)\right) \\ &\geq \mathrm{dist}(-\nabla \bar{f}(\mathbf{x}_k), \overline{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)).\end{aligned}$$

It then follows that

$$\int_{\tau_k}^{\tau_{k+1}} \left\|\Pi_{T_{\mathcal{X}}(\mathbf{x}_t^{C_i})}\left(-\nabla \bar{f}(\mathbf{x}_t^A)\right)\right\|^2 dt \geq \alpha_k \mathrm{dist}(-\nabla \bar{f}(\mathbf{x}_k), \overline{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k))^2.$$

Plugging this lower bound into the integrals on the left of (17) and using that $s_{i+1} - s_i \leq 1$ gives the bound on the expected value.

Now, we turn to the special case that Assumption A1 holds, and give a bound in high probability.

Plugging the definition of $\mathbf{b}_k$ into (16) and using the bound on the Goldstein subdifferentials above gives

$$\sum_{k=0}^{N-1} \alpha_k \mathrm{dist}(-\nabla \bar{f}(\mathbf{x}_k), \overline{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k))^2 \leq (u + 2u\ell) \sum_{i=0}^{\chi(N)} \max_{k \in [\mathcal{K}(i), \mathcal{K}(i+1)-1]} \mathbf{b}_k + \mathrm{osc}(\bar{f}). \quad (18)$$

Applying Lemma 4, and using a union bound gives that with probability at least $1 - \sum_{i=0}^{\chi(N)} \delta_i$, we have $\mathbf{b}_k \leq h_{\zeta(k)}(\delta_{\zeta(k)})$ for all $k = 0, \ldots, N-1$. Recall that $\zeta(k)$ was defined in Section 4.1. Using the bound $\mathbf{b}_k \leq h_{\zeta(k)}(\delta_{\zeta(k)})$ on the left and right now gives the result. ■

Proof of Corollary 8

Firstly, we know $s_{i+1} - s_i \geq \frac{1}{2}$ for all $i \in [0, \chi(N) - 1]$. Then $\tau_N > s_{\chi(N)} = \sum_{i=0}^{\chi(N)-1}(s_{i+1} - s_i) \geq \frac{1}{2}\chi(N)$. Furthermore, from the condition that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we have $\lim_{m\to\infty} \sum_{k=m}^{\infty} \alpha_k^2 = 0$. Therefore, $\lim_{i\to\infty} \sum_{\{j|s_i \leq \tau_j < s_{i+1}\}} \alpha_j^2 = 0$. This implies that if we choose $\epsilon > 0$, there exists $i_1 \in \mathbb{N}$ such that for all $i \geq i_1$, $\sum_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \alpha_j^2 \leq \epsilon^2$, so $\sqrt{\sum_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \alpha_j^2} \leq \epsilon$. Since $\alpha_j \leq \sqrt{\sum_{\{j|s_i \leq \tau_j \leq s_i\}} \alpha_j^2}$ for all j such that $s_i \leq \tau_j \leq s_{i+1}$, then $\alpha_j \leq \epsilon$ for all j such that $s_i \leq \tau_j \leq s_{i+1}$ and $i \geq i_1$. Therefore, $\max_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \sqrt{\alpha_j} \leq \sqrt{\epsilon}$ for $i \geq i_1$.

Without loss of generality, we can ignore the constant factors, since the right of (17) is arbitrarily small, if in only if the following quantity is arbitrarily small:

$$\begin{aligned}&\frac{\sum_{i=0}^{\chi(N)} \left\{\sqrt{\sum_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \alpha_j^2} + \max_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \sqrt{\alpha_j}\right\}}{s_{\chi(N)}} \\ \leq &\frac{\sum_{i=0}^{i_1} \left\{\sqrt{\sum_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \alpha_j^2} + \max_{\{j|s_i \leq \tau_j \leq s_{i+1}\}} \sqrt{\alpha_j}\right\}}{\frac{1}{2}\chi(N)} + \frac{(\epsilon + \sqrt{\epsilon})(\chi(N) - i_1)}{\frac{1}{2}(\chi(N) - i_1)}.\end{aligned}$$

The first term converges to zero as $\chi(N) \to \infty$ (i.e. $N \to \infty$) and the second term is $2(\epsilon + \sqrt{\epsilon})$, which is arbitrarily small. Therefore, we obtain asymptotic convergence for the expected value.

Now consider the case that Assumption A1 holds. Equation 18 implies that:

$$\frac{1}{\tau_N}\sum_{k=0}^{N-1}\alpha_k \mathrm{dist}(-\nabla \bar{f}(\mathbf{x}_k), \bar{\partial}_{\mathbf{b}_k}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k))^2 \leq \frac{(u+2u\ell)}{\tau_N}\sum_{i=0}^{\chi(N)}\max_{k\in[\mathcal{K}(i),\mathcal{K}(i+1)-1]}\mathbf{b}_k + \frac{\mathrm{osc}(\bar{f})}{\tau_N}.$$

Using again that $\tau_N \geq \frac{1}{2}\chi(N)$, it suffices to show that

$$\lim_{N\to\infty}\frac{\sum_{i=0}^{\chi(N)}\max_{k\in[\mathcal{K}(i),\mathcal{K}(i+1)-1]}\mathbf{b}_k}{\chi(N)}$$

Propostion 5 implies that there is an integer i_1 such that if $i \geq i_1$, then $\max_{k\in[\mathcal{K}(i),\mathcal{K}(i+1)-1]}\mathbf{b}_k \leq h_i(\delta_i)$, where $\delta_i = \frac{\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1}\alpha_j^2}{\sum_{k=0}^{\infty}\alpha_k^2}$.

Furthermore, Proposition 5 implies that $h_i(\delta_i) \to 0$. In particular, given any $\epsilon > 0$, there is number $i_2 \geq i_1$ such that if $i \geq i_2$, then $h_i(\delta_i) \leq \epsilon$. In particular, for all $i \geq i_2$, we have $\max_{k\in[\mathcal{K}(i),\mathcal{K}(i+1)-1]}\mathbf{b}_k \leq \epsilon$. So, similar to the expected value case, we have:

$$\frac{\sum_{i=0}^{\chi(N)}\max_{k\in[\mathcal{K}(i),\mathcal{K}(i+1)-1]}\mathbf{b}_k}{\chi(N)} \leq \frac{\sum_{i=0}^{i_2}\max_{k\in[\mathcal{K}(i),\mathcal{K}(i+1)-1]}\mathbf{b}_k}{\chi(N)} + \epsilon\frac{\chi(N)-i_2}{\chi(N)}.$$

The first term on the right converges to 0, while the second is arbitrarily small.

Additionally, Proposition 5 implies that $\mathbf{b}_k \to 0$ with probability 1. ■

Proof of Corollary 9

For a constant step size, the construction in Section 4.1 reduces to: $s_{i+1} - s_i = \alpha\lfloor\frac{1}{\alpha}\rfloor \leq 1$. Thus, we have $\chi(N) + 1 = \lceil\frac{N}{\lfloor 1/\alpha\rfloor}\rceil < \frac{N}{\lfloor 1/\alpha\rfloor} + 1$. If $\alpha \leq \frac{1}{2}$, $\alpha\lfloor\frac{1}{\alpha}\rfloor > \alpha(\frac{1}{\alpha} - 1) > \frac{1}{2}$, so $\lfloor\frac{1}{\alpha}\rfloor > \frac{1}{2\alpha}$ and $\chi(N) < 2\alpha N$. Therefore, the general bound in (17) can be simplified as

$$\begin{aligned}
&\frac{1}{\tau_N}\sum_{i=0}^{\chi(N)}\mathbb{E}\left[\int_{s_i}^{s_{i+1}}\left\|\Pi_{T_{\mathcal{X}}(\mathbf{x}_t^{C_i})}\left(-\nabla\bar{f}(\mathbf{x}_t^A)\right)\right\|^2 dt\right] \\
&\leq \frac{1}{\alpha N}\left(\left(\frac{N}{\lfloor 1/\alpha\rfloor} + 1\right)(u+2u\ell)\left(g_1(1)\sqrt{\lfloor\frac{1}{\alpha}\rfloor\alpha^2} + g_2(1)\sqrt{\alpha}\right) + osc(\bar{f})\right) \\
&< 2(1+2\ell)u(g_1(1)+g_2(1))\alpha^{\frac{1}{2}} + \frac{(osc(\bar{f}) + (1+2\ell)u(g_1(1)+g_2(1)))}{\alpha N}
\end{aligned}$$

where functions g_1 and g_2 were defined in the proof of Theorem 6 and the last inequality holds because $\alpha^{-1/2} < \alpha^{-1}$ for any $0 < \alpha < 1$.

To derive the bound on $\mathbf{b}_k$, note that $\mathcal{K}(i+1) - \mathcal{K}(i) = \lfloor\frac{1}{\alpha}\rfloor \leq \frac{1}{\alpha}$. Similar to above, we have $\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1}\alpha_j^2 \leq \alpha$. The bound then follows from Lemma 4.

For the high probability bound, we apply the Theorem 6 with $\delta_i = \frac{\delta}{\chi(N)+1}$. Then, we have with probability at least δ

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{dist}\left(-\nabla\bar{f}(\mathbf{x}_k), \bar{\partial}_{h_{\zeta(k)}\left(\frac{\delta}{\chi(N)+1}\right)}\mathcal{I}_{\mathcal{X}}(\mathbf{x}_k)\right)^2 \leq \frac{u+2u\ell}{\alpha N}\sum_{i=0}^{\chi(N)}h_i\left(\frac{\delta}{\chi(N)+1}\right) + \frac{Du}{\alpha N}.$$

Similar to the bound on $\mathbf{b}_k$ above, we use that $\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2 \leq \alpha$. So, we can bound:

$$h_i(\delta_i) \leq \left(c_3 \sqrt{\alpha \log(2\delta_i^{-1})} + \left(c_4 \log(2\delta_i^{-1}) + c_5 \right) \alpha \right)^{1/2} + c_7 \sqrt{\alpha}$$
$$\leq \underbrace{\left(\left(c_3 \sqrt{\log(2\delta_i^{-1})} + \left(c_4 \log(2\delta_i^{-1}) + c_5 \right) \right)^{1/2} + c_7 \right)}_{q(\delta_i)/(u+2u\ell)} \alpha^{1/4}.$$

Using again that $\chi(N) + 1 \leq 2\alpha N + 1$ gives:

$$\frac{1}{N} \sum_{k=0}^{N-1} \operatorname{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \overline{\partial}_{q(\frac{\delta}{2\alpha N+1})\alpha^{1/4}} \mathcal{I}_{\mathcal{X}}(\mathbf{x}_k) \right)^2$$
$$\leq \frac{1}{N} \sum_{k=0}^{N-1} \operatorname{dist}\left(-\nabla \bar{f}(\mathbf{x}_k), \overline{\partial}_{q\left(\frac{\delta}{\chi(N)+1}\right)\alpha^{1/4}} \mathcal{I}_{\mathcal{X}}(\mathbf{x}_k) \right)^2$$
$$\leq 2q\left(\frac{\delta}{\chi(N)+1} \right) \alpha^{1/4} + \frac{q\left(\frac{\delta}{\chi(N)+1} \right) + Du}{\alpha N}$$
$$\leq 2q\left(\frac{\delta}{2\alpha N+1} \right) \alpha^{1/4} + \frac{q\left(\frac{\delta}{2\alpha N+1} \right) + Du}{\alpha N}.$$

■

Appendix C. Supporting Lemmas

This section collects supporting lemmas which bound a series of intermediate processes.

The following lemma is directly used to prove Theorem 6.

Lemma 10 *Assume* $\mathbf{x}_{\tau_{k_0}}^A = \mathbf{x}_{\tau_{k_0}}^C \in \mathcal{X}$ *and* $\alpha_k < \frac{1}{2}$ *for all* $k \in \mathbb{N}$, $k \geq k_0$ *and* $t \in [\tau_k, \tau_{k+1})$*, the following bounds hold*

(i) If $\mathbf{z}_k$ *satisfies Assumption A2, then*

$$\mathbb{E}\left[\left\| \mathbf{x}_t^A - \mathbf{x}_t^C \right\| \right] \leq \sigma e^{\ell(\tau_k - \tau_{k_0})} \sqrt{\sum_{j=k_0}^{k-1} \alpha_j^2}$$
$$+ e^{\ell(\tau_k - \tau_{k_0})} \left(u + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2 \right)} \right) \max_{j \in [k_0, k]} \sqrt{\alpha_j}.$$

(ii) If $\mathbf{z}_k$ *satisfies Assumption A3, then*

$$\mathbb{E}\left[\left\| \mathbf{x}_t^A - \mathbf{x}_t^C \right\| \right] \leq 2\ell \Psi_2(\mathbf{z}) e^{\ell(\tau_k - \tau_{k_0})} \sqrt{\sum_{j=k_0}^{k-1} \alpha_j^2}$$
$$+ e^{\ell(\tau_k - \tau_{k_0})} \left(u + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2 \right)} \right) \max_{j \in [k_0, k]} \sqrt{\alpha_j}.$$

(iii) If $\mathbf{z}_k$ satisfies Assumption A1 and $\epsilon > 0$, then with probability at least $(1 - e^{-\epsilon})^2$,

$$\sup_{s \in [\tau_{k_0}, \tau_k)} \left\| \mathbf{x}_s^A - \mathbf{x}_s^C \right\| \leq$$

$$e^{\ell(\tau_k - \tau_{k_0})} \left(2\sqrt{2}\hat{\sigma} D \sqrt{\epsilon} \sqrt{\sum_{j=k_0}^{k-1} \alpha_j^2} + \left(4\hat{\sigma}^2 \epsilon + \hat{\sigma}^2 + n\hat{\sigma}^2\right) \sum_{j=k_0}^{k-1} \alpha_j^2 \right)^{1/2}$$

$$+ e^{\ell(\tau_k - \tau_{k_0})} \left(u + \sqrt{2r^{-1} u \left((\tau_k - \tau_{k_0}) D u + D^2 \right)} \right) \max_{j \in [k_0, k]} \sqrt{\alpha_j}.$$

Before proving Lemma 10, we will show how it can be used to prove Lemma 4 and Proposition 5 from the main text.

Proof of Lemma 4

Recall that Assumption A1 implies Assumption A2 with $\sigma = \sqrt{n}\hat{\sigma}$. Let $k_0 = \mathcal{K}(i)$ and assume that $k \leq \mathcal{K}(i+1) - 1$. In this case, $\tau_k - \tau_{k_0} \leq s_{i+1} - s_i \leq 1$. So, we can plug the upper bound of 1 into all of the $\tau_k - \tau_{k_0}$ terms in Lemma 10. Furthermore, $k \leq \mathcal{K}(i+1) - 1$ implies that

$$\sum_{j=k_0}^{k-1} \alpha_j^2 \leq \sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2 \quad \text{and} \quad \max_{j \in [k_0, k]} \sqrt{\alpha_j} \leq \max_{j \in [\mathcal{K}(i), \mathcal{K}(i+1))} \sqrt{\alpha_j}.$$

Plugging these bounds into Lemma 10 gives the bounds in expectation.

To get the bounds in high probability, we do the substitutions above. Furtheremore, note that $(1 - e^{-\epsilon})^2 \geq 1 - 2e^{-\epsilon}$. Set $\delta = 2e^{-\epsilon}$, which gives $\epsilon = \log(2\delta^{-1})$. Substituting this value for ϵ gives result. ∎

Proof of Proposition 5

By Lemma 4, the event $\max_{k \in [\mathcal{K}(i), \mathcal{K}(i+1))} \mathbf{b}_k > h_i(\delta_i)$ occurs with probability at most δ_i. By construction,

$$\sum_{i=0}^{\infty} \delta_i = 1,$$

So, the Borel-Cantelli Lemma implies that $\max_{k \in [\mathcal{K}(i), \mathcal{K}(i+1))} \mathbf{b}_k > h_i(\delta_i)$ can occur at most finitely many times.

To complete the proof, it suffices to show that when $i \to \infty$, $h_i(\delta_i) \to 0$. Note that $\alpha_k \to 0$ and $\mathcal{K}(i) \to \infty$. Thus,

$$\lim_{i \to \infty} \max_{k \in [\mathcal{K}(i), \mathcal{K}(i+1))} \sqrt{\alpha_k} = 0.$$

Similarly, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ and $\mathcal{K}(i) \to \infty$ implies that

$$\lim_{i \to \infty} \sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2 = 0.$$

Thus, to show that $h_i(\delta_i) = 0$, using $\delta_i = \frac{\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2}{\sum_{k=0}^{\infty} \alpha_k^2}$, it suffices to show that

$$\log\left(\frac{1}{\sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2} \right) \sum_{j=\mathcal{K}(i)}^{\mathcal{K}(i+1)-1} \alpha_j^2 \to 0.$$

This is now a special case of $\lim_{t\downarrow 0} t\log(t^{-1}) = 0$. ■

The following lemmas support the proof of Lemma 10.

Lemma 11 *Assume $\mathbf{x}^A_{\tau_{k_0}} = \mathbf{x}^M_{\tau_{k_0}} \in \mathcal{X}$ and $\mathbf{z}_k$ satisfies assumption A2, for all $k \in \mathbb{N}$, $k \geq k_0$, the following bound holds:*

$$\mathbb{E}\left[\|\mathbf{x}^A_{\tau_k} - \mathbf{x}^M_{\tau_k}\|\right] \leq \sigma e^{\ell(\tau_k - \tau_{k_0})}\sqrt{\sum_{s=k_0}^{k-1}\alpha_s^2}.$$

Lemma 12 *Assume $\mathbf{x}^A_{\tau_{k_0}} = \mathbf{x}^M_{\tau_{k_0}} \in \mathcal{X}$ and $\mathbf{z}_k$ satisfies assumption A3, for all $k \in \mathbb{N}$, $k \geq k_0$, the following bound holds:*

$$\mathbb{E}\left[\left\|\mathbf{x}^A_{\tau_k} - \mathbf{x}^M_{\tau_k}\right\|\right] \leq 2\ell\Psi_2(\mathbf{z})e^{\ell(\tau_k - \tau_{k_0})}\sqrt{\sum_{s=k_0}^{k-1}\alpha_s^2}.$$

Lemma 13 *Assume $\mathbf{x}^A_{\tau_{k_0}} = \mathbf{x}^M_{\tau_{k_0}} \in \mathcal{X}$, $\mathbf{z}_k$ satisfies assumption A1, $\alpha_k \leq \frac{1}{2}$, for all $k \in \mathbb{N}$, $k \geq k_0$ and $\epsilon > 0$, then with probability at least $(1 - e^{-\epsilon})^2$, the following bound holds:*

$$\max_{s\in[k_0,k]}\left\|\mathbf{x}^A_{\tau_s} - \mathbf{x}^M_{\tau_s}\right\| \leq e^{\ell(\tau_k - \tau_{k_0})}\left(2\sqrt{2}\hat{\sigma}D\sqrt{\epsilon}\sqrt{\sum_{j=k_0}^{k-1}\alpha_j^2} + \left(4\hat{\sigma}^2\epsilon + \hat{\sigma}^2 + n\hat{\sigma}^2\right)\sum_{j=k_0}^{k-1}\alpha_j^2\right)^{1/2}.$$

Lemma 14 *Assume that $\mathbf{x}^C_{\tau_{k_0}} = \mathbf{x}^D_{\tau_{k_0}}$, for all $k \in \mathbb{N}$, $k \geq k_0$, the following bound holds*

$$\left\|\mathbf{x}^C_{\tau_k} - \mathbf{x}^D_{\tau_k}\right\| \leq \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right)\max_{j\in[k_0,k]}\alpha_j}.$$

Lemma 15 *For all $t \in [\tau_k, \tau_{k+1})$, the following bound holds*

$$\left\|\mathbf{x}^C_t - \mathbf{x}^C_{\tau_k}\right\| \leq \alpha_k u.$$

Lemma 16 *Assume $\mathbf{x}^C_{\tau_{k_0}} = \mathbf{x}^D_{\tau_{k_0}} \in \mathcal{X}$, for all $t \in [\tau_k, \tau_{k+1})$ where $k \in \mathbb{N}$, $k \geq k_0$, the following bound holds*

$$\left\|\mathbf{x}^C_t - \mathbf{x}^D_t\right\| \leq \alpha_k u + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right)\max_{j\in[k_0,k]}\alpha_j}.$$

Lemma 17 *Assume $\mathbf{x}^M_{\tau_{k_0}} = \mathbf{x}^D_{\tau_{k_0}} \in \mathcal{X}$, $\alpha_k \leq \frac{1}{2}$ for all $k \in \mathbb{N}$, $k \geq k_0$, the following bound holds*

$$\left\|\mathbf{x}^M_{\tau_k} - \mathbf{x}^D_{\tau_k}\right\| \leq \left(e^{\ell(\tau_k - \tau_{k_0})} - 1\right)\max_{s\in[k_0,k]}\sqrt{\alpha_s}\left(u + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right)}\right).$$

Proof of Lemma 10

For $t \in [\tau_k, \tau_{k+1})$, $\mathbf{x}^A_t = \mathbf{x}^A_{\tau_k}$, then the triangle inequality gives

$$\|\mathbf{x}^A_t - \mathbf{x}^C_t\| \leq \|\mathbf{x}^A_{\tau_k} - \mathbf{x}^M_{\tau_k}\| + \|\mathbf{x}^M_{\tau_k} - \mathbf{x}^D_{\tau_k}\| + \|\mathbf{x}^D_{\tau_k} - \mathbf{x}^C_{\tau_k}\| + \|\mathbf{x}^C_t - \mathbf{x}^C_{\tau_k}\|.$$

For part (i), under Assumption A2, combininig Lemma 11, Lemma 17, Lemma 14 and Lemma 15 gives

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}_t^A - \mathbf{x}_t^C\right\|\right] \\
&\leq \sigma e^{\ell(-\tau_{k_0}+\tau_k)} \sqrt{\sum_{j=k_0}^{k-1} \alpha_j^2} + \left(e^{\ell(\tau_k-\tau_{k_0})} - 1\right) \max_{j\in[k_0,k]} \sqrt{\alpha_j} \left(u + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right)}\right) \\
&\qquad + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right) \max_{j\in[k_0,k]} \alpha_j} + \alpha_k u \\
&\leq \sigma e^{\ell(\tau_k-\tau_{k_0})} \sqrt{\sum_{j=0}^{k-1} \alpha_j^2} + e^{\ell(\tau_k-\tau_{k_0})} \left(u + \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right)}\right) \max_{j\in[k_0,k]} \sqrt{\alpha_j}
\end{aligned}
$$

where the last inequality uses that $\alpha_k \leq \sqrt{\alpha_k}$ for all $\alpha_k \leq \frac{1}{2}$.

For part (ii), under Assumption A3, combining Lemma 12, Lemma 17, Lemma 14 and Lemma 15 gives the desired result.

For part (iii), under Assumption A1, combining Lemma 13, Lemma 17, Lemma 14 and Lemma 15 gives the desired result. ■

Proof of Lemma 11

We introduce another intermediate process where $\mathbf{x}_{\tau_{k_0}}^B = \mathbf{x}_{\tau_{k_0}}^M$:

$$
\mathbf{x}_{\tau_{k+1}}^B = \Pi_{\mathcal{X}}\left(\mathbf{x}_{\tau_k}^B - \alpha_k \nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k)\right). \tag{19}
$$

The triangle inequality gives

$$
\left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^M\right\| \leq \left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\right\| + \left\|\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\right\|. \tag{20}
$$

Bound the first term on the RHS of (20) as:

$$
\begin{aligned}
\left\|\mathbf{x}_{\tau_{k+1}}^A - \mathbf{x}_{\tau_{k+1}}^B\right\| &\leq \left\|\mathbf{x}_{\tau_k}^A - \alpha_k \nabla f(\mathbf{x}_{\tau_k}^A, \mathbf{z}_k) - \left(\mathbf{x}_{\tau_k}^B - \alpha_k \nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k)\right)\right\| \\
&\leq \left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\right\| + \alpha_k \left\|\nabla f(\mathbf{x}_{\tau_k}^A, \mathbf{z}_k) - \nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k)\right\| \\
&\leq \left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\right\| + \alpha_k \ell \|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^M\| \\
&\leq (1 + \alpha_k \ell)\left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\right\| + \alpha_k \ell \|\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\|.
\end{aligned} \tag{21}
$$

Bound the second term on the RHS of (20) as:

$$
\begin{aligned}
\left\|\mathbf{x}_{\tau_{k+1}}^B - \mathbf{x}_{\tau_{k+1}}^M\right\|^2 &\leq \left\|\mathbf{x}_{\tau_k}^B - \alpha_k \nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k) - \left(\mathbf{x}_{\tau_k}^M - \alpha_k \nabla \bar{f}(\mathbf{x}_{\tau_k}^M)\right)\right\|^2 \\
&= \left\|\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\right\|^2 + \alpha_k^2 \|\nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k) - \nabla \bar{f}(\mathbf{x}_{\tau_k}^M)\|^2 \\
&\quad - 2\alpha_k \left(\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\right)^\top \left(\nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k) - \nabla \bar{f}(\mathbf{x}_{\tau_k}^M)\right).
\end{aligned} \tag{22}
$$

Since $\mathbf{x}_0$ is independent of all $\mathbf{z}_k$ and all $\mathbf{z}_k$ are independent, taking the expectation of the cross term of (22) gives

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\right)^\top \left(\nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k) - \nabla \bar{f}(\mathbf{x}_{\tau_k}^M)\right)\right] \\
=&\mathbb{E}\left[\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\right]^\top \mathbb{E}\left[\nabla f(\mathbf{x}_{\tau_k}^M, \mathbf{z}_k) - \nabla \bar{f}(\mathbf{x}_{\tau_k}^M)\right] = 0.
\end{aligned}
$$

Therefore, taking expectation over (22) gives

$$\mathbb{E}\left[\left\|\mathbf{x}^B_{\tau_{k+1}} - \mathbf{x}^M_{\tau_{k+1}}\right\|^2\right] \le \mathbb{E}\left[\left\|\mathbf{x}^B_{\tau_k} - \mathbf{x}^M_{\tau_k}\right\|^2\right] + \alpha_k^2\sigma^2. \tag{23}$$

Iterating and Jensen's inequality gives

$$\mathbb{E}\left[\left\|\mathbf{x}^B_{\tau_k} - \mathbf{x}^M_{\tau_k}\right\|\right] \le \sigma\sqrt{\sum_{j=k_0}^{k-1}\alpha_j^2}. \tag{24}$$

Taking expectation over (21) and plugging (24), we get

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}^A_{\tau_k} - \mathbf{x}^B_{\tau_k}\|] &\le (1+\alpha_{k-1}\ell)\mathbb{E}[\|\mathbf{x}^A_{\tau_{k-1}} - \mathbf{x}^B_{\tau_{k-1}}\|] + \alpha_{k-1}\ell\sqrt{\sum_{j=k_0}^{k-2}\alpha_j^2}\sigma \\
&\le \sum_{i=k_0+1}^{k-1}\Pi_{j=i+1}^{k-1}(1+\alpha_j\ell)\alpha_i\ell\sqrt{\sum_{s=k_0}^{i-1}\alpha_s^2}\sigma \\
&\le \sum_{i=k_0+1}^{k-1} e^{\ell(\tau_k-\tau_{i+1})}\alpha_i\ell\sqrt{\sum_{s=k_0}^{i-1}\alpha_s^2}\sigma \\
&\le e^{\ell\tau_k}\ell\int_{\tau_{k_0+1}}^{\tau_k} e^{-\ell w}dw\sqrt{\sum_{s=k_0}^{k-2}\alpha_s^2}\sigma \\
&\le (e^{\ell(-\tau_{k_0+1}+\tau_k)}-1)\sqrt{\sum_{s=k_0}^{k-2}\alpha_s^2}\sigma \\
&\le (e^{\ell(\tau_k-\tau_{k_0})}-1)\sqrt{\sum_{s=k_0}^{k-1}\alpha_s^2}\sigma
\end{aligned} \tag{25}$$

where the third inequality is because $1+x \le e^x$ for all $x \ge 0$ and the second to the last inequality uses a Riemann sum bound.

Combining (24) and (25) completes the proof.

■

Proof of Lemma 12

To obtain the desired bound, we further introduce the following two intermediate processes:

$$\mathbf{x}^{M,s}_{\tau_{k+1}} = \Pi_{\mathcal{X}}\left(\mathbf{x}^{M,s}_{\tau_k} - \alpha_k\mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s}\vee\mathcal{G}\right]\right) \tag{26a}$$

$$\mathbf{x}^{B,s}_{\tau_{k+1}} = \Pi_{\mathcal{X}}\left(\mathbf{x}^{B,s}_{\tau_k} - \alpha_k\mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1}\vee\mathcal{G}\right]\right) \tag{26b}$$

where $\mathcal{G} = \sigma(\{\mathbf{x}_0\})$. We set $\mathcal{F}_j = \{\emptyset, \mathcal{Z}\}$ for all $j < 0$. $\mathbf{x}^{M,s}_{\tau_{k_0}} = \mathbf{x}^{B,s}_{\tau_{k_0}} = \mathbf{x}^A_{\tau_{k_0}}$ for all $s \ge 0$. For $s = 0$, $\mathbf{x}^{M,0}_{\tau_k} = \mathbf{x}^A_{\tau_k}$ and for $s > k$, $\mathbf{x}^{M,s}_{\tau_k} = \mathbf{x}^M_{\tau_k}$.

Therefore, using the triangle inequality, we have

$$\left\|\mathbf{x}^A_{\tau_k} - \mathbf{x}^M_{\tau_k}\right\| \le \sum_{s=0}^{k} \left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{M,s+1}_{\tau_k}\right\| \quad \le \sum_{s=0}^{k} \left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k}\right\| + \sum_{s=0}^{k} \left\|\mathbf{x}^{B,s}_{\tau_k} - \mathbf{x}^{M,s+1}_{\tau_k}\right\|. \tag{27}$$

In the following, we want to bound $\mathbb{E}\left[\left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k}\right\|\right]$ and $\mathbb{E}\left[\left\|\mathbf{x}^{B,s}_{\tau_k} - \mathbf{x}^{M,s+1}_{\tau_k}\right\|\right]$.

$$\begin{aligned}
&\left\|\mathbf{x}^{M,s}_{\tau_{k+1}} - \mathbf{x}^{B,s}_{\tau_{k+1}}\right\|^2 \\
&\le \left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k} - \alpha_k \left(\mathbb{E}[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}] - \mathbb{E}[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}]\right)\right\|^2 \\
&= \left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k}\right\|^2 + \alpha_k^2 \left\|\mathbb{E}[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}] - \mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}\right]\right\|^2 \\
&- 2\alpha_k \left(\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k}\right)^\top \left(\mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}\right] - \mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}\right]\right)
\end{aligned} \tag{28}$$

We can show that the cross term has zero mean. By definition, $\mathbf{x}^{M,s}_{\tau_k}$ is $\mathcal{F}_{k-s-1} \vee \mathcal{G}$-measurable and $\mathbf{x}^{B,s}_{\tau_k}$ is $\mathcal{F}_{k-s-2} \vee \mathcal{G}$-measurable. Therefore, we have the following

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

For the second term of (28),

$$\begin{aligned}
&\left\|\mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}\right] - \mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}\right]\right\|^2 \\
&\le 2\left\|\mathbb{E}\left[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}\right] - \mathbb{E}\left[\nabla f\left(\mathbf{x}^{M,s}_{\tau_k}, \mathbb{E}\left[\mathbf{z}_k|\mathcal{F}^+_{k-s}\right]\right)|\mathcal{F}_{k-s} \vee \mathcal{G}\right]\right\|^2 \\
&+ 2\left\|\mathbb{E}[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}] - \mathbb{E}[\nabla f\left(\mathbf{x}^{M,s}_{\tau_k}, \mathbb{E}[\mathbf{z}_k|\mathcal{F}^+_{k-s}]\right)|\mathcal{F}_{k-s} \vee \mathcal{G}]\right\|^2 \\
&\le 2\ell^2 \mathbb{E}\left[\|\mathbf{z}_k - \mathbb{E}[\mathbf{z}_k|\mathcal{F}^+_{k-s}]\|^2|\mathcal{F}_{k-s} \vee \mathcal{G}\right] + 2\ell^2 \mathbb{E}\left[\|\mathbf{z}_k - \mathbb{E}[\mathbf{z}_k|\mathcal{F}^+_{k-s}]\|^2|\mathcal{F}_{k-s-1} \vee \mathcal{G}\right].
\end{aligned}$$

Taking expectation and plugging in the L-mixing property gives

$$\mathbb{E}\left[\left\|\mathbb{E}[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s} \vee \mathcal{G}] - \mathbb{E}[\nabla f(\mathbf{x}^{M,s}_{\tau_k}, \mathbf{z}_k)|\mathcal{F}_{k-s-1} \vee \mathcal{G}]\right\|^2\right] \le 4\ell^2 \psi_2(s, \mathbf{z})^2. \tag{29}$$

Therefore, taking expectation of (28) and plugging in (29), we have

$$\mathbb{E}\left[\left\|\mathbf{x}^{M,s}_{\tau_{k+1}} - \mathbf{x}^{B,s}_{\tau_{k+1}}\right\|^2\right] \le \mathbb{E}\left[\left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k}\right\|^2\right] + 4\ell^2 \psi_2(s, \mathbf{z})^2 \alpha_k^2.$$

Iterating gives

$$\mathbb{E}\left[\left\|\mathbf{x}^{M,s}_{\tau_k} - \mathbf{x}^{B,s}_{\tau_k}\right\|^2\right] \le 4\ell^2 \psi_2(s, \mathbf{z})^2 \sum_{j=k_0}^{k-1} \alpha_j^2.$$

Jensen's inequality gives

$$\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{M,s}-\mathbf{x}_{\tau_k}^{B,s}\right\|\right] \leq 2\ell\psi_2(s,\mathbf{z})\sqrt{\sum_{j=k_0}^{k-1}\alpha_j^2}. \tag{30}$$

Now, we proceed to bound $\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right]$.

$$\begin{aligned}
&\left\|\mathbf{x}_{\tau_{k+1}}^{B,s}-\mathbf{x}_{\tau_{k+1}}^{M,s+1}\right\| \\
&\leq \left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}-\alpha_k\left(\mathbb{E}\left[\nabla f(\mathbf{x}_{\tau_k}^{M,s},\mathbf{z}_k)|\mathcal{F}_{k-s-1}\vee\mathcal{G}\right]-\mathbb{E}\left[\nabla f(\mathbf{x}_{\tau_k}^{M,s+1},\mathbf{z}_k)|\mathcal{F}_{k-s-1}\vee\mathcal{G}\right]\right)\right\| \\
&\leq \left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|+\alpha_k\ell\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{M,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\middle|\mathcal{F}_{k-s-1}\vee\mathcal{G}\right]
\end{aligned}$$

Taking expectation gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_{\tau_{k+1}}^{B,s}-\mathbf{x}_{\tau_{k+1}}^{M,s+1}\right\|\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right]+\alpha_k\ell\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{M,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right] \\
&\leq \mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right]+\alpha_k\ell\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{M,s}-\mathbf{x}_{\tau_k}^{B,s}\right\|\right]+\alpha_k\ell\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right] \\
&\leq (1+\alpha_k\ell)\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right]+\alpha_k\ell\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{M,s}-\mathbf{x}_{\tau_k}^{B,s}\right\|\right].
\end{aligned}$$

Plugging (30) and iterating gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_{\tau_k}^{B,s}-\mathbf{x}_{\tau_k}^{M,s+1}\right\|\right] &\leq \sum_{i=k_0}^{k-1}\Pi_{j=i+1}^{k-1}(1+\alpha_j\ell)\alpha_i\ell\mathbb{E}\left[\left\|\mathbf{x}_{\tau_i}^{M,s}-\mathbf{x}_{\tau_i}^{B,s}\right\|\right] \\
&\leq \sum_{i=k_0}^{k-1}e^{\ell\sum_{j=i+1}^{k-1}\alpha_j}\alpha_i\ell 2\ell\psi_2(s,\mathbf{z})\sqrt{\sum_{s=k_0}^{i-1}\alpha_s^2} \\
&\leq 2\ell\psi_2(s,\mathbf{z})\sqrt{\sum_{s=k_0}^{k-2}\alpha_s^2}\sum_{i=k_0}^{k-1}e^{\tau_k}e^{-\tau_{i+1}}\alpha_i\ell \\
&\leq 2\ell\psi_2(s,\mathbf{z})\sqrt{\sum_{s=k_0}^{k-2}\alpha_s^2}e^{\tau_k}\ell\int_{\tau_{k_0+1}}^{\tau_k}e^{-\ell w}dw \\
&\leq 2\ell\psi_2(s,\mathbf{z})\sqrt{\sum_{s=k_0}^{k-2}\alpha_s^2}(e^{\ell(\tau_k-\tau_{k_0+1})}-1) \\
&\leq 2\ell\psi_2(s,\mathbf{z})(e^{\ell(\tau_k-\tau_{k_0})}-1)\sqrt{\sum_{s=k_0}^{k-1}\alpha_s^2}.
\end{aligned} \tag{31}$$

Plugging the bounds from (30) and (31) into (27) gives the desired bound. ■

Proof of Lemma 13

$\mathbf{x}_{\tau_k}^{B}$ is defined in Lemma 11. Recall

$$\mathbf{x}_{\tau_{k+1}}^{B}=\Pi_{\mathcal{X}}\left(\mathbf{x}_{\tau_k}^{B}-\alpha_k\nabla f(\mathbf{x}_{\tau_k}^{M},\mathbf{z}_k)\right).$$

Triangle inequality gives

$$\left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^M\right\| \le \left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\right\| + \left\|\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\right\|. \tag{32}$$

So the goal is to bound $\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\|$ and $\|\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\|$.

Similar to (25) but without taking expectaion, we have

$$\begin{aligned}\left\|\mathbf{x}_{\tau_k}^A - \mathbf{x}_{\tau_k}^B\right\| &\le \sum_{i=k_0+1}^{k-1} \Pi_{j=i+1}^{k-1}(1+\alpha_j\ell)\alpha_i\ell \max_{i\in[k_0,k-1]} \left\|\mathbf{x}_{\tau_i}^B - \mathbf{x}_{\tau_i}^M\right\| \\ &\le \left(e^{\ell(\tau_k-\tau_{k_0})} - 1\right) \max_{i\in[k_0,k-1]} \left\|\mathbf{x}_{\tau_i}^B - \mathbf{x}_{\tau_i}^M\right\|.\end{aligned} \tag{33}$$

Thus, we want to bound $\|\mathbf{x}_{\tau_i}^B - \mathbf{x}_{\tau_i}^M\|$ for all $i \in [k_0, k-1]$.

Iterating (22) gives

$$\|\mathbf{x}_{\tau_k}^B - \mathbf{x}_{\tau_k}^M\|^2 \le \sum_{i=k_0}^{k-1} 2\alpha_i(\mathbf{x}_{\tau_i}^M - \mathbf{x}_{\tau_i}^B)^\top \mathbf{z}_i + \sum_{i=k_0}^{k-1} \alpha_i^2\|\mathbf{z}_i\|^2. \tag{34}$$

In the following, we show how to bound the two terms on the RHS respectively.

Let $\mathbf{v}_i = \mathbf{x}_{\tau_i}^M - \mathbf{x}_{\tau_i}^B$ and we have $\|\mathbf{v}_i\| \le D$ for all i from the assumption on $\mathcal{X}$. First, we want to show $\max_{s\in[k_0,k-1]} 2\sum_{i=k_0}^{s} \alpha_i \mathbf{v}_i^\top \mathbf{z}_i$ is sub-Gaussian. From the uniform sub-Gaussian Assumption A1, we can obtain that for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}\left[e^{\lambda 2\sum_{i=k_0}^{k-1} \alpha_i \mathbf{v}_i^\top \mathbf{z}_i}\right] \le e^{\frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2 \sum_{i=k_0}^{k-1}\alpha_i^2}.$$

By definition, $\mathbf{v}_i^\top \mathbf{z}_i$ is $\mathcal{F}_i \vee \mathcal{G}$-measurable, where $\mathcal{G}$ is defined in Lemma 12. Then,

$$\mathbb{E}\left[e^{\lambda 2\sum_{i=k_0}^{s} \alpha_i \mathbf{v}_i^\top \mathbf{z}_i}\middle|\mathcal{F}_{s-1}\vee\mathcal{G}\right] \le e^{\lambda 2\sum_{i=k_0}^{s-1} \alpha_i \mathbf{v}_i^\top \mathbf{z}_i + \frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2\alpha_s^2}.$$

Let $M_s(\lambda) = e^{\sum_{i=k_0}^{s}\left(2\lambda\alpha_i\mathbf{v}_i^\top\mathbf{z}_i - \frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2\alpha_i^2\right)}$. We can show that $M_s(\lambda)$ is supermartingale:

$$\begin{aligned}\mathbb{E}\left[M_s(\lambda)|\mathcal{F}_{s-1}\vee\mathcal{G}\right] &\le e^{\sum_{i=k_0}^{s-1}\left(2\lambda\alpha_i\mathbf{v}_i^\top\mathbf{z}_i - \frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2\alpha_i^2\right)}\mathbb{E}\left[e^{2\lambda\alpha_s\mathbf{v}_s^\top\mathbf{z}_s - \frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2\alpha_s^2}\middle|\mathcal{F}_{s-1}\vee\mathcal{G}\right] \\ &\le M_{s-1}(\lambda).\end{aligned} \tag{35}$$

almost surely for all $s \ge k_0 + 1$.

By iterating (35), we have for all $s \in [k_0, k-1]$,

$$\mathbb{E}\left[M_s(\lambda)\right] \le 1.$$

Using Doob's maximal inequality (see Lattimore and Szepesvári, 2020, Theorem 3.9) and choosing an $\epsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{s\in[k_0,k-1]} M_s(\lambda) \geq e^{\epsilon}\right) \leq e^{-\epsilon}\mathbb{E}\left[M_{k_0}(\lambda)\right] \leq e^{-\epsilon} \\
\Leftrightarrow\, &\mathbb{P}\left(\max_{s\in[k_0,k-1]} \sum_{i=k_0}^{s}\left(2\lambda\alpha_i \mathbf{v}_i^\top \mathbf{z}_i - \frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2\alpha_i^2\right) \geq \epsilon\right) \leq e^{-\epsilon} \\
\Rightarrow\, &\mathbb{P}\left(\max_{s\in[k_0,k-1]} \sum_{i=k_0}^{s} 2\lambda\alpha_i \mathbf{v}_i^\top \mathbf{z}_i \geq \epsilon + \sum_{i=k_0}^{k-1}\frac{1}{2}\lambda^2 4D^2\hat{\sigma}^2\alpha_i^2\right) \leq e^{-\epsilon} \\
\Leftrightarrow\, &\mathbb{P}\left(\max_{s\in[k_0,k-1]} \sum_{i=k_0}^{s} 2\alpha_i \mathbf{v}_i^\top \mathbf{z}_i \geq \frac{\epsilon}{\lambda} + \sum_{i=k_0}^{k-1}\frac{1}{2}\lambda 4D^2\hat{\sigma}^2\alpha_i^2\right) \leq e^{-\epsilon}.
\end{aligned} \tag{36}$$

The RHS of the inequality inside the probability of (36) is minimized at $\lambda^* = \sqrt{\frac{\epsilon}{\frac{1}{2}4D^2\hat{\sigma}^2\sum_{i=k_0}^{k-1}\alpha_i^2}}$. Then plugging λ^* into (36) gives

$$\mathbb{P}\left(\max_{s\in[k_0,k-1]} \sum_{i=k_0}^{s} 2\alpha_i \mathbf{v}_i^\top \mathbf{z}_i \geq 2\sqrt{2}D\hat{\sigma}\sqrt{\epsilon}\sqrt{\sum_{i=k_0}^{k-1}\alpha_i^2}\right) \leq e^{-\epsilon}. \tag{37}$$

Next, we want to bound $\max_{s\in[k_0,k-1]}\sum_{i=k_0}^{s}\alpha_i^2\|\mathbf{z}_i\|^2$.

The following is the modification of the proof of (Wainwright, 2019, Theorem 2.6, IV).

Multiplying both sides of the definition of sub-Gaussian random vectors (3) by $e^{-\frac{1}{2t}\|v\|^2\hat{\sigma}^2}$ with $t \in (0,1)$ gives

$$\mathbb{E}\left[e^{v^\top \mathbf{z} - \frac{1}{2t}\|v\|^2\hat{\sigma}^2}\right] \leq e^{-\frac{1}{2}(\frac{1}{t}-1)\hat{\sigma}^2\|v\|^2}. \tag{38}$$

Integrating both sides over v gives

$$\int e^{-\frac{1}{2}(\frac{1}{t}-1)\hat{\sigma}^2\|v\|^2}dv = \frac{(2\pi)^{\frac{n}{2}}}{\left((\frac{1}{t}-1)\hat{\sigma}^2\right)^{\frac{n}{2}}} \tag{39}$$

and

$$\begin{aligned}
\int e^{v^\top \mathbf{z} - \frac{1}{2t}\hat{\sigma}^2\|v\|^2}dz &= e^{\frac{1}{2}\frac{t}{\hat{\sigma}^2}\|\mathbf{z}\|^2}\int e^{-\frac{1}{2}\frac{\hat{\sigma}^2}{t}\|v-\frac{t}{\hat{\sigma}^2}\mathbf{z}\|^2}dv \\
&= e^{\frac{1}{2}\frac{t}{\hat{\sigma}^2}\|\mathbf{z}\|^2}\frac{(2\pi)^{\frac{n}{2}}}{\left(\frac{\hat{\sigma}^2}{t}\right)^{\frac{n}{2}}}.
\end{aligned} \tag{40}$$

Plugging (39) and (40) into (38), we have for all $t \in (0,1)$,

$$\mathbb{E}\left[e^{\frac{1}{2}\frac{t}{\hat{\sigma}^2}\|\mathbf{z}\|^2}\right] \leq \frac{1}{(1-t)^{\frac{n}{2}}}.$$

Let $\lambda = \frac{1}{2}\frac{t}{\hat{\sigma}^2}$. Then for $0 < \lambda < \frac{1}{2\hat{\sigma}^2}$,

$$\mathbb{E}\left[e^{\lambda\|\mathbf{z}\|^2}\right] \leq \frac{1}{(1-2\hat{\sigma}^2\lambda)^{\frac{n}{2}}}.$$

Let $g(\lambda) = \log \frac{1}{(1-2\hat{\sigma}^2\lambda)^{\frac{n}{2}}} = -\frac{n}{2}\log(1-2\lambda\hat{\sigma}^2)$. Then, applying the Taylor expansion gives

$$\begin{aligned} g(\lambda) = \frac{n}{2}\sum_{k=1}^{\infty}\frac{1}{k!}(2\lambda\hat{\sigma}^2)^k &= n\lambda\hat{\sigma}^2 + \sum_{k=2}^{\infty}\frac{1}{k!}(2\lambda\hat{\sigma}^2)^k \\ &\leq n\lambda\hat{\sigma}^2 + \frac{1}{2}\frac{(2\lambda\hat{\sigma}^2)^2}{1-(2\lambda\hat{\sigma}^2)}. \end{aligned}$$

If $2\lambda\hat{\sigma}^2 \leq \frac{1}{2}$, then $\lambda \leq \frac{1}{4\hat{\sigma}^2}$. Therefore, $g(\lambda) \leq n\lambda\hat{\sigma}^2 + 4\lambda^2\hat{\sigma}^4$ and

$$\mathbb{E}\left[e^{\lambda\|\mathbf{z}\|^2}\right] \leq e^{n\lambda\hat{\sigma}^2+4\lambda^2\hat{\sigma}^4}.$$

If $0 < \lambda \leq \frac{1}{4\alpha_i^2\hat{\sigma}^2}$ for all $i \in [k_0, k-1]$, then we can show that

$$M_s(\lambda) = e^{\sum_{i=k_0}^{s}\left(\lambda\alpha_i^2\|\mathbf{z}_i\|^2 - n\lambda\alpha_i^2\hat{\sigma}^2 - 4\lambda^2\alpha_i^4\hat{\sigma}^4\right)}$$

is also supermartingale with $\mathbb{E}\left[M_s(\lambda)\right] \leq 1$ for all $s \in [k_0, k-1]$.

Similar to the process of getting (36) and choosing the same ϵ, we have

$$\begin{aligned} &\mathbb{P}\left(\max_{s\in[k_0,k-1]} M_s(\lambda) \geq e^{\epsilon}\right) \leq e^{-\epsilon}\mathbb{E}\left[M_{k_0}(\lambda)\right] \leq e^{-\epsilon} \\ \Rightarrow &\mathbb{P}\left(\max_{s\in[k_0,k-1]}\sum_{i=k_0}^{s}\alpha_i^2\|\mathbf{z}_i\|^2 \geq \frac{\epsilon}{\lambda} + \lambda 4\sum_{i=k_0}^{k-1}\alpha_i^4\hat{\sigma}^4 + \sum_{i=k_0}^{k-1}\alpha_i^2 n\hat{\sigma}^2\right) \leq e^{-\epsilon} \end{aligned} \tag{41}$$

We can choose $\lambda = \frac{1}{4\hat{\sigma}^2\max_{i\in[k_0,k-1]}\alpha_i^2}$ and plugging it into the RHS of the inequality inside the probability in (41). Then, the following holds:

$$\begin{aligned} &\mathbb{P}\left(\max_{s\in[k_0,k-1]}\sum_{i=k_0}^{s}\alpha_i^2\|\mathbf{z}_i\|^2 \geq 4\hat{\sigma}^2\epsilon\max_{i\in[k_0,k-1]}\alpha_i^2 + \hat{\sigma}^2\frac{1}{\max_{i\in[k_0,k-1]}\alpha_i^2}\sum_{i=k_0}^{k-1}\alpha_i^4 + \sum_{i=k_0}^{k-1}\alpha_i^2 n\hat{\sigma}^2\right) \leq e^{-\epsilon} \\ \Rightarrow &\mathbb{P}\left(\max_{s\in[k_0,k-1]}\sum_{i=k_0}^{s}\alpha_i^2\|\mathbf{z}_i\|^2 \geq 4\hat{\sigma}^2\epsilon\max_{i\in[k_0,k-1]}\alpha_i^2 + \left(\hat{\sigma}^2 + n\hat{\sigma}^2\right)\sum_{i=k_0}^{k-1}\alpha_i^2\right) \leq e^{-\epsilon} \\ \Rightarrow &\mathbb{P}\left(\max_{s\in[k_0,k-1]}\sum_{i=k_0}^{s}\alpha_i^2\|\mathbf{z}_i\|^2 \geq \left(4\hat{\sigma}^2\epsilon + \hat{\sigma}^2 + n\hat{\sigma}^2\right)\sum_{i=k_0}^{k-1}\alpha_i^2\right) \leq e^{-\epsilon}. \end{aligned} \tag{42}$$

where the first arrow holds because $\sum_{i=k_0}^{k-1}\alpha_i^4 \leq \max_{j\in[k_0,k-1]}\alpha_j^2\sum_{i=k_0}^{k-1}\alpha_i^2$.

The intersection of the respective complements of the events in (37) and (42) is the event that $\max_{i\in[k_0,k-1]} \|\mathbf{x}^B_{\tau_i} - \mathbf{x}^M_{\tau_i}\|$ is upper bounded by

$$\left(2\sqrt{2}\hat{\sigma}D\sqrt{\epsilon}\sqrt{\sum_{j=k_0}^{k-1}\alpha_j^2} + \left(4\hat{\sigma}^2\epsilon + \hat{\sigma}^2 + n\hat{\sigma}^2\right)\sum_{j=k_0}^{k-1}\alpha_j^2\right)^{1/2}.$$

Such an event occurs with probability $(1-e^{-\epsilon})^2$.

Further combining (32) and (33) completes the proof.

■

In the following proof, we follow the notation in (Rockafellar, 2015). Let $\gamma(x|\mathcal{X})$ denote the gauge function:

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

and let $\delta(x|\mathcal{X})$ be the support function:

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

Proof of Lemma 14

Applying Lemma 2.2 (i) in (Tanaka, 1979) gives

$$\begin{aligned}
\left\|\mathbf{x}^C_{\tau_k} - \mathbf{x}^D_{\tau_k}\right\|^2 &\le \left\|\mathbf{y}^C_{\tau_k} - \mathbf{y}^D_{\tau_k}\right\|^2 + 2\int_{\tau_{k_0}}^{\tau_k}(\mathbf{y}^C_{\tau_k} - \mathbf{y}^D_{\tau_k} - \mathbf{y}^C_s + \mathbf{y}^D_s)^\top(\mathbf{v}^D_s d\boldsymbol{\mu}^D(s) - \mathbf{v}^C_s d\boldsymbol{\mu}^C(s)) \\
&\le 2\int_{\tau_{k_0}}^{\tau_k}(\mathbf{y}^C_s - \mathbf{y}^D_s)^\top \mathbf{v}^C_s d\boldsymbol{\mu}(s) \\
&\le 2\int_{\tau_{k_0}}^{\tau_k}\gamma(\mathbf{y}^C_s - \mathbf{y}^D_s|\mathcal{X})\delta(\mathbf{v}^C_s|\mathcal{X})d\boldsymbol{\mu}^C(s) \\
&\le 2\sup_{s\in[\tau_{k_0},\tau_k]}\gamma(\mathbf{y}^C_s - \mathbf{y}^D_s|\mathcal{X})\int_{\tau_{k_0}}^{\tau_k}\delta(\mathbf{v}^C_s|\mathcal{X})d\boldsymbol{\mu}^C(s). \qquad (43)
\end{aligned}$$

The second inequality is because $\mathbf{y}^D_s = \mathbf{y}^C_{\tau_k}$ for all $s \in [\tau_k, \tau_{k+1})$, $\boldsymbol{\mu}^D$ is supported on the discrete set $\{\tau_0, \tau_1, \tau_2, \cdots\}$ and the integrand is zero on this set . The third inequality uses the inequality $x^\top y \le \gamma(x|\mathcal{X})\delta(y|\mathcal{X})$ and the last inequality follows Hölder's inequality.

Since $\mathcal{X}$ contains a ball of radius r around the origin, we have $\gamma(x|\mathcal{X}) \le r^{-1}\|x\|$. Then, the following holds

$$\begin{aligned}
\sup_{s\in[\tau_{k_0},\tau_k]}\gamma(\mathbf{y}^C_s - \mathbf{y}^D_s|\mathcal{X}) &\le r^{-1}\sup_{s\in[\tau_{k_0},\tau_k]}\|\mathbf{y}^C_s - \mathbf{y}^D_s\| \\
&\le r^{-1}\max_{j\in[k_0,k]}\int_{\tau_j}^{\tau_{j+1}}\left\|\nabla\bar{f}(\mathbf{x}^C_s)\right\|ds \\
&\le r^{-1}u\max_{j\in[k_0,k]}\alpha_j. \qquad (44)
\end{aligned}$$

To bound the integral in (43), we take the following derivative

$$
\begin{aligned}
d\|\mathbf{x}_t^C\|^2 &= 2(\mathbf{x}_t^C)^\top \left(-\nabla \bar{f}(\mathbf{x}_t^C)dt - \mathbf{v}_t^C d\boldsymbol{\mu}^C(t)\right) \\
\Leftrightarrow 2(\mathbf{x}_t^C)^\top \mathbf{v}_t^C d\boldsymbol{\mu}^C(t) &= -2(\mathbf{x}_t^C)^\top \nabla \bar{f}(\mathbf{x}_t^C)dt - d\|\mathbf{x}_t^C\|^2 \qquad (45)
\end{aligned}
$$

By construction, $(\mathbf{x}_t^C)^\top \mathbf{v}_t^C = \sup\{x^\top \mathbf{v}_t^C | x \in \mathcal{X}\} = \delta(\mathbf{v}_t^C|\mathcal{X})$. Therefore, taking the integral of (45) gives

$$
\begin{aligned}
2\int_{\tau_{k_0}}^{\tau_k} \delta(\mathbf{v}_s^C|\mathcal{X})d\mu(s) &= -2\int_{\tau_{k_0}}^{\tau_k} (\mathbf{x}_s^C)^\top \nabla \bar{f}(\mathbf{x}_s^C)ds + \|\mathbf{x}_{\tau_{k_0}}^C\|^2 - \|\mathbf{x}_{\tau_k}^C\|^2 \\
\Leftrightarrow \int_{\tau_{k_0}}^{\tau_k} \delta(\mathbf{v}_s^C|\mathcal{X})d\mu(s) &= -\int_{\tau_{k_0}}^{\tau_k} (\mathbf{x}_s^C)^\top \nabla \bar{f}(\mathbf{x}_s^C)ds + \frac{1}{2}\|\mathbf{x}_{\tau_{k_0}}^C\|^2 - \frac{1}{2}\|\mathbf{x}_{\tau_k}^C\|^2 \\
&\le (\tau_k - \tau_{k_0})Du + D^2. \qquad (46)
\end{aligned}
$$

Plugging (44) and (46) into (43), we have

$$
\left\|\mathbf{x}_{\tau_k}^C - \mathbf{x}_{\tau_k}^D\right\|^2 \le 2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right) \max_{j\in[k_0,k]} \alpha_j
$$

which gives

$$
\left\|\mathbf{x}_{\tau_k}^C - \mathbf{x}_{\tau_k}^D\right\| \le \sqrt{2r^{-1}u\left((\tau_k - \tau_{k_0})Du + D^2\right) \max_{j\in[k_0,k]} \alpha_j}.
$$

■

Proof of Lemma 15

$$
\begin{aligned}
\left\|\frac{d\mathbf{x}_t^C}{dt}\right\| &= \left\|\Pi_{T_\mathcal{X}(\mathbf{x}_t^C)}\left(-\nabla \bar{f}(\mathbf{x}_t^C)\right)\right\| \\
&= \left\|\Pi_{T_\mathcal{X}(\mathbf{x}_t^C)}\left(-\nabla \bar{f}(\mathbf{x}_t^C)\right) - \Pi_{T_\mathcal{X}(\mathbf{x}_t^C)}(0)\right\| \\
&\le \left\|\nabla \bar{f}(\mathbf{x}_t^C)\right\|
\end{aligned}
$$

where the first equality uses $0 \in T_\mathcal{X}(x)$ and the inequality uses the non-expansiveness of convex projection.

Therefore,

$$
\begin{aligned}
\left\|\mathbf{x}_t^C - \mathbf{x}_{\tau_k}^C\right\| &= \left\|\int_{\tau_k}^t \Pi_{T_\mathcal{X}(\mathbf{x}_s^C)}\left(-\nabla \bar{f}(\mathbf{x}_s^C)\right) ds\right\| \\
&\le \alpha_k u.
\end{aligned}
$$

■

Proof of Lemma 16

For $t \in [\tau_k, \tau_{k+1})$, the triangle inequality gives

$$
\begin{aligned}
\|\mathbf{x}_t^C - \mathbf{x}_t^D\| &\le \|\mathbf{x}_t^C - \mathbf{x}_{\tau_k}^C + \mathbf{x}_{\tau_k}^C - \mathbf{x}_t^D\| \\
&\le \|\mathbf{x}_t^C - \mathbf{x}_{\tau_k}^C\| + \|\mathbf{x}_{\tau_k}^C - \mathbf{x}_{\tau_k}^D\|
\end{aligned}
$$

Plugging Lemma 15 and Lemma 14 gives the desired bound. ■

Proof of Lemma 17

Define $\boldsymbol{\rho}_t = \mathbf{x}_t^M + \mathbf{y}_t^M - \mathbf{y}_{\tau_k}^M - (\mathbf{x}_t^D + \mathbf{y}_t^C - \mathbf{y}_t^D)$ for all $t \in [\tau_k, \tau_{k+1})$. This gives $\boldsymbol{\rho}_{\tau_k} = \mathbf{x}_{\tau_k}^M - \mathbf{x}_{\tau_k}^D$. Then calculate

$$
\begin{aligned}
d\|\boldsymbol{\rho}_t\| &= \left(\frac{\boldsymbol{\rho}_t}{\|\boldsymbol{\rho}_t\|}\right)^\top d\boldsymbol{\rho}_t \\
&= \left(\frac{\boldsymbol{\rho}_t}{\|\boldsymbol{\rho}_t\|}\right)^\top \left(\nabla \bar{f}(\mathbf{x}_t^C) - \nabla \bar{f}(\mathbf{x}_t^M)\right) dt \\
&\leq \left\|\nabla \bar{f}(\mathbf{x}_t^C) - \nabla \bar{f}(\mathbf{x}_t^M)\right\| dt \\
&\leq \ell \left\|\mathbf{x}_t^M - \mathbf{x}_t^C\right\| dt \\
&\leq \ell \left(\left\|\mathbf{x}_t^M - \mathbf{x}_t^D\right\| + \left\|\mathbf{x}_t^D - \mathbf{x}_t^C\right\|\right) dt
\end{aligned}
\tag{47}
$$

where the second inequality is because $\nabla \bar{f}(x)$ is ℓ-Lipschitz.

Taking the integral gives

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

where the second equality is from Lemma 20 in (Lamperski, 2021) and the inequality uses (47).

Setting $t = \tau_{k+1}$ gives

$$
\|\boldsymbol{\rho}_{\tau_{k+1}}\| \leq (1 + \ell\alpha_k)\|\boldsymbol{\rho}_{\tau_k}\| + \ell \int_{\tau_k}^{\tau_{k+1}} \left\|\mathbf{x}_s^C - \mathbf{x}_s^D\right\| ds. \tag{48}
$$

Using the assumption that $\boldsymbol{\rho}_{k_0} = \mathbf{x}_{k_0}^M - \mathbf{x}_{k_0}^D = 0$ and iterating gives

$$\begin{aligned}
\|\boldsymbol{\rho}_{\tau_k}\| &\leq \sum_{i=k_0}^{k-1} \Pi_{j=i+1}^{k-1}(1+\ell\alpha_j)\ell \int_{\tau_i}^{\tau_{i+1}} \left\|\mathbf{x}_s^C - \mathbf{x}_s^D\right\| ds \\
&\leq \sum_{i=k_0}^{k-1} \Pi_{j=i+1}^{k-1}(1+\ell\alpha_j)\ell\alpha_i \left(\max_{s\in[i,i+1]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_{i+1}-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,i+1]} \alpha_j} \right) \\
&\leq \sum_{i=k_0}^{k-1} e^{\ell(\tau_k - \tau_{i+1})}\ell\alpha_i \left(\max_{s\in[i,i+1]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_{i+1}-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,i+1]} \alpha_j} \right) \\
&= \ell e^{\ell\tau_k} \sum_{i=k_0}^{k-1} e^{-\ell\tau_{i+1}}\alpha_i \left(\max_{s\in[i,i+1]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_{i+1}-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,i+1]} \alpha_j} \right) \\
&\leq \ell e^{\ell\tau_k} \sum_{i=k_0}^{k-1} \int_{\tau_i}^{\tau_{i+1}} e^{-\ell w} dw \left(\max_{s\in[i,i+1]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_{i+1}-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,i+1]} \alpha_j} \right) \\
&\leq \ell e^{\ell\tau_k} \int_{\tau_{k_0}}^{\tau_k} e^{-\ell w} dw \left(\max_{s\in[k_0,k]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_k-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,k]} \alpha_j} \right) \\
&\leq \ell e^{\ell\tau_k} \frac{1}{\ell}(e^{-\ell\tau_{k_0}} - e^{-\ell\tau_k}) \left(\max_{s\in[k_0,k]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_k-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,k]} \alpha_j} \right) \\
&\leq (e^{\ell(\tau_k-\tau_{k_0})} - 1) \left(\max_{s\in[k_0,k]} \alpha_s u + \sqrt{2r^{-1}u\left((\tau_k-\tau_{k_0})Du + D^2\right) \max_{j\in[k_0,k]} \alpha_j} \right) \\
&\leq (e^{\ell(\tau_k-\tau_{k_0})} - 1) \max_{s\in[k_0,k]} \sqrt{\alpha_s} \left(u + \sqrt{2r^{-1}u\left((\tau_k-\tau_{k_0})Du + D^2\right)} \right)
\end{aligned}$$

where the second inequality uses Lemma 16 and the last inequality uses that $\alpha_s \leq \frac{1}{2}$ for all $s \in \mathbb{N}$. ■

Appendix D. Supporting Results on Variational Geometry

The following lemmas are standard in the field of optimization and variational analysis. We present the proofs to support the results in the main paper.

Lemma 18 *For any $x \in \mathbb{R}^n$ and convex set $\mathcal{X}$, $y^* = \Pi_{\mathcal{X}}(x)$ iff $x - y^* \in \mathcal{N}_{\mathcal{X}}(y^*)$ and $y^* \in \mathcal{X}$.*

Proof First, the definition of the convex projection is equivalent to

$$\Pi_{\mathcal{X}}(x) = \arg\min_{y\in\mathcal{X}} \frac{1}{2}\|y - x\|^2.$$

Set $f(y) = \frac{1}{2}\|y - x\|^2$ which is strongly convex thus has a unique minimizer.

($\Rightarrow$)

Let y^* be the minimizer of f, i.e. $y^* = \Pi_{\mathcal{X}}(x)$. From the necessary optimality condition, we have $-\nabla f(y^*) \in \mathcal{N}_{\mathcal{X}}(y^*)$, i.e. $x - y^* \in \mathcal{N}_{\mathcal{X}}(y^*)$.

($\Leftarrow$)

Let $y^* \in \mathcal{X}$ and $x - y^* \in \mathcal{N}_{\mathcal{X}}(y^*)$.

From the definition of normal cone, $x - y^* \in \mathcal{N}_{\mathcal{X}}(y^*) \Leftrightarrow \langle x - y^*, y - y^* \rangle \leq 0, \ \forall y \in \mathcal{X}$. Besides,

$$\begin{aligned} \|x-y\|^2 - \|x-y^*\|^2 &= \|x - y^* + y^* - y\|^2 - \|x-y^*\|^2 \\ &= \|x-y^*\|^2 + \|y^*-y\|^2 + 2(x-y^*)^\top (y^*-y) - \|x-y^*\|^2 \\ &\geq 0 \end{aligned}$$

which implies that y^* is the minimizer of f, i.e. $y^* = \Pi_{\mathcal{X}}(x)$. ■

Lemma 19 *For all $x \in \mathcal{X}$, $g \in \mathbb{R}^n$, we have*

$$g^\top \Pi_{T_{\mathcal{X}}(x)}(g) = \|\Pi_{T_{\mathcal{X}}(x)}(g)\|^2$$

Proof It suffices to show that $\left(g - \Pi_{T_{\mathcal{X}}(x)}(g)\right)^\top \Pi_{T_{\mathcal{X}}(x)}(g) = 0$.

Firstly, from Lemma 18, we have $g - \Pi_{T_{\mathcal{X}}(x)}(g) \in \mathcal{N}_{T_{\mathcal{X}}(x)}(\Pi_{T_{\mathcal{X}}(x)}(g))$, i.e.

$$\left(g - \Pi_{T_{\mathcal{X}}(x)}(g)\right)^\top \Pi_{T_{\mathcal{X}}(x)}(g) \geq \left(g - \Pi_{T_{\mathcal{X}}(x)}(g)\right)^\top y, \ \forall y \in T_{\mathcal{X}}(x). \tag{49}$$

For notation simplicity, set $\phi = g - \Pi_{T_{\mathcal{X}}(x)}(g)$ for the analysis below.

Note that $0 \in T_{\mathcal{X}}(x)$, then we have $\phi^\top \Pi_{T_{\mathcal{X}}(x)}(g) \geq 0$. Furthermore, from the definition of tangent cone, if $y \in T_{\mathcal{X}}(x)$, then $ty \in T_{\mathcal{X}}(x)$ for all $t \geq 0$. For the sake of contradiction, suppose $\phi^\top y > 0$. Then, there exists $t > 0$, such that $\phi^\top ty \geq \phi^\top \Pi_{T_{\mathcal{X}}(x)}(g)$, which contradicts (49). Therefore, we conclude that $\phi^\top y \leq 0$, which further implies that $\phi^\top \Pi_{T_{\mathcal{X}}(x)}(g) \leq 0$ since $\Pi_{T_{\mathcal{X}}(x)}(g) \in T_{\mathcal{X}}(x)$. Therefore, we have $\phi^\top \Pi_{T_{\mathcal{X}}(x)}(g) = 0$ as desired. ■

The following lemma is a special case of the Moreau decomposition, and enables us to use the Skorokhod problem framework. See Hiriart-Urruty and Lemaréchal (2004).

Lemma 20 *For all $x \in \mathcal{X}$, $g \in \mathbb{R}^n$, the following holds*

$$\Pi_{T_{\mathcal{X}}(x)}(g) = g - \Pi_{\mathcal{N}_{\mathcal{X}}(x)}(g). \tag{50}$$

Appendix E. Background on the Skorokhod Problem

This appendix presents background on the Skorokhod problem needed for the paper.

The Skorokhod problem is a classical framework for constraining stochastic processes to remain in a set. It is a useful tool to analyze projection-based algorithms in continous time.

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

- $x_t = y_t + \phi_t \in \mathcal{X}$ 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 $v_s \in \mathcal{N}_{\mathcal{X}}(x_s)$ for all $s \in [0, T)$, while the measure, μ, satisfies $\mu([0, T)) < \infty$ for any $T > 0$.

It is shown in (Tanaka, 1979) that a solution exists and is unique when y is right-continuous with left limits and $\mathcal{X}$ is convex. The existence and uniqueness of the solution implies that we can view the Skorokhod solution as a mapping: $x = \mathcal{S}(y)$. And we are often interested in x_t, thus we will call x_t as the solution of the Skorokhod problem corresponding to y_t.

In the following, we present the connection between Skorokhod problems and projected algorithms assuming y_t is piecewise constant. Specifically, assuming that $0 = \tau_0 < \tau_1 < \cdots < \tau_{N-1} \leq T$ are the jump points of y_t, and let $S_k = [\tau_k, \tau_{k+1})$ for $k < N - 1$ and $S_{N-1} = [\tau_{N-1}, T]$. Then y_t can be represented as

$$y_t = \sum_{k=0}^{N-1} y_{\tau_k} \mathbb{1}_{S_k}(t).$$

Then, the solution of the Skorokhod problem has the form

$$x_{\tau_{k+1}} = \Pi_{\mathcal{X}}(x_{\tau_k} + y_{\tau_{k+1}} - y_{\tau_k}).$$