Online Covariance Estimation in Nonsmooth Stochastic Approximation

Liwei Jiang* LJIANG306@GATECH.EDU
Georgia Institute of Technology

Abhishek Roy* ABHISHEKROY@TAMU.EDU
Texas A&M University

Krishnakumar Balasubramanian KRIZNAKUMAR@GMAIL.COM
University of California, Davis

Damek Davis DAMEK@WHARTON.UPENN.EDU
University of Pennsylvania

Dmitriy Drusvyatskiy DDRUSV@UW.EDU
University of Washington

Sen Na SENNA@GATECH.EDU
Georgia Institute of Technology

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider applying stochastic approximation (SA) methods to solve nonsmooth variational inclusion problems. Existing studies have shown that the averaged iterates of SA methods exhibit asymptotic normality, with an optimal limiting covariance matrix in the local minimax sense of Hájek and Le Cam. However, no methods have been proposed to estimate this covariance matrix in a nonsmooth and potentially non-monotone (nonconvex) setting. In this paper, we study an online batch-means covariance matrix estimator introduced in Zhu et al. (2023). The estimator groups the SA iterates appropriately and computes the sample covariance among batches as an estimate of the limiting covariance. Its construction does not require prior knowledge of the total sample size, and updates can be performed recursively as new data arrives. We establish that, as long as the batch size sequence is properly specified (depending on the stepsize sequence), the estimator achieves a convergence rate of order $O(\sqrt{d}n^{-1/8+\varepsilon})$ for any $\varepsilon > 0$, where d and n denote the problem dimensionality and the number of iterations (or samples) used. Although the problem is nonsmooth and potentially non-monotone (nonconvex), our convergence rate matches the best-known rate for covariance estimation methods using only first-order information in smooth and strongly-convex settings. The consistency of this covariance estimator enables asymptotically valid statistical inference, including constructing confidence intervals and performing hypothesis testing.

Keywords: Stochastic approximation, nonsmoothness, asymptotic normality, covariance estimation

*. Equal Contribution

1. Introduction

A landmark result by Polyak and Juditsky (1992) shows that for smooth and strongly convex optimization, Stochastic Gradient Descent (SGD) exhibits a central limit theorem: the averaged SGD iterates with a proper scaling factor converge to a normal distribution; see Toulis and Airoldi (2017); Duchi and Ruan (2021) for extensions and Anastasiou et al. (2019); Shao and Zhang (2022); Samsonov et al. (2024) for quantitative non-asymptotic bounds. Recently, Davis et al. (2024) extended this result to nonsmooth problems, showing that when solutions vary smoothly with respect to perturbations, the averaged generic stochastic approximation (SA) iterates remain asymptotically normal. This limiting distribution paves the way for constructing confidence intervals and statistical tests, critical tools for uncertainty quantification in machine learning and optimization. However, to perform (asymptotically) valid statistical inference, we need to estimate the covariance matrix of the limiting distribution. While efficient online estimators are well understood in the smooth setting, estimation in the nonsmooth setting has remained completely open. In this paper, we develop an online estimator with computation and memory scaling quadratically in dimension, and establish its rate of convergence in expectation (matching the smooth setting).

The theory encompasses many important problems in machine learning and operations research. Consider a two-player zero-sum game. To find the Nash equilibrium, the two players aim to solve:

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} \mathop{\mathbb{E}}_{\nu \sim \mathcal{P}} [f(x_1, x_2, \nu)],$$

where $f(x_1, x_2, \nu)$ is a random payoff function and $\mathcal{X}_1, \mathcal{X}_2$ are strategy sets. Players update their strategies based on noisy observations, projecting onto their respective strategy sets. Another example is stochastic nonlinear programming; we solve:

$$\min_{x} \mathop{\mathbb{E}}_{\nu \sim \mathcal{P}} [f(x, \nu)] \quad \text{subject to} \quad g_i(x) \le 0, \quad i = 1, \ldots, m, \tag{1.1}$$

where the objective depends on random data. Both settings, along with many others, can be unified through stochastic variational inequalities of the form:

$$0 \in F(x) := \mathop{\mathbb{E}}_{\nu \sim \mathcal{P}} [A(x, \nu)] + N_{\mathcal{X}}(x), \tag{1.2}$$

where $A(\cdot, \nu)$ is a smooth operator for each ν, and $N_{\mathcal{X}}$ denotes the normal cone to the constraint set $\mathcal{X}$. Throughout, we fix a solution $x^\star$ of this inclusion.

To solve the above problems in an online fashion, we consider SA algorithms based on a generalized gradient mapping, $G : \mathbb{R}_{++} \times \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$, of F. Given x_0, the algorithm iterates as

$$x_{k+1} = x_k - \eta_{k+1} G_{\eta_{k+1}}(x_k, \nu_{k+1}), \tag{1.3}$$

where $\eta_{k+1} > 0$ is a stepsize sequence and ν_k is stochastic noise. As we show in Section 5, this framework unifies many online algorithms – in games it captures simultaneous gradient play; in constrained optimization it yields projected gradient methods; and more generally, it encompasses stochastic forward-backward splitting.

Davis et al. (2024) showed that when solutions to the perturbed system vary smoothly – that is, when the graph of the solution map $S(v) = \{x : v \in F(x)\}$ locally coincides with the graph of some smooth function $\sigma(\cdot)$ – the averaged iterates of (1.3) are asymptotically normal:

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{D} N(0, \Sigma),$$

where $\bar{x}_k = \sum_{i=1}^{k} x_i/k$ and $\Sigma = \nabla\sigma(0) \cdot \mathrm{Cov}(A(x^\star, z)) \cdot \nabla\sigma(0)^\top$. For example, in stochastic nonlinear programming (1.1), $A(x^\star, \nu) = \nabla f(x^\star, \nu)$ and $\nabla\sigma(0)$ takes a particularly elegant form

$$\nabla\sigma(0) = (P_T \nabla^2_{xx}\mathcal{L}(x^\star, y^\star) P_T)^\dagger,$$

where $(x^\star, y^\star)$ is the primal-dual solution of (1.1), $\mathcal{L}(x, y) = f(x) + \sum_{i=1}^{n+m} y_i g_i(x)$ is the Lagrangian function, and P_T projects onto the tangent space of active constraints at the solution $x^\star$.

In order to leverage the aforementioned result in practice to construct confidence sets, it is required to estimate the asymptotic covariance matrix Σ. The batch-means estimator (Lahiri, 2003; Flegal and Jones, 2010) from the larger Markov chain literature has been recently adapted in the literature for developing *online* estimators of Σ; see, for example, Zhu et al. (2023) and Roy and Balasubramanian (2023). The key idea is to divide the iterates into blocks of increasing size, with each block providing an approximately independent estimate of the covariance matrix. The block sizes are carefully chosen to balance the bias-variance tradeoff while maintaining the desirable convergence rate. Specifically, let $\{a_m\}_m$ be a strictly increasing sequence of integers with $a_1 = 1$. For any $k = 1, 2, \ldots$, we construct a block B_k consisting of the iterates $\{x_{t_k}, x_{t_k+1}, \ldots, x_k\}$ where $t_k = a_m$ for $k \in [a_m, a_{m+1})$. Let $l_k = |B_k|$ denote the size of the block B_k. After n iterations, the batch-means covariance estimator is given by:

$$\hat{\Sigma}_n = \frac{\sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} x_k - l_i \bar{x}_n\right) \left(\sum_{k=t_i}^{i} x_k - l_i \bar{x}_n\right)^\top}{\sum_{i=1}^{n} l_i}. \tag{1.4}$$

Zhu et al. (2023) showed that for SGD with i.i.d. data stream, $\hat{\Sigma}_n$ (asymptotically) consistently estimates Σ with a convergence rate of order $O(n^{-1/8})$. Subsequently, Roy and Balasubramanian (2023) extended this result to Markovian data. However, these limited existing works on online covariance estimation for first-order methods apply only to smooth and strongly convex problems, and their analyses do not apply to generic iterations as in (1.3).

Main Contribution. Our main contribution is to show that, despite significant complexity introduced by nonsmooth geometry, we can achieve the same convergence rate as in the smooth case using the same covariance estimator (1.4). In particular, we establish that under reasonable conditions and with a properly chosen batch size control sequence $\{a_m\}_m$, the online batch-means estimator $\hat{\Sigma}_n$ in (1.4) with generic SA iterates (1.3) satisfies

$$\mathbb{E}\|\hat{\Sigma}_n - \Sigma\|_2 = O(\sqrt{d}n^{-1/8+\varepsilon}) \quad \text{for any } \varepsilon > 0.$$

We also emphasize that when applying our result to stochastic optimization problems, the objective does not need to be strongly convex or even convex. This is in contrast with all existing works that heavily rely on global strong convexity (Chen et al., 2020; Zhu et al., 2023; Roy and Balasubramanian, 2023). Our analysis addresses the following main challenges:

1. Due to the nonsmooth nature of problem (1.2), Taylor's theorem – on which all existing methods (Chen et al., 2020; Zhu et al., 2023; Roy and Balasubramanian, 2023) are based – is no longer applicable. Our key insight is that, despite the problem being nonsmooth, typical instances exhibit partial smoothness near the solution. In other words, there exists a distinctive manifold containing the solution and capturing the hidden smoothness of the map F. In a local neighborhood around the solution, we project all iterates onto this manifold, forming what we

call the *shadow sequence*. We then prove that the shadow sequence behaves almost as if it were generated by a smooth dynamic.

2. Our analysis of the shadow sequence builds on prior work on nonsmooth asymptotic normality (Davis et al., 2024); however, their asymptotic guarantees are insufficient for our non-asymptotic study. In this work, we provide a more refined analysis and establish a tighter bound on the distance between the original iterates and their shadows. Our results show that the hypothetical batch-means estimator constructed from the shadow sequence converges to the same limit – and at the same rate – as the estimator based on the original sequence (1.4). Consequently, the problem reduces to analyzing the estimator derived from smooth dynamics.

3. Due to the local nature of both the manifold and the shadow sequence, the above argument holds only when the iterates remain within a local neighborhood of the solution. To address this, we introduce a stopping time. Under light-tailed noise, we apply a martingale concentration inequality to show that, with high probability, the original iterates stay within the local neighborhood after a certain number of iterations. Consequently, the shadow sequence always exists, and the stopping time can finally be dropped in the convergence guarantee.

We should mention that our above techniques extend beyond the covariance estimation problem, offering a template for analyzing other nonsmooth SA algorithms whose dynamics are implicitly governed by an underlying local smooth structure.

Paper organization. In Section 2, we introduce the notations and preliminaries, including smooth manifold and nonsmooth analysis. In Section 3, we present the assumptions and main results. In Section 4, we address the issue of the stopping time involved in our main results by providing a high-probability guarantee. In Section 5, we present specific examples of SA algorithms for nonsmooth problems, and we conclude and discuss future work in Section 6. Concrete examples of nonsmooth variational inclusion problems satisfying our assumptions, as well as the proofs of theoretical results, are deferred to the appendix.

2. Notations and preliminaries

Notations. Throughout the paper, the symbol $\mathbb{R}^d$ denotes a Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. The symbol $\mathbb{B}$ denotes the closed unit ball in $\mathbb{R}^d$, while $B_r(x)$ denotes the closed ball of radius r around a point x. When $A \in \mathbb{R}^{m \times n}$ is a matrix, $\|A\|_2$ denotes the spectral norm of A. For any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its *domain* is defined as $\operatorname{dom} f := \{x \in \mathbb{R}^d : f(x) < \infty\}$. We say f is *closed* if its epigraph is a closed set, or equivalently if f is lower-semicontinuous. The *proximal map of f with parameter* $\alpha > 0$ is given by

$$\operatorname{prox}_{\alpha f}(x) := \operatorname*{argmin}_{y} \left\{ f(y) + \frac{1}{2\alpha} \|y - x\|_2^2 \right\}.$$

The *distance* and the *projection* of a point $x \in \mathbb{R}^d$ onto a set $Q \subset \mathbb{R}^d$ are, respectively,

$$d(x, Q) := \inf_{y \in Q} \|y - x\|_2 \qquad \text{and} \qquad P_Q(x) := \operatorname*{argmin}_{y \in Q} \|y - x\|_2.$$

The indicator function of Q, denoted by $\delta_Q(\cdot)$, is defined to be zero on Q and $+\infty$ off it. The symbol $o(h)$ stands for any function $o(\cdot)$ satisfying $o(h)/h \to 0$ as $h \searrow 0$.

Smooth manifold. To be self-contained, we make a few definitions for smooth manifold; we refer the reader to Lee (2013); Boumal (2020) for details. Throughout the paper, all smooth manifolds $\mathcal{M}$ are assumed to be embedded in $\mathbb{R}^d$, and we consider the tangent and normal spaces to $\mathcal{M}$ as subspaces of $\mathbb{R}^d$. In particular, for any $x \in \mathcal{M}$, we denote the tangent and normal spaces of $\mathcal{M}$ at x by $T_{\mathcal{M}}(x)$ and $N_{\mathcal{M}}(x)$, respectively. A map $F\colon \mathcal{M} \to \mathbb{R}^m$ is called C^p ($p \geq 1$) smooth near a point x if there exists a C^p-smooth map $\hat{F}\colon U \to \mathbb{R}^d$ defined on some neighborhood $U \subset \mathbb{R}^d$ of x that agrees with F on $\mathcal{M}$ near x. In this case, we define the *covariant Jacobian* $\nabla_{\mathcal{M}} F(x)\colon T_{\mathcal{M}}(x) \to \mathbb{R}^m$ by the expression $\nabla_{\mathcal{M}} F(x)(u) = \nabla \hat{F}(x)u$ for all $u \in T_{\mathcal{M}}(x)$.

Nonsmooth analysis. Next, we introduce a few terminologies used in nonsmooth and variational analysis. The introduction follows Rockafellar and Wets (2009). Consider a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and a point $x \in \operatorname{dom} f$. The *Fréchet subdifferential of f at x*, denoted $\hat{\partial} f(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying the approximation property:

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \quad \text{as} \quad y \to x.$$

The *limiting subdifferential of f at x*, denoted $\partial f(x)$, consists of all vectors $v \in \mathbb{R}^d$ such that there exist sequences $x_i \in \mathbb{R}^d$ and Fréchet subgradients $v_i \in \hat{\partial} f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (x, f(x), v)$ as $i \to \infty$. A point x satisfying $0 \in \partial f(x)$ is called *critical* for f. For any set Q and $x \in Q$, the *Fréchet normal cone of Q at x* is defined by $\hat{N}_Q(x) := \hat{\partial}\delta_Q(x)$, where δ_Q is the indicator function of Q. Similarly, the *limiting normal cone of Q at x* is defined by $N_Q(x) := \partial\delta_Q(x)$.

3. Assumptions and main results

Setting the stage, our goal is to find a point x satisfying the inclusion

$$0 \in F(x), \tag{3.1}$$

where $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued map. Throughout, we fix one such solution $x^\star$ of (3.1). We assume the existence of a distinctive manifold $\mathcal{M}$ that contains $x^\star$ and satisfies the property that the map $x \mapsto P_{T_{\mathcal{M}}(x)} F(x)$ is single-valued and C^p-smooth on $\mathcal{M}$ near $x^\star$. The following assumption provides a precise statement of this assumption.

Assumption 1 (Smooth structure) Suppose that there exists a C^p ($p \geq 1$) manifold $\mathcal{M} \subset \mathbb{R}^d$ such that the map $F_{\mathcal{M}}\colon \mathcal{M} \to \mathbb{R}^d$ defined by $F_{\mathcal{M}}(x) := P_{T_{\mathcal{M}}(x)} F(x)$ is single-valued and C^p smooth on some neighborhood V of $x^\star$ in $\mathcal{M}$. Moreover, there exists $\gamma > 0$ and $L_{\mathcal{M}} > 0$ such that $F_{\mathcal{M}}$ is $L_{\mathcal{M}}$-Lipschitz in $V \cap \mathcal{M}$, and for any $x \in V \cap \mathcal{M}$,

$$\langle F_{\mathcal{M}}(x), x - x^\star\rangle \geq \gamma\|x - x^\star\|^2. \tag{3.2}$$

Note that in the case when $F = \nabla f$ for some smooth function f, the manifold $\mathcal{M}$ is simply $\mathbb{R}^d$, and the condition (3.2) is equivalent to the local quadratic growth condition (Davis and Jiang, 2022). To illustrate the role of manifold $\mathcal{M}$ for nonsmooth map F, we consider the following two examples: ℓ_1-regularization problems and nonlinear programming. A detailed discussion of these and more examples can be found in Appendix A.

Example 1 (ℓ_1-regularization) Consider the stochastic optimization problem with ℓ_1 regularization

$$\min_x \ g(x) = f(x) + \lambda\|x\|_1,$$

where $f(x) = \mathbb{E}_{\nu\in\mathcal{P}}[f(x,\nu)]$ is a C^p-smooth function in $\mathbb{R}^d$. Consider now $x^\star \in \mathbb{R}^d$, a critical point of the function g, and define the index set $\mathcal{I} = \{i\colon x_i^\star = 0\}$. Then, the set $\mathcal{M} = \{x\colon x_i = 0,\ \forall i \in \mathcal{I}\}$ is an affine space, hence a smooth manifold. It is easy to show that when $\nabla^2 f(x^\star)$ is positive definite restricted onto $T_{\mathcal{M}}(x^\star)$, the map $F = \partial g$ satisfies Assumption 1 with manifold $\mathcal{M}$.

Example 2 (Nonlinear programming) Consider the problem of nonlinear programming

$$\begin{aligned}
&\min_x \ f(x), \\
&\text{s.t.}\ \ g_i(x) \le 0 \qquad \text{for } i = 1,\ldots,m, \\
&\qquad\ g_i(x) = 0 \qquad \text{for } i = m+1,\ldots,n,
\end{aligned}$$

where f and g_i are C^p-smooth functions on $\mathbb{R}^d$. Let $\mathcal{X}$ denote the set of all feasible points to the problem. Consider now a point $x^\star \in \mathcal{X}$ that is critical for the function $f + \delta_{\mathcal{X}}$, and define the active index set $\mathcal{I} = \{i : g_i(x^\star) = 0\}$. Suppose the Linear Independence Constraint Qualification (LICQ) condition holds, i.e., the gradients $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent. Then, the set $\mathcal{M} = \{x : g_i(x) = 0\ \forall i \in \mathcal{I}\}$ is a C^p smooth manifold locally around $x^\star$. In the literature on nonlinear programming, the manifold $\mathcal{M}$ is also referred to as the active set (Nocedal and Wright, 2006). Define the Lagrangian function

$$\mathcal{L}(x,y) := f(x) + \sum_{i=1}^{n+m} y_i g_i(x).$$

The criticality of $x^\star$ and LICQ ensure that there exists a (unique) Lagrange multiplier vector $y^\star \in \mathbb{R}^m_+ \times \mathbb{R}^n$ satisfying $\nabla_x \mathcal{L}(x^\star, y^\star) = 0$ and $y_i^\star = 0$ for all $i \notin \mathcal{I}$. Assume in addition that $\nabla^2_{xx}\mathcal{L}(x^\star, y^\star)$ is positive definite when restricted onto $T_{\mathcal{M}}(x^\star)$, often called the Second-Order Sufficient Condition (SOSC); we can then show that $F = \nabla f + N_{\mathcal{X}}$ satisfies Assumption 1 with the manifold $\mathcal{M}$.

The stochastic approximation (SA) algorithms we consider in this work assume access to a *generalized gradient mapping* $G : \mathbb{R}_{++} \times \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$. As stated in Section 1, given x_0, our generic SA algorithm iterates as

$$x_{k+1} = x_k - \eta_{k+1} G_{\eta_{k+1}}(x_k, \nu_{k+1}), \quad \forall k \ge 0, \tag{3.3}$$

where $\eta_{k+1} > 0$ is a stepsize sequence and ν_k is stochastic noise. We now state two assumptions on G that are required in Davis et al. (2024) for establishing the asymptotic normality of the averaged iterates of (3.3). The first assumption is similar to classical Lipschitz assumptions and ensures that the stepsize length can only scale linearly in $\|\nu\|$.

Assumption 2 (Steplength) We suppose there exist a constant $C > 0$ and a neighborhood $\mathcal{U}$ of $x^\star$ such that the map G satisfies $\sup_{x\in\mathcal{U}_F} \|G_\eta(x,\nu)\| \le C(1 + \|\nu\|)$ for any $\nu \in \mathbb{R}^d$ and $\eta > 0$, where we set $\mathcal{U}_F := \mathcal{U} \cap \mathrm{dom} F$.

The second assumption precisely characterizes the relationship between two mappings, G and $F_{\mathcal{M}}$. For simplicity, we abuse the notation C to denote a general upper bound.

Assumption 3 We suppose that there exist constants $C, \mu > 0$, a manifold $\mathcal{M}$ containing $x^\star$, and a neighborhood $\mathcal{U}$ of $x^\star$ such that the following hold for any $\nu \in \mathbb{R}^d$ and $\eta > 0$, where we set $\mathcal{U}_F := \mathcal{U} \cap \mathrm{dom} F$:

1. **(Tangent comparison)** For any $x \in \mathcal{U}_F$, we have
$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(G_\eta(x,\nu) - F(P_{\mathcal{M}}(x)) - \nu)\| \leq C(1 + \|\nu\|)^2(\mathrm{dist}(x, \mathcal{M}) + \eta).$$
2. **(Proximal Aiming)** For any $x \in \mathcal{U}_F$, we have
$$\langle G_\eta(x,\nu) - \nu, x - P_{\mathcal{M}}(x)\rangle \geq \mu \cdot \mathrm{dist}(x, \mathcal{M}) - (1 + \|\nu\|)^2(o(\mathrm{dist}(x, \mathcal{M})) + C\eta).$$

In the above assumption, Item 1 asserts that in the tangent directions of $\mathcal{M}$, the gradient map G accurately approximates the map F; while Item 2 asserts that in the normal directions, the gradient map G points outward from $\mathcal{M}$. In the context of stochastic optimization, Assumptions 1–3 neither imply global strong convexity nor global convexity. See Example 4 in Appendix A for a concrete example. These broader and weaker assumptions extend the scope of existing online inference works, which have focused solely on strongly convex problems (Chen et al., 2020; Zhu et al., 2023; Roy and Balasubramanian, 2023).

In the next two assumptions, we consider the choice of stepsize and the conditions on stochastic noise for online covariance estimation.

Assumption 4 We assume the following conditions hold.

1. The map G_η is measurable.
2. The stepsize $\eta_k = \eta k^{-\alpha}$ for some $\eta > 0$ and $\alpha \in (\frac{1}{2}, 1)$.
3. $\{\nu_{k+1}\}$ is a martingale difference sequence w.r.t. to the increasing sequence of σ-fields $\mathcal{F}_k = \sigma(x_{0:k}, \nu_{1:k})$. Furthermore, there exists a function $q \colon \mathbb{R}^d \to \mathbb{R}_+$ that is bounded on bounded sets satisfying $\mathbb{E}_k[\|\nu_{k+1}\|^8] \leq q(x_k)$, where $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$.
4. The inclusion $x_k \in \mathrm{dom} F$ holds for all $k \geq 0$.

Assumption 4 on the stepsize and noise is almost identical to (Davis et al., 2024, Assumption I) for establishing asymptotic normality guarantees. The only difference is the requirement of the eighth moment of $\|\nu_k\|$, whereas Davis et al. (2024) requires only the fourth moment. A stricter noise moment condition appears to be natural for the covariance estimation problem. For example, the noise moment condition for covariance estimation of simple SGD method is also stricter than the moment condition needed for asymptotic normality; see Polyak and Juditsky (1992) and Chen et al. (2020); Zhu et al. (2023) for comparisons.

We next impose an additional assumption concerning the covariance of the stochastic noise ν_k. Similar assumptions also widely appear in the literature on both first-order methods (Duchi and Ruan, 2021; Davis et al., 2024; Chen et al., 2020; Zhu et al., 2023; Roy and Balasubramanian, 2023) and second-order methods (Bercu et al., 2020; Na and Mahoney, 2022).

Assumption 5 Fix $x^\star \in \mathrm{dom} F$ at which Assumption 1 holds and let U be a matrix whose columns form an orthogonal basis of $T_{\mathcal{M}}(x^\star)$. We assume the gradient noise can be decomposed as $\nu_{k+1} = \nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k)$, where $\nu_{k+1}^{(2)} \colon \mathrm{dom} F \to \mathbb{R}^d$ is a random function satisfying for some $C > 0$,
$$\mathbb{E}_k[\|\nu_{k+1}^{(2)}(x)\|^2] \leq C\|x - x^\star\|^2 \qquad \text{for all } x \in \mathrm{dom} F,$$

and $\mathbb{E}_k[\nu_{k+1}^{(2)}(x)] = \mathbb{E}_k[\nu_{k+1}^{(1)}] = 0$. In addition, we assume the following covariance matrix is constant for all $k \geq 1$:

$$S := \mathbb{E}_k[U^\top \nu_k^{(1)} {\nu_k^{(1)}}^\top U]. \tag{3.4}$$

Note that all the previous assumptions regulate only the local behavior of the maps F and G. To control the behavior of the iterates far from $x^\star$, we impose the following mild assumption and rigorously show that it holds for a variety of nonsmooth SA methods in Appendix E.

Assumption 6 (Bounded sequence in expectation) There exists a constant $C_{\text{ub}} > 0$ such that $\mathbb{E}[\|x_k - x^\star\|^2] \leq C_{\text{ub}}$.

Let U be a matrix whose columns form an orthonormal basis of $T_{\mathcal{M}}(x^\star)$. We recall that the limiting covariance matrix in the nonsmooth asymptotic normality result takes the following form (Davis et al., 2024, Theorem 5.1):

$$\Sigma := U(U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(x^\star)U)^{-1} S (U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(x^\star)U)^{-\top} U^\top, \tag{3.5}$$

where $\nabla_{\mathcal{M}} F_{\mathcal{M}}(x^\star)$ is the covariant Jacobian of $F_{\mathcal{M}}$, and S is defined in (3.4).

We are now ready to state our main result on the convergence of the online batch-means covariance estimator (1.4). The formal statement of our result crucially relies on local arguments and frequently refers to the following stopping time: given an index $k \geq 0$ and a constant $\delta \in (0,1)$, we define

$$\tau_{k,\delta} := \inf\{l \geq k \colon x_l \notin B_\delta(x^\star)\},$$

which is the first time after k that the iterate leaves $B_\delta(x^\star)$. The following is our main convergence theorem, with its proof provided in Appendix B.

Theorem 1 *Under Assumptions 1–6, let us set $a_m = \lfloor Cm^\beta \rfloor$ for some constant $C \geq 1$ and $\beta > \frac{1}{1-\alpha}$. Then, for the iteration scheme (3.3) and any $k_s \leq n$, we have*[1]

$$\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta} > n}] \lesssim k_s^3 (dn^{\frac{(\alpha-1)+\beta}{\beta}} + \sqrt{d}n^{\frac{(\alpha-1)+\beta}{2\beta}} + \sqrt{d}n^{-\frac{1}{2\beta}}).$$

Remark 2 *Choosing $\beta = \frac{2}{1-\alpha}$, we have*

$$\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta} > n}] \lesssim k_s^3 (dn^{-\frac{1-\alpha}{2}} + \sqrt{d}n^{\frac{1-\alpha}{4}}).$$

Further choosing $\alpha = \frac{1}{2} + 4\varepsilon$ for some arbitrarily small $\varepsilon > 0$, we have

$$\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta} > n}] \lesssim k_s^3 (dn^{-\frac{1}{4}+2\varepsilon} + \sqrt{d}n^{-\frac{1}{8}+\varepsilon}). \tag{3.6}$$

A comparison of Theorem 1 with related settings is in order. In particular, (3.6) shows that as long as k_s is a constant, we recover the convergence rate in the smooth case with an i.i.d data stream (Zhu et al., 2023). In Section 4, we show that under mild assumptions, the probability that the iterates leave the local neighborhood after k_s decays exponentially in k_s. Moreover, by allowing $k_s \asymp \log^2 n$, we recover the best-known convergence rate $O(n^{-1/8})$ in the smooth case up to logarithmic factors. More interestingly, this rate also matches the rate obtained in the smooth case for exponentially mixing Markovian data streams (Roy and Balasubramanian, 2023).

1. In the rest of the paper, we use $a_n \lesssim b_n$ to denote $a_n \leq Cb_n$ for some constant C independent of k_s (if applicable), d and n, and $a_n \asymp b_n$ to denote $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Proof ideas. Our key insight is that, by Item 2 of Assumption 3, the iteration sequence x_k generated by the dynamics (3.3) can be locally but closely approximated by its projection onto $\mathcal{M}$, namely, the "shadow sequence" defined as

$$y_k = P_{\mathcal{M}}(x_k).$$

By carefully quantifying the distance between x_k and y_k, we show that this error decays sufficiently fast so that the hypothetical batch-means estimator constructed with the shadow sequence y_k, similar to (1.4), converges to the same limit – and at the same rate – as the estimator constructed with x_k. Consequently, it suffices to analyze the convergence of the batch-means estimator applied to y_k.

Another crucial implication of Assumption 3 is that the update rule of y_k can be interpreted as an inexact Riemannian SA algorithm operating on the restriction of F to the manifold $\mathcal{M}$. More precisely, we show that the shadow sequence exhibits the recursion

$$y_{k+1} = y_k - \eta_{k+1} F_{\mathcal{M}}(y_k) - \eta_{k+1} P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \text{Error}_k.$$

For the sake of illustration, let us first assume that $\text{Error}_k = 0$. Due to Assumption 1, the dynamics of y_k are smooth, allowing us to adapt the analysis of batch-means estimators developed in the context of stochastic smooth optimization (Chen et al., 2020; Zhu et al., 2023). In the more general setting, we derive sharp upper bounds on the error terms and demonstrate that their contribution to the covariance estimation error is dominated by the convergence rate established in the smooth case.

Note that our main result is local and relies on the stopping time $\tau_{k_s,\delta}$. In this regard, we show in the following section that, under sub-Gaussian noise conditions, the iterates remain near the solution with high probability. Our analysis leverages martingale concentration inequalities applied to (3.3).

4. High probability guarantee

So far, we have only made assumptions on F and G locally near $x^\star$, except for assuming the sequence x_k is bounded in expectation (as proved in Appendix E). To establish global convergence guarantees, we require the following assumption.

Assumption 7 We assume that there are constants $\gamma, C > 0$ such that:

1. **(Aiming towards solution)** For any $x \in \mathbb{R}^d$, we have $\langle G_\eta(x,\nu) - \nu, x - x^\star \rangle \geq \gamma \|x - x^\star\|_2^2 - C\eta(1 + \|x - x^\star\|_2^2 + \|\nu\|_2^2)$.
2. **(Global steplength)** For any $x \in \mathbb{R}^d$, we have $\|G_\eta(x,\nu)\|_2^2 \leq C(1 + \|x - x^\star\|_2^2 + \|\nu\|_2^2)$.

Assumption 7 extends the standard strong convexity and Lipschitz gradient conditions commonly assumed in stochastic smooth optimization. In particular, we have $G_\eta(x,\nu) = \nabla f(x) + \nu$ in the case of minimizing a γ-strongly convex function f. Therefore, Item 1 is ensured by the γ-strong convexity, since $\langle G_\eta(x,\nu) - \nu,\ x - x^\star \rangle = \langle \nabla f(x),\ x - x^\star \rangle \geq \gamma \|x - x^\star\|_2^2$. Moreover, the Lipschitz gradient condition implies Item 2, as we observe that $\|G_\eta(x,\nu)\|_2 = \|\nabla f(x) + \nu\| \lesssim \|x - x^\star\| + \|\nu\|$. Beyond the smooth case, we show in Appendix E that Assumption 7 holds for various nonsmooth SA methods.

We additionally impose the following light-tail assumption on the noise.

Assumption 8 (Light tail) The noise ν_{k+1} is mean-zero norm sub-Gaussian conditioned on $\mathcal{F}_k$ with parameter $\sigma/2$, i.e., $\mathbb{E}_k[\nu_{k+1}] = 0$ and $\mathbb{P}_k\{\|\nu_{k+1}\| \geq \tau\} \leq 2\exp(-2\tau^2/\sigma^2)$ for all $\tau > 0$.

By standard results in high-dimensional statistics (Jin et al., 2019, Lemma 3), we know that $\|\nu_{k+1}\|^2$ is sub-exponential with parameter $c\sigma^2$ conditioned on $\mathcal{F}_k$, where c is some absolute constant. Below is a high-probability guarantee demonstrating that x_k stays within $B_\delta(x^\star)$ for all sufficiently large k. We present its proof in Appendix C.

Proposition 3 *Suppose Assumptions 7 and 8 hold. Let c be the universal constant defined above. Suppose also $\eta \le \min\left\{\frac{\gamma}{3C}, \frac{1}{3c\gamma C}\right\}$. Then, for any radius δ and any k such that*

$$k \ge \max\left\{\left(\frac{\log(4\|x_0-x^\star\|^2/\delta)}{C_\alpha\gamma\eta}\right)^{1/(1-\alpha)}, \left(\frac{\log\left(\frac{16\tilde{C}\alpha\eta^2}{(2\alpha-1)\delta}\right)}{C_\alpha\gamma\eta}\right)^{1/(1-\alpha)}, \left(\frac{2^{2\alpha+2}\tilde{C}\eta^2}{(2\alpha-1)\delta}\right)^{1/(2\alpha-1)}\right\},$$

where $\tilde{C} = 3cC\sigma^2 + 3C$ and $C_\alpha = \frac{1-0.5^{1-\alpha}}{2(1-\alpha)}$, we have

$$\mathbb{P}(\|x_i - x^\star\| < \delta, \forall i \ge k) \ge 1 - \frac{32\eta^2\sigma^4\exp\left(-\frac{\gamma\delta\sqrt{k}}{4\eta\sigma^2}\right)}{\gamma^2\delta^2} - \frac{8\eta\delta\sqrt{k}\exp\left(-\frac{\gamma\delta\sqrt{k}}{4\eta\sigma^2}\right)}{\gamma}.$$

With the above high-probability guarantee, we strengthen the local result in Theorem 1 to a global result by suppressing the stopping time involved in the theorem statement. Our global result is stated in Theorem 4. The proof can be found in Appendix D.

Theorem 4 *Under the assumptions of Theorem 1 along with Assumptions 7 and 8, for the SA update of (3.3), for $a_M \le n \le a_{M+1}$, we have*

$$\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_{op}] \lesssim_{\log} \sqrt{d}M^{-\frac{1}{2}} + \sqrt{d}M^{\frac{(\alpha-1)\beta+1}{2}} \lesssim \sqrt{d}n^{-\frac{1}{2\beta}} + \sqrt{d}n^{-\frac{(\alpha-1)\beta+1}{2\beta}},$$

where $\|\cdot\|_{op}$ is the operator norm, and "$\lesssim_{\log}$" hides logarithmic terms of n.

Taking $\beta = \frac{2}{1-\alpha}$ in Theorem 4, we have $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_{op}] \lesssim_{\log} \sqrt{d}n^{-\frac{1-\alpha}{4}}$. Ignoring the logarithmic factors, this matches the best-known rate in the smooth case (Chen et al., 2020; Zhu et al., 2023).

5. Examples of stochastic approximation algorithms

In this section, we illustrate the broad applicability of our generic SA update in (3.3) and the mildness of our required assumptions. In particular, we consider solving nonsmooth problems using different SA algorithms and provide sufficient conditions for Assumptions 1–3 to hold. More concretely, let us consider the variational inclusion problem:

$$0 \in A(x) + \partial g(x) + \partial f(x), \tag{5.1}$$

where $A : \mathbb{R}^d \to \mathbb{R}^d$ is any single-valued continuous map, $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed function, and $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed function that is bounded from below[2]. The problem (5.1) is a special case of (3.1) since one can take $F(x) := A(x) + \partial g(x) + \partial f(x)$. First, the local boundedness condition of G in Assumption 2 is widely used in the literature, with a variety of known sufficient conditions. The following lemma describes several such conditions, which we will use in what follows.

2. In particular, $\operatorname{prox}_{\alpha f}(x)$ is nonempty for all $x \in \mathbb{R}^d$ and all $\alpha > 0$.

Lemma 5 (Lemma 4.2 in Davis et al. (2024)) *Suppose $A(\cdot)$ and $s_g(\cdot)$ are locally bounded around $x^\star$. Then Assumption 2 holds in any of the following settings.*

1. *f is the indicator function of a closed set $\mathcal{X}$.*
2. *f is convex and the function $x \mapsto$ dist$(0, \partial f(x))$ is bounded on $\operatorname{dom} f$ near $x^\star$.*
3. *f is Lipschitz continuous on $\operatorname{dom} g \cap \operatorname{dom} f$.*

Then, we investigate Assumptions 1 and 3. Recall that both assumptions require the existence of a distinctive manifold $\mathcal{M}$ that captures the hidden smoothness of the problem. One candidate of such a manifold is the *active manifold*, which has been modeled in various ways, including identifiable surfaces (Wright, 1993), partial smoothness (Lewis, 2002), $\mathcal{UV}$-structures (Lemaréchal et al., 2000; Mifflin and Sagastizábal, 2005), $g \circ F$ decomposable functions (Shapiro, 2003), and minimal identifiable sets (Drusvyatskiy and Lewis, 2014). In this work, we adopt the characterization of active manifold used in Drusvyatskiy and Lewis (2014).

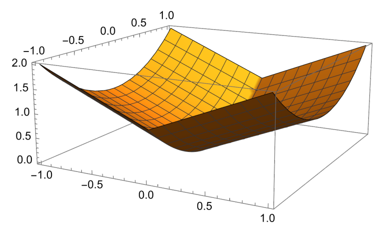

Figure 1: $f(x_1, x_2) = |x_1| + x_2^2$ with x_2-axis as an active manifold.

Definition 6 (Active manifold) Consider a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and fix a set $\mathcal{M} \subset \operatorname{dom} f$ that contains a critical point $x^\star$ with $0 \in \partial f(x^\star)$. Then $\mathcal{M}$ is called an *active C^p-manifold around $x^\star$* if there exists a constant $\chi > 0$ satisfying the following conditions.

- (smoothness) Near $x^\star$, the set $\mathcal{M}$ is a C^p manifold and the restriction of f to $\mathcal{M}$ is C^p-smooth.
- (sharpness) The lower bound holds:

$$\inf\{\|v\| : v \in \partial f(x),\ x \in U \setminus \mathcal{M}\} > 0$$

  where $U = \{x \in B_\chi(x^\star) : |f(x) - f(x^\star)| < \chi\}$.

More generally, we say $\mathcal{M}$ *is an active manifold for f at $x^\star$ for $\bar{v} \in \partial f(x^\star)$* if $\mathcal{M}$ is an active manifold for the tilted function $f_{\bar{v}}(x) = f(x) - \langle \bar{v}, x \rangle$ at $x^\star$.

The sharpness condition simply means that the subgradients of f remain uniformly bounded away from zero at points off the manifold that are sufficiently close to $x^\star$ in both distance and function value. The localization in function value can be omitted, for example, if f is weakly convex or if f is continuous on its domain; see Drusvyatskiy and Lewis (2014) for details. Figure 1 is an example of active manifold of a nonsmooth function.

To proceed, we introduce two extra conditions along the active manifold that tightly couple the subgradients of f on and off the manifold. These two conditions were first introduced in (Davis et al., 2025, Section 3) to prove saddle point avoidance in nonsmooth optimization. They are very mild conditions and hold for a wide range of examples. We verify these regularity conditions in detail for the cases of ℓ_1-regularization, nonlinear programming, and two-player game in Appendix A.

Definition 7 (($b_{\leq}$)-regularity and strong (a)-regularity) Consider a function $f\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ that is locally Lipschitz continuous on its domain. Fix a set $\mathcal{M} \subset \mathrm{dom} f$ that is a C^1 manifold around $x^\star$ and such that the restriction of f to $\mathcal{M}$ is C^1-smooth near $x^\star$. We say that f is ($b_{\leq}$)*-regular along* $\mathcal{M}$ *at* $x^\star$ if there exists $\chi > 0$ such that

$$f(y) \geq f(x) + \langle v, y - x\rangle + (1 + \|v\|) \cdot o(\|y - x\|)$$

holds for all $x \in \mathrm{dom} f \cap B_\chi(x^\star)$, $y \in \mathcal{M} \cap B_\chi(x^\star)$, and $v \in \partial f(x)$. Additionally, we say that f is *strongly* (a)*-regular along* $\mathcal{M}$ *near* $x^\star$ if there exist constants $C, \chi > 0$ satisfying

$$\|P_{T_{\mathcal{M}(y)}}(v - \nabla_{\mathcal{M}} f(y))\| \leq C(1 + \|v\|)\|x - y\|$$

for all $x \in \mathrm{dom} f \cap B_\chi(x^\star)$, $y \in \mathcal{M} \cap B_\chi(x^\star)$, and $v \in \partial f(x)$.

Roughly speaking, ($b_{\leq}$)-regularity condition is a weakening of Taylor's theorem for nonsmooth functions; strong (a)-regularity condition is a weakening of Lipschitz continuity of the gradient. We next provide sufficient conditions of Assumptions 1 and 3 in several popular settings, including projected SGD (hence Subgradient Descent) and projected Stochastic Gradient Descent Ascent methods.

5.1. Stochastic (projected) forward algorithm ($f = \delta_{\mathcal{X}}$)

First, we focus on the particular instance of (5.1) where f is an indicator function of a closed set $\mathcal{X}$. In this case, the iteration (3.3) reduces to a stochastic projected forward algorithm:

$$x_{k+1} \in P_{\mathcal{X}}(x_k - \eta_{k+1}(A(x_k) + s_g(x_k) + \nu_{k+1})).$$

The map G takes the form $G_\eta(x, \nu) := (x - s_{\mathcal{X}}(x - \eta(A(x) + s_g(x) + \nu)))/\eta$, where $s_{\mathcal{X}}(x)$ is any selection of the projection map $P_{\mathcal{X}}(x)$.

The following proposition shows that Assumptions 1 and 3 hold when $g + f$ admits an active manifold at $x^\star$ with certain regularity conditions. Its proof is a combination of Corollary 4.7 and Lemma 10.3 in Davis et al. (2024).

Proposition 8 *Suppose f is the indicator function of a closed set $\mathcal{X}$ and both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $x^\star$. Moreover, suppose the inclusion $-A(x^\star) \in \hat{\partial}(g + f)(x^\star)$ holds, $g + f$ admits a C^2 active manifold around $x^\star$ for the vector $\bar{v} = -A(x^\star)$, and both g and f are ($b_{\leq}$)-regular and strongly (a)-regular along $\mathcal{M}$ at $x^\star$. Then Assumption 3 holds. Furthermore, if there exists $\gamma > 0$ such that $\langle \nabla_{\mathcal{M}}(A + \partial g)(x^\star)v, v\rangle \geq \gamma\|v\|_2^2$, for all $v \in T_{\mathcal{M}}(x^\star)$, then Assumption 1 holds with manifold $\mathcal{M}$.*

5.2. Stochastic forward-backward method ($g = 0$)

Second, we focus on the particular instance of (5.1) where $g = 0$. In this case, the iteration (3.3) reduces to a stochastic forward-backward algorithm:

$$x_{k+1} \in \mathrm{prox}_{\eta_{k+1} f}(x_k - \eta_{k+1}(A(x_k) + \nu_{k+1})).$$

The map G becomes $G_\eta(x, \nu) := (x - s_f(x - \eta(A(x) + \nu))/\eta)$, where s_f is any selection of the proximal map $\mathrm{prox}_{\eta f}(x)$ (cf. Section 2).

The following proposition shows that Assumptions 1 and 3 hold when f admits an active manifold at $x^\star$ with certain regularity conditions. Its proof is a combination of Corollary 4.9 and Lemma 10.3 in Davis et al. (2024).

Proposition 9 *Suppose $g = 0$ and both f and $A(\cdot)$ are Lipschitz continuous on $\mathrm{dom} f$ near $x^\star$. Moreover, suppose the inclusion $-A(x^\star) \in \hat{\partial} f(x^\star)$ holds, f admits a C^2 active manifold around $x^\star$ for $\bar{v} = -A(x^\star)$, and f is both $(b)_\leq$-regular and strongly (a)-regular along $\mathcal{M}$ at $x^\star$. Then Assumption 3 holds. Furthermore, if there exists $\gamma > 0$ such that*

$$\langle \nabla_{\mathcal{M}}(A + \partial f)(x^\star)v, v\rangle \geq \gamma \|v\|_2^2, \qquad \textit{for all } v \in T_{\mathcal{M}}(x^\star),$$

then Assumption 1 holds with manifold $\mathcal{M}$.

6. Conclusion and future work

In this paper, we studied covariance estimation for nonsmooth stochastic approximation (SA) methods. The estimator was initially proposed for SGD in Zhu et al. (2023) for smooth, strongly convex optimization problems. The key idea is to group iterates into blocks of increasing size, with each block providing an approximately independent estimate of the covariance matrix. This estimator can be computed fully online, with both computation and memory scaling quadratically in dimension. Our work demonstrated that, with a properly chosen batch size control sequence, the same estimator achieves the expected convergence rate of order $O(\sqrt{d}n^{-1/8+\varepsilon})$ for any $\varepsilon > 0$ in nonsmooth and potentially non-monotone (nonconvex) setting. Our analysis involves highly nontrivial extensions of Zhu et al. (2023), where we developed a localization technique and constructed a shadow sequence to address the challenges arising from the lack of smoothness. Additionally, we established high-probability guarantees on the stopping time at which iterates leave the local neighborhood. The consistency of our covariance estimator enables asymptotically valid statistical inference for stochastic nonsmooth variational inclusion problems, covering numerous examples as provided in Appendix A.

One future research direction is studying covariance estimation for nonsmooth SA methods under Markovian noise, inspired by reinforcement learning applications. In addition, an open and challenging question is establishing the lower bound of covariance estimation and investigating whether the estimator (1.4) for first-order methods is minimax optimal. Finally, designing non-asymptotically optimal (nonsmooth) SA methods along with suitable covariance estimators is also a promising topic for future research.

Acknowledgments

Liwei Jiang is partly supported by the Gatech ARC postdoc fellowship. Krishna Balasubramanian is supported by NSF DMS-2413426 and NSF DMS 2053918 awards. Damek Davis is supported by an Alfred P. Sloan research fellowship and NSF DMS award 2047637. Dmitriy Drusvyatskiy is supported by NSF DMS-2306322, NSF CCF 1740551, AFOSR FA9550-24-1-0092 awards.

References

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*, pages 115–137. PMLR, 2019.

Bernard Bercu, Antoine Godichon, and Bruno Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1): 348–367, 2020.

Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, Aug*, 2020.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1):251–273, 2020.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of machine learning research*, 24(147):1–56, 2023.

Damek Davis and Liwei Jiang. A nearly linearly convergent first-order method for nonsmooth functions with quadratic growth. *arXiv preprint arXiv:2205.00064*, 2022.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *The Annals of Statistics*, 52(4):1485–1508, 2024.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization. *Foundations of Computational Mathematics*, pages 1–83, 2025.

Dmitriy Drusvyatskiy and Adrian S Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, 147(1):467–498, 2014.

John C Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

James M Flegal and Galin L Jones. Batch means and spectral variance estimators in markov chain monte carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2003.

John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.

Claude Lemaréchal, François Oustry, and Claudia Sagastizábal. The $\mathcal{U}$-lagrangian of a convex function. *Transactions of the American mathematical Society*, 352(2):711–729, 2000.

Adrian S Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3): 702–725, 2002.

Chris Junchi Li, Yaodong Yu, Nicolas Loizou, Gauthier Gidel, Yi Ma, Nicolas Le Roux, and Michael Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pages 9793–9826. PMLR, 2022.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Robert Mifflin and Claudia Sagastizábal. A VU-algorithm for convex minimization. *Mathematical programming*, 104(2):583–608, 2005.

Sen Na and Michael W Mahoney. Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *arXiv preprint arXiv:2205.13687*, 2022.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

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

Abhishek Roy and Krishnakumar Balasubramanian. Online covariance estimation for stochastic gradient descent under markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning. *arXiv preprint arXiv:2405.16644*, 2024.

Qi-Man Shao and Zhuo-Song Zhang. Berry–esseen bounds for multivariate nonlinear statistics with applications to m-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3): 1548–1576, 2022.

Alexander Shapiro. On a class of nonsmooth composite functions. *Mathematics of Operations Research*, 28(4):677–692, 2003. doi: 10.1287/moor.28.4.677.20512. URL https://doi.org/10.1287/moor.28.4.677.20512.

Panos Toulis and Edoardo Maria Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 2017.

Stephen J Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

Appendix A. Concrete Examples

In this section, we expand on the discussion in Section 3 and provide some concrete examples that satisfy Assumptions 1–3.

Example 1 (ℓ_1-regularization) Consider the stochastic optimization problem with ℓ_1 regularization

$$\min_x \; g(x) = f(x) + \lambda\|x\|_1,$$

where $f(x) = \mathbb{E}_{\nu\in\mathcal{P}}[f(x,\nu)]$ is a C^p-smooth function in $\mathbb{R}^d$. Consider now a point $x^\star \in \mathbb{R}^d$ that is critical for the function g and define the index set $\mathcal{I} = \{i\colon x_i^\star = 0\}$. Then, the set

$$\mathcal{M} = \{x\colon x_i = 0,\ \forall i \in \mathcal{I}\}$$

is an affine space, hence a smooth manifold. Note that the definition of criticality ensures that $0 \in \partial g(x^\star)$, so we always have

$$-(\nabla f(x^\star))_i \in [-\lambda, \lambda], \qquad \forall i \in \mathcal{I}.$$

Suppose the following condition is true:

- **(Strict complementarity)** $-(\nabla f(x^\star))_i \in (-\lambda, \lambda)$ for all $i \in \mathcal{I}$.

Then $\mathcal{M}$ is indeed an active C^p manifold of g at $x^\star$. Moreover, $(b_{\le})$-regularity and strong (a)-regularity hold trivially for g along $\mathcal{M}$ at $x^\star$. If, in addition, $\nabla^2 f(x^\star)$ is positive definite when restricted to the tangent space of $\mathcal{M}$, then Proposition 8 and Lemma 5 imply that Assumptions 1–3 hold for the stochastic subgradient method; similarly, Proposition 5.2 and Lemma 5 imply that these assumptions also hold for the stochastic proximal gradient method. We mention that there is typically a bias between the center of the asymptotic normality, $x^\star$, and the minimizer of f due to the presence of the regularization term.

Example 2 (Nonlinear programming) Consider the problem of nonlinear programming

$$\begin{aligned} &\min_x \; f(x), \\ &\text{s.t. } g_i(x) \le 0 \qquad \text{for } i = 1, \dots, m, \\ &\qquad\; g_i(x) = 0 \qquad \text{for } i = m+1, \dots, n, \end{aligned}$$

where f and g_i are C^p-smooth functions on $\mathbb{R}^d$. Let us denote the set of all feasible points to the problem as

$$\mathcal{X} = \{x : g_i(x) \le 0 \;\text{ for }\; 1 \le i \le m \quad \text{and} \quad g_i(x) = 0 \;\text{ for }\; m+1 \le i \le n\}.$$

Consider now a point $x^\star \in \mathcal{X}$ that is critical for the function $f + \delta_{\mathcal{X}}$ and define the active index set

$$\mathcal{I} = \{i : g_i(x^\star) = 0\}.$$

Suppose the following is true:

- **(LICQ)** the gradients $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent.

Then the set

$$\mathcal{M} = \{x : g_i(x) = 0,\ \forall i \in \mathcal{I}\}$$

is a C^p smooth manifold locally around $x^\star$. Moreover, all three functions f, $\delta_{\mathcal{X}}$, and $f + \delta_{\mathcal{X}}$ are $(b_{\leq})$-regular and strongly (a)-regular along $\mathcal{M}$ near $x^\star$. To ensure that $\mathcal{M}$ is an active manifold of $f + \delta_{\mathcal{X}}$, an extra condition is required. Define the Lagrangian function

$$\mathcal{L}(x, y) := f(x) + \textstyle\sum_{i=1}^{n+m} y_i g_i(x).$$

The criticality of $x^\star$ and LICQ ensure that there exists a (unique) Lagrange multiplier vector $y^\star \in \mathbb{R}^m_+ \times \mathbb{R}^n$ satisfying $\nabla_x \mathcal{L}(x^\star, y^\star) = 0$ and $y_i^\star = 0$ for all $i \notin \mathcal{I}$. Suppose the following standard assumption is true:

- **(Strict complementarity)** $y_i^\star > 0$ for all $i \in \mathcal{I} \cap \{1, \ldots, m\}$.

Then $\mathcal{M}$ is indeed an active C^p manifold for $f + \delta_{\mathcal{X}}$ at $x^\star$. Assume in addition that $\nabla^2_{xx}\mathcal{L}(x^\star, y^\star)$ is positive definite when restricted onto $T_{\mathcal{M}}(x^\star)$, often called the Second-Order Sufficient Condition (SOSC) in nonlinear programming literature (Nocedal and Wright, 2006); Proposition 8 and Lemma 5 imply that Assumptions 1–3 hold for stochastic projected gradient method.

Example 3 (Entropy-regularized zero-sum two-player matrix game) Consider the following optimization problem that arises in an zero-sum two-player matrix game (Cen et al., 2021; Li et al., 2022)

$$\operatorname*{argmin}_{z \in \Delta^{d-1}} \operatorname*{argmax}_{w \in \Delta^{d-1}} f(z, w) := z^\top \mathbb{E}\left[\mathcal{A}_\xi\right] w + \lambda \mathcal{H}(z) - \lambda \mathcal{H}(w), \tag{A.1}$$

where Δ^{d-1} is the d-dimensional probability simplex, λ is the regularization parameter, and $\mathcal{H}(\mu) = -\sum_{i=1}^d \mu_i \log \mu_i$ is the entropy regularization. The regularization is often imposed to account for the imperfect knowledge about the payoff matrix $\mathcal{A} = \mathbb{E}\left[\mathcal{A}_\xi\right]$ (Mertikopoulos and Sandholm, 2016). The solution of the above problem is known as the Quantal Response Equilibrium (QRE) in game theory (McKelvey and Palfrey, 1995). In particular, the solution of (A.1) turns out to be the solution of the following fixed point equation:

$$z_i^\star \propto \exp([Aw^*]_i/\lambda) \qquad w_i^\star \propto \exp(-[Az^\star]_i/\lambda) \qquad \forall 1 \leq i \leq d.$$

Let $\mathcal{X} = \Delta^{d-1} \times \Delta^{d-1} \subset \mathbb{R}^{2d}$, then problem (A.1) can be reformulated as the following variational inclusion problem:

$$0 \in \begin{bmatrix} \nabla_z f(z, w) \\ -\nabla_w f(z, w) \end{bmatrix} + N_{\mathcal{X}}(z, w).$$

Observe that $(z^\star, w^\star)$ lies in the relative interior of $\Delta^{d-1} \times \Delta^{d-1}$. Consequently,

$$\mathcal{M} := \left\{(z, w) : \textstyle\sum_{i=1}^d z_i = 1, \sum_{i=1}^d w_i = 1\right\}$$

is an active manifold of $\delta_{\mathcal{X}}$ at $(z^\star, w^\star)$ for $-\begin{bmatrix} \nabla_z f(z, w) \\ -\nabla_w f(z, w) \end{bmatrix}$. Also, it is trivial to show that $\delta_{\mathcal{X}}$ is both $(b_{\leq})$-regular and strong (a)-regular along $\mathcal{M}$ at $(z^\star, w^\star)$. Moreover, Cen et al. (2021) showed that f is strongly-convex strongly-concave locally near $(z^\star, w^\star)$, so a combination of Proposition 8 and Lemma 5 implies that Assumptions 1–3 hold for stochastic projected forward method.

The following is a nonconvex and nonsmooth function satisfying Assumptions 1–3 for the stochastic subgradient method.

Example 4 (Nonconvex example) Consider the function with the origin as the minimizer:

$$f(x,y) = |x - y^2| + \frac{x^2+y^2}{2}.$$

Note that for any $0 < t < 1$, we have

$$f(t^2,t) + f(t^2,-t) = (t^4+t^2) < 2t^2 + t^4 = 2f(t^2,0),$$

which implies that f is not convex in any local neighborhood of the origin. Meanwhile, one can easily check that $\mathcal{M} = \{(x,y)\colon x = y^2\}$ is an active manifold of f at the origin, and f is both $(b_{\leq})$-regularity and strong (a)-regularity along $\mathcal{M}$ at the origin. Moreover, $\nabla_{\mathcal{M}} f(0,0)$ is positive definite on the y-axis. A combination of Proposition 8 and Lemma 5 implies that Assumptions 1–3 hold for stochastic subgradient method.

Appendix B. Proof of Theorem 1

We introduce some more notations for the rest of this section. First, by our choice that $a_m = \lfloor Cm^\beta \rfloor$, we have $n_m \asymp m^{\beta-1}$. Let M be an integer such that $a_M \leq n < a_{M+1}$. Let $H := U^\top \nabla_{\mathcal{M}} F_{\mathcal{M}}(x^\star) U$. Note that H is not necessarily a symmetric matrix Davis et al. (2024). Define

$$W_i^j := \textstyle\prod_{k=i+1}^j (\mathrm{I} - \eta_k H) \quad \text{for } j > i \quad \text{with} \quad W_i^i := \mathrm{I},$$
$$S_i^j := \textstyle\sum_{k=i+1}^j W_i^k \quad \text{for } j > i \quad \text{with} \quad S_i^i := 0.$$

Let $\delta > 0$ be small enough so that Assumption 1 – 3 hold inside $B_\delta(x^\star)$. We consider the shadow sequence

$$y_k = \begin{cases} P_{\mathcal{M}}(x_k) & \text{if } x_k \in B_{2\delta}(x^\star) \\ x^\star & \text{otherwise.} \end{cases}$$

By Proposition 6.3 in Davis et al. (2024), there exists $\mathcal{F}_{k+1}$-measurable random vectors $E_k \in \mathbb{R}^d$ such that the shadow sequence satisfies $y_k \in B_{4\delta}(x^\star) \cap \mathcal{M}$ for all k and the recursion holds:

$$y_{k+1} = y_k - \eta_{k+1} F_{\mathcal{M}}(y_k) - \eta_{k+1} P_{T_{\mathcal{M}}(y_k)}(\nu_{k+1}) + \eta_{k+1} E_k \qquad \text{for all } k \geq 1.$$

Define an auxiliary sequence $z_k = x^\star + U\Delta_k$ where $\Delta_k := U^\top (y_k - x^\star)$. Consider the following two estimators defined in terms of z_k and Δ_k respectively.

$$\Sigma_n' = \frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i (z_k - x^\star) - l_i(\bar{z}_n - x^\star)\right)\left(\sum_{k=t_i}^i (z_k - x^\star) - l_i(\bar{z}_n - x^\star)\right)^\top}{\sum_{i=1}^n l_i}.$$

$$\tilde{\Sigma}_n = \frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i \Delta_k - l_i\bar{\Delta}_n\right)\left(\sum_{k=t_i}^i \Delta_k - l_i\bar{\Delta}_n\right)^\top}{\sum_{i=1}^n l_i}. \tag{B.2}$$

Observing that $\Sigma'_n = U\tilde{\Sigma}_n U^\top$ and recalling from (3.5) that $\Sigma = UH^{-1}SH^{-\top}U^\top$, we have,

$$\begin{aligned}\mathbb{E}[\|\Sigma'_n - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &= \mathbb{E}[\|U(\tilde{\Sigma}_n - H^{-1}SH^{-\top})U^\top\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\le \mathbb{E}[\|\tilde{\Sigma}_n - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]. \end{aligned} \tag{B.3}$$

Using triangle inequality and (B.3), we have

$$\begin{aligned}\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &\le \mathbb{E}[\|\Sigma'_n - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[\|\hat{\Sigma}_n - \Sigma'_n\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\le \underbrace{\boxed{\mathbb{E}[\|\tilde{\Sigma}_n - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]}}_{\text{Lemma 10}} + \underbrace{\boxed{\mathbb{E}[\|\hat{\Sigma}_n - \Sigma'_n\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]}}_{\text{Lemma 11}}.\end{aligned}$$

On the one hand, by Lemma 10 and the assumption that $\beta > \frac{1}{1-\alpha}$,

$$\begin{aligned}&\mathbb{E}[\|\tilde{\Sigma}_n - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim dk_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d}k_s^2 M^{-\frac{1}{2}} + \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} + k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}} + k_s^{2\alpha+1} M^{-1} \\ &\lesssim k_s^3 (dM^{(\alpha-1)\beta+1} + \sqrt{d}M^{\frac{(\alpha-1)\beta+1}{2}} + \sqrt{d}M^{-\frac{1}{2}})\end{aligned} \tag{B.4}$$

On the other hand, by Lemma 11 and the assumption that $\beta > \frac{1}{1-\alpha}$,

$$\begin{aligned}&\mathbb{E}[\|\hat{\Sigma}_n - \Sigma'_n\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim \sqrt{d}k_s^{\frac{3}{2}+\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta}{2}} + d^{\frac{1}{4}} k_s^{\frac{5}{2}} M^{-\frac{3}{4}} + d^{\frac{1}{4}} k_s^{\frac{\alpha}{4}+\frac{3}{2}} M^{\frac{(\alpha-1)\beta-1}{4}} + k_s^{\frac{3}{2}} M^{-\frac{1}{2}} + k_s^3 M^{-1} \\ &\lesssim k_s^3 \sqrt{d} M^{-\frac{1}{2}}\end{aligned} \tag{B.5}$$

Combining (B.4), and (B.5) and using the fact that $n \asymp M^\beta$, we conclude the proof of Theorem 1.

Lemma 10 *Let the conditions of Theorem 1 be true. We have,*

$$\begin{aligned}&\mathbb{E}[\|\tilde{\Sigma}_n - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim dk_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d}k_s^2 M^{-\frac{1}{2}} + \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} + k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}} + k_s^{2\alpha+1} M^{-1}.\end{aligned}$$

Proof Following the proof of Lemma 10.7 in Davis et al. (2024), we have

$$\Delta_{k+1} = (I - \eta_{k+1}H)\Delta_k - \eta_{k+1}\Big(U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k))\Big) - \eta_{k+1}\Big(R(y_k) + \zeta_{k+1} - U^\top E_k\Big), \tag{B.7}$$

where $\zeta_{k+1} = U^\top P_{T_\mathcal{M}(y_k)}(\nu_{k+1}) - U^\top P_{T_\mathcal{M}(x^\star)}(\nu_{k+1})$, and

$$R(y) = U^\top F_\mathcal{M}(y) - U^\top \nabla_\mathcal{M} F_\mathcal{M}(x^\star) UU^\top (y - x^\star).$$

Summing both sides of (B.7) from $k = i$ to j, we get

$$\begin{aligned}\textstyle\sum_{k=i}^j \Delta_k &= S_{i-1}^j \Delta_{i-1} + \textstyle\sum_{k=i}^j (\mathrm{I} + S_k^j)\eta_k \Big(U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k)) + R(y_k) - U^\top E_k + \zeta_{k+1}\Big) \\ &= \lambda_i^j + e_i^j,\end{aligned} \tag{B.8}$$

where we define

$$\lambda_i^j := S_{i-1}^j \Delta_{i-1} + \textstyle\sum_{k=i}^j (\mathrm{I} + S_k^j)\eta_k \left(U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k))\right),$$
$$e_i^j := \textstyle\sum_{k=i}^j (\mathrm{I} + S_k^j)\eta_k \left(R(y_k) - U^\top E_k + \zeta_{k+1}\right).$$

Plugging (B.8) into the definition of $\tilde{\Sigma}_n$ in (B.2), we write and divide $\tilde{\Sigma}_n$ into four parts.

$$\tilde{\Sigma}_n = (\textstyle\sum_{i=1}^n l_i)^{-1}[\underbrace{\textstyle\sum_{i=1}^n(\lambda_{t_i}^i - n^{-1}l_i\lambda_1^n)(\lambda_{t_i}^i - n^{-1}l_i\lambda_1^n)^\top}_{\mathsf{I}} + \underbrace{\textstyle\sum_{i=1}^n(e_{t_i}^i - n^{-1}l_ie_1^n)(\lambda_{t_i}^i - n^{-1}l_i\lambda_1^n)^\top}_{\mathsf{II}}]$$
$$+ (\textstyle\sum_{i=1}^n l_i)^{-1}[\underbrace{\textstyle\sum_{i=1}^n(\lambda_{t_i}^i - n^{-1}l_i\lambda_1^n)(e_{t_i}^i - n^{-1}l_ie_1^n)^\top}_{\mathsf{III}} + \underbrace{\textstyle\sum_{i=1}^n(e_{t_i}^i - n^{-1}l_ie_1^n)(e_{t_i}^i - n^{-1}l_ie_1^n)^\top}_{\mathsf{IV}}].$$

In what follows, we will provide upper bounds on $\mathbb{E}[\|(\sum_{i=1}^n l_i)^{-1}\mathsf{I} - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$, $\mathbb{E}[\|(\sum_{i=1}^n l_i)^{-1}\mathsf{II}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$, $\mathbb{E}[(\sum_{i=1}^n l_i)^{-1}\mathsf{III}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$, and $\mathbb{E}[(\sum_{i=1}^n l_i)^{-1}\mathsf{IV}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$ separately. The lemma then follows from the triangle inequality.

Analysis of term I: Note that the goal is to bound

$$\begin{aligned}
&\mathbb{E}[\|(\textstyle\sum_{i=1}^n l_i)^{-1}\mathsf{I} - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&= \mathbb{E}[\|(\textstyle\sum_{i=1}^n l_i)^{-1}\sum_{i=1}^n \lambda_{t_i}^i\lambda_{t_i}^{i\,\top} - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1}n^{-1}\|\sum_{i=1}^n l_i\lambda_{t_i}^i\lambda_1^{n\,\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\quad + \mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1}n^{-1}\|\sum_{i=1}^n l_i\lambda_1^n\lambda_{t_i}^{i\,\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1}n^{-2}\sum_{i=1}^n l_i^2\|\lambda_1^n\lambda_1^{n\,\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}].
\end{aligned}$$

We bound terms on the RHS one by one.

- The first term $\mathbb{E}[\|(\sum_{i=1}^n l_i)^{-1}\sum_{i=1}^n \lambda_{t_i}^i\lambda_{t_i}^{i\,\top} - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$. To this end, we rewrite

$$\textstyle\sum_{i=1}^n \lambda_{t_i}^i\lambda_{t_i}^{i\,\top} = \sum_{i=1}^n(\upsilon_i + \omega_i)(\upsilon_i + \omega_i)^\top,$$

where

$$\upsilon_i := S_{t_i-1}^i\Delta_{t_i-1} + \sum_{k=t_i}^{i}(\eta_k\,\mathrm{I} + \eta_kS_k^i - H^{-1})(U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k)))$$

and

$$\omega_i := \sum_{k=t_i}^{i} H^{-1}U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k))).$$

Note that

$$\mathbb{E}[\|\upsilon_i\upsilon_i^\top\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \le \mathrm{tr}\left(\mathbb{E}[\upsilon_i\upsilon_i^\top \mathbb{1}_{\tau_{k_s,\delta}>n}]\right) \le d\|\mathbb{E}[\upsilon_i\upsilon_i^\top]\mathbb{1}_{\tau_{k_s,\delta}>n}\|_2. \tag{B.10}$$

On the other hand, direct calculation shows

$$\begin{aligned}
&\|\mathbb{E}[\upsilon_i\upsilon_i^\top \mathbb{1}_{\tau_{k_s,\delta}>n}]\|_2 \\
&\le \|\mathbb{E}[\upsilon_i\upsilon_i^\top \mathbb{1}_{\tau_{k_s,\delta}>t_i-1}]\|_2 \\
&\le \|S_{t_i-1}^i\|_2^2\|\mathbb{E}[\Delta_{t_i-1}\Delta_{t_i-1}^\top \mathbb{1}_{\tau_{k_s,\delta}>t_i-1}]\|_2
\end{aligned} \tag{B.11}$$

$$+\textstyle\sum_{k=t_i}^{i}\|\eta_k\,\mathrm{I}+\eta_k S_k^i - H^{-1}\|_2^2\|U^\top\mathbb{E}[(\nu_{k+1}^{(1)}+\nu_{k+1}^{(2)}(x_k))(\nu_{k+1}^{(1)}+\nu_{k+1}^{(2)}(x_k))^\top]U\|_2,$$

where the first inequality follows from the definition of the stopping time, and the second inequality follows from Assumption 5 that $\{\nu_{k+1}^{(1)}\}$ and $\{\nu_{k+1}^{(2)}(x_k)\}$ are martingale difference sequences. We then bound the RHS of (B.11). For the first term in (B.11), we consider two cases:

1. $t_i - 1 \ge k_s$. Using Lemma 19 and Lemma 25, we have,

$$\begin{aligned}\|S_{t_i-1}^i\|_2^2\|\mathbb{E}[\Delta_{t_i-1}\Delta_{t_i-1}^\top\mathbb{1}_{\tau_{k_s,\delta}>t_i-1}]\|_2 &\le \|S_{t_i-1}^i\|_2^2\mathbb{E}[\|\Delta_{t_i-1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>t_i-1}]\\ &\lesssim k_s^\alpha t_i^\alpha.\end{aligned} \tag{B.12}$$

2. $t_i - 1 < k_s$. By the definition of y_i, we always have $\|\Delta_{t_i-1}\| \le 4\delta$. Applying Lemma 25, we have

$$\|S_{t_i-1}^i\|_2^2\|\mathbb{E}[\Delta_{t_i-1}\Delta_{t_i-1}^\top\mathbb{1}_{\tau_{k_s,\delta}>t_i-1}]\|_2 \lesssim t_i^{2\alpha} \lesssim k_s^{2\alpha}.$$

Next, we consider the second term on the RHS of (B.11). By Assumption 5 and 6, we have

$$\begin{aligned}\|U^\top\mathbb{E}[(\nu_{k+1}^{(1)}+\nu_{k+1}^{(2)}(x_k))(\nu_{k+1}^{(1)}+\nu_{k+1}^{(2)}(x_k))^\top]U\|_2 &\lesssim \mathbb{E}[\|\nu_{k+1}^{(1)}\|_2^2]+\mathbb{E}[\|\nu_{k+1}^{(2)}(x_k)\|_2^2]\\ &\lesssim \mathbb{E}[\|x_k - x^\star\|_2^2]\\ &\lesssim C_{\text{ub}}.\end{aligned}$$

In addition, following the same proof of (Zhu et al., 2023, Lemma B.3) we obtain,

$$\textstyle\sum_{k=t_i}^{i}\|\eta_k\,\mathrm{I}+\eta_k S_k^i - H^{-1}\|_2^2 \lesssim l_i t_i^{2\alpha-2} + i^\alpha.$$

Combining, we have

$$\begin{aligned}&\sum_{k=t_i}^{i}\|\eta_k\,\mathrm{I}+\eta_k S_k^i - H^{-1}\|_2^2\|U^\top\mathbb{E}[(\nu_{k+1}^{(1)}+\nu_{k+1}^{(2)}(x_k))(\nu_{k+1}^{(1)}+\nu_{k+1}^{(2)}(x_k))^\top]U\|_2\\ &\lesssim l_i t_i^{2\alpha-2} + i^\alpha.\end{aligned} \tag{B.14}$$

By basic calculus and our choice of a_m and n_m, we can easily verify the following three inequalities:

$$\textstyle\sum_{i=1}^{n} l_i \asymp \sum_{m=1}^{M} n_m^2 \asymp \sum_{m=1}^{M} m^{2\beta-2} \asymp M^{2\beta-1}; \tag{B.15}$$

$$\textstyle\sum_{i=1}^{n} l_i^2 \asymp \sum_{m=1}^{M} n_m^3 \asymp \sum_{m=1}^{M} m^{3\beta-3} \asymp M^{3\beta-2}; \tag{B.16}$$

$$\textstyle\sum_{m=1}^{M} a_m^{-2\alpha} n_m^3 \asymp \sum_{m=1}^{M} m^{3\beta-2\alpha\beta-3} \asymp M^{3\beta-2\alpha\beta-2}. \tag{B.17}$$

Combining (B.10), (B.11), (B.14), (B.12), and (B.15), we have

$$\textstyle\sum_{i=1}^{n}\mathbb{E}[\|\upsilon_i\upsilon_i^\top\|_2\mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim d\left[\sum_{i=1}^{n}(l_i t_i^{2\alpha-2} + i^\alpha + k_s^\alpha t_i^\alpha + k_s^{2\alpha})\right]$$

$$
\begin{aligned}
&= d\left[\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1}(l_i a_m^{2\alpha-2} + i^{\alpha} + k_s^{\alpha} a_m^{\alpha} + k_s^{2\alpha})\right] \\
&\lesssim d\left[M^{2\alpha\beta-1} + M^{\beta(1+\alpha)} + k_s^{\alpha} M^{\beta(1+\alpha)} + k_s^{2\alpha} M^{\beta}\right].
\end{aligned}
$$

Then, by (B.15) and the assumption that $n \geq k_s$,

$$
\begin{aligned}
(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \mathbb{E}[\|v_i v_i^{\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &\lesssim d k_s^{\alpha} M^{(\alpha-1)\beta+1} + d k_s^{2\alpha} M^{1-\beta} \\
&\lesssim d k_s^{\alpha} M^{(\alpha-1)\beta+1}.
\end{aligned} \tag{B.18}
$$

Define $\hat{\omega}_i = \sum_{k=t_i}^{i} H^{-1} U^{\top} \nu_{k+1}^{(1)}$. Using the same proof of Step 1 of (Zhu et al., 2023, Lemma B.2), we have,

$$
\mathbb{E}[\|(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \hat{\omega}_i \hat{\omega}_i^{\top} - H^{-1} S H^{-\top}\|_2] \lesssim \sqrt{d} M^{-\frac{1}{2}}. \tag{B.19}
$$

Following the proof of Step 2 of (Zhu et al., 2023, Lemma B.2), we have

$$
\begin{aligned}
&\mathbb{E}[\|(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \hat{\omega}_i \hat{\omega}_i^{\top} - (\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \omega_i \omega_i^{\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\leq 2 \cdot \underbrace{\mathbb{E}[\|(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} H^{-1} U^{\top} \left(\sum_{k=t_i}^{i} \nu_{k+1}^{(1)}\right) \left(\sum_{k=t_i}^{i} \nu_{k+1}^{(2)}(x_k)\right)^{\top} U H^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]}_{(i)} \\
&\quad + \underbrace{\mathbb{E}[\|(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} H^{-1} U^{\top} \left(\sum_{k=t_i}^{i} \nu_{k+1}^{(2)}(x_k)\right) \left(\sum_{k=t_i}^{i} \nu_{k+1}^{(2)}(x_k)\right)^{\top} U H^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]}_{(ii)}
\end{aligned} \tag{B.20}
$$

By Cauchy-Schwarz inequality,

$$
(i) \leq \sqrt{\mathbb{E}[\|(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \hat{\omega}_i \hat{\omega}_i^{\top}\|_2]} \cdot \sqrt{(ii)}. \tag{B.21}
$$

By (B.19), we have

$$
\mathbb{E}[\|(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \hat{\omega}_i \hat{\omega}_i^{\top}\|_2] \lesssim 1.
$$

Therefore, it suffices to bound (ii). By triangle inequality and the inequality that $\|C\|_2 \leq \operatorname{tr}(C)$, for any positive semi-definite matrix C,

$$
\begin{aligned}
(ii) &\leq (\textstyle\sum_{i=1}^{n} l_i)^{-1} \mathbb{E}\left[\operatorname{tr}\left(\sum_{i=1}^{n} H^{-1} U^{\top} \left(\sum_{k=t_i}^{i} \nu_{k+1}^{(2)}(x_k)\right) \left(\sum_{k=t_i}^{i} \nu_{k+1}^{(2)}(x_k)\right)^{\top} U H^{-\top}\right) \mathbb{1}_{\tau_{k_s,\delta}>n}\right] \\
&= (\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \mathbb{E}[\|\sum_{k=t_i}^{i} H^{-1} U^{\top} \nu_{k+1}^{(2)}(x_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}].
\end{aligned} \tag{B.22}
$$

Since $\nu_k^{(2)}$ is a martingale difference sequence and $\tau_{k_s,\delta}$ is a stopping time, for any $i \leq n$, we have

$$
\begin{aligned}
&\mathbb{E}[\|\textstyle\sum_{k=t_i}^{i} H^{-1} U^{\top} \nu_{k+1}^{(2)}(x_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\leq \mathbb{E}[\|\textstyle\sum_{k=t_i}^{i} H^{-1} U^{\top} \nu_{k+1}^{(2)}(x_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>i}] \\
&= \mathbb{E}[\|\textstyle\sum_{k=t_i}^{i-1} H^{-1} U^{\top} \nu_{k+1}^{(2)}(x_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>i}] + \mathbb{E}[\|H^{-1} U^{\top} \nu_{i+1}^{(2)}(x_i)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>i}] \\
&\;\;\vdots
\end{aligned}
$$

$$\leq \sum_{k=t_i}^{i} \mathbb{E}[\|H^{-1}U^\top \nu_{k+1}^{(2)}(x_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}].$$

When $k \geq k_s$, by Lemma 17 and Lemma 19, we have $\mathbb{E}[\|x_k - x^\star\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] \lesssim k_s^\alpha k^{-\alpha}$; on the other hand, when $k < k_s$, by Assumption 6, we always have $\mathbb{E}[\|x_k - x^\star\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] \leq C_{\text{ub}}$. Combining, we have

$$\mathbb{E}[\| \sum_{k=t_i}^{i} H^{-1}U^\top \nu_{k+1}^{(2)}(x_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim \begin{cases} \sum_{k=t_i}^{i} k_s^\alpha k^{-\alpha} & t_i \geq k_s \\ l_i C_{\text{ub}} & t_i < k_s. \end{cases}$$

By (B.15), (B.22), and $\beta > \frac{1}{1-\alpha}$, we have

$$\begin{aligned} (ii) &\lesssim (\textstyle\sum_{i=1}^n l_i)^{-1} \left(\sum_{i=1}^n \sum_{k=t_i}^{i} k_s^\alpha k^{-\alpha} + k_s^2 C_{\text{ub}} \right) \\ &\lesssim k_s^\alpha M^{-\alpha\beta} + k_s^2 M^{1-2\beta} \\ &\lesssim k_s^2 M^{-\alpha\beta}. \end{aligned} \tag{B.23}$$

Combining (B.19), (B.20), (B.21), and (B.23), we have

$$\begin{aligned} &\mathbb{E}[\|(\textstyle\sum_{i=1}^n l_i)^{-1} \sum_{i=1}^n \omega_i \omega_i^\top - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim \sqrt{d}M^{-\frac{1}{2}} + ((d/M)^{\frac{1}{4}} + 1)k_s M^{-\alpha\beta/2} + k_s^2 M^{-\alpha\beta} \\ &\lesssim \sqrt{d}M^{-\frac{1}{2}} + k_s^2 M^{-\frac{\alpha\beta}{2}}. \end{aligned} \tag{B.24}$$

Then by triangle inequality,

$$\mathbb{E}[\|(\textstyle\sum_{i=1}^n l_i)^{-1} \sum_{i=1}^n \omega_i \omega_i^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim \sqrt{d}M^{-\frac{1}{2}} + k_s^2 M^{-\alpha\beta} + 1. \tag{B.25}$$

Combining (B.18), and (B.25), and using Cauchy-Schwarz inequality, we have,

$$(\textstyle\sum_{i=1}^n l_i)^{-1} \sum_{i=1}^n \mathbb{E}[\|\upsilon_i \omega_i^\top\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} (d^{\frac{1}{4}} M^{-\frac{1}{4}} + k_s M^{-\frac{\alpha\beta}{2}} + 1). \tag{B.26}$$

Similarly,

$$(\textstyle\sum_{i=1}^n l_i)^{-1} \sum_{i=1}^n \mathbb{E}[\|\omega_i \upsilon_i^\top\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} (d^{\frac{1}{4}} M^{-\frac{1}{4}} + k_s M^{-\frac{\alpha\beta}{2}} + 1). \tag{B.27}$$

Then, combining (B.18), (B.24), (B.26), and (B.27), we have

$$\begin{aligned} &\mathbb{E}[\|(\textstyle\sum_{i=1}^n l_i)^{-1} \sum_{i=1}^n \lambda_{t_i}^i {\lambda_{t_i}^i}^\top - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s}>n}] \\ &\lesssim dk_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d}M^{-\frac{1}{2}} + k_s^2 M^{-\frac{\alpha\beta}{2}} + \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} (d^{\frac{1}{4}} M^{-\frac{1}{4}} + k_s M^{-\frac{\alpha\beta}{2}} + 1) \\ &\lesssim dk_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d}M^{-\frac{1}{2}} + k_s^2 M^{-\frac{\alpha\beta}{2}} + \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}}. \end{aligned} \tag{B.28}$$

- The fourth term $\mathbb{E}[(\sum_{i=1}^n l_i)^{-1} n^{-2} \sum_{i=1}^n l_i^2 \|\lambda_1^n {\lambda_1^n}^\top\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$. We have

$$\begin{aligned} &\mathbb{E}[\|\lambda_1^n {\lambda_1^n}^\top\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\leq \mathbb{E}[\|\lambda_1^n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &= \mathbb{E}[\|S_0^n \Delta_0 + \textstyle\sum_{k=1}^n (\mathrm{I} + S_k^n)\eta_k (U^\top (\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k)))\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \end{aligned}$$

$$\leq \mathbb{E}[\|S_0^n\Delta_0\|_2^2] + \textstyle\sum_{k=1}^n \|\mathrm{I} + S_k^n\|_2^2\eta_k^2\mathbb{E}[\|U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k))\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] \quad \text{(B.29)}$$
$$\lesssim 1 + \textstyle\sum_{k=1}^n \mathbb{E}[\|U^\top(\nu_{k+1}^{(1)} + \nu_{k+1}^{(2)}(x_k))\|_2^2] \quad \text{(B.30)}$$
$$\lesssim n. \quad \text{(B.31)}$$

where the estimate (B.29) follows from the martingale difference property of $\nu_k^{(1)}$ and $\nu_k^{(2)}$, the estimate (B.30) follows from Lemma 25, and the estimate (B.31) follows from Assumption 5 and 6.

Then, by (B.15) and (B.16), we have

$$\mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1} n^{-2} \sum_{i=1}^n l_i^2 \|\lambda_1^n \lambda_1^{n\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \leq (\sum_{i=1}^n l_i)^{-1} n^{-2} \sum_{i=1}^n l_i^2 \mathbb{E}[\|\lambda_1^n \lambda_1^{n\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$$
$$\lesssim n^{-1}(\sum_{i=1}^n l_i)^{-1} \sum_{i=1}^n l_i^2 \lesssim M^{-1}. \quad \text{(B.32)}$$

- The second term $\mathbb{E}[(\sum_{i=1}^n l_i)^{-1} n^{-1} \| \sum_{i=1}^n l_i \lambda_{t_i}^i \lambda_1^{n\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$. Note that

$$\mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1} n^{-1} \|\sum_{i=1}^n l_i \lambda_{t_i}^i \lambda_1^{n\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$$
$$\leq \mathbb{E}\left[\sqrt{(\textstyle\sum_{i=1}^n l_i)^{-1}\|\sum_{i=1}^n \lambda_{t_i}^i \lambda_{t_i}^{i\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}} \cdot \sqrt{(\sum_{i=1}^n l_i)^{-1} n^{-2} \|\sum_{i=1}^n l_i^2 \lambda_1^n \lambda_1^{n\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}}\right]$$
$$\leq \sqrt{(\textstyle\sum_{i=1}^n l_i)^{-1}\mathbb{E}[\|\sum_{i=1}^n \lambda_{t_i}^i \lambda_{t_i}^{i\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]} \cdot \sqrt{(\sum_{i=1}^n l_i)^{-1} n^{-2}\mathbb{E}[\|\sum_{i=1}^n l_i^2 \lambda_1^n \lambda_1^{n\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]}$$
$$\lesssim \sqrt{dk_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d}M^{-\frac{1}{2}} + k_s M^{-\alpha\beta/2} + \sqrt{d}M^{\frac{(\alpha-1)\beta}{2}+\frac{1}{2}} + 1} \cdot M^{-\frac{1}{2}}$$
$$\lesssim d^{\frac{1}{2}} k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta}{2}} + d^{\frac{1}{4}} M^{-\frac{3}{4}} + k_s^{\frac{1}{2}} M^{-\frac{\alpha\beta}{2}-\frac{1}{2}} + d^{\frac{1}{4}} k_s^{\frac{\alpha}{4}} M^{\frac{(\alpha-1)\beta-1}{4}} + M^{-\frac{1}{2}}, \quad \text{(B.33)}$$

where the first inequality follows from Cauchy-Schwarz, the second inequality follows from Holder's inequality, and the third inequality follows from (B.28) and (B.32).

- The third term $\mathbb{E}[(\sum_{i=1}^n l_i)^{-1} n^{-1} \|\sum_{i=1}^n l_i \lambda_1^n \lambda_{t_i}^{i\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$. By the same calculation as the second term, we have

$$\mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1} n^{-1} \|\sum_{i=1}^n l_i \lambda_1^n \lambda_{t_i}^{i\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$$
$$\lesssim d^{\frac{1}{2}} k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta}{2}} + d^{\frac{1}{4}} M^{-\frac{3}{4}} + k_s^{\frac{1}{2}} M^{-\frac{\alpha\beta}{2}-\frac{1}{2}} + d^{\frac{1}{4}} k_s^{\frac{\alpha}{4}} M^{\frac{(\alpha-1)\beta}{4}-\frac{1}{4}} + M^{-\frac{1}{2}} \quad \text{(B.34)}$$

Combining (B.28), (B.32), (B.33), and (B.34), we have,

$$\mathbb{E}[\|(\textstyle\sum_{i=1}^n l_i)^{-1}\mathsf{I} - H^{-1}SH^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$$
$$\lesssim dk_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d}k_s^2 M^{-\frac{1}{2}} + \sqrt{d}k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}}. \quad \text{(B.35)}$$

Next, we bound term IV. We then bound terms II, and III using Cauchy-Schwarz inequality and the bounds on term I, and term IV.

Bound on term IV: Note that

$$\mathbb{E}[\|\mathsf{IV}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \leq \textstyle\sum_{i=1}^n \mathbb{E}[\|e_{t_i}^n - n^{-1} l_i e_1^n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$$

$$\leq 2\sum_{i=1}^{n}(\mathbb{E}[\|e_{t_i}^n\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[n^{-2}l_i^2\|e_1^n\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}]). \tag{B.36}$$

First, we bound the first term in the RHS of (B.36). Note that for any $j > i \geq k_s$, using Lemma 18, we have

$$\begin{aligned}\sum_{k=i}^{j}\mathbb{E}[\|E_k\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] &\leq \sum_{k=i}^{j}\mathbb{E}[\|E_k\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>k}] \\ &\lesssim k_s^{2\alpha}\sum_{k=i}^{j}\eta_k^2.\end{aligned} \tag{B.37}$$

We also have, by Claim 2 in the proof of Lemma D.5 of Davis et al. (2024) and Lemma 19,

$$\mathbb{E}[\|R(y_k)\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] \leq \mathbb{E}[\|R(y_k)\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>k}] \leq \mathbb{E}[\|y_k - x^\star\|_2^4\mathbb{1}_{\tau_{k_s,\delta}>k}] \lesssim k_s^{2\alpha}\eta_k^2. \tag{B.38}$$

Direct calculation shows

$$\begin{aligned}
&\mathbb{E}[\|\textstyle\sum_{k=i}^{j}(\mathrm{I}+S_k^j)\eta_k\zeta_{k+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&= \mathbb{E}[\|(\mathrm{I}+S_j^j)\eta_j\zeta_{j+1} + \textstyle\sum_{k=i}^{j-1}(\mathrm{I}+S_k^j)\eta_k\zeta_{k+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>j}] \\
&= \mathbb{E}[\|(\mathrm{I}+S_j^j)\eta_j\zeta_{j+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>j}] + \mathbb{E}[\|\textstyle\sum_{k=i}^{j-1}(\mathrm{I}+S_k^j)\eta_k\zeta_{k+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>j}] \\
&\quad + 2\mathbb{E}[(\mathrm{I}+S_j^j)\eta_j\zeta_{j+1}^\top\textstyle\sum_{k=i}^{j-1}(\mathrm{I}+S_k^j)\eta_{k+1}\zeta_{k+1}\mathbb{1}_{\tau_{k_s,\delta}>j}] \\
&= \mathbb{E}[\|(\mathrm{I}+S_j^j)\eta_j\zeta_{j+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>j}] + \mathbb{E}[\|\textstyle\sum_{k=i}^{j-1}(\mathrm{I}+S_k^j)\eta_k\zeta_{k+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>j}] \\
&\ \ \vdots \\
&= \textstyle\sum_{k=i}^{j}\mathbb{E}[\|(\mathrm{I}+S_k^j)\eta_k\zeta_{k+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>k}] \\
&\lesssim \textstyle\sum_{k=i}^{j}\mathbb{E}[\|\zeta_{k+1}\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>k}] && \text{(B.39)} \\
&\lesssim \textstyle\sum_{k=i}^{j}\mathbb{E}[\|y_k - x^\star\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>k}] && \text{(B.40)} \\
&\lesssim k_s^{\alpha}\textstyle\sum_{k=i}^{j}\eta_k && \text{(B.41)} \\
&\leq (j-i+1)k_s^{\alpha}i^{-\alpha}, && \text{(B.42)}
\end{aligned}$$

where the first several equalities follows from the fact that $\{\zeta_k\}_k$ is a martingale-difference sequence, and we have $\mathbb{E}[\zeta_i^\top\zeta_{j+1}] = 0$ for $i \neq j$, the estimate (B.39) follows from Lemma 25, the estimate (B.40) follows from the definition of ζ_k and Lipschitz continuity of $P_{T_\mathcal{M}}(\cdot)$, and the estimate (B.41) follows from Lemma 19. Combining (B.37), (B.38), and (B.42), and using Lemma 25, for i such that $t_i \geq k_s$, we have

$$\begin{aligned}
\mathbb{E}[\|e_{t_i}^i\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] &\lesssim l_i\textstyle\sum_{k=t_i}^{i}\mathbb{E}[\|(\mathrm{I}+S_k^i)\eta_k\left(R(y_k) - U^\top E_k\right)\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] + k_s^{\alpha}\sum_{k=t_i}^{i}\eta_k \\
&\lesssim l_i\textstyle\sum_{k=t_i}^{i}(\mathbb{E}[\|R(y_k)\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[\|U^\top E_k\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}]) + l_ik_s^{\alpha}t_i^{-\alpha} \\
&\lesssim l_ik_s^{2\alpha}\textstyle\sum_{k=t_i}^{i}\eta_k^2 + l_ik_s^{\alpha}t_i^{-\alpha} \\
&\leq l_i^2k_s^{2\alpha}t_i^{-2\alpha} + l_ik_s^{\alpha}t_i^{-\alpha},
\end{aligned}$$

where the third inequality follows from $\|R(y_k)\|_2 \lesssim \|y_k - x^\star\|_2^2$, Lemma 19, and Lemma 18. On the other hand, for i such that $t_i < k_s$, we have

$$\begin{aligned}
\mathbb{E}[\|e_{t_i}^i\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] &\leq l_i\textstyle\sum_{k=t_i}^{i}\mathbb{E}[\|(\mathrm{I}+S_k^i)\eta_k\left(R(y_k) - U^\top E_k + \zeta_{k+1}\right)\|_2^2\mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\lesssim l_i\textstyle\sum_{k=t_i}^{i}(\mathbb{E}[\|R(y_k)\|_2^2] + \mathbb{E}[\|E_k\|_2^2] + \mathbb{E}[\|\zeta_{k+1}\|^2])
\end{aligned}$$

$$
\begin{aligned}
&\leq l_i \textstyle\sum_{k=t_i}^{i}(1+k^{2\alpha}) \\
&\lesssim k_s^{2\alpha+2},
\end{aligned}
$$

where the first inequality follows from Lemma 25, the third inequality follows from Lemma 18, and the last inequality follows from $l_i \leq i \lesssim k_s$. As a result, for $\beta > (1-\alpha)^{-1}$,

$$
\begin{aligned}
\textstyle\sum_{i=1}^{n} \mathbb{E}[\|e_{t_i}^i\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &\lesssim \textstyle\sum_{m=1}^{M}\sum_{i=a_m+1}^{a_{m+1}}(l_i^2 k_s^{2\alpha} a_m^{-2\alpha} + l_i k_s^{\alpha} a_m^{-\alpha}) + k_s^{2\alpha+3} \\
&\lesssim \textstyle\sum_{m=1}^{M}(n_m^3 k_s^{2\alpha} a_m^{-2\alpha} + n_m^2 k_s^{\alpha} a_m^{-\alpha}) + k_s^{2\alpha+3} \\
&\lesssim k_s^{2\alpha} M^{3\beta-2\alpha\beta-2} + k_s^{\alpha} M^{2\beta-\alpha\beta-1} + k_s^{2\alpha+3} \\
&\lesssim k_s^{2\alpha+3} M^{3\beta-2\alpha\beta-2}.
\end{aligned} \tag{B.43}
$$

Combining (B.43) and (B.15), we have

$$
(\textstyle\sum_{i=1}^{n} l_i)^{-1} \sum_{i=1}^{n} \mathbb{E}[\|e_{t_i}^i\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim k_s^{2\alpha+3} M^{(1-2a)\beta-1}. \tag{B.44}
$$

Next, we look at the second term in the RHS of (B.36). Note that

$$
\begin{aligned}
&\mathbb{E}[\|e_1^n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&= \mathbb{E}[\|\textstyle\sum_{k=1}^{n}(I+S_k^i)\eta_k(R(y_k) - U^\top E_k + \zeta_{k+1})\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\lesssim n\mathbb{E}[\textstyle\sum_{k=1}^{n}\|(I+S_k^i)\eta_k(R(y_k) - U^\top E_k)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[\|\sum_{k=1}^{n}(I+S_k^i)\eta_k\zeta_{k+1}\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\leq n\textstyle\sum_{k=1}^{n}\|(I+S_k^i)\|_2^2\eta_k^2\mathbb{E}[\|R(y_k) - U^\top E_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[\|\sum_{k=1}^{n}(I+S_k^i)\eta_k\zeta_{k+1}\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\lesssim n\left(\textstyle\sum_{k=1}^{n}\mathbb{E}[\|R(y_k) - U^\top E_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}]\right) + n \\
&\lesssim n(\textstyle\sum_{k=1}^{n} k_s^{2\alpha}\eta_k^2 + k_s^{1+2\alpha}) + n \\
&\lesssim n k_s^{1+2\alpha},
\end{aligned}
$$

where the first inequality follows from Young's inequality and Cauchy-Schwarz, the second inequality follows from Jensen's inequality, the third inequality follows by Lemma 25, and the same calculation as (B.42), and the fourth inequality follows from Lemma 25, Lemma 19, and Lemma 18. By (B.16), we have

$$
n^{-2}\textstyle\sum_{i=1}^{n} l_i^2\mathbb{E}[\|e_1^n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim k_s^{2\alpha+1} n^{-1}\sum_{i=1}^{n} l_i^2 \lesssim k_s^{2\alpha+1} n^{-1} M^{3\beta-2}.
$$

Using the fact $n \asymp M^\beta$, and (B.15), we get,

$$
n^{-2}(\textstyle\sum_{i=1}^{n} l_i)^{-1}\sum_{i=1}^{n} l_i^2\mathbb{E}[\|e_1^n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim k_s^{2\alpha+1} M^{-1}. \tag{B.45}
$$

Combining (B.44), and (B.45), we have

$$
\mathbb{E}[(\textstyle\sum_{i=1}^{n} l_i)^{-1}\|IV\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim k_s^{2\alpha+1} M^{-1}. \tag{B.46}
$$

Bound on term II: Combining (B.35), and (B.44), and using Cauchy-Schwarz inequality, we obtain,

$$
\begin{aligned}
&\mathbb{E}[(\textstyle\sum_{i=1}^{n} l_i)^{-1}\|II\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\leq \left(\mathbb{E}[(\textstyle\sum_{i=1}^{n} l_i)^{-1}\|I\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]\right)^{1/2}\left(\mathbb{E}[(\textstyle\sum_{i=1}^{n} l_i)^{-1}\|IV\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]\right)^{1/2}
\end{aligned}
$$

$$\lesssim k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}}. \tag{B.47}$$

Bound on term III: Similar to term II we have,

$$\mathbb{E}[(\textstyle\sum_{i=1}^n l_i)^{-1} \|III\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}}. \tag{B.48}$$

Combining (B.35), (B.46), (B.47), and (B.48), we have

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\Sigma} - H^{-1} S H^{-\top}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\lesssim d k_s^{\alpha} M^{(\alpha-1)\beta+1} + \sqrt{d} k_s^2 M^{-\frac{1}{2}} + \sqrt{d} k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} + k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}} + k_s^{2\alpha+1} M^{-1}.
\end{aligned}$$

■

Lemma 11 *Let the conditions of Theorem 1 be true. Then,*

$$\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_n'\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim \sqrt{d} k_s^{\frac{3}{2}+\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta}{2}} + d^{\frac{1}{4}} k_s^{\frac{5}{2}} M^{-\frac{3}{4}} + d^{\frac{1}{4}} k_s^{\frac{\alpha}{4}+\frac{3}{2}} M^{\frac{(\alpha-1)\beta-1}{4}} + k_s^{\frac{3}{2}} M^{-\frac{1}{2}} + k_s^3 M^{-1}.$$

Proof Define $\rho_k \coloneqq x_k - z_k$. We have the following expansion:

$$\begin{aligned}
\hat{\Sigma}_n - \Sigma_n' = &\underbrace{\frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)\left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)^\top}{\sum_{i=1}^n l_i}}_{\mathsf{V}} + \underbrace{\frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)\left(\sum_{k=t_i}^i z_k - l_i \bar{z}_n\right)^\top}{\sum_{i=1}^n l_i}}_{\mathsf{VI}} \\
&+ \underbrace{\frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i z_k - l_i \bar{z}_n\right)\left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)^\top}{\sum_{i=1}^n l_i}}_{\mathsf{VII}}.
\end{aligned}$$

In what follows, we bound them separately.
Bound on Term V: First, we calculate

$$\begin{aligned}
&\mathbb{E}[\|\textstyle\sum_{i=1}^n \left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)\left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\leq \textstyle\sum_{i=1}^n \mathbb{E}[\|\left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)\left(\sum_{k=t_i}^i \rho_k - l_i \bar{\rho}_n\right)^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&= \textstyle\sum_{i=1}^n \mathbb{E}[\|\sum_{k=t_i}^i (\rho_k - \bar{\rho}_n)\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\
&\lesssim \textstyle\sum_{i=1}^n l_i \sum_{k=t_i}^i (\mathbb{E}[\|\rho_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[\|\bar{\rho}_n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}]),
\end{aligned} \tag{B.50}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Using Equation 10.6 in Davis et al. (2024), we have

$$\|\rho_k\|_2 \leq \|x_k - y_k\|_2 + \|y_k - z_k\|_2 \lesssim \|D_k\|_2 + \|y_k - x^\star\|_2^2.$$

Applying Lemma 16 and Lemma 19, for i such that $t_i \geq k_s$, we have

$$\begin{aligned}
\textstyle\sum_{k=t_i}^i \mathbb{E}[\|\rho_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &\lesssim \textstyle\sum_{k=t_i}^i (\mathbb{E}[\|D_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] + \mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}]) \\
&\lesssim k_s^{2\alpha} \textstyle\sum_{k=t_i}^i \eta_k^2
\end{aligned}$$

$$\lesssim k_s^{2\alpha} l_i t_i^{-2\alpha}. \tag{B.52}$$

On the other hand, for i such that $l_i < t_i < k_s$, we have

$$\begin{aligned}\textstyle\sum_{k=t_i}^{i} \mathbb{E}[\|\rho_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &\lesssim \textstyle\sum_{k=t_i}^{i} (\mathbb{E}[\|D_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] + \mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}]) \\ &\lesssim k_s, ,\end{aligned} \tag{B.53}$$

where the second inequality follows from Assumption 6 and the definition of y_k. Similar to (B.52) and (B.53), we have

$$\mathbb{E}[\|\bar{\rho}_n\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \le n^{-1} \textstyle\sum_{k=1}^{n} \mathbb{E}[\|\rho_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] \lesssim n^{-1}(k_s^{2\alpha} \sum_{k=k_s}^{n} \eta_k^2 + k_s) \lesssim k_s n^{-1}. \tag{B.54}$$

Combining (B.50), (B.52), (B.53), and (B.54), and using the fact that $t_i \asymp i$, we have

$$\begin{aligned}&\mathbb{E}[\|\textstyle\sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} \rho_k - l_i\bar{\rho}_n\right)\left(\sum_{k=t_i}^{i} \rho_k - l_i\bar{\rho}_n\right)^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim \textstyle\sum_{i=1}^{n} l_i^2 (k_s^{2\alpha} t_i^{-2\alpha} + k_s n^{-1}) + k_s^3 \\ &= \textstyle\sum_{m=1}^{M} \sum_{i=a_m+1}^{a_{m+1}} l_i^2 k_s^{2\alpha} a_m^{-2\alpha} + k_s n^{-1} \sum_{i=1}^{n} l_i^2 + k_s^3 \\ &\le k_s^{2\alpha} \textstyle\sum_{m=1}^{M} a_m^{-2\alpha} n_m^3 + k_s n^{-1} \sum_{i=1}^{n} l_i^2 + k_s^3\end{aligned} \tag{B.55}$$

Combining (B.55), (B.16), and (B.17), and observing $n \asymp M^\beta$, we obtain,

$$\mathbb{E}[\|\textstyle\sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} \rho_k - l_i\bar{\rho}_n\right)\left(\sum_{k=t_i}^{i} \rho_k - l_i\bar{\rho}_n\right)^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \lesssim k_s^{2\alpha} M^{3\beta-2\alpha\beta-2} + k_s M^{2\beta-2} + k_s^3.$$

Then, by (B.15),

$$\begin{aligned}\mathbb{E}[\|V\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] &= (\textstyle\sum_{i=1}^{n} l_i)^{-1} \mathbb{E}[\|\sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} \rho_k - l_i\bar{\rho}_n\right)\left(\sum_{k=t_i}^{i} \rho_k - l_i\bar{\rho}_n\right)^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim k_s^{2\alpha} M^{\beta-2\alpha\beta-1} + k_s M^{-1} + k_s^3 M^{1-2\beta} \\ &\lesssim k_s^3 M^{-1}.\end{aligned} \tag{B.56}$$

Bound on Term VI: By Lemma 10, we have,

$$\begin{aligned}&\mathbb{E}[\|\Sigma_n' - \Sigma\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim d k_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d} k_s^2 M^{-\frac{1}{2}} + \sqrt{d} k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} + k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}} + k_s^{2\alpha+1} M^{-1}.\end{aligned}$$

Then, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\mathbb{E}[\|\mathsf{VI}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\le \left(\mathbb{E}[\|V\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]\right)^{1/2} \left((\textstyle\sum_{i=1}^{n} l_i)^{-1} \mathbb{E}[\| \sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} z_k - l_i\bar{z}_n\right)\left(\sum_{k=t_i}^{i} z_k - l_i\bar{z}_n\right)^\top \|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]\right)^{1/2} \\ &\lesssim k_s^{\frac{3}{2}} M^{-1/2} \sqrt{d k_s^\alpha M^{(\alpha-1)\beta+1} + \sqrt{d} k_s^2 M^{-\frac{1}{2}} + \sqrt{d} k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} + k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}} + k_s^{2\alpha+1} M^{-1} + 1}\end{aligned} \tag{B.57}$$

Bound on Term VII: Similar to Term VI, we have,

$$\mathbb{E}[\|\mathsf{VII}\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}]$$

$$\lesssim k_s^{\frac{3}{2}} M^{-\frac{1}{2}} \sqrt{dk_s^{\alpha} M^{(\alpha-1)\beta+1} + \sqrt{d} k_s^2 M^{-\frac{1}{2}} + \sqrt{d} k_s^{\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta+1}{2}} + k_s^{\alpha+\frac{1}{2}} M^{-\frac{1}{2}} + k_s^{2\alpha+1} M^{-1}}. \tag{B.58}$$

Combining (B.56), (B.57), and (B.58), we obtain,

$$\begin{aligned} &\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_n'\|_2 \mathbb{1}_{\tau_{k_s,\delta}>n}] \\ &\lesssim \sqrt{d} k_s^{\frac{3}{2}+\frac{\alpha}{2}} M^{\frac{(\alpha-1)\beta}{2}} + d^{\frac{1}{4}} k_s^{\frac{5}{2}} M^{-\frac{3}{4}} + d^{\frac{1}{4}} k_s^{\frac{\alpha}{4}+\frac{3}{2}} M^{\frac{(\alpha-1)\beta-1}{4}} + k_s^{\frac{3}{2}} M^{-\frac{1}{2}} + k_s^3 M^{-1}. \end{aligned}$$

■

Appendix C. Proof of Proposition 3

The basic probabilistic tool we use to achieve high probability bound was originally developed by Harvey et al. Harvey et al. (2019) and then generalized by Cutler et al. (2023).

Proposition 12 (Proposition 29 in Cutler et al. (2023)) *Consider scalar stochastic processes (V_k), (D_k), and (X_k) on a probability space with Filtration $(\mathcal{H}_k)$ such that V_k is nonnegative and $\mathcal{H}_k$ measurable and the inequality*

$$V_{k+1} \leq \alpha_k V_k + D_k \sqrt{V_k} + X_k + \kappa_k$$

holds for for some deterministic constants $\alpha_k \in (-\infty, 1]$ and $\kappa_k \in \mathbb{R}$. Suppose that the moment generating functions of D_k and X_k conditioned on $\mathcal{H}_k$ satisfy the following inequalities for some deterministic constants $\sigma_k, \nu_k > 0$:

- *$\mathbb{E}[\exp(\lambda D_k) \mid \mathcal{H}_k] \leq \exp(\lambda^2 \sigma_k^2/2)$ for all $\lambda \geq 0$. (e.g., D_k is mean-zero sub-Gaussian conditioned on $\mathcal{H}_k$ with parameter σ_k).*
- *$\mathbb{E}[\exp(\lambda X_k) \mid \mathcal{H}_k] \leq \exp(\lambda \nu_k)$ for all $0 \leq \lambda \leq \frac{1}{\nu_k}$. (e.g., X_k is nonnegative and subexponential conditioned on $\mathcal{H}_k$ with parameter ν_k).*

Then, the inequality

$$\mathbb{E}[\exp(\lambda V_{k+1})] \leq \exp(\lambda(\nu_k + \kappa_k)) \mathbb{E}\left[\exp\left(\lambda \left(\frac{1+\alpha_k}{2} V_k\right)\right)\right]$$

holds for all $0 \leq \lambda \leq \min\left\{\frac{1-\alpha_k}{2\sigma_k^2}, \frac{1}{2\nu_k}\right\}$.

Now we prove Proposition 3. Recall that we let $v_k = G_{\eta_{k+1}}(x_k, \nu_{k+1})$. We have

$$\begin{aligned} \|x_{k+1} - x^\star\|^2 &= \|x_k - \eta_{k+1} v_k - x^\star\|^2 \\ &= \|x_k - x^\star\|^2 - 2\eta_{k+1} \langle v_k, x_k - x^\star \rangle + \eta_{k+1}^2 \|v_k\|^2 \\ &\leq \|x_k - x^\star\|^2 - 2\gamma\eta_{k+1} \|x_k - x^\star\|^2 + 2C\eta_{k+1}^2 (1 + \|x_k - x^\star\|^2 + \|\nu_{k+1}\|^2) \\ &\quad - 2\eta_{k+1} \langle \nu_{k+1}, x_k - x^\star \rangle + C\eta_{k+1}^2 (1 + \|x_k - x^\star\|^2 + \|\nu_{k+1}\|^2) \\ &\leq (1 - \gamma\eta_{k+1}) \|x_k - x^\star\|^2 - 2\eta_{k+1} \langle \nu_{k+1}, x_k - x^\star \rangle \\ &\quad + 3C\eta_{k+1}^2 \|\nu_{k+1}\|^2 + 3C\eta_{k+1}^2, \end{aligned}$$

where the first inequality follows from Assumption 7 and the second inequality follows from the upper bound on η. Define

$$\psi_k = \begin{cases} \frac{x_k - x^\star}{\|x_k - x^\star\|} & x_k \neq x^\star \\ 0 & \text{otherwise} \end{cases}.$$

Note that $2\eta_{k+1} \langle \nu_{k+1}, \psi_k \rangle$ is mean-zero sub-Gaussian conditioned on $\mathcal{F}_k$ with parameter $\eta_{k+1}\sigma$, and $3C\eta_{k+1}^2 \|\nu_{k+1}\|^2$ is sub-exponential with parameter $3cC\eta_{k+1}^2\sigma^2$. We can apply Proposition 12 with

$$V_k = \|x_k - x^\star\|, \quad \alpha_k = 1 - \gamma\eta_{k+1}, \quad D_k = -2\eta_{k+1} \langle \nu_{k+1}, \psi_k \rangle$$

and

$$X_k = 3C\eta_{k+1}^2 \|\nu_{k+1}\|^2, \quad \kappa_k = 3C\eta_{k+1}^2.$$

Recalling $\tilde{C} = 3cC\sigma^2 + 3C$ we have from Proposition 12 that

$$\mathbb{E}[\exp(\lambda \|x_{k+1} - x^\star\|^2)] \leq \exp(\lambda \tilde{C} \eta_{k+1}^2) \mathbb{E}[\exp(\lambda (1 - \gamma\eta_{k+1}/2) \|x_k - x^\star\|)] \tag{C.1}$$

for all $0 \leq \lambda \leq \min\left\{ \frac{\gamma}{2\eta_{k+1}\sigma^2}, \frac{1}{6C\eta_{k+1}^2\sigma^2} \right\} = \frac{\gamma}{2\eta_{k+1}\sigma^2}$. Define

$$p_i^j := \begin{cases} \prod_{k=i}^j \left(1 - \frac{\gamma\eta_i}{2}\right) & i \leq j \\ 1 & i = j+1. \end{cases}$$

Applying (C.1) recursively, we deduce

$$\mathbb{E}[\exp(\lambda \|x_k - x^\star\|)] \leq \exp\left(\lambda p_1^k \|x_0 - x^\star\|^2 + \lambda \tilde{C} \left(\sum_{i=1}^k p_{i+1}^k \eta_i^2\right)\right) \tag{C.2}$$

Recall that $C_\alpha = \frac{1-(1/2)^{1-\alpha}}{2(1-\alpha)}$. By Lemma 20, for $1 \leq i \leq \lfloor k/2 \rfloor$,

$$p_i^k \leq \exp(-C_\alpha \gamma\eta (k+1)^{1-\alpha}). \tag{C.3}$$

Consequently, we have

$$\begin{aligned} \lambda\tilde{C}\left(\sum_{i=1}^k p_{i+1}^k \eta_i^2\right) &= \lambda\tilde{C}\left(\sum_{i=1}^{\lfloor k/2 \rfloor} p_{i+1}^k \eta_i^2 + \sum_{i=\lfloor k/2 \rfloor+1}^k p_{i+1}^k \eta_i^2\right) \\ &\leq \lambda\tilde{C}\left(\exp\left(-C_\alpha\gamma\eta(k+1)^{1-\alpha}\right) \sum_{i=1}^\infty \eta_i^2 + \sum_{i=\lfloor k/2 \rfloor+1}^k \eta_i^2\right) \\ &\leq \lambda\tilde{C}\eta^2 \underbrace{\left(\left(1 + \frac{1}{2\alpha - 1}\right)\exp\left(-C_\alpha\gamma\eta(k+1)^{1-\alpha}\right) + \frac{1}{(2\alpha-1)2^{1-2\alpha}} k^{1-2\alpha}\right)}_{:=H_k}, \end{aligned}$$

where the first inequality follows from (C.3) and the fact that $p_{i+1}^k \leq 1$, and the second inequality follows from Lemma 21. By (C.2), we have

$$\mathbb{E}[\exp(\lambda \|x_k - x^\star\|)] \leq \exp\left(\lambda \exp(-C_\alpha\gamma\eta k^{1-\alpha}) \|x_0 - x^\star\|^2 + \lambda\tilde{C}\eta^2 H_k\right)$$

By our assumption on k, we have

$$\exp(-C_\alpha\gamma\eta k^{1-\alpha}) \|x_0 - x^\star\|^2 \leq \frac{\delta}{4} \qquad \text{and} \quad \tilde{C}\eta^2 H_k \leq \frac{\delta}{4}.$$

Then, by Markov's inequality, we have

$$\begin{aligned}\mathbb{P}(\|x_k - x^\star\| \geq \delta) &\leq \exp(-\lambda\delta)\mathbb{E}[\exp(\lambda\|x_k - x^\star\|)] \\ &\leq \exp(-\lambda\delta/2)\end{aligned}$$

Note that by taking $\lambda = \frac{\gamma}{2\eta_{k+1}\sigma^2}$, we have

$$\mathbb{P}(\|x_k - x^\star\| \geq \delta) \leq \exp\left(-\frac{\gamma(k+1)^\alpha\delta}{4\eta\sigma^2}\right), \tag{C.4}$$

which is summable. Combining, we have

$$\begin{aligned}\mathbb{P}(\|x_i - x^\star\| < \delta, \forall i \geq k) &\geq 1 - \textstyle\sum_{i=k}^\infty \mathbb{P}(\|x_i - x^\star\| \geq \delta) \\ &\geq 1 - \textstyle\sum_{i=k}^\infty \exp\left(-\frac{\gamma(k+1)^\alpha\delta}{4\eta\sigma^2}\right) \\ &\geq 1 - \frac{32\eta^2\sigma^4\exp\left(-\frac{\gamma\delta\sqrt{k}}{4\eta\sigma^2}\right)}{\gamma^2\delta^2} - \frac{8\eta\delta^2\sqrt{k}\exp\left(-\frac{\gamma\delta\sqrt{k}}{4\eta\sigma^2}\right)}{\gamma\delta},\end{aligned}$$

where the first inequality follows from the union bound, the second inequality follows from (C.4), and the last inequality follows from Lemma 23.

Appendix D. Proofs of Theorem 4

Lemma 13 *Let $\hat{\Sigma}_n$ be defined as in (1.4). Suppose that Assumption 6 holds. Then we have*

$$\mathbb{E}[\|\hat{\Sigma}_n\|_{op}] \leq 4C_{ub}n.$$

Proof Note that

$$\begin{aligned}\hat{\Sigma}_n &= \frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i x_k - l_i\bar{x}_n\right)\left(\sum_{k=t_i}^i x_k - l_i\bar{x}_n\right)^\top}{\sum_{i=1}^n l_i} \\ &= \frac{\sum_{i=1}^n \left(\sum_{k=t_i}^i (x_k - x^\star) - l_i(\bar{x}_n - x^\star)\right)\left(\sum_{k=t_i}^i (x_k - x^\star) - l_i(\bar{x}_n - x^\star)\right)^\top}{\sum_{i=1}^n l_i},\end{aligned}$$

we can without loss of generality assume that $x^\star = 0$ and $\mathbb{E}[\|x_k\|_2^2] \leq C_{\text{ub}}$ for all $k \geq 0$. Note that by Jensen's inequality, $\mathbb{E}[\|x_n^\star\|_2^2] \leq C_{\text{ub}}$. We have

$$\begin{aligned}\mathbb{E}[\|\hat{\Sigma}_n\|_{op}] &\leq \frac{\sum_{i=1}^n \mathbb{E}[\|\sum_{k=t_i}^i x_k - l_i\bar{x}_n\|_2^2]}{\sum_{i=1}^n l_i} \\ &\leq \frac{\sum_{i=1}^n l_i \sum_{k=t_i}^i \mathbb{E}[\|x_k - \bar{x}_n\|_2^2]}{\sum_{i=1}^n l_i} \\ &\leq \frac{4C_{\text{ub}}\sum_{i=1}^n l_i^2}{\sum_{i=1}^n l_i} \\ &\leq 4C_{\text{ub}}n,\end{aligned}$$

where the first inequality follows from triangle inequality, the second inequality follows from Jensen's inequality, the third inequality follows from $\mathbb{E}[\|x_n^\star\|_2^2] \le C_{\text{ub}}$, and the last inequality follows from $l_i \le i \le n$. The conclusion then follows. ■

Now we prove Theorem 4. By Proposition 3, for any $k_s \gtrsim 1$ and $n \ge k_s$, we have

$$P(\tau_{k_s,\delta} \le n) \le \frac{32\eta^2\sigma^4 \exp\left(-\frac{\gamma\delta\sqrt{k_s}}{4\eta\sigma^2}\right)}{\gamma^2\delta^2} + \frac{8\eta\delta\sqrt{k_s}\exp\left(-\frac{\gamma\delta\sqrt{k_s}}{4\eta\sigma^2}\right)}{\gamma}$$

For $n \gtrsim 1$, taking $k_s \asymp \log^2 n$ so that $P(\tau_{k_s,\delta} \le n) \lesssim n^{-2}$, we have

$$\begin{aligned}
\mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_{op}] &= \mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_{op}\mathbb{1}_{\tau_{k_s,\delta}>n}] + \mathbb{E}[\|\hat{\Sigma}_n - \Sigma\|_{op}\mathbb{1}_{\tau_{k_s,\delta}\le n}] \\
&\lesssim_{\log} \sqrt{d}M^{-\frac{1}{2}} + \sqrt{d}M^{\frac{(\alpha-1)\beta+2}{2}} + nP(\tau_{k_s,\delta} \le n) \\
&\lesssim \sqrt{d}M^{-\frac{1}{2}} + \sqrt{d}M^{\frac{(\alpha-1)\beta+2}{2}} \\
&\lesssim \sqrt{d}n^{-\frac{1}{2\beta}} + \sqrt{d}n^{-\frac{(\alpha-1)\beta+1}{2\beta}},
\end{aligned}$$

where the first inequality follows from Theorem 1 and Lemma 13, the second inequality follows from $P(\tau_{k_s,\delta} \le n) \lesssim n^{-2}$, and the last inequality follows from $n \approx M^\beta$.

Appendix E. Extra assumption verification for stochastic approximation

The following proposition shows that under convexity (monotonicity), Assumption 6 holds for all the stochastic approximation algorithms in Section 5.

Proposition 14 *Suppose that the variational inclusion problem takes the form of (5.1), and Assumption 4 and 5 holds. Moreover, suppose that A is a Lipschitz and monotone map and we are in one of the following scenarios:*

1. *One applies the stochastic forward algorithm to the case $f = 0$ and g is Lipschitz and convex.*
2. *One applies the stochastic projected forward algorithm to the case f is the indicator function of a closed convex set $\mathcal{X}$ and g is Lipschitz and convex.*
3. *One applies the stochastic forward-backward algorithm to the case f is Lipschitz in its domain and $g = 0$.*

Then Assumption 6 holds.

Proof Note that the first scenario is a special case of the second one, we only prove it for the second and third cases.

Stochastic projected forward algorithm. By the definition of $x^\star$, there exists $v^\star \in \partial g(x^\star)$ and $w^\star \in N_{\mathcal{X}}(x^*)$ such that

$$0 = A(x^\star) + v^\star + w^\star.$$

By monotonicity of A and convexity of g, for any x_k and $s_g(x_k) \in \partial g(x_k)$, we have

$$\langle A(x_k) + s_g(x_k) + w^\star, x_k - x^\star\rangle = \langle A(x_k) + s_g(x_k) - A(x^\star) - v^\star, x_k - x^\star\rangle \ge 0.$$

Note also that $w^\star \in N_{\mathcal{X}}(x^\star)$, we have

$$\langle A(x_k) + s_g(x_k), x_k - x^\star \rangle \geq -\langle w^\star, x_k - x^\star \rangle \geq 0. \tag{E.1}$$

As a result, there exists some constant $C > 0$ such that

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^\star\|_2^2] &= \mathbb{E}[\|P_{\mathcal{X}}(x_k - \eta_{k+1}(A(x_k) + s_g(x_k) + \nu_{k+1})) - x^\star\|_2^2] \\
&\leq \mathbb{E}[\|x_k - \eta_{k+1}(A(x_k) + s_g(x_k) + \nu_{k+1}) - x^\star\|_2^2] \\
&\leq \mathbb{E}[\|x_k - x^\star\|^2] - 2\eta_{k+1}\mathbb{E}[\langle A(x_k) + s_g(x_k) + \nu_{k+1}, x_k - x^*\rangle] + C\eta_{k+1}^2(1 + \mathbb{E}[\|x_k - x^\star\|_2^2]) \\
&\leq (1 + C\eta_{k+1}^2)\mathbb{E}[\|x_k - x^\star\|^2] + C\eta_{k+1}^2,
\end{aligned}$$

where the first inequality follows from the fact that $P_{\mathcal{X}}$ is 1-Lipschitz, and the last inequality follows from (E.1). The results then follow from Lemma 26.

Stochastic forward-backward algorithm. By definition of $x^\star$, there exists $w \in \partial f(x^\star)$ such that

$$0 = A(x^\star) + w^\star.$$

For any x_k, we denote $x_k - \eta_{k+1}(A(x_k) + \nu_{k+1})$ by x_k^+ and $\frac{x_k^+ - \mathrm{prox}_{\eta_{k+1} f}(x_k^+)}{\eta_{k+1}}$ by w_k^+. By the property of the proximal operator, we have $w_k^+ \in \partial f(x_k^+)$. Moreover, by monotonicity of A and convexity of f, we have

$$\left\langle A(x_k^+) + w_k^+, x_k^+ - x^* \right\rangle = \left\langle A(x_k^+) + w_k^+ - A(x^\star) - w^\star, x_k^+ - x^* \right\rangle \geq 0. \tag{E.2}$$

Next, we bound $\|x_{k+1} - x_k\|$. By definition of x_{k+1} and Lipschitz property of f and A, there exists some constant $C > 0$ (may change from line to line) such that

$$\begin{aligned}
\frac{1}{2\eta_{k+1}}\|x_{k+1} - x_k\|_2^2 &\leq f(x_k) - f(x_{k+1}) - \langle A(x_k) + \nu_{k+1}, x_{k+1} - x_k\rangle \\
&\leq C(1 + \|x_k - x^\star\|_2 + \|\nu_{k+1}\|_2)\|x_{k+1} - x_k\|_2.
\end{aligned}$$

As a consequence, there exists a constant $C > 0$ such that

$$\|x_{k+1} - x_k\|_2 \leq C\eta_{k+1}(1 + \|x_k - x^\star\|_2 + \|\nu_{k+1}\|_2). \tag{E.3}$$

In addition, by Lipschitz continuity of A and f, there exists some constant $C > 0$) such that

$$\|x_k - x_k^+\|_2 \leq C\eta_{k+1}(1 + \|x_k - x^\star\|_2 + \|\nu_{k+1}\|_2). \tag{E.4}$$

As a result of (E.3) and (E.4), there exists a constant $C > 0$ such that

$$\|w_k^+\|_2 \leq \frac{1}{\eta_{k+1}}(\|x_k - x_k^+\|_2 + \|x_{k+1} - x_k\|_2) \leq C(1 + \|x_k - x^\star\|_2 + \|\nu_{k+1}\|_2)$$

Consequently, there exists a constant $C > 0$ such that

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^\star\|_2^2] &= \mathbb{E}[\|x_k - x^\star - \eta_{k+1}(A(x_k) + \nu_{k+1} + w_k^+)\|_2^2] \\
&= \mathbb{E}[\|x_k - x^\star\|_2^2] - 2\eta_{k+1}\mathbb{E}[\langle x_k - x^\star, A(x_k) + w_k^+\rangle] + C\eta_{k+1}^2(1 + \mathbb{E}[\|x_k - x^\star\|_2^2]).
\end{aligned}$$

Next, we show that $2\eta_{k+1}\mathbb{E}[\langle x_k - x^\star, A(x_k) + w_k^+\rangle]$ is lower bound. By (E.2), we have

$$\begin{aligned}
&2\eta_{k+1}\mathbb{E}[\langle x_k - x^\star, A(x_k) + w_k^+\rangle] \\
&= 2\eta_{k+1}(\mathbb{E}[\langle x_k - x_k^+, A(x_k) + w_k^+\rangle] + \mathbb{E}[\langle x_k^+ - x^\star, A(x_k) - A(x_k^+)\rangle] + \mathbb{E}[\langle x_k^+ - x^\star, A(x_k^+) + w_k^+\rangle]) \\
&\geq -2\eta_{k+1}(\underbrace{\mathbb{E}[\|x_k - x_k^+\|_2\|A(x_k) + w_k^+\|_2]}_{(I)} + \underbrace{\mathbb{E}[\|x_k^+ - x^\star\|_2\|A(x_k) - A(x_k^+)\|_2]}_{(II)}).
\end{aligned}$$

We bound (I) and (II) separately. By Holder's inequality,

$$\begin{aligned}
(I) &\leq (\mathbb{E}[\|x_k - x^+\|_2^2])^{\frac{1}{2}}(\mathbb{E}[\|A(x_k) + w_k^+\|_2^2])^{\frac{1}{2}} \\
&\leq C\eta_{k+1}(1 + \mathbb{E}[\|x_k - x^\star\|^2]),
\end{aligned}$$

where the second inequality follows from (E.4). On the other hand,

$$\begin{aligned}
(II) &\leq C \cdot \mathbb{E}[\|x_k^+ - x^\star\|_2\|x_k - x_k^+\|_2] \\
&\leq C(\mathbb{E}[\|x_k - x_k^+\|_2^2] + \mathbb{E}[\|x_k - x^\star\|_2\|x_k - x_k^+\|_2]) \\
&\leq C\left(\mathbb{E}[\|x_k - x_k^+\|_2^2] + (\mathbb{E}[\|x_k - x^\star\|_2^2])^{\frac{1}{2}}(\mathbb{E}[\|x_k - x_k^+\|_2^2])^{\frac{1}{2}}\right) \\
&\leq C\eta_{k+1}(1 + \mathbb{E}[\|x_k - x^\star\|_2^2]).
\end{aligned}$$

Combining, We have

$$2\eta_{k+1}\mathbb{E}[\langle x_k - x^\star, A(x_k) + w_k^+\rangle] \geq -C\eta_{k+1}(1 + \mathbb{E}[\|x_k - x^\star\|_2^2]).$$

Consequently, there exists constant $C > 0$ such that

$$\mathbb{E}[\|x_{k+1} - x^\star\|_2^2] \leq (1 + C\eta_{k+1}^2)\mathbb{E}[\|x_k - x^\star\|^2] + C\eta_{k+1}^2.$$

The results then follow from Lemma 26. ∎

The following proposition shows that under strong convexity (monotonicity), Assumption 7 holds for all the stochastic approximation algorithms in Section 5.

Proposition 15 *Suppose that the variational inclusion problem takes the form of (5.1). Assume that A is strongly monotone and Lipschitz. Suppose we are in one of the following scenarios:*

1. *One applies the stochastic forward algorithm to the case $f = 0$ and g is Lipschitz and convex.*
2. *One applies the stochastic projected forward algorithm to the case f is the indicator function of a closed set $\mathcal{X}$ and g is Lipschitz and convex.*
3. *One applies the stochastic forward-backward algorithm to the case f is Lipschitz in its domain and $g = 0$.*

Then Assumption 7 holds.

Proof Since the stochastic forward algorithm is a special case of the stochastic projected forward algorithm, it suffices to prove the result for both the stochastic projected forward algorithm (case 2) and the stochastic forward-backward algorithm (case 3.)

Stochastic projected forward algorithm. Recall s_g is a selection of ∂g. There exists some constant $C > 0$ (it may change from line to line through the proof) such that

$$\begin{aligned}
\|G_\eta(x,\nu)\|_2 &= \left\| \frac{x - P_{\mathcal{X}}(x - \eta(A(x) + s_g(x) + \nu))}{\eta} \right\|_2 \\
&\le \|A(x) + s_g(x) + \nu\|_2 \\
&\le C(1 + \|x - x^\star\|_2),
\end{aligned}$$

where the first inequality follows from the fact that $P_{\mathcal{X}}$ is 1-Lipschitz and the second inequality follows from the Lipschitz continuity of A and g. Item 2 follows. On the other hand, by the definition of $x^\star$, there exists $v^\star \in \partial g(x^\star)$ and $w^\star \in N_{\mathcal{X}}(x^*)$ such that

$$0 = A(x^\star) + v^\star + w^\star.$$

By strong monotonicity of A and convexity of g, there exists $\gamma > 0$ such that for any x and $s_g(x) \in \partial g(x)$, we have

$$\begin{aligned}
\langle A(x) + s_g(x) + w^\star, x - x^\star \rangle &= \langle A(x) + s_g(x) - A(x^\star) - v^\star, x - x^\star \rangle \\
&\ge \gamma \|x - x^\star\|_2^2.
\end{aligned}$$

As a result of $w^\star \in N_{\mathcal{X}}(x^*)$, we have

$$\langle A(x) + s_g(x), x - x^\star \rangle \ge \gamma \|x_k - x^\star\|_2^2.$$

Next, we denote $x - \eta(A(x) + s_g(x) + \nu)$ by x^+ and $\frac{x^+ - P_{\mathcal{X}}(x^+)}{\eta}$ by w. Note that $w \in N_{\mathcal{X}}(x^+)$ and $G_\eta(x,\nu) = w + A(x) + s_g(x) + \nu$, so we have

$$\|w\|_2 \le C(1 + \|x - x^\star\|_2 + \nu).$$

Therefore,

$$\begin{aligned}
\langle G_\eta(x,\nu) - \nu, x - x^\star \rangle &= \langle w + A(x) + s_g(x), x - x^\star \rangle \\
&\ge \gamma \|x - x^\star\|^2 + \left\langle w, x - x^+ \right\rangle + \left\langle w, x^+ - x^\star \right\rangle \\
&\ge \gamma \|x - x^\star\|^2 + \left\langle w, x - x^+ \right\rangle,
\end{aligned}$$

where the first inequality follows from E and the second inequality follows from $w \in N_{\mathcal{X}}(x^+)$. Note also that

$$\|x - x^+\|_2 \le C\eta(1 + \|x - x^\star\|_2 + \|\nu\|_2),$$

we have

$$|\left\langle w, x - x^+ \right\rangle| \le \|w\|_2 \|x - x_+\|_2 \le C\eta(1 + \|x - x^\star\|_2^2 + \|\nu\|_2^2).$$

Combining, we have

$$\langle G_\eta(x,\nu) - \nu, x - x^\star \rangle \ge \gamma \|x - x^\star\|^2 - C\eta(1 + \|x - x^\star\|_2^2 + \|\nu\|_2^2).$$

Stochastic forward-backward algorithm. First, we bound $\|G_\eta(x,\nu)\|_2$. By definition of proximal operator and Lipschitz property of f and A, there exists some constant $C>0$ (may change from line to line) such that

$$\begin{aligned}\frac{\eta}{2}\|G_\eta(x,\nu)\|_2^2 &\le f(x)-f(x-\eta G_\eta(x,\nu))+\eta\left\langle A(x)+\nu, G_\eta(x,\nu)\right\rangle\\ &\le C\eta\|G_\eta(x,\nu)\|_2(1+\|x-x^\star\|_2+\|\nu\|_2).\end{aligned}$$

As a consequence, there exists a constant $C>0$ such that

$$\|G_\eta(x,\nu)\|_2 \le C(1+\|x-x^\star\|_2+\|\nu\|_2). \tag{E.5}$$

Therefore, item 2 follows. Next, by the definition of $x^\star$, there exists $w^\star\in\partial f(x^\star)$ such that

$$0=A(x^\star)+w^\star.$$

For any x, we denote $x-\eta(A(x)+\nu)$ by x^+ and $\frac{x^+-\mathrm{prox}_{\eta f}(x^+)}{\eta}$ by w^+. By the property of the proximal operator, we have $w^+\in\partial f(x^+)$. Moreover, by strong monotonicity of A and convexity of f, we have

$$\begin{aligned}\left\langle A(x^+)+w^+, x^+-x^\star\right\rangle &= \left\langle A(x^+)+w^+-A(x^\star)-w^\star, x^+-x^\star\right\rangle\\ &\ge \gamma\|x^+-x^\star\|^2.\end{aligned} \tag{E.6}$$

In addition, by Lipschitz continuity of A and f, there exists some constant $C>0$) such that

$$\|x-x^+\|_2 \le C\eta(1+\|x-x^\star\|_2+\|\nu\|_2). \tag{E.7}$$

As a result of (E.5) and (E.7), there exists constant $C>0$ such that

$$\|w^+\|_2 \le \frac{1}{\eta}\|x-x^+\|_2+\|G_\eta(x,\nu)\|_2 \le C(1+\|x-x^\star\|_2+\|\nu\|_2)$$

Note that $G_\eta(x,\nu)=w^++A(x)+\nu$, we have

$$\begin{aligned}\langle G_\eta(x,\nu)-\nu, x-x^\star\rangle &= \left\langle w^++A(x), x-x^\star\right\rangle\\ &= \underbrace{\left\langle A(x)-A(x^+), x-x^\star\right\rangle}_{(I)}+\underbrace{\left\langle w^++A(x^+), x^+-x^\star\right\rangle}_{(II)}+\underbrace{\left\langle w^++A(x^+), x-x^+\right\rangle}_{(III)}.\end{aligned}$$

We lower-bound each term separately. By Lipschitz continuity of A and (E.7), we have

$$\|(I)\|_2 \le C(1+\|x-x^\star\|_2^2+\|\nu\|_2^2).$$

By (E.6), we have

$$\begin{aligned}(II) &\ge \gamma\|x^+-x^\star\|^2\\ &\ge \gamma\|x-x^\star\|_2^2-2\|x_+-x\|_2\|x-x^\star\|_2\\ &\ge \gamma\|x-x^\star\|_2^2-C(1+\|x-x^\star\|_2^2+\|\nu\|_2^2).\end{aligned}$$

Moreover,

$$\begin{aligned}\|(III)\|_2 &\le (\|w^+\|_2+\|A(x_+)\|_2)\|x-x^+\|_2\\ &\le C(1+\|x-x^\star\|_2^2+\|\nu\|_2^2),\end{aligned}$$

where the last inequality follows from the Lipschitz continuity of A and (E.5). The results then follows by combining (I), (II), and (III). ∎

Appendix F. Technical lemmas

Recall that for a given index $k \geq 0$ and a constant $\delta \in (0,1)$, the stopping time is defined as

$$\tau_{k,\delta} := \inf\{l \geq k \colon x_l \notin B_\delta(x^\star)\},$$

which is the first time after k that the iterate leaves $B_\delta(x^\star)$. Now, define $D_k := \operatorname{dist}(x_k, \mathcal{M})$, $v_k := G_{\eta_{k+1}}(x_k, \nu_{k+1})$ for all $k \geq 0$. In what follows, C denotes constant and may change from line to line.

Lemma 16 *Suppose that Assumptions 2, 3, and 4 hold. If $\alpha \in (1/2, 1)$, then for any sufficiently small $\delta > 0$, any $k_s \geq 0$, there exists a constant C depending on δ, k_s and α such that for any $l \geq s \geq k_s$,*

$$\textstyle\sum_{k=s}^{l} \mathbb{E}[D_k^2 1_{\tau_{k_s,\delta} > k}] \leq C k_s^{2\alpha} \sum_{k=s}^{l} \eta_k^2.$$

Proof First, we note that it suffices to show the result for all $k_s \geq \left(\frac{4\alpha}{\mu\eta}\right)^{1/(1-\alpha)}$ since the cases when $k_s \leq \left(\frac{4\alpha}{\mu\eta}\right)^{1/(1-\alpha)}$ can be handled by enlarging C properly. Define $A_k := \{\tau_{k_s,\delta} > k\}$ for all $k \geq k_s$. We require that δ is small enough so that $B_\delta(x^\star)$ is contained in the neighborhood where Assumption 3 holds with probability 1. Note that we require k_s (or η) to be large enough so the conclusions of Lemma 22 holds for all $k \geq k_s$. We first prove a recurrence relation satisfied by the sequence D_k. To that end, recall the update rule (3.3), for all $k \geq 0$, when $x_k \in B_\delta(x^\star)$, we have

$$\begin{aligned}
D_{k+1}^2 &\leq \|x_{k+1} - P_{\mathcal{M}}(x_k)\|^2 \\
&= \|x_k - \eta_{k+1} v_k - P_{\mathcal{M}}(x_k)\|^2 \\
&= \|x_k - P_{\mathcal{M}}(x_k)\|^2 - 2\eta_{k+1}\langle v_k, x_k - P_{\mathcal{M}}(x_k)\rangle + \eta_{k+1}^2 \|v_k\|^2 \\
&\leq D_k^2 - 2\eta_{k+1}\mu D_k + 2\eta_{k+1}(1 + \|\nu_{k+1}\|)^2 o(D_k) \\
&\qquad - 2\eta_{k+1}\langle \nu_{k+1}, x_k - P_{\mathcal{M}}(x_k)\rangle + \underbrace{C(1 + \|\nu_{k+1}\|)^2}_{:=B_{k+1}} \eta_{k+1}^2,
\end{aligned} \tag{F.1}$$

where the second inequality follows from Assumption 2 and Condition 2 of Assumption 3. Note that the bound $\mathbb{E}_k[\|\nu_{k+1}\|^4] 1_{A_k} \leq q(x_k) 1_{A_k}$ implies that there exists $C > 0$ such that

$$\mathbb{E}_k[B_{k+1}] 1_{A_k} \leq C,$$

meaning the conditional expectation is bounded for all i. Moreover, by shrinking δ if necessary, we have

$$\mathbb{E}_k[(1 + \|\nu_{k+1}\|)^2 o(D_k) 1_{A_k}] \leq \frac{\mu}{2} D_k 1_{A_k}.$$

Thus, for each $k \geq k_s$, we have

$$\begin{aligned}
\mathbb{E}_k[D_{k+1}^2 1_{A_{k+1}}] &\leq \mathbb{E}_k[D_{k+1}^2 1_{A_k}] \\
&\leq D_k^2 1_{A_k} - \mu\eta_{k+1} D_k 1_{A_k} + C\eta_{k+1}^2
\end{aligned} \tag{F.2}$$

where the first inequality follows from $1_{A_{k+1}} \leq 1_{A_k}$, the second inequality the assumption that $\{\nu_k\}$ is a martingale difference sequence and A_k is $\mathcal{F}_k$ measurable. Taking expectations on both sides, we have

$$\mathbb{E}[D_{k+1}^2 1_{A_{k+1}}] \leq \mathbb{E}[D_k^2 1_{A_k}] - \mu\eta_{k+1}\mathbb{E}[D_k 1_{A_k}] + C\eta_{k+1}^2. \tag{F.3}$$

Summing (F.3) from $k = s$ to l and using Lemma 17, we have

$$\sum_{k=s}^{l} \eta_{k+1}\mathbb{E}[D_k 1_{A_k}] \lesssim \mathbb{E}[D_s^2 1_{A_s}] + \sum_{k=s}^{l} \eta_{k+1}^2 \lesssim k_s^{\alpha}\eta_s + \sum_{k=s}^{l} \eta_{k+1}^2 \lesssim k_s^{\alpha} \sum_{k=s}^{l} \eta_{k+1}^2. \tag{F.4}$$

On the other hand, when $x_k \in B_\delta(x^\star)$, we have

$$\begin{aligned}
D_{k+1}^4 &\leq \|x_k - \eta_{k+1}v_k - P_{\mathcal{M}}(x_k)\|^4 \\
&= D_k^4 - 4\eta_{k+1}\langle v_k, x_k - P_{\mathcal{M}}(x_k)\rangle D_k^2 + \eta_{k+1}^4\|v_k\|^4 + 2\eta_{k+1}^2 D_k^2\|v_k\|_2^2 + 4\eta_{k+1}^2\langle v_k, x_k - P_{\mathcal{M}}(x_k)\rangle^2 \\
&\quad - 4\eta_{k+1}^3\langle v_k, x_k - P_{\mathcal{M}}(x_k)\rangle \|v_k\|^2 &\text{(F.5)}\\
&\leq D_k^4 - 4\mu\eta_{k+1}D_k^3 + 4\eta_{k+1}D_k^2(1+\|\nu_{k+1}\|^2)o(D_k) - 4\eta_{k+1}\langle \nu_{k+1}, x_k - P_{\mathcal{M}}(x_k)\rangle D_k^2 + \eta_{k+1}^4\|v_k\|^4 \\
&\quad + 6\eta_{k+1}^2 D_k^2\|v_k\|^2 + 4\eta_{k+1}^3 D_k\|v_k\|^3, &\text{(F.6)}
\end{aligned}$$

where the equality (F.5) follows from expanding the fourth power directly and the estimate (F.6) follows from Conditioning 2 of Assumption 3 and Cauchy-Schwarz inequality. Thus, there exists constant $C > 0$ such that for each $i \geq 0$, we have

$$\begin{aligned}
\mathbb{E}_k[D_{k+1}^4 1_{A_{k+1}}] &\leq \mathbb{E}_k[D_{k+1}^4 1_{A_k}] \\
&\leq D_k^4 1_{A_k} - 2\mu\eta_{k+1}D_k^3 1_{A_k} + C\eta_{k+1}^2 D_k^2 1_{A_k} + C\eta_{k+1}^3 D_k 1_{A_k} + C\eta_{k+1}^4 \\
&\leq (1-\mu\eta_{k+1})D_k^4 1_{A_k} - \mu\eta_{k+1}D_k^3 1_{A_k} + C\eta_{k+1}^2 D_k^2 1_{A_k} + C\eta_{k+1}^3 D_k 1_{A_k} + C\eta_{k+1}^4,
\end{aligned} \tag{F.7}$$

where the first inequality follows from $1_{A_{k+1}} \leq 1_{A_k}$, the second inequality follows from the assumption that $\{\nu_{k+1}\}$ is a martingale difference sequence, our choice of δ, and the bound on the fourth moment of ν_i, and the third inequality follows from the assumption that $\delta < 1$. By Lemma 22, for all $k \geq k_s$, we have

$$\frac{1-\mu\eta_{k+1}}{\eta_{k+1}^2} \leq \frac{1}{\eta_k^2}.$$

Taking expectation and dividing both sides of (F.7) by η_{k+1}^2, we have

$$\frac{\mathbb{E}[D_{k+1}^4 1_{A_{k+1}}]}{\eta_{k+1}^2} \leq \frac{\mathbb{E}[D_k^4 1_{A_k}]}{\eta_k^2} - 2\frac{\mu}{\eta_{k+1}}\mathbb{E}[D_k^3 1_{A_k}] + C\mathbb{E}[D_k^2 1_{A_k}] + C\eta_{k+1}\mathbb{E}[D_k 1_{A_k}] + C\eta_{k+1}^2. \tag{F.8}$$

For any index $l \geq s \geq k_s$, summing (F.8) from s to l, we have

$$\begin{aligned}
\sum_{k=s}^{l} \frac{1}{\eta_{k+1}}\mathbb{E}[D_k^3 1_{A_k}] &\leq \frac{\mathbb{E}[D_s^4 1_{A_s}]}{\eta_s^2} + C\left(\sum_{k=s}^{l}\mathbb{E}[D_k^2 1_{A_k}] + \sum_{k=s}^{l}\eta_{k+1}\mathbb{E}[D_k 1_{A_k}] + \sum_{k=s}^{l}\eta_{k+1}^2\right) \\
&\lesssim \sum_{k=s}^{l}\mathbb{E}[D_k^2 1_{A_k}] + k_s^{3\alpha}\sum_{k=s}^{l}\eta_{k+1}^2, &\text{(F.9)}
\end{aligned}$$

where the second inequality follows from Lemma 17 and the estimate (F.4). Combining (F.4) and (F.9), we have

$$\begin{aligned}
\sum_{k=s}^{l}\mathbb{E}[D_k^2 1_{A_k}] &\leq \sqrt{\sum_{k=s}^{l}\eta_{k+1}\mathbb{E}[D_k 1_{A_k}] \cdot \sum_{k=s}^{l}\frac{1}{\eta_{k+1}}\mathbb{E}[D_k^3 1_{A_k}]} \\
&\lesssim \sqrt{k_s^{\alpha}\sum_{k=s}^{l}\eta_{k+1}^2 \cdot (k_s^{3\alpha}\sum_{k=s}^{l}\eta_{k+1}^2 + \sum_{k=s}^{l}\mathbb{E}[D_k^2 1_{A_k}])},
\end{aligned}$$

where the first inequality follows from Holder's inequality. Simple calculation yields

$$\textstyle\sum_{k=s}^{l} \mathbb{E}[D_k^2 1_{A_k}] \lesssim k_s^{2\alpha} \sum_{k=s}^{l} \eta_{k+1}^2,$$

as desired. ■

Lemma 17 *Suppose that Assumption 2, 3, and 4 hold. Then for any sufficiently small $\delta > 0$, there exists a constant $C > 0$ such that for any $k_s \geq 1$, and any $k \geq k_s$,*

$$\mathbb{E}[D_k^2 1_{\tau_{k_s,\delta} > k}] \leq C k_s^{\alpha} \eta_k, \quad \mathbb{E}[D_k^4 1_{\tau_{k_s,\delta} > k}] \leq C k_s^{3\alpha} \eta_k^3.$$

Proof We require that δ is small enough so that $\delta \leq 1$ and $B_\delta(x^\star)$ is contained in the neighborhood where Assumption 3 holds with probability 1. Define $A_k := \{\tau_{k_s,\delta} > k\}$ for all $k \geq k_s$. Following the calculation in (F.1), we obtain (F.2). Consequently,

$$\begin{aligned}\mathbb{E}_k[D_{k+1}^2 1_{A_{k+1}}] &\leq D_k^2 1_{A_k} - \mu\eta_{k+1} D_k 1_{A_k} + C\eta_{k+1}^2 \\ &\leq (1 - \mu\eta_{k+1}) D_k^2 1_{A_k} + C\eta_{k+1}^2,\end{aligned}$$

where the second inequality follows from $\delta \leq 1$. Taking expectations, we have

$$\mathbb{E}[D_{k+1}^2 1_{A_{k+1}}] \leq (1 - \mu\eta_{k+1}) \mathbb{E}[D_k^2 1_{A_k}] + C\eta_{k+1}^2.$$

By Lemma 24, there exists a constant C such that for any $k \geq k_s$,

$$\mathbb{E}[D_k^2 1_{A_k}] \leq C k_s^{\alpha} \eta_k.$$

On the other hand, by the same argument of the proof of Lemma 16, we have (F.7), which reads

$$\mathbb{E}_k[D_{k+1}^4 1_{A_{k+1}}] \leq (1 - \mu\eta_{k+1}) D_k^4 1_{A_k} - \mu\eta_{k+1} D_k^3 1_{A_k} + C\eta_{k+1}^2 D_k^2 1_{A_k} + C\eta_{k+1}^3 D_k 1_{A_k} + C\eta_{k+1}^4.$$

Note that there exists a constant $\tilde{C}$ depending only on μ and C such that when $D_k \geq \tilde{C}\eta_{k+1}$, we have

$$-\mu\eta_{k+1} D_k^3 1_{A_k} + C\eta_{k+1}^2 D_k^2 1_{A_k} + C\eta_{k+1}^3 D_k 1_{A_k} \leq 0.$$

When $D_k \leq \tilde{C}\eta_{k+1}$, we have

$$-\mu\eta_{k+1} D_k^3 1_{A_k} + C\eta_{k+1}^2 D_k^2 1_{A_k} + C\eta_{k+1}^3 D_k 1_{A_k} \leq (C\tilde{C}^2 + C\tilde{C} + C)\eta_{k+1}^4.$$

Therefore, by enlarging C if necessary, we always have

$$\mathbb{E}_k[D_{k+1}^4 1_{A_{k+1}}] \leq (1 - \mu\eta_{k+1}) D_k^4 1_{A_k} + C\eta_{k+1}^4.$$

Taking expectations, we have

$$\mathbb{E}[D_{k+1}^4 1_{A_{k+1}}] \leq (1 - \mu\eta_{k+1}) \mathbb{E}[D_k^4 1_{A_k}] + C\eta_{k+1}^4.$$

By Lemma 24, there exists a constant C such that for any $k \geq k_s$,

$$\mathbb{E}[D_k^4 1_{A_k}] \leq C k_s^{3\alpha} \eta_k^3.$$

■

We have the following lemma for the size of E_k.

Lemma 18 *Suppose that Assumption 1 – 6 hold. Let $\delta > 0$ be small enough so that Assumption 1 – 3 hold inside $B_\delta(x^\star)$. For any $k \geq 0$, we have*

$$\mathbb{E}[\|E_k\|_2^2] \lesssim \left(\frac{\delta}{\eta_{k+1}}\right)^2 + \delta^2 + C_{ub}$$

Additionally, for any $k_s \geq 0$, and $j \geq i \geq k_s$, we have

$$\textstyle\sum_{k=i}^{j} \mathbb{E}[\|E_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] \lesssim \sum_{k=i}^{j} k_s^{2\alpha}\eta_k^2.$$

Proof By definition, we always have

$$\begin{aligned}\mathbb{E}[\|E_k\|_2^2] &= \mathbb{E}\left[\left\|\frac{y_{k+1}-y_k}{\eta_{k+1}} + F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k)\right\|_2^2\right] \\ &\lesssim \left[\left(\frac{\delta}{\eta_{k+1}}\right)^2 + \delta^2 + C_{\text{ub}}\right],\end{aligned}$$

where the last inequality follows from $y_k \in B_{4\delta}(x^\star)$, the smoothness of $F_{\mathcal{M}}$, and Assumption 5 and 6. On the other hand, by (Davis et al., 2025, Proposition 6.3, item 2(a)), we have

$$\|E_k\|_2 1_{\tau_{k_s,\delta}>k} \lesssim (1+\|\nu_k\|^2)(D_k+\eta_k). \tag{F.11}$$

The estimate F.10 then follows from Assumption 4 and Lemma 16. ■

Lemma 19 *Suppose that Assumption 1 – 6 hold. Let $\delta > 0$ be small enough so that Assumption 1 – 3 hold inside $B_\delta(x^\star)$. For any $k_s \geq 0$ and $k \geq k_s$, we have*

$$\mathbb{E}[\|y_k - x^\star\|_2^p \mathbb{1}_{\tau_{k_s,\delta}>k}] \lesssim k_s^{p\alpha/2}\eta_k^{p/2}, \quad p = 1,2,4.$$

Proof Note that

$$\begin{aligned}
&\|y_{k+1}-x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} \\
&= \|y_k - \eta_{k+1}F_{\mathcal{M}}(y_k) - \eta_{k+1}P_{T_{\mathcal{M}}(y_k)}(\nu_k) + \eta_{k+1}E_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} \\
&\leq \|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} - 4\eta_{k+1}\left\langle y_k - x^\star, F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\right\rangle \|y_k - x^\star\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k} \\
&\quad + 6\eta_{k+1}^2 \|y_k - x^\star\|_2^2 \|F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k} \\
&\quad + 4\eta_{k+1}^3 \|y_k - x^\star\|_2 \|F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\|_2^3 \mathbb{1}_{\tau_{k_s,\delta}>k} + \eta_{k+1}^4 \|F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} \\
&\leq \underbrace{\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} - 4\eta_{k+1}\left\langle y_k - x^\star, F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\right\rangle \|y_k - x^\star\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}}_{(I)} \\
&\quad + \underbrace{8\eta_{k+1}^2 \|y_k - x^\star\|_2^2 \|F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}}_{(II)} + \underbrace{3\eta_{k+1}^4 \|F_{\mathcal{M}}(y_k) + P_{T_{\mathcal{M}}(y_k)}(\nu_k) - E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}}_{(III)},
\end{aligned}$$

where the first inequality follows from direct expansion and Cauchy-Schwarz inequality, and the second inequality follows from Young's inequality. Taking expectations, we have

$$\mathbb{E}[\|y_{k+1}-x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] \leq \mathbb{E}[(I)] + \mathbb{E}[(II)] + \mathbb{E}[(III)].$$

We bound the terms separately. By Assumption 1 and the inequality of arithmetic and geometric means, we have

$$\begin{aligned}\mathbb{E}_k[(I)] &= \mathbb{E}_k[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} - 4\eta_{k+1}\left\langle y_k - x^\star, F_{\mathcal{M}}(y_k)\mathbb{1}_{\tau_{k_s,\delta}>k} - \mathbb{E}_k[E_k]\right\rangle \|y_k - x^\star\|_2^2 \mathbb{1}_{\tau_{k_s,\delta}>k}] \\ &\le (1-4\gamma\eta_{k+1})\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} + 4\eta_{k+1}\mathbb{E}_k[\|E_k\|_2\|y_k - x^\star\|_2^3 \mathbb{1}_{\tau_{k_s,\delta}>k}] \\ &\le (1-4\gamma\eta_{k+1})\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} + \gamma\eta_{k+1}\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} + \frac{81}{\gamma^3}\eta_{k+1}\mathbb{E}_k[\|E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] \\ &\le (1-3\gamma\eta_{k+1})\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} + \frac{81}{\gamma^3}\eta_{k+1}\mathbb{E}_k[\|E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}],\end{aligned}$$

Taking expectation, using (F.11), and applying Lemma 17, there exists constant C such that

$$\begin{aligned}\mathbb{E}[(I)] &\le (1-3\gamma\eta_{k+1})\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + \frac{81}{\gamma^3}\eta_{k+1}\mathbb{E}[\|E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] \\ &\le (1-3\gamma\eta_{k+1})\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + Ck_s^{3\alpha}\eta_{k+1}^4.\end{aligned}$$

Similarly,

$$\begin{aligned}(II) &\le 24\eta_{k+1}^2\|y_k - x^\star\|_2^2(\|F_{\mathcal{M}}(y_k)\|_2^2 + \|P_{T_{\mathcal{M}}(y_k)}(\nu_k)\|_2^2 + \|E_k\|_2^2)\mathbb{1}_{\tau_{k_s,\delta}>k} \\ &\le (24L_{\mathcal{M}}^2 + 12)\eta_{k+1}^2\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} + 12\eta_{k+1}^2\|E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} \\ &\quad + \gamma\eta_{k+1}\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k} + \frac{144}{\gamma}\eta_{k+1}^3\|P_{T_{\mathcal{M}}(y_k)}(\nu_k)\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k},\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from Lipschitz continuity of $F_{\mathcal{M}}$ and Young's inequality. Taking expectation, and using (F.11), and Lemma 17 and Assumption 5, there exists a constant C such that

$$\mathbb{E}[(II)] \le (24L_{\mathcal{M}}^2 + 12)\eta_{k+1}^2\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + Ck_s^3\eta_{k+1}^5 + \gamma\eta_{k+1}\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + C\eta_{k+1}^3.$$

Moreover, using Jensen's inequality, (F.11), and applying Lemma 17, there exists a constant C such that

$$\begin{aligned}\mathbb{E}[(III)] &\le 81\eta_{k+1}^4\mathbb{E}[\|F_{\mathcal{M}}(y_k)\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + 81\eta_{k+1}^4\mathbb{E}[\|P_{T_{\mathcal{M}}(y_k)}(\nu_k)\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + 81\eta_{k+1}^4\mathbb{E}[\|E_k\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] \\ &\le C\eta_{k+1}^4 + Ck_s^{3\alpha}\eta_{k+1}^7.\end{aligned}$$

Combining and using the fact that $k_s\eta_{k+1} \lesssim 1$, we have

$$\mathbb{E}[\|y_{k+1} - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k+1}] \le (1 - 2\gamma\eta_{k+1} + (24L_{\mathcal{M}}^2 + 12)\eta_{k+1}^2)\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + Ck_s^{2\alpha}\eta_{k+1}^3.$$

As a result, for any $k \ge \max\left\{k_s, \left(\frac{\eta(24L_{\mathcal{M}}^2+12)}{\gamma}\right)^{1/\alpha}\right\}$, we have

$$\mathbb{E}[\|y_{k+1} - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] \le (1-\gamma\eta_{k+1})\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] + Ck_s^{2\alpha}\eta_{k+1}^3.$$

By Lemma 24 and the fact that $\left(\frac{\eta(24L_{\mathcal{M}}^2+12)}{\gamma}\right)^{1/\alpha}$ is a constant, there exists constant C such that

$$\mathbb{E}[\|y_k - x^\star\|_2^4 \mathbb{1}_{\tau_{k_s,\delta}>k}] \le Ck_s^{2\alpha}\eta_k^2, \quad \forall k \ge k_s.$$

This resolves the case when $p = 4$. The other two cases follow from Holder's inequality. ∎

Appendix G. Auxiliary lemmas

Lemma 20 *Define*

$$p_i^j = \begin{cases} \prod_{k=i}^j \left(1 - \frac{\gamma\eta_i}{2}\right) & i \le j \\ 1 & i = j+1. \end{cases}$$

Then for any $j \ge i$,

$$p_i^j \le \exp\left(-\frac{\gamma\eta((j+1)^{1-\alpha} - i^{1-\alpha})}{2(1-\alpha)}\right)$$

Proof Note that

$$\begin{aligned} \log(p_i^j) &= \sum_{k=i}^j \log\left(1 - \frac{\gamma\eta_i}{2}\right) \\ &\le -\frac{\gamma\eta}{2} \sum_{k=i}^j k^{-\alpha} \\ &\le -\frac{\gamma\eta}{2} \int_i^{j+1} x^{-\alpha} dx \\ &= -\frac{\gamma\eta((j+1)^{1-\alpha} - i^{1-\alpha})}{2(1-\alpha)}. \end{aligned}$$

∎

Lemma 21 *For any $\alpha \in \left(\frac{1}{2}, 1\right)$ and $1 \le i \le j$, we have*

$$\sum_{k=i}^j k^{-2\alpha} \le \sum_{k=i}^\infty k^{-2\alpha} \le 1 + \frac{1}{2\alpha-1} i^{1-2\alpha}$$

Proof Note that

$$\begin{aligned} \sum_{k=i}^\infty k^{-2\alpha} &\le 1 + \sum_{k=i+1}^\infty k^{-2\alpha} \\ &\le 1 + \int_i^\infty x^{-2\alpha} dx \\ &= 1 + \frac{1}{2\alpha - 1} i^{1-2\alpha}. \end{aligned}$$

∎

Lemma 22 *If $\alpha \in (\frac{1}{2}, 1)$, then for all $k \ge \left(\frac{4\alpha}{\mu\eta}\right)^{1/(1-\alpha)}$, we have*

$$\frac{1 - \mu\eta_{k+1}}{\eta_{k+1}^2} \le \frac{1}{\eta_k^2}.$$

If $\alpha = 1$ and $\eta \ge \frac{4}{\mu}$, then the same inequality holds for all $k \ge 1$.

Proof Note that $\eta_k = \eta k^{-\alpha}$ by Assumption 4, it suffices to show that

$$\frac{1 - \mu\eta(k+1)^{-\alpha}}{(k+1)^{-2\alpha}} \le \frac{1}{k^{-2\alpha}}.$$

Equivalently, we show that

$$(k+1)^{2\alpha} - \mu\eta(k+1)^{\alpha} \le k^{2\alpha} \tag{G.1}$$

Note that

$$\begin{aligned}\left(1+\frac{1}{k}\right)^{2\alpha} &\le 1 + \frac{4\alpha}{k} \\ &\le 1 + \frac{\mu\eta(k+1)^{\alpha}}{k^{2\alpha}},\end{aligned}$$

where the first inequality follows from the fact that $(1+x)^{2\alpha} \le 1 + 4\alpha x$ for all $\alpha \in (\frac{1}{2}, 1]$ and $x \in (0,1]$, and the second inequality follows from our assumption on k and η, for the cases $\alpha \in (\frac{1}{2}, 1)$ and $\alpha = 1$, respectively. Rearranging it, we obtain (G.1). ■

Lemma 23 *Let $\alpha \in \left(\frac{1}{2}, 1\right)$ and $C > 0$. Then for any $k \ge 0$, we have*

$$\textstyle\sum_{i=k}^{\infty} \exp(-C(i+1)^{\alpha}) \le \frac{2\exp(-C\sqrt{k})}{C^2} + \frac{2\sqrt{k}\exp(-C\sqrt{k})}{C}$$

Proof Note that

$$\begin{aligned}\textstyle\sum_{i=k}^{\infty} \exp(-C(i+1)^{\alpha}) &\le \textstyle\sum_{i=k}^{\infty} \exp(-C(i+1)^{1/2}) \\ &\le \int_k^{\infty} \exp(-Cx^{1/2})dx \\ &\le \int_{\sqrt{k}}^{\infty} 2u\exp(-Cu)du \\ &= \frac{2\exp(-C\sqrt{k})}{C^2} + \frac{2\sqrt{k}\exp(-C\sqrt{k})}{C},\end{aligned}$$

where the equality follows from the standard calculus calculation using integration by parts. ■

Lemma 24 *Let $\alpha \in (0, \frac{1}{2})$, $\theta > \alpha$, $c_1 > 0$, $c_2 > 0$, and $c_3 > 0$ be constants. Let $\{s_k\}$ be a sequence such that $0 \le s_k \le c_3$ for all $k \ge 0$. Suppose that there exists $k_0 \ge 0$ such that*

$$s_{k+1} \le (1 - c_1(k+1)^{-\alpha})s_k + c_2(k+1)^{-\theta}, \quad \forall k \ge k_0. \tag{G.2}$$

Let $C = \max\left\{c_3, c_3\left(\frac{2(\theta-\alpha)}{c_1}\right)^{\frac{\theta-\alpha}{1-\alpha}}, \frac{2c_2}{c_1(k_0+1)^{\theta-\alpha}}\right\}$. We have

$$s_k \le C(k_0+1)^{\theta-\alpha}(k+1)^{-(\theta-\alpha)}, \forall k \ge k_0.$$

Proof We first show that the desired bound holds for all the $k_0 \le k \le \max\left\{\left(\frac{2(\theta-\alpha)}{c_1}\right)^{\frac{1}{1-\alpha}}, k_0\right\}$. Note that $s_k \le c_3$, it suffices to show

$$C(k_0+1)^{\theta-\alpha}\left(\max\left\{\left(\frac{2(\theta-\alpha)}{c_1}\right)^{\frac{1}{1-\alpha}}, k_0\right\} + 1\right)^{-(\theta-\alpha)} \ge c_3,$$

which holds by our assumption on C. Next, we apply induction to prove the bound for all $k \geq \max\left\{\left(\frac{2(\theta-\alpha)}{c_1}\right)^{\frac{1}{1-\alpha}}, k_0\right\}$. Suppose that the bound holds for some $k \geq \max\left\{\left(\frac{2(\theta-\alpha)}{c_1}\right)^{\frac{1}{1-\alpha}}, k_0\right\}$. By (G.2), we have

$$\begin{aligned} s_{k+1} &\leq C(k_0+1)^{\theta-\alpha}(k+1)^{-(\theta-\alpha)} - c_1 C(k_0+1)^{\theta-\alpha}(k+1)^{-\theta} + c_2(k+1)^{-\theta} \\ &\leq C(k_0+1)^{\theta-\alpha}(k+1)^{-(\theta-\alpha)} - \frac{c_1 C}{2}(k_0+1)^{\theta-\alpha}(k+1)^{-\theta} \\ &= C(k_0+1)^{\theta-\alpha}(k+1)^{-(\theta-\alpha)}\left(1-\frac{c_1}{2}(k+1)^{-\alpha}\right), \end{aligned}$$

where the second inequality follows from the lower bound on C. In addition, simple calculus shows that for any $x \in [0,1/2]$, $(1-x)^{\theta-\alpha} \geq 1-2(\theta-\alpha)x$. Therefore, $\left(1-\frac{1}{k+2}\right)^{\theta-\alpha} \geq 1-\frac{2(\theta-\alpha)}{k+2}$. By the lower bound on k, we have

$$1-\frac{2(\theta-\alpha)}{k+2} \geq 1-\frac{c_1}{2}(k+1)^{-\alpha}.$$

Combining, we have

$$s_{k+1} \leq C(k_0+1)^{\theta-\alpha}(k+2)^{-(\theta-\alpha)}.$$

The result follows. ∎

Lemma 25 ((Zhu et al., 2023, Lemma A.2)) *For any $j > i$, we have*

$$\|S_i^j\|_2 \lesssim i^{\alpha}.$$

Lemma 26 *Let $\{x_k\}_{k=0}^{\infty}$ be a nonnegative sequence satisfying*

$$x_{k+1} \leq (1+C_1(k+1)^{-2\alpha})x_k + C_2(k+1)^{-2\alpha},$$

where C_1 and C_2 are positive constants, and $\alpha \in (1/2, 1)$. Then, there exists a constant C depending on C_1, C_2, α and x_0 such that $x_k \leq C$ holds for any $k \geq 0$.

Proof We begin by unrolling the recurrence. For any $k \geq 0$, iterating the inequality gives

$$x_{k+1} \leq \prod_{j=0}^{k}\left(1+C_1(j+1)^{-2\alpha}\right)x_0 + \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}\left(1+C_1(j+1)^{-2\alpha}\right)\right)C_2(i+1)^{-2\alpha}.$$

To bound the products, we use the inequality $\log(1+u) \leq u$ for all $u > -1$. Hence,

$$\prod_{j=i+1}^{k}\left(1+C_1(j+1)^{-2\alpha}\right) \leq \exp\left(C_1\sum_{j=i+1}^{k}(j+1)^{-2\alpha}\right) \leq \exp\left(C_1\sum_{j=1}^{\infty} j^{-2\alpha}\right) =: M.$$

Since $\alpha > \frac{1}{2}$, we have $2\alpha > 1$, and the series $\sum_{j=1}^{\infty} j^{-2\alpha}$ converges; hence, $M < \infty$.

Using this bound, we deduce that

$$x_{k+1} \leq Mx_0 + MC_2\sum_{i=0}^{k}(i+1)^{-2\alpha}.$$

Setting $C = M(x_0 + C_2\sum_{i=0}^{k}(i+1)^{-2\alpha})$ concludes the proof. ∎