Online Covariance Matrix Estimation in Sketched Newton Methods

Wei Kuang[1], Sen Na[2], and Mihai Anitescu[1,3]

[1]Department of Statistics, The University of Chicago
[2]School of Industrial and Systems Engineering, Georgia Institute of Technology
[3]Mathematics and Computer Science Division, Argonne National Laboratory

Abstract

Given the ubiquity of streaming data, online algorithms have been widely used for parameter estimation, with second-order methods particularly standing out for their efficiency and robustness. In this paper, we study an online sketched Newton method that leverages a randomized sketching technique to perform an approximate Newton step in each iteration, thereby eliminating the computational bottleneck of second-order methods. While existing studies have established the asymptotic normality of sketched Newton methods, a consistent estimator of the limiting covariance matrix remains an open problem. We propose a fully online covariance matrix estimator that is constructed entirely from the Newton iterates and requires no matrix factorization. Compared to covariance estimators for first-order online methods, our estimator for second-order methods is *batch-free*. We establish the consistency and convergence rate of our estimator, and coupled with asymptotic normality results, we can then perform online statistical inference for the model parameters based on sketched Newton methods. We also discuss the extension of our estimator to constrained problems, and demonstrate its superior performance on regression problems as well as benchmark problems in the CUTEst set.

1 Introduction

We consider the following stochastic optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) = \mathbb{E}_{\mathcal{P}}[f(\boldsymbol{x};\xi)], \tag{1}$$

where $F:\mathbb{R}^d\rightarrow\mathbb{R}$ is a stochastic, strongly convex objective function, $f(\cdot;\xi)$ is its noisy observation, and $\xi\sim\mathcal{P}$ is a random variable. Problems of form (1) appear in various decision-making applications in statistics and data science, including online recommendation (Li et al., 2010), precision medicine (Kosorok and Laber, 2019), energy control (Wallace and Ziemba, 2005), portfolio allocation (Fan et al., 2012), and e-commerce (Chen et al., 2022). In these applications, (1) is often interpreted as a model parameter estimation problem, where $\boldsymbol{x}$ denotes the model parameter and ξ denotes a random data sample. The true model parameter $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x})$ is the minimizer of the expected population loss F.

The classic offline approach to solving (1) is *sample average approximation* or M-estimation, which generates t i.i.d. samples $\xi_1,\ldots,\xi_t\sim\mathcal{P}$ and approximates the population loss F by the empirical loss:

$$\widehat{\boldsymbol{x}}_t = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d}\left\{\widehat{F}_t(\boldsymbol{x}) \coloneqq \frac{1}{t}\sum_{i=1}^{t} f(\boldsymbol{x};\xi_i)\right\}. \tag{2}$$

The statistical properties, e.g., $\sqrt{t}$-consistency and asymptotic normality, of M-estimators $\widehat{\boldsymbol{x}}_t$ are well-known in the literature (Vaart, 1998; Hastie et al., 2009), and numerous deterministic optimization methods can be applied to solve Problem (2), such as gradient descent and Newton's method (Boyd and Vandenberghe, 2004). However, deterministic methods are not appealing for large datasets due to their significant computation and memory costs. In contrast, online methods via *stochastic approximation* have recently attracted much attention. These methods efficiently process each sample once received and then discard, making them well-suited for modern streaming data. Thus, it is particularly critical to quantify the uncertainty of online methods and leverage the methods to perform *online statistical inference* for model parameters.

One of the most fundamental online methods is stochastic gradient descent (SGD) (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952), which takes the form

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \nabla f(\boldsymbol{x}_t; \xi_t), \quad t \geq 1.$$

There exists a long sequence of literature that quantifies the uncertainty of SGD and its many variants. Early works established almost sure convergence and asymptotic normality results of SGD in restricted settings (Sacks, 1958; Fabian, 1968; Robbins and Siegmund, 1971; Fabian, 1973; Ljung, 1977; Ermoliev, 1983; Lai, 2003). Later on, Ruppert (1988); Polyak and Juditsky (1992) proposed averaging SGD iterates as $\bar{\boldsymbol{x}}_t = \sum_{i=1}^{t} \boldsymbol{x}_i / t$ and established generic asymptotic normality results for $\bar{\boldsymbol{x}}_t$. This seminal asymptotic study has then been generalized to other gradient-based methods, including implicit SGD (Toulis et al., 2014; Toulis and Airoldi, 2017), constant-stepsize SGD (Li et al., 2018; Mou et al., 2020), moment-adjusted SGD (Liang and Su, 2019), momentum-accelerated SGD (Tang et al., 2023), and projected SGD (Duchi and Ruan, 2021; Davis et al., 2023). Additionally, studies under non-i.i.d. settings have also been reported (Chen et al., 2020a; Liu et al., 2023b; Li et al., 2023).

With the asymptotic normality result for the averaged iterate $\bar{\boldsymbol{x}}_t$ (see (18) for the definition of $\Omega^\star$):

$$\sqrt{t}(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(0, \Omega^\star), \tag{3}$$

estimating the limiting covariance matrix $\Omega^\star$ is the crucial next step to perform online statistical inference. While some inferential procedures may bypass the need for this estimation, such as bootstrapping (Fang et al., 2018; Liu et al., 2023a; Zhong et al., 2023; Lam and Wang, 2023) and random scaling (Li et al., 2021; Lee et al., 2022), many works have focused on (3) and proposed different online covariance matrix estimators for their simplicity and directness. In particular, Chen et al. (2020b) proposed two estimators: a plug-in estimator and a batch-means estimator. Compared to the plug-in estimator, which averages the estimated objective Hessians and then computes its inverse — resulting in significant computational costs — the batch-means estimator is obtained simply through the SGD iterates. Chen et al. (2020b) investigated the choice of batch sizes given a fixed total sample size, while Zhu et al. (2021) refined that estimator by not requiring the total sample size being fixed in advance. The two aforementioned works utilized increasing batch sizes, which has been relaxed to equal batch sizes recently by Zhu and Dong (2021) and Singh et al. (2023). Combining the asymptotic normality with covariance estimation, we can then construct online confidence intervals for model parameters $\boldsymbol{x}^\star$ based on SGD iterates.

Along with SGD, stochastic Newton methods multiply the gradient direction by a Hessian inverse to incorporate the objective's curvature information, leading to improved and more robust performance, particularly when dealing with Hessian matrices that have eigenvalues on significantly different scales (Byrd et al., 2016; Kovalev et al., 2019; Bercu et al., 2020). The online updating scheme takes the form:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha_t \Delta \boldsymbol{x}_t \qquad \text{with} \qquad B_t \Delta \boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t; \xi_t), \tag{4}$$

where $B_t \approx \nabla^2 F(\boldsymbol{x}_t)$ is an estimate of the objective Hessian. A growing body of literature focuses on performing (online) statistical inference based on (4). Leluc and Portier (2023) considered B_t as a general preconditioning matrix and established the asymptotic normality for the *last* iterate $\boldsymbol{x}_t$ assuming the convergence of B_t. The authors showed that $\boldsymbol{x}_t$ achieves asymptotic efficiency (i.e., minimal covariance) when $B_t \rightarrow \nabla^2 F(\boldsymbol{x}^\star)$, corresponding to online Newton methods. Bercu et al. (2020) developed an online Newton method for logistic regression and established similar asymptotic normality for $\boldsymbol{x}_t$. Cénac et al. (2020) and Boyer and Godichon-Baggioni (2022) expanded that approach to more general regression problems and investigated statistical inference on weighted Newton iterates $\bar{\boldsymbol{x}}_t = \sum_{i=1}^t \boldsymbol{x}_i / t$. The above studies revolved around regression problems where the estimated Hessian B_t can be expressed as an average of rank-one matrices, allowing its inverse B_t^{-1} to be updated online by the Sherman-Morrison formula (Sherman and Morrison, 1950). However, computing the inverse of a general estimated Hessian can be computationally demanding, with an $O(d^3)$ time complexity.

To address the above computational bottleneck, Na and Mahoney (2022) introduced an *online sketched Newton method* that leverages a randomized sketching technique to approximately solve the Newton system (4), without requiring the approximation error to vanish. Specifically, the time complexity can be reduced to $O(\texttt{nnz(S)}d)$, where $S \in \mathbb{R}^{d \times q}$ is the sketching matrix with $q \ll d$. For instance, when S is a sparse sketching vector, the time complexity is $O(d)$. Na and Mahoney (2022) quantified the uncertainty of both sampling and sketching and established the asymptotic normality for the last iterate $\boldsymbol{x}_t$ of the sketched Newton method (see (19) for the definition of $\Xi^\star$):

$$1/\sqrt{\alpha_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(0, \Xi^\star), \tag{5}$$

where the limiting covariance $\Xi^\star \neq \Omega^\star$ depends on the underlying sketching distribution in a complex manner. Due to the challenges of estimating the sketching components in $\Xi^\star$, the authors proposed a plug-in estimator for $\Omega^\star$ instead. That estimator raises two major concerns. First, the plug-in estimator is generally not asymptotically consistent, although the bias is controlled by the approximation error. It is only consistent when solving the Newton system exactly (so that the approximation error is zero). The bias significantly compromises the performance of statistical inference. Second, the plug-in estimator involves the inversion of the estimated Hessian, leading to an $O(d^3)$ time complexity that contradicts the spirit of using sketching solvers.

Motivated by the limitations of plug-in estimators and the success of batch-means estimators in first-order methods, we propose a novel *weighted sample covariance* estimator for $\Xi^\star$. Our estimator is constructed entirely from the sketched Newton iterates with varying weights, and does not involve any matrix inversion, making it computationally efficient. Additionally, our estimator has a simple recursive form, aligning well with the online nature of the method. Unlike batch-means estimators in first-order methods, our estimator is *batch-free*. We establish the consistency and convergence rate of our estimator, and coupled with the asymptotic normality in (5), we can then construct asymptotically valid confidence intervals for the true model parameters $\boldsymbol{x}^\star$ based on the Newton iterates $\{\boldsymbol{x}_t\}$. The challenge in our analysis lies in quantifying multiple sources of randomness (sampling, sketching, and adaptive stepsize introduced later); all of them affect the asymptotic behavior of online Newton methods. We emphasize that our analysis naturally holds for degenerate designs where the Newton systems are exactly solved and/or the stepsizes are deterministic. To our knowledge, the proposed estimator is the first online construction of a consistent limiting covariance matrix estimator for online second-order methods. We demonstrate its superior empirical performance through extensive experiments on regression problems and benchmark problems from the CUTEst test set.

Structure of the paper: We introduce the online sketched Newton method in Section 2, and present assumptions and some preliminary theoretical results in Section 3. In Section 4, we introduce the weighted sample covariance matrix estimator and present its theoretical guarantees. The numerical experiments are provided in Section 5, followed by conclusions and future work in Section 6.

Notation: Throughout the paper, we use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the spectral norm for matrices, $\|\cdot\|_F$ to denote the Frobenius norm for matrices, and $\mathrm{Tr}(\cdot)$ to denote the trace of a matrix. We use $O(\cdot)$ and $o(\cdot)$ to denote the big and small O notation in the usual sense. In particular, for two positive sequences $\{a_t, b_t\}$, $a_t = O(b_t)$ (also denoted as $a_t \lesssim b_t$) if $a_t \leq cb_t$ for a positive constant c and all large enough t. Analogously, $a_t = o(b_t)$ if $a_t/b_t \to 0$ as $t \to \infty$. For a random variable sequence $\{X_t\}$, $X_t = O_p(a_t)$ indicates that X_t/a_t is stochastically bounded. For two scalars a and b, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. We let I denote the identity matrix, $\mathbf{0}$ denote zero vector or matrix, and $\boldsymbol{e}_i$ denote the vector with i-th entry being 1 and 0 otherwise; their dimensions are clear from the context. For a sequence of compatible matrices $\{A_i\}$, $\prod_{k=i}^{j} A_k = A_j A_{j-1} \cdots A_i$ if $j \geq i$ and I if $j < i$. For a matrix A, $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) represents the smallest (largest) eigenvalue of A. We also let $F_t = F(\boldsymbol{x}_t)$ and $F^\star = F(\boldsymbol{x}^\star)$ (similar for $\nabla F_t, \nabla^2 F_t$, etc.), and let $\mathbf{1}_{\{\cdot\}}$ denote the indicator function.

2 Online Sketched Newton Method

At a high level, the online sketched Newton method takes the following update scheme:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \bar{\alpha}_t \bar{\Delta} \boldsymbol{x}_t, \tag{6}$$

where $\bar{\Delta}\boldsymbol{x}_t$ *approximately* solves the Newton system $B_t \Delta \boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t; \xi_t)$ via the sketching solver (see (10)) and $\bar{\alpha}_t$ is an *adaptive*, *potentially random* stepsize (see (11)).

More precisely, given the current iterate $\boldsymbol{x}_t$, we randomly generate a sample $\xi_t \sim \mathcal{P}$ and obtain the gradient and Hessian estimates:

$$\bar{g}_t = \nabla f(\boldsymbol{x}_t; \xi_t) \qquad \text{and} \qquad \bar{H}_t = \nabla^2 f(\boldsymbol{x}_t; \xi_t).$$

Then, we define B_t to be the Hessian average using samples $\{\xi_i\}_{i=0}^{t-1}$, expressed as

$$B_t = \frac{1}{t}\sum_{i=0}^{t-1} \bar{H}_i \quad \overset{\text{online update}}{\Longrightarrow} \quad B_t = \frac{t-1}{t} B_{t-1} + \frac{1}{t}\bar{H}_{t-1}. \tag{7}$$

In this paper, we use $\bar{(\cdot)}$ to denote a random quantity that depends on the current sample ξ_t. Note that the estimate $\bar{H}_t$ is only used in the $(t+1)$-th iteration; thus B_t is deterministic conditional on $\boldsymbol{x}_t$ (this is why we do not use the notation $\bar{B}_t$). The Hessian average is widely used in Newton methods to accelerate the convergence rate (Na et al., 2022). In certain problems, B_t can be expressed as the sum of rank-1 matrices, allowing its inverse to be updated online in a manner similar to (7) (Bercu et al., 2020; Cénac et al., 2020; Boyer and Godichon-Baggioni, 2022; Leluc and Portier, 2023). However, solving the Newton system $B_t \Delta \boldsymbol{x}_t = -\bar{g}_t$ for a generic stochastic function can be expensive.

We now employ the sketching solver to approximately solve $B_t \Delta \boldsymbol{x}_t = -\bar{g}_t$. At each inner iteration j, we generate a sketching matrix/vector $S_{t,j} \in \mathbb{R}^{d \times q} \sim S$ for some $q \geq 1$ and solve the subproblem:

$$\Delta \boldsymbol{x}_{t,j+1} = \underset{\Delta \boldsymbol{x}}{\operatorname{argmin}} \|\Delta \boldsymbol{x} - \Delta \boldsymbol{x}_{t,j}\|^2, \qquad \text{s.t.} \qquad S_{t,j}^T B_t \Delta \boldsymbol{x} = -S_{t,j}^T \bar{g}_t. \tag{8}$$

In particular, we only aim to solve the sketched Newton system $S_{t,j}^T B_t \Delta\boldsymbol{x} = -S_{t,j}^T \bar{g}_t$ at the j-th inner iteration, but we prefer the solution that is as close as possible to the current solution approximation $\Delta\boldsymbol{x}_{t,j}$. The closed-form recursion of (8) is ($\Delta\boldsymbol{x}_{t,0} = \boldsymbol{0}$):

$$\Delta\boldsymbol{x}_{t,j+1} = \Delta\boldsymbol{x}_{t,j} - B_t S_{t,j}(S_{t,j}^T B_t^2 S_{t,j})^{\dagger} S_{t,j}^T (B_t \Delta\boldsymbol{x}_{t,j} + \bar{g}_t), \tag{9}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudoinverse. When we employ a sketching vector ($q = 1$), the pseudoinverse reduces to the reciprocal, meaning that solving the Newton system is *matrix-free*—no matrix factorization is needed. We refer to Strohmer and Vershynin (2008); Gower and Richtárik (2015) for specific examples of choosing dense/sparse sketching matrices and their trade-offs.

For a deterministic integer τ, we let

$$\bar{\Delta}\boldsymbol{x}_t = \Delta\boldsymbol{x}_{t,\tau}, \tag{10}$$

and then we update the iterate $\boldsymbol{x}_t$ as in (6) with a potentially random stepsize $\bar{\alpha}_t$ satisfying

$$\beta_t \leq \bar{\alpha}_t \leq \beta_t + \chi_t \qquad \text{with} \qquad \beta_t = \frac{c_\beta}{(t+1)^\beta} \quad \text{and} \quad \chi_t = \frac{c_\chi}{(t+1)^\chi}. \tag{11}$$

The motivation for using a well-controlled random stepsize is to enhance the adaptivity of the method without compromising the asymptotic normality guarantee. Particularly, different directions may prefer different stepsizes, so that $\bar{\alpha}_t$ depends on $\bar{\Delta}\boldsymbol{x}_t$ and is random. Berahas et al. (2021, 2023); Curtis et al. (2024) have proposed various adaptive stepsize selection schemes for Newton methods on constrained problems that precisely satisfy the condition in (11).

3 Assumptions and Asymptotic Normality

In this section, we introduce assumptions and present the asymptotic normality guarantee for sketched Newton methods. Throughout the paper, we let $\mathcal{F}_t = \sigma(\{\xi_i\}_{i=0}^t)$, for any $t \geq 0$, be the filtration of σ-algebras generated by the sample sequence $\xi_0, \xi_1, \xi_2 \ldots$.

3.1 Assumptions

We first impose a Lipschitz continuity condition on the objective Hessian $\nabla^2 F(\boldsymbol{x})$, which is standard in existing literature (Bercu et al., 2020; Cénac et al., 2020; Na and Mahoney, 2022).

Assumption 3.1. We assume $F(\boldsymbol{x})$ is twice continuously differentiable and its Hessian $\nabla^2 F(\boldsymbol{x})$ is Υ_L-Lipschitz continuous. In particular, for any $\boldsymbol{x}$ and $\boldsymbol{x}'$, we have

$$\|\nabla^2 F(\boldsymbol{x}) - \nabla^2 F(\boldsymbol{x}')\| \leq \Upsilon_L \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

The next assumption regards the noise in stochastic gradients. We assume that the fourth conditional moment of the gradient noise satisfies a growth condition. This assumption aligns with existing literature on covariance estimation for SGD (Chen et al., 2020b; Zhu et al., 2021).

Assumption 3.2. We assume the function $f(\boldsymbol{x};\xi)$ is twice continuously differentiable with respect to $\boldsymbol{x}$ for any ξ, and $\|\nabla f(\boldsymbol{x};\xi)\|$ is uniformly integrable for any $\boldsymbol{x}$. This implies $\mathbb{E}[\bar{g}_t \mid \mathcal{F}_{t-1}] = \nabla F_t$. Furthermore, there exist constants $C_{g,1}, C_{g,2} > 0$ such that

$$\mathbb{E}[\|\bar{g}_t - \nabla F_t\|^4 \mid \mathcal{F}_{t-1}] \leq C_{g,1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4 + C_{g,2}, \qquad \forall t \geq 0. \tag{12}$$

The above growth condition is weaker than the bounded moment condition $\mathbb{E}[\|\bar{g}_t - \nabla F_t\|^4 \mid \mathcal{F}_{t-1}] \leq C$ assumed for the plug-in estimator in Na and Mahoney (2022). In fact, the bounded fourth moment can be relaxed to a bounded $(2+\epsilon)$-moment for establishing asymptotic normality of SGD and Newton methods, but it is widely imposed for limiting covariance estimation.

The next assumption imposes lower and upper bounds for stochastic Hessians, with a growth condition on the Hessian noise.

Assumption 3.3. There exist constants $\Upsilon_H > \gamma_H > 0$ such that for any ξ and any $\boldsymbol{x}$,

$$\gamma_H \leq \lambda_{\min}(\nabla^2 f(\boldsymbol{x};\xi)) \leq \lambda_{\max}(\nabla^2 f(\boldsymbol{x};\xi)) \leq \Upsilon_H, \tag{13}$$

which implies $\mathbb{E}[\bar{H}_t \mid \mathcal{F}_{t-1}] = \nabla^2 F_t$. Furthermore, there exist constants $C_{H,1}, C_{H,2} > 0$ such that

$$\mathbb{E}\left[\|\bar{H}_t - \nabla^2 F_t\|^4 \mid \mathcal{F}_{t-1}\right] \leq C_{H,1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4 + C_{H,2}, \qquad \forall t \geq 0. \tag{14}$$

The condition (13) is widely used in the literature on stochastic second-order methods (Byrd et al., 2016; Berahas et al., 2016; Moritz et al., 2016). By the averaging structure of B_t, we know (13) implies

$$\gamma_H \leq \lambda_{\min}(B_t) \leq \lambda_{\max}(B_t) \leq \Upsilon_H. \tag{15}$$

Moreover, as shown in (Chen et al., 2020b, Lemma 3.1), the condition (13) together with the bounded $\mathbb{E}[\|\nabla f(\boldsymbol{x}^\star;\xi)\|^4]$ implies (12). The growth condition on the Hessian noise in (14) is analogous to that on the gradient noise in (12).

We finally require the following assumption regarding the sketching distribution.

Assumption 3.4. For $t \geq 0$, we assume the sketching matrix $S_{t,j} \overset{iid}{\sim} S$ satisfies $\mathbb{E}[B_t S(S^T B_t^2 S)^\dagger S^T B_t \mid \mathcal{F}_{t-1}] \succeq \gamma_S I$ and $\mathbb{E}[\|S\|^2\|S^\dagger\|^2] \leq \Upsilon_S$ for some constants $\gamma_S, \Upsilon_S > 0$.

The above two expectations are taken over the randomness of the sketching matrix S. The lower bound of the projection matrix $B_t S(S^T B_t^2 S)^\dagger S^T B_t$ is commonly required by sketching solvers to ensure convergence (Gower and Richtárik, 2015). We trivially have $\gamma_S \leq 1$. The bounded second moment of the condition number of S is necessary to analyze the difference between two projection matrices, $\|B_t S(S^T B_t^2 S)^\dagger S^T B_t - B^\star S(S^T (B^\star)^2 S)^\dagger S^T B^\star\|$ (Na and Mahoney, 2022, Lemma 5.2). Under (15), both conditions easily hold for various sketching distributions, such as Gaussian sketching $S \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and Uniform sketching $S \sim \text{Unif}(\{\boldsymbol{e}_i\}_{i=1}^d)$, where $\boldsymbol{e}_i$ is the i-th canonical basis of $\mathbb{R}^d$ (called randomized Kaczmarz method (Strohmer and Vershynin, 2008)).

3.2 Almost sure convergence and asymptotic normality

We review the almost sure convergence and asymptotic normality of the sketched Newton method as established in (Na and Mahoney, 2022, Theorems 4.7 and 5.6). We emphasize that our growth conditions on the gradient and Hessian noises are weaker than those assumed in Na and Mahoney (2022), and by a sharper analysis, our assumption on χ_t is relaxed from $\chi > 1$ to $\chi > 0.5(\beta+1)$. We show that all their results still hold, with proofs deferred to Appendix D for completeness.

Theorem 3.5 (Almost sure convergence)**.** Consider the iteration scheme (6). Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \geq \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1-\gamma_S$, and the stepsize parameters satisfy $\beta \in (0.5, 1]$, $\chi > 0.5(\beta+1)$, and $c_\beta, c_\chi > 0$. Then, we have $\boldsymbol{x}_t \to \boldsymbol{x}^\star$ as $t \to \infty$ almost surely.

To present the normality result, we need to introduce some additional notation. Let $B^\star = \nabla^2 F(\boldsymbol{x}^\star)$. For $S_1, \dots, S_\tau \overset{iid}{\sim} S$, we define the product of τ projection matrices as

$$\widetilde{C}^\star = \prod_{j=1}^{\tau} (I - B^\star S_j (S_j^T (B^\star)^2 S_j)^\dagger S_j^T B^\star) \tag{16}$$

and let $C^\star = \mathbb{E}[\widetilde{C}^\star]$. Then, we denote the eigenvalue decomposition of $I - C^\star$ as

$$I - C^\star = U\Sigma U^T \qquad \text{with} \qquad \Sigma = \text{diag}(\sigma_1, \dots, \sigma_d). \tag{17}$$

We also define

$$\Omega^\star = (B^\star)^{-1}\mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)](B^\star)^{-1}. \tag{18}$$

With the above notation, we have the following normality guarantee for the scheme (6).

Theorem 3.6 (Asymptotic normality)**.** Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \geq \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1 - \gamma_S$, and the stepsize parameters satisfy $\beta \in (0.5, 1]$, $\chi > 1.5\beta$, and $c_\beta > 1/\{1.5(1-\rho^\tau)\}$ for $\beta = 1$. Then, we have

$$\sqrt{1/\bar{\alpha}_t}(\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Xi^\star),$$

where $\Xi^\star$ is the solution to the following Lyapunov equation:

$$\left(\left\{1 - \frac{\mathbf{1}_{\{\beta=1\}}}{2c_\beta}\right\} I - C^\star\right)\Xi^\star + \Xi^\star\left(\left\{1 - \frac{\mathbf{1}_{\{\beta=1\}}}{2c_\beta}\right\} I - C^\star\right) = \mathbb{E}\big[(I - \widetilde{C}^\star)\Omega^\star(I - \widetilde{C}^\star)^T\big]. \tag{19}$$

In fact, the limiting covariance $\Xi^\star$ has an explicit form as:

$$\Xi^\star = U\big(\Theta \circ U^T\mathbb{E}[(I - \widetilde{C}^*)\Omega^\star(I - \widetilde{C}^*)^T]U\big)U^T \quad \text{with} \quad [\Theta]_{k,l} = \frac{1}{\sigma_k + \sigma_l - \mathbf{1}_{\{\beta=1\}}/c_\beta}, \tag{20}$$

where $\circ$ denotes the matrix Hadamard product. There exists a degenerate case. When the Newton systems are exactly solved ($\tau = \infty$), then $\widetilde{C}^\star = C^\star = \mathbf{0}$, $\Sigma = I$, and $\Xi^\star = \Omega^\star/(2 - \mathbf{1}_{\{\beta=1\}}/c_\beta)$. In this case, we have $\Xi^\star = \Omega^\star/2$ for $\beta \in (0.5, 1)$ and $\Xi^\star = \Omega^\star$ for $\beta = c_\beta = 1$. For the latter setup, we know $\Xi^\star = \Omega^\star$ achieves the asymptotic minimax lower bound (Duchi and Ruan, 2021).

4 Online Covariance Matrix Estimation

In this section, we build upon the results in Section 3 and construct an online estimator for the limiting covariance matrix $\Xi^\star$. With the covariance estimator, we are then able to perform online statistical inference, such as constructing asymptotically valid confidence intervals for model parameters.

4.1 Weighted sample covariance estimator

Let $\varphi_t = \beta_t + \chi_t/2$ be the centered stepsize. Our weighted sample covariance estimator is defined as

$$\widehat{\Xi}_t = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_t)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_t)^T \qquad \text{with} \qquad \bar{\boldsymbol{x}}_t = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{x}_i. \tag{21}$$

Algorithm 1 Construction of Weighted Sample Covariance Estimator

1: **Input:** initial iterate $\boldsymbol{x}_0$, positive sequences $\{\beta_t, \chi_t\}$, an integer $\tau > 0$, $B_0 = I$;
2: **Initialize:** $W_0 = \mathbf{0} \in \mathbb{R}^{d\times d}$, $\boldsymbol{v}_0 = \bar{\boldsymbol{x}}_0 = \mathbf{0} \in \mathbb{R}^d$, $a_0 = 0$
3: **for** $t = 0, 1, 2, \dots$ **do**
4: Obtain the sketched Newton iterate $\boldsymbol{x}_{t+1}$ and let $\varphi_t = \beta_t + \chi_t/2$;
5: Update the quantities as:

$$W_{t+1} = \frac{t}{t+1}W_t + \frac{1/\varphi_t}{(t+1)}\,\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^T, \qquad \boldsymbol{v}_{t+1} = \frac{t}{t+1}\boldsymbol{v}_t + \frac{1/\varphi_t}{t+1}\,\boldsymbol{x}_{t+1};$$
$$\bar{\boldsymbol{x}}_{t+1} = \frac{t}{t+1}\bar{\boldsymbol{x}}_t + \frac{1}{t+1}\boldsymbol{x}_{t+1}, \qquad a_{t+1} = \frac{t}{t+1}a_t + \frac{1/\varphi_t}{t+1};$$

6: **end for**
7: **Output:** Covariance estimator $\widehat{\Xi}_t = W_t - \boldsymbol{v}_t\bar{\boldsymbol{x}}_t^T - \bar{\boldsymbol{x}}_t\boldsymbol{v}_t^T + a_t\bar{\boldsymbol{x}}_t\bar{\boldsymbol{x}}_t^T$.

This estimator can be rewritten as

$$\widehat{\Xi}_t = W_t - \boldsymbol{v}_t\bar{\boldsymbol{x}}_t^T - \bar{\boldsymbol{x}}_t\boldsymbol{v}_t^T + a_t\bar{\boldsymbol{x}}_t\bar{\boldsymbol{x}}_t^T,$$

where

$$W_t = \frac{1}{t}\sum_{i=1}^t \frac{1}{\varphi_{i-1}}\boldsymbol{x}_i\boldsymbol{x}_i^T, \qquad \boldsymbol{v}_t = \frac{1}{t}\sum_{i=1}^t \frac{1}{\varphi_{i-1}}\boldsymbol{x}_i, \qquad a_t = \frac{1}{t}\sum_{i=1}^t \frac{1}{\varphi_{i-1}}. \tag{22}$$

We mention that $W_t, \boldsymbol{v}_t, \bar{\boldsymbol{x}}_t, a_t$ can all be updated recursively, meaning that $\widehat{\Xi}_t$ can be computed in a fully online fashion. The detailed steps are shown in Algorithm 1.

We note that the estimator $\widehat{\Xi}_t$ is in a similar flavor to batch-means covariance estimators designed for first-order online methods. In particular, Chen et al. (2020b); Zhu et al. (2021) grouped SGD iterates into multiple batches and estimated the covariance $\Omega^\star$ in (3) by computing the sample covariance among batches. However, a significant difference from those batch-means estimators is that our estimator $\widehat{\Xi}_t$ is *batch-free*. Specifically, batch-means estimators rely on additional batch size sequences that must satisfy certain conditions and largely affect both the theoretical and empirical performance of the estimators. In contrast, we assign proper weights to the iterates based on the stepsizes, eliminating the need to tune any extra parameters beyond those required by the online method itself. That being said, we observe that the memory and computational complexities of inference based on sketched Newton methods are comparable to those of first-order methods. The memory complexity is dominated by storing B_t and W_t, incurring a cost of $O(d^2)$, independent of the sample size t. The computational complexity includes $O(\tau\cdot\texttt{nnz(S)}\,d)$ flops for computing the sketched Newton direction and $O(d^2)$ flops for updating $\widehat{\Xi}_t$. For instance, when $S \sim \text{Unif}(\{\boldsymbol{e}_i\}_{i=1}^d)$, Na and Mahoney (2022) showed $\tau = O(d)$, suggesting that the overall computational complexity of sketched Newton inference is $O(d^2)$. This order precisely matches the complexity in Chen et al. (2020b); Zhu et al. (2021).

Remark 4.1. We compare $\widehat{\Xi}_t$ with the plug-in estimator proposed in Na and Mahoney (2022). Due to significant challenges in estimating sketch-related quantities in (19), Na and Mahoney (2022) simply neglected all those quantities and estimated $\Omega^\star/(2-\mathbf{1}_{\{\beta=1\}}/c_\beta)$ instead. Their plug-in estimator is defined as:

$$\widetilde{\Xi}_t = \frac{1}{2-\mathbf{1}_{\{\beta=1\}}/c_\beta}\cdot B_t^{-1}\Big(\frac{1}{t}\sum_{i=1}^t \bar{g}_i\bar{g}_i^T\Big)B_t^{-1}. \tag{23}$$

Comparing $\widetilde{\Xi}_t$ with $\widehat{\Xi}_t$, we clearly see that $\widetilde{\Xi}_t$ is not matrix-free as it involves the inverse of B_t (i.e., $O(d^3)$ flops), which contradicts the spirit of using sketching solvers. Furthermore, $\widetilde{\Xi}_t$ is a biased estimator of $\Xi^\star$, leading to invalid confidence coverage even as $t \to \infty$.

4.2 Convergence rate of the estimator

To establish the convergence rate of $\widehat{\Xi}_t$, we first present a preparation result that provides error bounds for the Newton iterate $\boldsymbol{x}_t$ and the averaged Hessian B_t. We show that the fourth moments of $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$ and $\|B_t - B^\star\|$ scale as $O(\beta_t^2 + \chi_t^4/\beta_t^4)$. When $\chi_t \gtrsim \beta_t^{1.5}$, the error χ_t^4/β_t^4 incurred by the adaptivity of stepsize dominates. In contrast, when $\chi_t \lesssim \beta_t^{1.5}$, adaptive stepsizes lead only to a higher-order error. A matching error bound (for the iterate $\boldsymbol{x}_t$) has been established for SGD methods with $\chi_t = 0$ (Chen et al., 2020b). That said, our analysis is more involved due to higher-order methods, sketching components, and randomness in stepsizes.

Lemma 4.2 (Error bounds of $\boldsymbol{x}_t$ and B_t)**.** Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \geq \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1-\gamma_S$, and the stepsize parameters satisfy $\beta \in (0,1)$, $\chi > \beta$, and $c_\beta, c_\chi > 0$. Then, we have

$$\mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4\big] \lesssim \beta_t^2 + \frac{\chi_t^4}{\beta_t^4} \qquad \text{and} \qquad \mathbb{E}\big[\|B_t - B^\star\|^4\big] \lesssim \beta_t^2 + \frac{\chi_t^4}{\beta_t^4}.$$

With the above lemma, we show the convergence rate of $\widehat{\Xi}_t$ in the following theorem.

Theorem 4.3. Under the conditions in Lemma 4.2, except for strengthening $\chi > \beta$ to $\chi > 1.5\beta$, the covariance estimator $\widehat{\Xi}_t$ defined in (21) satisfies

$$\mathbb{E}\left[\left\|\widehat{\Xi}_t - \Xi^\star\right\|\right] \lesssim \begin{cases} \sqrt{\beta_t} + \chi_t/\beta_t^{1.5} & \text{for } 0 < \beta \leq 0.5, \\ 1/\sqrt{t\beta_t} + \chi_t/\beta_t^{1.5} & \text{for } 0.5 < \beta < 1. \end{cases}$$

Since $\sqrt{\beta_t} \vee \chi_t/\beta_t^{1.5} \to 0$ and $t\beta_t \to \infty$ as $t \to \infty$ (because $\chi > 1.5\beta$), Theorem 4.3 states that $\widehat{\Xi}_t$ is an (asymptotically) consistent estimator of $\Xi^\star$. Note that $\chi > 1.5\beta$ is already required by the asymptotic normality guarantee (cf. Theorem 3.6). Similar to Lemma 4.2, $\chi \geq 2\beta$ (i.e., $\chi_t \lesssim \beta_t^2$) makes the adaptivity error $\chi_t/\beta_t^{1.5}$ higher order.

When we suppress the sketching solver, the limiting covariance $\Xi^\star = \Omega^\star/2$, meaning that $2\widehat{\Xi}_t$ is a consistent estimator of $\Omega^\star$. Notably, this result suggests that we can estimate the optimal covariance matrix $\Omega^\star$ without grouping the iterates, computing the batch means, and tuning batch size sequences, which significantly differs and simplifies the estimation procedure in first-order methods. This advantage is indeed achieved by leveraging Hessian estimates; however, we preserve the computation and memory costs as low as those of first-order methods. We defer a comprehensive discussion of Theorem 4.3 to Section 4.3.

Theorem 4.3 immediately implies the following corollary, which demonstrates the construction of confidence intervals/regions.

Corollary 4.4. Let us set the coverage probability as $1-q$ with $q \in (0,1)$. Consider performing the online scheme (6) and computing the covariance estimator (21). Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \geq \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1-\gamma_S$, and the stepsize parameters satisfy $\beta \in (0.5,1)$, $\chi > 1.5\beta$, and $c_\beta, c_\chi > 0$. Then, we have

$$P\left(\boldsymbol{x}^\star \in \mathcal{E}_{t,q}\right) \to 1-q \qquad \text{as} \qquad t \to \infty,$$

where $\mathcal{E}_{t,q} = \{\boldsymbol{x} \in \mathbb{R}^d : (\boldsymbol{x}-\boldsymbol{x}_t)^T \widehat{\Xi}_t^{-1}(\boldsymbol{x}-\boldsymbol{x}_t)/\bar{\alpha}_t \leq \chi^2_{d,1-q}\}$. Furthermore, for any direction $\boldsymbol{w} \in \mathbb{R}^d$,

$$P\left(\boldsymbol{w}^T\boldsymbol{x}^\star \in \left[\boldsymbol{w}^T\boldsymbol{x}_t \pm z_{1-q/2}\sqrt{\bar{\alpha}_t \cdot \boldsymbol{w}^T\widehat{\Xi}_t\boldsymbol{w}}\right]\right) \to 1-q \qquad \text{as} \quad t\to\infty.$$

Here, $\chi^2_{d,1-q}$ is the $(1-q)$-quantile of χ^2_d distribution, while $z_{1-q/2}$ is the $(1-q/2)$-quantile of standard Gaussian distribution.

We would like to emphasize that the above statistical inference procedure is fully online and matrix-free. In particular, $\boldsymbol{x}_t$ is updated with online nature; for confidence intervals, $\boldsymbol{w}^T\widehat{\Xi}_t\boldsymbol{w}$ is computed online as introduced in Section 4.1; for confidence region, $\widehat{\Xi}_t^{-1}$ can also be updated online:

$$\widehat{\Xi}_{t+1}^{-1} = \frac{t+1}{t}\widehat{\Xi}_t^{-1} - \frac{t+1}{t}\widehat{\Xi}_t^{-1}R_t\left(\Pi_t + R_t^T\widehat{\Xi}_t^{-1}R_t\right)^{-1}R_t^T\widehat{\Xi}_t^{-1},$$

where $R_t = (\boldsymbol{v}_t - a_t\bar{\boldsymbol{x}}_t; \bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}}_{t+1}; \boldsymbol{x}_{t+1} - \bar{\boldsymbol{x}}_{t+1}) \in \mathbb{R}^{d\times 3}$ and $\Pi_t = (a_t, 1, 0; 1, 0, 0; 0, 0, t\varphi_t) \in \mathbb{R}^{3\times 3}$. See Appendix A for the derivation of the above recursion.

4.3 Comparison and generalization of existing studies

In this section, we first compare our weighted sample covariance estimator $\widehat{\Xi}_t$ with other existing covariance estimators for both first- and second-order online methods. Then, we discuss the generalization of our estimator to other methods.

• **Plug-in estimator of online sketched Newton.** Recall from Remark 4.1 that, due to the challenges of estimating sketch-related quantities $\widetilde{C}^\star$ and $C^\star$ in (19), a recent work Na and Mahoney (2022) simply neglected these quantities and designed a plug-in covariance estimator $\widetilde{\Xi}_t$ in (23). In addition to concerns about excessive computation, (Na and Mahoney, 2022, Theorem 5.8) indicated that for $\beta \in (0.5, 1)$,

$$\|\widetilde{\Xi}_t - \Xi^\star\| = O(\sqrt{\beta_t\log(1/\beta_t)}) + O((1-\gamma_S)^\tau).$$

Here, the second term accounts for the oversight in estimating sketch-related quantities. It decays exponentially with the sketching steps τ but *does not vanish* for any finite τ. Thus, $\widetilde{\Xi}_t$ is not a consistent estimator of $\Xi^\star$. In the degenerate case where the Newton systems are solved exactly ($\tau = \infty$), $\widetilde{\Xi}_t$ converges to $\Xi^\star$ at a rate of $O(\sqrt{\beta_t\log(1/\beta_t)})$, which is faster than that of our estimator $\widehat{\Xi}_t$. In this case, choosing between $\widehat{\Xi}_t$ and $\widetilde{\Xi}_t$ involves a trade-off between faster convergence and computational efficiency. It is worth noting that the faster convergence of the plug-in estimator is anticipated (see Chen et al. (2020b) for a comparison of plug-in and batch-means estimators in SGD methods), since its convergence rate is fully tied to that of the iterates. In contrast, the convergence rate of our sample covariance is additionally confined by the slow decay of correlations among the iterates.

• **Batch-means estimator of SGD.** As introduced in Section 4.1, Chen et al. (2020b); Zhu et al. (2021) grouped SGD iterates into batches and estimated the limiting covariance $\Omega^\star$ by the sample covariance among batches (each batch mean is treated as one sample). Singh et al. (2023) further relaxed their conditions from increasing batch sizes to equal batch sizes. Compared to this type of estimators, our estimator $\widehat{\Xi}_t$ is batch-free, requiring no additional parameters beyond those of the algorithm itself. Furthermore, the aforementioned works all showed that the convergence rate of the batch-means estimators is $O(1/\sqrt[4]{t\beta_t})$, which is slower than that of $\widehat{\Xi}_t$ in Theorem 4.3. Intuitively, the batch-means estimators require a long batch of iterates to obtain a single sample, while our batch-free estimator treats each individual iterate as a single sample, making it more efficient in utilizing (correlated) iterates.

• **Generalization to conditioned SGD.** We point out that $\widehat{\Xi}_t$ can also serve as a consistent covariance estimator for conditioned SGD methods, which follow the update form (4) though B_t may not approximate the objective Hessian $\nabla^2 F_t$. Leluc and Portier (2023) established asymptotic normality for conditioned SGD methods under the assumption of convergence of the conditioning matrix B_t. These methods include AdaGrad (Duchi et al., 2011), RMSProp (Tieleman, 2012), and quasi-Newton methods (Byrd et al., 2016) as special cases. Notably, Theorem 4.3 does not requireB_t to converge to the true Hessian $\nabla^2 F(\boldsymbol{x}^\star)$, making our analysis directly applicable to conditioned SGD methods.

• **Generalization to sketched Sequential Quadratic Programming.** We consider a constrained stochastic optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) = \mathbb{E}_{\mathcal{P}}[f(\boldsymbol{x};\xi)] \quad \text{s.t.} \quad c(\boldsymbol{x}) = \mathbf{0},$$

where $F:\mathbb{R}^d\to\mathbb{R}$ is a stochastic objective with $f(\cdot;\xi)$ as the noisy observation, and $c:\mathbb{R}^d\to\mathbb{R}^m$ imposes deterministic constraints on the model parameters $\boldsymbol{x}$. Such problems appear widely in statistical machine learning, including constrained M-estimation and algorithmic fairness. Na and Mahoney (2022) designed an online sketched Sequential Quadratic Programming (SQP) method for solving the problem. Define $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = F(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$ as the Lagrangian function, where $\boldsymbol{\lambda}\in\mathbb{R}^m$ are the dual variables. The sketched SQP method can be regarded as applying the sketched Newton method to $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})$, leading to the update $(\boldsymbol{x}_{t+1},\boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t,\boldsymbol{\lambda}_t) + \bar{\alpha}_t(\bar{\Delta}\boldsymbol{x}_t,\bar{\Delta}\boldsymbol{\lambda}_t)$, where $(\bar{\Delta}\boldsymbol{x}_t,\bar{\Delta}\boldsymbol{\lambda}_t)$ is the sketched solution to the primal-dual Newton system:

$$\begin{pmatrix} B_t & G_t^T \\ G_t & \mathbf{0} \end{pmatrix}\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} = -\begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}.$$

Here, analogous to (4), $B_t \approx \nabla^2_{\boldsymbol{x}}\mathcal{L}_t$ is an estimate of the Lagrangian Hessian with respect to $\boldsymbol{x}$, $G_t = \nabla c_t \in \mathbb{R}^{m\times d}$ is the constraints Jacobian, and $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t = \nabla F(\boldsymbol{x}_t;\xi_t) + G_t^T\boldsymbol{\lambda}_t$ is the estimate of the Lagrangian gradient with respect to $\boldsymbol{x}$. Na and Mahoney (2022) established asymptotic normality for the SQP iterate $(\boldsymbol{x}_t,\boldsymbol{\lambda}_t)$. We observe that the constraints are not essential in the SQP analysis; therefore, our construction of $\widehat{\Xi}_t$ is naturally applied to the covariance estimation of the sketched SQP method. An empirical demonstration of $\widehat{\Xi}_t$ for constrained problems is presented in Section 5.3.

5 Numerical Experiment

In this section, we demonstrate the empirical performance of the weighted sample covariance matrix $\widehat{\Xi}_t$ on both regression problems and benchmark CUTEst problems (Gould et al., 2014). We compare $\widehat{\Xi}_t$ with two other online covariance estimators: the plug-in estimator $\widetilde{\Xi}_t$ in (23) (based on sketched Newton) and the batch-means estimator $\bar{\Xi}_t$ (based on SGD) (Zhu et al., 2021, Algorithm 2). We evaluate the performance of each estimator by both the (relative) covariance estimation error and the coverage rate of constructed confidence intervals. We defer some implementation details and results to Appendix F due to the space limit.

5.1 Linear regression

We consider the linear regression model:

$$\xi_b = \boldsymbol{\xi}_a^T\boldsymbol{x}^\star + \varepsilon,$$

where $\xi = (\boldsymbol{\xi}_a, \xi_b) \in \mathbb{R}^d \times \mathbb{R}$ is the feature-response vector and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise. For this model, we use the squared loss defined as $f(\boldsymbol{x}; \xi) = \frac{1}{2}(\xi_b - \boldsymbol{\xi}_a^T \boldsymbol{x})^2$. In our experiment, we apply Gaussian features $\boldsymbol{\xi}_a \sim \mathcal{N}(0, \Sigma_a)$ with different dimensions and covariance matrix Σ_a. In particular, we vary $d \in \{20, 40, 60, 100\}$, and for each d, we consider three types of covariance matrices. (i) Identity matrix: $\Sigma_a = I$. (ii) Toeplitz matrix: $[\Sigma_a]_{i,j} = r^{|i-j|}$ with $r \in \{0.4, 0.5, 0.6\}$. (iii) Equi-correlation matrix: $[\Sigma_a]_{i,i} = 1$ and $[\Sigma_a]_{i,j} = r$ for $i \neq j$, with $r \in \{0.1, 0.2, 0.3\}$. The true model parameter is set as $\boldsymbol{x}^\star = (1/d, \ldots, 1/d)^T \in \mathbb{R}^d$. This simulation setting follows directly from existing studies (Chen et al., 2020b; Zhu et al., 2021; Na and Mahoney, 2022).

For the batch-means estimator $\bar{\Xi}_t$, we adopt the setup in Zhu et al. (2021) by setting the stepsize of SGD as $\beta_t = t^{-\beta}/2$ and the batch size as $a_m = \lfloor m^{2/(1-\beta)} \rfloor$ (in their notation) with $\beta = 0.505$. For both the plug-in estimator $\tilde{\Xi}_t$ and our sample covariance estimator $\widehat{\Xi}_t$, we implement sketched Newton methods with varying sketching steps $\tau \in \{10, 20, 40, \infty\}$. When $\tau = \infty$, the scheme reduces to standard Newton method. We perform the Kaczmarz method, where the sketching distribution in (9) is $S \sim \mathrm{Unif}(\{\boldsymbol{e}_i\}_{i=1}^d)$ (cf. Section 3.1). We let $\beta_t = t^{-\beta}$ and $\bar{\alpha}_t \sim \mathrm{Unif}[\beta_t, \beta_t + \beta_t^2]$. For all estimators, we initialize the method at $\boldsymbol{x}_0 = \boldsymbol{0}$, run 2×10^5 iterations, and aim to construct confidence intervals for the averaged parameters $\sum_{i=1}^d \boldsymbol{x}_i^\star / d$. All the results are averaged over 200 independent runs.

Our results are summarized in Table 1. From the table 1, we observe that, in most cases, the empirical coverage rate of the confidence interval based on $\widehat{\Xi}_t$ is approximately 95%. In contrast, confidence interval based on $\bar{\Xi}_t$ shows a tendency for undercoverage, which is probably due to the slower convergence of $\bar{\Xi}_t$ than $\widehat{\Xi}_t$. Moreover, $\widehat{\Xi}_t$ works well under the cases where it is an unbiased estimator ($\Sigma_a = I$ or $\tau = \infty$), but the coverage rate falls under 95% in other scenarios. This shows that the bias greatly compromises the performance of confidence intervals. Comprehensive exploration of varying r is reported in Appendix F.1.

To further visualize the table, we use the case of Topelitz Σ_a with $r = 0.5$ under $d = 5$ as an example, plotting the trajectories of the average covariance estimation error and empirical coverage rate. The plots are presented in Figure 1. It is important to note that, under the same parameter setting, the limiting covariance $\Xi^\star$ is different for ASGD and the sketched Newton methods with varying τ. To ensure a fair comparison, we evaluate the relative error $\|\widehat{\Xi}_t - \Xi^\star\| / \|\Xi^\star\|$. From the first row of plots, we see that the absolute value of the slope of the green lines exceeds $(1 - 0.505)/2$, which aligns with the theoretical upper bound established in Theorem 4.3. The plug-in estimator exhibits rapid convergence; however, its bias is evident in the sketched Newton plot. Furthermore, we note $\bar{\Xi}_t$ converges slower than $\widehat{\Xi}_t$, consistent with the theoretical results. Additionally, we observe oscillations in the estimation error of $\bar{\Xi}_t$, corresponding to the batching process, which is undesirable as increasing the sample size does not always guarantee a reduction in estimation error. This is because the limited sample size in a newly created batch introduces errors. The second row of plots demonstrates that $\widehat{\Xi}_t$ performs well in statistical inference for model parameters.

5.2 Logistic regression

Next, we consider the logistic regression model:

$$\mathbb{P}(\xi_b \mid \boldsymbol{\xi}_a) = \frac{\exp(\xi_b \cdot \boldsymbol{\xi}_a^T \boldsymbol{x}^\star)}{1 + \xi_b \cdot \exp(\xi_b \cdot \boldsymbol{\xi}_a^T \boldsymbol{x}^\star)} \qquad \text{with} \qquad \xi_b \in \{-1, 1\}.$$

For this model, we use the log loss defined as $f(\boldsymbol{x}; \xi) = \log\left(1 + \exp(-\xi_b \cdot \boldsymbol{\xi}_a^T \boldsymbol{x})\right)$. We follow the same experimental setup as in the linear regression model in Section 5.1.

Σ_a	Dim	ASGD	Sketched Newton							
			$\tau = \infty$		$\tau = 10$		$\tau = 20$		$\tau = 40$	
		$\bar{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
Identity	20	89.50	95.00	95.50	95.50	96.00	95.50	94.50	96.00	95.50
	40	86.00	97.00	96.50	96.00	95.50	94.50	94.50	94.00	92.50
	60	87.50	95.00	95.00	94.50	93.50	96.50	95.00	95.50	95.00
	100	92.00	100.0	100.0	95.00	93.00	95.00	91.50	95.50	95.50
Toeplitz $r = 0.5$	20	90.00	96.00	96.50	89.50	95.50	94.00	97.00	90.50	96.00
	40	88.50	94.50	94.50	94.00	97.00	89.00	96.00	90.00	96.50
	60	91.00	96.00	96.00	91.00	97.00	85.00	97.00	87.00	93.50
	100	90.50	100.0	100.0	92.50	97.50	86.50	94.50	88.50	93.50
Equi-corr $r = 0.2$	20	91.50	96.00	96.50	82.50	94.50	84.00	96.00	89.50	96.00
	40	88.50	98.00	98.50	71.00	92.50	74.50	94.00	79.50	98.00
	60	86.00	97.00	96.00	70.00	96.00	75.00	94.50	77.00	96.00
	100	83.50	100.0	100.0	71.00	95.50	70.00	96.00	68.50	93.50

Table 1: *The empirical coverage rate (%) for linear regression at confidence level* 95%.

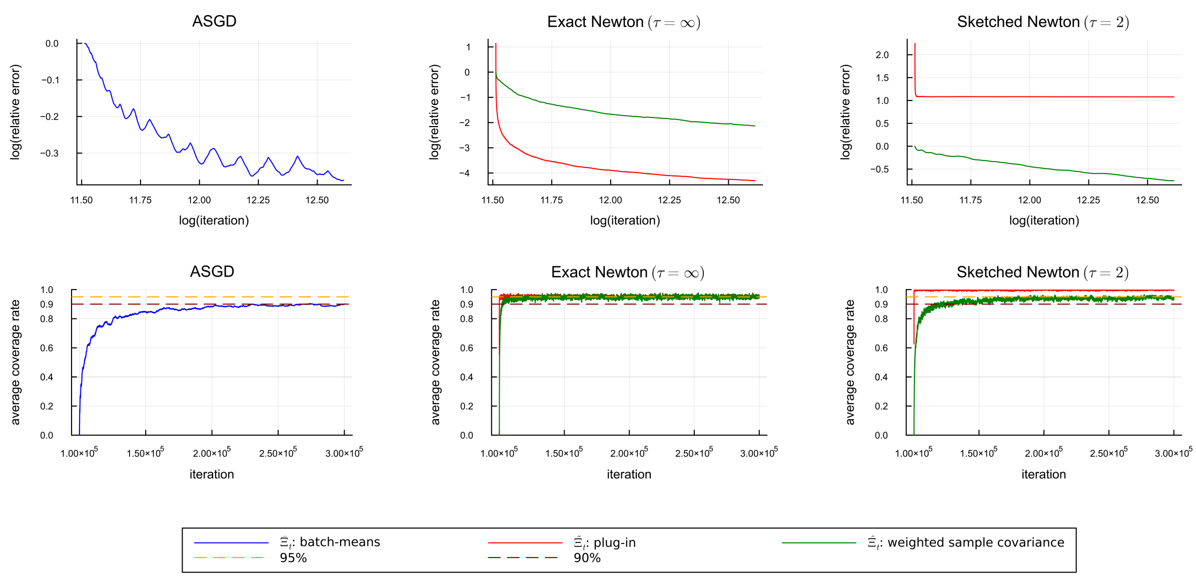

Figure 1: *The average relative covariance estimation error and empirical coverage rate (%). The first row displays the relative error for three methods: ASGD, exact Newton, and sketched Newton (from left to right). For ASGD, the batch-means estimator* $\bar{\Xi}_t$ *is used, while for both Newton methods, the plug-in estimator* $\widetilde{\Xi}_t$ *and the weighted sample covariance matrix* $\widehat{\Xi}_t$ *are compared. The second row similarly presents the empirical coverage rate across various settings for these methods. The plots reflect the consistency of* $\widehat{\Xi}_t$ *and demonstrate its superior performance in statistical inference.*

Our results are summarized in Table 2 and Figure 2, where we observe similar patterns as in the linear regression case. The performance of $\widehat{\Xi}_t$ in logistic regression further demonstrates the effectiveness and reliability of our estimator across different settings. A detailed comparison under more comprehensive settings can be found in Appendix F.1.

Σ_a	Dim	ASGD	Sketched Newton							
			$\tau = \infty$		$\tau = 10$		$\tau = 20$		$\tau = 40$	
		$\bar{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
Identity	20	85.00	96.50	96.00	97.00	97.00	93.50	93.50	97.50	97.00
	40	84.00	95.00	94.00	96.50	94.00	97.50	96.00	97.00	96.50
	60	83.50	95.00	95.00	95.00	95.00	93.00	91.50	93.50	93.00
	100	83.50	94.00	94.50	95.00	92.00	96.00	93.00	96.00	95.00
Toeplitz $r = 0.5$	20	85.50	95.00	95.00	90.50	96.00	94.50	97.00	95.00	96.50
	40	85.50	93.50	93.50	89.50	96.00	93.00	95.50	90.00	94.00
	60	88.00	94.50	94.00	93.00	95.50	87.50	93.50	91.00	96.00
	100	85.00	93.00	93.00	93.00	96.00	89.00	93.50	92.00	95.50
Equi-corr $r = 0.2$	20	87.50	96.50	96.00	85.50	96.00	89.00	95.00	92.50	96.50
	40	86.00	96.00	95.00	81.50	92.50	86.50	97.00	84.50	95.50
	60	85.00	93.50	93.50	79.00	94.00	80.00	97.50	79.00	95.50
	100	76.00	96.50	96.00	74.00	92.50	77.00	92.50	73.00	94.50

Table 2: *The empirical coverage rate (%) for logistic regression at confidence level* 95%.

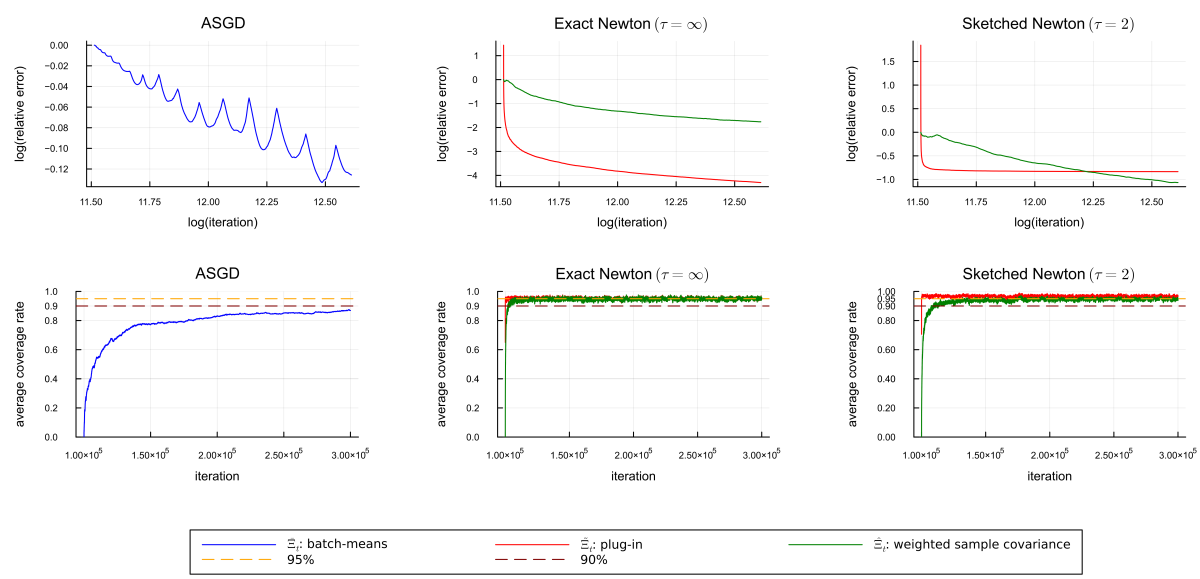

Figure 2: *The average relative covariance estimation error and empirical coverage rate (%). See Figure 1 for interpretation.*

Problem	$\sigma^2 = 1$		$\sigma^2 = 10^{-1}$		$\sigma^2 = 10^{-2}$		$\sigma^2 = 10^{-4}$	
	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
`HS48`	2.080	0.221	2.080	0.231	2.077	0.241	2.081	0.245
`HS78`	1.913	0.290	1.913	0.291	1.911	0.279	1.911	0.270
`HS7`	0.113	0.099	0.070	0.095	0.070	0.102	0.069	0.102
`MARATOS`	0.132	0.096	0.093	0.091	0.089	0.089	0.087	0.083

Table 3: *The average covariance estimation error for CUTEst problems.*

5.3 CUTEst

In this section, we explore the empirical performance of $\widehat{\Xi}_t$ in constrained optimization, as discussed in Section 4.3. We perform four equality-constrained problems from the CUTEst test set: `HS7`, `MARATOS`, `HS48`, and `HS78` (Gould et al., 2014). Due to the presence of constraints, SGD is not applicable, so we compare $\widehat{\Xi}_t$ and $\widetilde{\Xi}_t$ only. For each problem and at each iteration, the CUTEst package provides true evaluations of the function, gradient, and Hessian. With those quantities, we generate our estimates by letting $\bar{g}_t \sim \mathcal{N}(\nabla F_t, \sigma^2(I + \mathbf{1}\mathbf{1}^T))$ and $[\bar{H}_t]_{i,j} = [\bar{H}_t]_{j,i} \sim \mathcal{N}([\nabla^2 F_t]_{i,j}, \sigma^2)$. We vary the sampling variance σ^2 from $\sigma^2 \in \{10^{-4}, 10^{-2}, 10^{-1}, 1\}$ and set $\tau = 5$. The other parameters are set as in Section 5.1, while the problem initialization is provided by the CUTEst package. The true solution is computed using the IPOPT solver (Wächter and Biegler, 2005).

The results are summarized in Table 3. In the previous subsections, we have demonstrated the consistency of $\widehat{\Xi}_t$ leads to its superior performance in statistical inference. For simplicity, we report only the covariance estimation error here. The results show that the estimation error of $\widehat{\Xi}_t$ is consistently small across all settings, showing its robustness to noise levels. In contrast, $\widetilde{\Xi}_t$ performs well as an estimator for problems `HS7` and `MARATOS`, where the bias is small; however, for problems `HS48` and `HS78`, where the bias is more evident, the estimation error of $\widetilde{\Xi}_t$ is significantly larger. The plots of the trajectory of estimation error are provided in Appendix F.2.

6 Conclusion and Future Work

In this paper, we designed a limiting covariance matrix estimator for sketched stochastic Newton methods. Our estimator is fully online and constructed entirely from the Newton iterates. We established the consistency and convergence rate of the estimator. Compared to plug-in estimators for second-order methods, our estimator is asymptotically consistent and more computationally efficient, requiring no matrix factorization. Compared to batch-means estimators for first-order methods, our estimator is batch-free and exhibits faster convergence. Based on our study, we can then construct asymptotically valid confidence intervals/regions for the model parameters using sketched Newton methods. We also discussed the generalization of our estimator to constrained stochastic problems. Extensive experiments on regression problems demonstrate the superior performance of our estimator.

For future research, it would be of interest to explore the lower bound for the online covariance estimation problem, as it provides insights into the statistical efficiency of our weighted sample covariance. Additionally, constructing difference test statistics with different asymptotic distributions based on (sketched) Newton iterates could be promising. In particular, although the normality achieved by Newton methods is asymptotically minimax optimal (Na and Mahoney, 2022), recent studies have observed that confidence intervals constructed using other test statistics, such as t-statistics (Zhu et al.,

2024) and their variants (Lee et al., 2022; Luo et al., 2022; Chen et al., 2024), may exhibit better coverage rates in some problems due to the absence of further covariance estimation. Lastly, performing inference based on Newton methods in non-asymptotic and high-dimensional settings, where the problem dimension grows with the sample size, would also be an interesting direction.

Acknowledgements

This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract DE-AC02-06CH11357.

References

A. S. Berahas, J. Nocedal, and M. Takac. A multi-batch l-bfgs method for machine learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2): 1352–1379, 2021.

A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient jacobians. *Mathematics of Operations Research*, 2023.

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

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

C. Boyer and A. Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972, 2022.

R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

H. Chen, W. Lu, and R. Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2020a.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251 – 273, 2020b.

X. Chen, Z. Owen, C. Pixton, and D. Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937, 2022.

X. Chen, Z. Lai, H. Li, and Y. Zhang. Online statistical inference for stochastic optimization via kiefer-wolfowitz methods. *Journal of the American Statistical Association*, pages 1–24, 2024.

F. E. Curtis, D. P. Robinson, and B. Zhou. A stochastic inexact sequential quadratic optimization algorithm for nonlinear equality-constrained optimization. *INFORMS Journal on Optimization*, 2024.

P. Cénac, A. Godichon-Baggioni, and B. Portier. An efficient averaged stochastic gauss-newton algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*, 2020.

D. Davis, D. Drusvyatskiy, and L. Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *arXiv preprint arXiv:2301.06632*, 2023.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1), 2021.

M. Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

Y. Ermoliev. Stochastic quasigradient methods and their application to system optimization†. *Stochastics*, 9(1–2):1–36, 1983.

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

V. Fabian. Asymptotically efficient stochastic approximation; the rm case. *The Annals of Statistics*, 1(3), 1973.

J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012.

Y. Fang, J. Xu, and L. Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.

N. I. M. Gould, D. Orban, and P. L. Toint. Cutest: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3):545–557, 2014.

R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

M. R. Kosorok and E. B. Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6(1):263–286, 2019.

D. Kovalev, K. Mishchenko, and P. Richtárik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*, 2019.

T. L. Lai. Stochastic approximation: invited paper. *The Annals of Statistics*, 31(2), 2003.

H. Lam and Z. Wang. Resampling stochastic gradient descent cheaply for efficient uncertainty quantification. *arXiv preprint arXiv:2310.11065*, 2023.

S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast and robust online inference with stochastic gradient descent via random scaling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 7381–7389, 2022.

R. Leluc and F. Portier. Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*, 2023.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10. ACM, 2010.

T. Li, L. Liu, A. Kyrillidis, and C. Caramanis. Statistical inference using sgd. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

X. Li, J. Liang, X. Chang, and Z. Zhang. Statistical estimation and inference via local sgd in federated learning. *arXiv preprint arXiv:2109.01326*, 2021.

X. Li, W. Yang, J. Liang, Z. Zhang, and M. I. Jordan. A statistical analysis of polyak-ruppert averaged q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2261. PMLR, 2023.

T. Liang and W. J. Su. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81 (2):431–456, 2019.

R. Liu, X. Chen, and Z. Shang. Statistical inference with stochastic gradient methods under ϕ-mixing data. *arXiv preprint arXiv:2302.12717*, 2023a.

W. Liu, J. Tu, Y. Zhang, and X. Chen. Online estimation and inference for robust policy evaluation in reinforcement learning. *arXiv preprint arXiv:2310.02581*, 2023b.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22 (4):551–575, 1977.

Y. Luo, X. Huo, and Y. Mei. Covariance estimators for the root-sgd algorithm in online learning. *arXiv preprint arXiv:2212.01259*, 2022.

P. Moritz, R. Nishihara, and M. Jordan. A linearly-convergent stochastic l-bfgs algorithm. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 249–258, Cadiz, Spain, 2016. PMLR.

W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2947–2997. PMLR, 2020.

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

S. Na, M. Dereziński, and M. W. Mahoney. Hessian averaging in stochastic newton methods achieves superlinear convergence. *Mathematical Programming*, 201(1–2):473–520, 2022.

Y. Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, 2018.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

E. Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2008.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. *Optimizing Methods in Statistics*, pages 233–257, 1971.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

J. Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405, 1958.

J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

R. Singh, A. Shukla, and D. Vats. On the utility of equal batch sizes for inference in stochastic gradient descent. *arXiv preprint arXiv:2303.07706*, 2023.

T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2008.

K. Tang, W. Liu, Y. Zhang, and X. Chen. Acceleration of stochastic gradient descent with momentum by averaging: finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023.

T. Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

P. Toulis, E. Airoldi, and J. Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 667–675. PMLR, 2014.

P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4), 2017.

A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

S. W. Wallace and W. T. Ziemba. *Applications of Stochastic Programming.* Society for Industrial and Applied Mathematics, 2005.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2005.

Y. Zhong, T. Kuffner, and S. Lahiri. Online bootstrap inference with nonconvex stochastic gradient descent estimator. *arXiv preprint arXiv:2306.02205*, 2023.

W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2021.

W. Zhu, Z. Lou, Z. Wei, and W. B. Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

Y. Zhu and J. Dong. On constructing confidence region for model parameters in stochastic gradient descent via batch means. In *2021 Winter Simulation Conference (WSC)*. IEEE, 2021.

A Online Update of $\widehat{\Xi}_t^{-1}$

We introduce how to online update $\widehat{\Xi}_t^{-1}$ for constructing the confidence region in Corollary 4.4. By the definition of $\widehat{\Xi}_t$ in (21), we have

$$
\begin{aligned}
\widehat{\Xi}_{t+1} &\overset{(21)}{=} \frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t+1})^T \\
&= \frac{1}{t+1}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t+1})^T + \frac{1/\varphi_t}{t+1}(\boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1})^T \\
&= \frac{t}{t+1}\left(\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t})(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t})^T + \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t})(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})^T\right. \\
&\quad \left. + \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{t})^T + \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})^T\right) \\
&\quad + \frac{1/\varphi_t}{t+1}(\boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1})^T \\
&\overset{(22)}{=} \frac{t}{t+1}\left(\widehat{\Xi}_t + (\boldsymbol{v}_t - a_t\bar{\boldsymbol{x}}_t)(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})^T + (\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{v}_t - a_t\bar{\boldsymbol{x}}_t)^T + a_t(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})(\bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1})^T\right) \\
&\quad + \frac{1/\varphi_t}{t+1}(\boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1})^T.
\end{aligned}
$$

Let us define two matrices $R_t \in \mathbb{R}^{d\times 3}$ and $\Lambda_t \in \mathbb{R}^{3\times 3}$ as

$$
R_t = (\boldsymbol{v}_t - a_t\bar{\boldsymbol{x}}_t; \bar{\boldsymbol{x}}_t-\bar{\boldsymbol{x}}_{t+1}; \boldsymbol{x}_{t+1}-\bar{\boldsymbol{x}}_{t+1}), \qquad \Lambda_t = \begin{pmatrix} 0 & 1 & 0 \\ 1 & a_t & 0 \\ 0 & 0 & 1/(t\varphi_t)\end{pmatrix}.
$$

Then, we have

$$
\widehat{\Xi}_{t+1} = \frac{t}{t+1}\left(\widehat{\Xi}_t + R_t\Lambda_t R_t^T\right).
$$

Thus, by Sherman–Morrison–Woodbury formula, we obtain

$$
\widehat{\Xi}_{t+1}^{-1} = \frac{t+1}{t}\widehat{\Xi}_t^{-1} - \frac{t+1}{t}\widehat{\Xi}_t^{-1}R_t\left(\Lambda_t^{-1} + R_t^T\widehat{\Xi}_t^{-1}R_t\right)^{-1}R_t^T\widehat{\Xi}_t^{-1}.
$$

B Preparation Lemmas

We introduce some preparation lemmas regarding the stepsize sequences and the update direction.

Lemma B.1 (Na and Mahoney (2022), Lemma B.1)**.** Suppose $\{\varphi_i\}_i$ is a positive sequence that satisfies $\lim_{i\rightarrow\infty} i(1-\varphi_{i-1}/\varphi_i) = \varphi$. Then, for any $p\geq 0$, we have $\lim_{i\rightarrow\infty} i\left(1-\varphi_{i-1}^p/\varphi_i^p\right) = p\cdot\varphi$.

Lemma B.2 (Na and Mahoney (2022), Lemma B.3(a))**.** Let $\{\phi_i\}_i, \{\varphi_i\}_i, \{\sigma_i\}_i$ be three positive sequences. Suppose[1]

$$
\lim_{i\rightarrow\infty} i\left(1-\phi_{i-1}/\phi_i\right) = \phi < 0, \qquad \lim_{i\rightarrow\infty}\varphi_i = 0, \qquad \lim_{i\rightarrow\infty} i\varphi_i = \widetilde{\varphi} \tag{B.1}
$$

[1]In fact, $\phi<0$ is only required by Lemma B.3(b) in Na and Mahoney (2022), and the statements in Lemma B.3(a) hold for any constant ϕ.

for a constant ϕ and a (possibly infinite) constant $\widetilde{\varphi} \in (0, \infty]$. For any $l \geq 1$, if we further have

$$\sum_{k=1}^{l} \sigma_k + \phi/\widetilde{\varphi} > 0, \tag{B.2}$$

then the following results hold as $t \to \infty$:

$$\frac{1}{\phi_t} \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \varphi_i \phi_i \longrightarrow \frac{1}{\sum_{k=1}^{l} \sigma_k + \phi/\widetilde{\varphi}}, \tag{B.3}$$

$$\frac{1}{\phi_t} \left\{ \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \varphi_i \phi_i a_i + b \cdot \prod_{j=0}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \right\} \longrightarrow 0, \tag{B.4}$$

where the second result holds for any constant b and any sequence $\{a_t\}_t$ such that $a_t \to 0$.

Lemma B.3. Suppose $\{\phi_i\}_i$ and $\{\sigma_i\}_i$ are two positive sequences, and $\{\phi_i\}_i$ satisfies $\lim_{i\to\infty} i(1 - \phi_{i-1}/\phi_i) = \phi < 0$ for a constant ϕ. Let $\varphi_i = c_\varphi/(i+1)^\varphi + o(1/(i+1)^\varphi)$ for constants $c_\varphi > 0$ and $\varphi \in (0, 1)$. For any $l \geq 1$, we have

$$\left| \frac{1}{\phi_t} \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \varphi_i \phi_i - \frac{1}{\sum_{k=1}^{l} \sigma_k} \right| \lesssim \begin{cases} \varphi_t, & \varphi \in (0, 0.5), \\ \left(0.5 - \dfrac{\phi/c_\varphi^2}{(\sum_{k=1}^{l} \sigma_k)^2}\right) \varphi_t, & \varphi = 0.5, \\ -\dfrac{\phi}{(\sum_{k=1}^{l} \sigma_k)^2} \cdot \dfrac{1}{t\varphi_t}, & \varphi \in (0.5, 1). \end{cases}$$

Lemma B.4. Suppose $\{\phi_i\}_i, \{\varphi_i\}_i, \{\sigma_i\}_i$ be three positive sequences that satisfy the assumptions in Lemma B.2. Let $\{\eta_i\}_i$ be a positive sequence such that $\lim_{i\to\infty} \eta_i/\varphi_i = 1$. For any $l \geq 1$, if $\sum_{k=1}^{l} \sigma_k/2 + \phi/\widetilde{\varphi} > 0$, then we have

$$\prod_{i=0}^{t} \prod_{k=1}^{l} |1 - \eta_i \sigma_k| + \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} |1 - \eta_j \sigma_k| \varphi_i \phi_i \lesssim \frac{1}{\sum_{k=1}^{l} \sigma_k/2 + \phi/\widetilde{\varphi}} \cdot \phi_t.$$

Lemma B.5. For the t-th iteration, let us define two sketching matrices

$$\widetilde{C}_{t,j} = I - \left(B_t S_{t,j} (S_{t,j}^T B_t^2 S_{t,j})^\dagger S_{t,j}^T B_t\right) \qquad \text{and} \qquad \widetilde{C}_t = \prod_{j=1}^{\tau} \widetilde{C}_{t,j}, \tag{B.5}$$

and we also let $C_t = \mathbb{E}[\widetilde{C}_t \mid \mathcal{F}_{t-1}]$. Then, under Assumptions 3.2 and 3.4, the following results hold (recall that $C^\star$ is defined in (16)):

(a) We have $\bar{\Delta} \boldsymbol{x}_t = (I - \widetilde{C}_t) \Delta \boldsymbol{x}_t = -(I - \widetilde{C}_t) B_t^{-1} \bar{g}_t$ for any $t \geq 0$.

(b) We have $\mathbb{E}[\bar{\Delta} \boldsymbol{x}_t \mid \mathcal{F}_{t-1}] = -(I - C_t) B_t^{-1} \nabla F_t$ for any $t \geq 0$.

(c) We have $\|C_t\| \leq \rho^\tau$ for any $t \geq 0$ with $\rho = 1 - \gamma_S$. When $\boldsymbol{x}_t \to \boldsymbol{x}^\star$, we also have $\|C^\star\| \leq \rho^\tau$.

(d) When $\boldsymbol{x}_t \to \boldsymbol{x}^\star$, we have $(1 - \rho^\tau) I \preceq I - C^\star \preceq I$.

C Proofs of Preparation Lemmas

C.1 Proof of Lemma B.3

We note that (B.1) is satisfied with $\widetilde{\varphi}=\infty$ and (B.2) holds as $\sum_{k=1}^l \sigma_k + \phi/\widetilde{\varphi} = \sum_{k=1}^l \sigma_k > 0$. Thus, Lemma B.2 holds and its proof (Na and Mahoney, 2022, (C.1)) suggests the following decomposition

$$\frac{1}{\phi_t}\sum_{i=0}^t\prod_{j=i+1}^t\prod_{k=1}^l(1-\varphi_j\sigma_k)\,\varphi_i\phi_i - \frac{1}{\sum_{k=1}^l\sigma_k} = \frac{1}{\phi_t}\prod_{j=1}^t\prod_{k=1}^l(1-\varphi_j\sigma_k)\cdot\phi_0\Big(\varphi_0-\frac{1}{\sum_{k=1}^l\sigma_k}\Big)$$
$$+\frac{1}{\phi_t}\sum_{i=1}^t\prod_{j=i+1}^t\prod_{k=1}^l(1-\varphi_j\sigma_k)\phi_i\Big\{\varphi_i-\frac{1}{\sum_{k=1}^l\sigma_k}\Big(1-\frac{\phi_{i-1}}{\phi_i}\prod_{k=1}^l(1-\varphi_i\sigma_k)\Big)\Big\} =: I+II. \quad (C.1)$$

We first calculate the rate of the term in the curly bracket in II. We note that

$$\prod_{k=1}^l(1-\varphi_i\sigma_k) = 1-\sum_{k=1}^l\sigma_k\varphi_i + 0.5\Big\{\big(\sum_{k=1}^l\sigma_k\big)^2-\big(\sum_{k=1}^l\sigma_k^2\big)\Big\}\varphi_i^2+o(\varphi_i^2),$$
$$\frac{\phi_{i-1}}{\phi_i} = 1-\phi\cdot\frac{1}{i+1}+o\Big(\frac{1}{i+1}\Big).$$

Thus, the rate of the multiplication of these two terms is

$$\frac{\phi_{i-1}}{\phi_i}\prod_{k=1}^l(1-\varphi_i\sigma_k) = \begin{cases} 1-\sum_{k=1}^l\sigma_k\varphi_i+0.5\Big\{\big(\sum_{k=1}^l\sigma_k\big)^2-\big(\sum_{k=1}^l\sigma_k^2\big)\Big\}\varphi_i^2+o(\varphi_i^2), & \varphi\in(0,0.5),\\ 1-\sum_{k=1}^l\sigma_k\varphi_i+\Big\{0.5\{\big(\sum_{k=1}^l\sigma_k\big)^2-\big(\sum_{k=1}^l\sigma_k^2\big)\}-\frac{\phi}{c_\varphi^2}\Big\}\varphi_i^2+o(\varphi_i^2), & \varphi=0.5,\\ 1-\sum_{k=1}^l\sigma_k\varphi_i-\frac{\phi}{i+1}+o\Big(\frac{1}{i+1}\Big), & \varphi\in(0.5,1). \end{cases} \quad (C.2)$$

Let us first consider the case $\varphi\in(0.5,1)$. We plug the above display above into II in (C.1) and get

$$II = -\frac{\phi}{\sum_{k=1}^l\sigma_k}\cdot\frac{1}{\phi_t}\sum_{i=1}^t\prod_{j=i+1}^t\prod_{k=1}^l(1-\varphi_j\sigma_k)\varphi_i\cdot\Big\{\frac{\phi_i}{(i+1)\varphi_i}+o\Big(\frac{\phi_i}{(i+1)\varphi_i}\Big)\Big\}.$$

We note that

$$\lim_{i\to\infty} i\left(1-\frac{\phi_{i-1}/(i\varphi_{i-1})}{\phi_i/((i+1)\varphi_i)}\right) = \lim_{i\to\infty} i\left(1-\frac{\phi_{i-1}}{\phi_i}+\frac{\phi_{i-1}}{\phi_i}\Big(1-\frac{1/(i\varphi_{i-1})}{1/((i+1)\varphi_i)}\Big)\right) = \phi+\varphi-1<0,$$

so we can apply Lemma B.2 and derive

$$\lim_{t\to\infty}\frac{1}{\phi_t/((t+1)\varphi_t)}\sum_{i=1}^t\prod_{j=i+1}^t\prod_{k=1}^l(1-\varphi_j\sigma_k)\varphi_i\cdot\frac{\phi_i}{(i+1)\varphi_i} \overset{(B.3)}{=} \frac{1}{\sum_{k=1}^l\sigma_k},$$
$$\lim_{t\to\infty}\frac{1}{\phi_t/((t+1)\varphi_t)}\sum_{i=1}^t\prod_{j=i+1}^t\prod_{k=1}^l(1-\varphi_j\sigma_k)\varphi_i\cdot o\Big(\frac{\phi_i}{(i+1)\varphi_i}\Big) \overset{(B.4)}{=} 0.$$

Combining the two displays, we have $|II| \lesssim -\frac{\phi}{(\sum_{k=1}^{l}\sigma_k)^2} \cdot \frac{1}{(t+1)\varphi_t}$. For the term I in (C.1), we have

$$\lim_{t\to\infty} \frac{1}{\phi_t/\big((t+1)\varphi_t\big)} \prod_{j=1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\cdot\phi_0\left(\varphi_0 - \frac{1}{\sum_{k=1}^{l}\sigma_k}\right) \overset{(B.4)}{=} 0.$$

This indicates $|I| = o(1/(t+1)\varphi_t)$. Combining the rates of $|I|$ and $|II|$ with (C.1), we complete the proof for the case $\varphi \in (0.5, 1)$. For the case $\varphi = 0.5$, we know from (C.2) and (C.1) that

$$II = \frac{0.5\{\big(\sum_{k=1}^{l}\sigma_k\big)^2 - \big(\sum_{k=1}^{l}\sigma_k^2\big)\} - \phi/c_\varphi^2}{\sum_{k=1}^{l}\sigma_k}\frac{1}{\phi_t}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i\cdot\{\varphi_i + o(\varphi_i)\}\,.$$

Following the same analysis as above and applying Lemma B.2, we obtain

$$|II| \lesssim \frac{0.5\{\big(\sum_{k=1}^{l}\sigma_k\big)^2 - \big(\sum_{k=1}^{l}\sigma_k^2\big)\} - \phi/c_\varphi^2}{(\sum_{k=1}^{l}\sigma_k)^2}\varphi_t \le \left(0.5 - \frac{\phi/c_\varphi^2}{(\sum_{k=1}^{l}\sigma_k)^2}\right)\varphi_t.$$

We also have $|I| = o(\varphi_t)$ and, hence, complete the proof for the case $\varphi = 0.5$. The proof for the case $\varphi \in (0, 0.5)$ can be done similarly by noting that

$$II = \frac{0.5\{\big(\sum_{k=1}^{l}\sigma_k\big)^2 - \big(\sum_{k=1}^{l}\sigma_k^2\big)\}}{\sum_{k=1}^{l}\sigma_k}\frac{1}{\phi_t}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i\cdot\{\varphi_i + o(\varphi_i)\}\,.$$

We complete the proof.

C.2 Proof of Lemma B.4

Since $\lim_{t\to\infty}\eta_t/\varphi_t = 1$ and $\lim_{t\to\infty}\varphi_t = 0$, there exists a fixed integer $\widetilde{t}$ such that for any $t \ge \widetilde{t}$ and $1 \le k \le l$, we have $\eta_t \ge \varphi_t/2$ and $0 < 1 - \eta_t\sigma_k \le 1 - \varphi_t\sigma_k/2$. Define a sequence $\{\widetilde{\phi}_t\}_{t=\widetilde{t}-1}^{\infty}$ as follows:

$$\widetilde{\phi}_t = \begin{cases} \phi_t + \sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{\widetilde{t}-1}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\phi_i, & t = \widetilde{t}-1, \\ \phi_t, & t \ge \widetilde{t}. \end{cases}$$

With the above sequence, we use the techniques in (Na and Mahoney, 2022, (E.19)) and rewrite the following series as

$$\begin{aligned}
&\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\phi_i = \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\phi_i + \sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{t}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\phi_i \\
&= \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\phi_i + \prod_{j=\widetilde{t}}^{t}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\cdot\sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{\widetilde{t}-1}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\phi_i \\
&= \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}|1-\eta_j\sigma_k|\varphi_i\widetilde{\phi}_i \le \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k/2)\varphi_i\widetilde{\phi}_i \\
&\overset{(B.3)}{\lesssim} \frac{1}{\sum_{k=1}^{l}\sigma_k/2 + \phi/\widetilde{\varphi}}\cdot\phi_t.
\end{aligned}$$

Additionally, we know

$$\prod_{i=0}^{t}\prod_{k=1}^{l}|1-\eta_i\sigma_k| = \prod_{i=\widetilde{t}}^{t}\prod_{k=1}^{l}|1-\eta_i\sigma_k|\cdot\prod_{i=0}^{\widetilde{t}-1}\prod_{k=1}^{l}|1-\eta_i\sigma_k|$$
$$\le \prod_{i=\widetilde{t}}^{t}\prod_{k=1}^{l}(1-\varphi_i\sigma_k/2)\cdot\prod_{i=0}^{\widetilde{t}-1}\prod_{k=1}^{l}|1-\eta_i\sigma_k| \overset{\text{(B.4)}}{=} o(\phi_t).$$

We complete the proof.

C.3 Proof of Lemma B.5

Recalling $B_t\Delta\boldsymbol{x}_t = -\bar{g}_t$, we subtract $\Delta\boldsymbol{x}_t$ from both sides of (9) and obtain

$$\Delta\boldsymbol{x}_{t,j+1}-\Delta\boldsymbol{x}_t = \big(I - B_tS_{t,j}(S_{t,j}^TB_t^2S_{t,j})^{\dagger}S_{t,j}^TB_t\big)(\Delta\boldsymbol{x}_{t,j}-\Delta\boldsymbol{x}_t) = \widetilde{C}_{t,j}(\Delta\boldsymbol{x}_{t,j}-\Delta\boldsymbol{x}_t).$$

Since $\Delta\boldsymbol{x}_{t,0}=\boldsymbol{0}$, we complete the proof of (a). By the independence between sketching and sampling and the unbiasedness of $\bar{g}_t$ in Assumption 3.2, we complete the proof of (b). (c) can be found in Lemma 4.4 and Corollary 5.4 in Na and Mahoney (2022). (d) is an immediate result from (c) by observing that $C^\star \succeq \boldsymbol{0}$.

D Proofs of Section 3.2

To ease the presentation, we assume throughout the proof and without loss of generality that all upper bound constants in the assumptions $\Upsilon_L, \Upsilon_S, \Upsilon_H, C_{g,1}, C_{g,2}, C_{H,1}, C_{H,2} \ge 1$, and the lower bound constant $0<\gamma_H\le 1$. The range of these constants is not crucial to the analysis; all results still hold by replacing γ_H by $\gamma_H\wedge 1$ (similar for other constants).

D.1 Proof of Theorem 3.5

By Assumption 3.3, $\|\nabla^2F(\boldsymbol{x})\|\le\Upsilon_H$. Applying Taylor's expansion, we have

$$\begin{aligned}
&F_{t+1}-F^\star\\
&\le F_t - F^\star + \bar{\alpha}_t\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t + \frac{\Upsilon_H}{2}\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2\\
&= F_t - F^\star + \bar{\alpha}_t\mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] + \bar{\alpha}_t\Big(\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t\mid\mathcal{F}_{t-1}]\Big) + \frac{\Upsilon_H}{2}\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2. \quad \text{(D.1)}
\end{aligned}$$

Then, we take expectation on both sides conditioning on $\mathcal{F}_{t-1}$ and obtain

$$\begin{aligned}
\mathbb{E}[F_{t+1}-F^\star\mid\mathcal{F}_{t-1}] &\le F_t - F^\star + \mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t\mid\mathcal{F}_{t-1}]\mid\mathcal{F}_{t-1}\Big]\\
&\quad + \mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t\mid\mathcal{F}_{t-1}]\Big\}\mid\mathcal{F}_{t-1}\Big] + \frac{\Upsilon_H}{2}\mathbb{E}\big[\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2\mid\mathcal{F}_{t-1}\big]. \quad \text{(D.2)}
\end{aligned}$$

For the second term on the right hand side, we apply Assumption 3.3, Lemma B.5(b, c), and have

$$\begin{aligned}
\mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] &= -\nabla F_t^T(I-C_t)B_t^{-1}\nabla F_t \cdot \mathbb{E}\Big[\bar{\alpha}_t \mid \mathcal{F}_{t-1}\Big] \\
&\leq \Big(-\frac{1}{\Upsilon_H}\|\nabla F_t\|^2 + \frac{\rho^\tau}{\gamma_H}\|\nabla F_t\|^2\Big)\cdot \mathbb{E}\Big[\bar{\alpha}_t \mid \mathcal{F}_{t-1}\Big] \\
&\leq -\frac{3}{4\Upsilon_H}\beta_t\|\nabla F_t\|^2 \quad (\text{by } \rho^\tau \leq \gamma_H/4\Upsilon_H \text{ and } \beta_t \leq \bar{\alpha}_t). \quad \text{(D.3)}
\end{aligned}$$

For the third term in (D.2), we note $\mathbb{E}\big[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\big] = 0$. Thus, we have (recall $\varphi_t = \beta_t + \chi_t/2$)

$$\begin{aligned}
\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] &= \mathbb{E}\big[(\bar{\alpha}_t-\varphi_t)\nabla F_t^T\big(\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\big) \mid \mathcal{F}_{t-1}\big] \\
&\leq \frac{\chi_t}{2}\|\nabla F_t\|\mathbb{E}\Big[\big\|\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\big\| \mid \mathcal{F}_{t-1}\Big] \quad (\text{by } |\bar{\alpha}_t - \varphi_t| \leq \chi_t/2). \quad \text{(D.4)}
\end{aligned}$$

By Lemma B.5(a, b, c), we obtain

$$\begin{aligned}
\big\|\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\big\| &= \big\|(I-\widetilde{C}_t)B_t^{-1}\bar{g}_t - (I-C_t)B_t^{-1}\nabla F_t\big\| \\
&\leq \|C_t - \widetilde{C}_t\|\|B_t^{-1}\|\|\nabla F_t\| + \|I-\widetilde{C}_t\|\|B_t^{-1}\|\|\bar{g}_t - \nabla F_t\| \overset{(15)}{\leq} \frac{2}{\gamma_H}\|\nabla F_t\| + \frac{2}{\gamma_H}\|\bar{g}_t - \nabla F_t\|.
\end{aligned}$$

Thus, applying Assumption 3.2, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\big\|\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\big\| \mid \mathcal{F}_{t-1}\Big] &\leq \frac{2}{\gamma_H}\|\nabla F_t\| + \frac{2C_{g,1}^{1/4}}{\gamma_H}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \frac{2C_{g,2}^{1/4}}{\gamma_H} \\
&\leq \frac{4C_{g,1}^{1/4}}{\gamma_H^2}\|\nabla F_t\| + \frac{2C_{g,2}^{1/4}}{\gamma_H} \quad (\text{by } C_{g,1} \geq 1), \qquad \text{(D.5)}
\end{aligned}$$

where the last inequality also uses the property of strong convexity of $F(\boldsymbol{x})$ (Nesterov, 2018)

$$\frac{\gamma_H}{2}\left\|\boldsymbol{x}_t - \boldsymbol{x}^\star\right\|^2 \leq F_t - F^\star \leq \frac{1}{2\gamma_H}\left\|\nabla F_t\right\|^2. \qquad \text{(D.6)}$$

Combining (D.4) and (D.5), we get

$$\begin{aligned}
&\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] \leq \frac{2C_{g,1}^{1/4}}{\gamma_H^2}\chi_t\|\nabla F_t\|^2 + \frac{C_{g,2}^{1/4}}{\gamma_H}\chi_t\|\nabla F_t\| \\
&\leq \frac{2C_{g,1}^{1/4}}{\gamma_H^2}\chi_t\|\nabla F_t\|^2 + \frac{1}{4\Upsilon_H}\beta_t\|\nabla F_t\|^2 + \frac{\Upsilon_H C_{g,2}^{1/2}}{\gamma_H^2}\cdot\frac{\chi_t^2}{\beta_t} \quad (\text{by Young's inequality}). \qquad \text{(D.7)}
\end{aligned}$$

Let $\eta_t = \beta_t + \chi_t$. We apply Lemma B.5(a) and bound the last term in (D.2) by

$$\begin{aligned}
\mathbb{E}\big[\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] &\leq \mathbb{E}\big[\bar{\alpha}_t^2\|(I+\widetilde{C}_t)\|^2\|B_t^{-1}\|^2\|\bar{g}_t\|^2 \mid \mathcal{F}_{t-1}\big] \\
&\leq \frac{8}{\gamma_H^2}\eta_t^2\Big(\|\nabla F_t\|^2 + \mathbb{E}[\|\bar{g}_t - \nabla F_t\|^2 \mid \mathcal{F}_{t-1}]\Big) \\
&\leq \frac{16C_{g,1}^{1/2}}{\gamma_H^4}\eta_t^2\|\nabla F_t\|^2 + \frac{8C_{g,2}^{1/2}}{\gamma_H^2}\eta_t^2 \quad (\text{by Assumption 3.2}). \qquad \text{(D.8)}
\end{aligned}$$

We plug (D.3), (D.7), and (D.8) into (D.2), and obtain

$$\mathbb{E}[F_{t+1}-F^\star \mid \mathcal{F}_{t-1}] \\ \leq F_t - F^\star - \left(\frac{1}{2\Upsilon_H}\beta_t - \frac{2C_{g,1}^{1/4}}{\gamma_H^2}\chi_t - \frac{8\Upsilon_H C_{g,1}^{1/2}}{\gamma_H^4}\eta_t^2\right)\|\nabla F_t\|^2 + \frac{4\Upsilon_H C_{g,2}^{1/2}}{\gamma_H^2}\left(\frac{\chi_t^2}{\beta_t}+\eta_t^2\right).$$

Since $\beta_t = c_\beta/(t+1)^\beta$ with $\beta \in (0.5, 1]$ and $\chi_t = c_\chi/(t+1)^\chi$ with $\chi > 0.5(\beta+1) \geq \beta$, there exists a fixed integer t_0 such that $\frac{2C_{g,1}^{1/4}}{\gamma_H^2}\chi_t + \frac{8\Upsilon_H C_{g,1}^{1/2}}{\gamma_H^4}\eta_t^2 \leq \frac{1}{4\Upsilon_H}\beta_t$ for all $t \geq t_0$. Thus, for $t \geq t_0$, we have

$$\mathbb{E}[F_{t+1}-F^\star \mid \mathcal{F}_{t-1}] \leq F_t - F^\star - \frac{1}{4\Upsilon_H}\beta_t\|\nabla F_t\|^2 + \frac{4\Upsilon_H C_{g,2}^{1/2}}{\gamma_H^2}\left(\frac{\chi_t^2}{\beta_t}+\eta_t^2\right).$$

Note that $\sum_{t=t_0}^\infty \chi_t^2/\beta_t < \infty$ and $\sum_{t=t_0}^\infty \eta_t^2 \lesssim \sum_{t=t_0}^\infty \beta_t^2 + \sum_{t=t_0}^\infty \chi_t^2 < \infty$. Thus, we apply the Robbins-Siegmund Theorem (Duflo, 2013, Theorem 1.3.12) and conclude that $F_t - F^\star$ converges to a finite random variable, and $\sum_{t=t_0}^\infty \beta_t \|\nabla F_t\|^2 < \infty$ almost surely. Furthermore, we have $\liminf_{t\to\infty}\|\nabla F_t\| = 0$ due to $\sum_{t=t_0}^\infty \beta_t = \infty$, which leads to $\liminf_{t\to\infty}(F_t - F^\star) = 0$ according to (D.6). Since $F_t - F^\star$ converges almost surely, the conclusion can be strengthened to $\lim_{t\to\infty} F_t - F^\star = 0$. Again, we apply (D.6) and obtain $\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^\star$ almost surely. This completes the proof.

D.2 Proof of Theorem 3.6

The proof of asymptotic normality is almost identical to the proof of Theorem 5.6 in Na and Mahoney (2022). Since $\chi > 1.5\beta \Rightarrow \chi > 0.5(\beta+1)$, we have $\boldsymbol{x}_t \to \boldsymbol{x}^\star$ almost surely, as proved in Theorem 3.5. Therefore, we only have to note that our growth conditions in Assumptions 3.2 and 3.3 on gradients and Hessians do not affect the proof of normality (though they affect the proof of convergence), since the term $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$ in the growth conditions converges to 0 almost surely.

E Proofs of Section 4.2

To clear up tedious constants, we assume $\eta_t = \beta_t + \chi_t \leq 1$, $\forall t \geq 0$, without loss of generality for the remainder of this paper. Note that this condition is non-essential, since $\eta_t \to 0$ and the condition will always hold for sufficiently large, fixed t.

E.1 Proof of Lemma 4.2

We separate the proof into two parts.

Part 1: Bound of $\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4]$. We take square on both sides of (D.1) and take expectation conditioning on $\mathcal{F}_{t-1}$, then we get

$$\mathbb{E}\big[(F_{t+1}-F^\star)^2 \mid \mathcal{F}_{t-1}\big] \leq (F_t - F^\star)^2 + \mathbb{E}\big[2\bar{\alpha}_t(F_t-F^\star)\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}\big] + \mathbb{E}\big[\bar{\alpha}_t^2\Upsilon_H(F_t-F^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] \\ + \mathbb{E}\big[\bar{\alpha}_t^2(\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t)^2 \mid \mathcal{F}_{t-1}\big] + \mathbb{E}\big[\bar{\alpha}_t^3\Upsilon_H\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] + \frac{1}{4}\mathbb{E}\big[\bar{\alpha}_t^4\Upsilon_H^2\|\bar{\Delta}\boldsymbol{x}_t\|^4 \mid \mathcal{F}_{t-1}\big]. \quad \text{(E.1)}$$

We rearrange these terms by the order of $\bar{\alpha}_t$ and analyze them one by one.

- **Term 1:** $\mathbb{E}[2\bar{\alpha}_t(F_t - F^\star)\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]$.

This term can be decomposed as

$$\mathbb{E}[2\bar{\alpha}_t(F_t - F^\star)\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] = 2(F_t - F^\star)\mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] + 2(F_t - F^\star)\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big].$$

For the first term on the right hand side, by (D.3) and (D.6), we have

$$2(F_t-F^\star)\mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] \le -\frac{3}{2\Upsilon_H}\beta_t(F_t-F^\star)\|\nabla F_t\|^2 \le -\frac{3\gamma_H}{\Upsilon_H}\beta_t(F_t-F^\star)^2. \tag{E.2}$$

For the second term on the right hand side, (D.4) and (D.5) give us

$$\begin{aligned}
&2(F_t - F^\star)\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] \\
&\le \chi_t(F_t - F^\star)\|\nabla F_t\|\Big(\frac{2}{\gamma_H}\|\nabla F_t\| + \frac{2C_{g,1}^{1/4}}{\gamma_H}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \frac{2C_{g,2}^{1/4}}{\gamma_H}\Big) \\
&\le \frac{4\Upsilon_H^{1/2}}{\gamma_H}\Big(\Upsilon_H^{1/2} + \frac{C_{g,1}^{1/4}}{\gamma_H^{1/2}}\Big)\chi_t(F_t - F^\star)^2 + \frac{2\sqrt{2}\Upsilon_H^{1/2}C_{g,2}^{1/4}}{\gamma_H}\chi_t(F_t - F^\star)^{3/2} \\
&\le \left(\frac{4\Upsilon_H^{1/2}}{\gamma_H}\Big(\Upsilon_H^{1/2} + \frac{C_{g,1}^{1/4}}{\gamma_H^{1/2}}\Big)\chi_t + \frac{\gamma_H}{2\Upsilon_H}\beta_t\right)(F_t - F^\star)^2 + \frac{3^4\Upsilon_H^5C_{g,2}}{\gamma_H^7}\cdot\frac{\chi_t^4}{\beta_t^3} \quad \text{(Young's inequality)}.
\end{aligned} \tag{E.3}$$

Here, the second inequality is due to (D.6) and the following Υ_H-Lipschitz continuity property of $\nabla F(\boldsymbol{x})$ (Nesterov, 2018):

$$\frac{1}{2\Upsilon_H}\|\nabla F_t\|^2 \le F_t - F^\star \le \frac{\Upsilon_H}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2. \tag{E.4}$$

• **Term 2:** $\mathbb{E}[\bar{\alpha}_t^2\Upsilon_H(F_t - F^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2 + \bar{\alpha}_t^2(\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t)^2 \mid \mathcal{F}_{t-1}]$.
Since $\bar{\alpha}_t \le \eta_t$, we bound this term by

$$\begin{aligned}
&\mathbb{E}[\,\bar{\alpha}_t^2\Upsilon_H(F_t - F^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2 + \bar{\alpha}_t^2(\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t)^2 \mid \mathcal{F}_{t-1}] \le \eta_t^2\mathbb{E}\big[\big(\Upsilon_H(F_t - F^\star) + \|\nabla F_t\|^2\big)\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] \\
&\overset{(E.4)}{\le} 3\Upsilon_H\eta_t^2\mathbb{E}\big[(F_t - F^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] \le \frac{\gamma_H}{2\Upsilon_H}(\beta_t + \chi_t)(F_t - F^\star)^2 + \frac{9\Upsilon_H^3}{2\gamma_H}\eta_t^3\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4 \mid \mathcal{F}_{t-1}\big],
\end{aligned} \tag{E.5}$$

where the last inequality is by Young's inequality.

• **Term 3:** $\mathbb{E}[\,\bar{\alpha}_t^3\Upsilon_H\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}]$.
Similarly, we use Young's inequality, apply (E.4), and have

$$\begin{aligned}
&\mathbb{E}[\,\bar{\alpha}_t^3\Upsilon_H\nabla F_t^T\bar{\Delta}\boldsymbol{x}_t\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}] \le \eta_t^3\Upsilon_H\mathbb{E}[\|\nabla F_t\|\|\bar{\Delta}\boldsymbol{x}_t\|^3 \mid \mathcal{F}_{t-1}] \\
&\le \frac{1}{4}\eta_t^3\|\nabla F_t\|^4 + \frac{3\Upsilon_H^{4/3}}{4}\eta_t^3\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4 \mid \mathcal{F}_{t-1}\big] \overset{(E.4)}{\le} \Upsilon_H^2\eta_t^3(F_t - F^\star)^2 + \frac{3\Upsilon_H^{4/3}}{4}\eta_t^3\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4 \mid \mathcal{F}_{t-1}\big].
\end{aligned} \tag{E.6}$$

Substituting (E.2), (E.3), (E.5), and (E.6) into (E.1), we obtain

$$\begin{aligned}
\mathbb{E}[(F_{t+1} - F^\star)^2 \mid \mathcal{F}_{t-1}] \le{}& \Big(1 - \frac{2\gamma_H}{\Upsilon_H}\beta_t + \frac{\gamma_H}{2\Upsilon_H}\chi_t + \frac{4\Upsilon_H^{1/2}}{\gamma_H}\Big(\Upsilon_H^{1/2} + \frac{C_{g,1}^{1/4}}{\gamma_H^{1/2}}\Big)\chi_t + \Upsilon_H^2\eta_t^3\Big)(F_t - F^\star)^2 \\
&+ \frac{3^4\Upsilon_H^5C_{g,2}}{\gamma_H^7}\cdot\frac{\chi_t^4}{\beta_t^3} + \frac{6\Upsilon_H^3}{\gamma_H}\eta_t^3\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4 \mid \mathcal{F}_{t-1}\big] \quad (\text{by } \eta_t \le 1).
\end{aligned}$$

Following the analysis in (D.8), we apply Assumption 3.2 and have

$$\begin{aligned}
\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4 \mid \mathcal{F}_{t-1}\big] &\leq \frac{2^4}{\gamma_H^4}\mathbb{E}[\|\bar{g}_t\|^4 \mid \mathcal{F}_{t-1}] \leq \frac{2^7}{\gamma_H^4}\left(\mathbb{E}[\|\bar{g}_t - \nabla F_t\|^4 \mid \mathcal{F}_{t-1}] + \|\nabla F_t\|^4\right) \\
&\overset{(E.4)}{\leq} \frac{2^9\Upsilon_H^2}{\gamma_H^4}(F_t - F^\star)^2 + \frac{2^7 C_{g,1}}{\gamma_H^4}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4 + \frac{2^7 C_{g,2}}{\gamma_H^4} \\
&\overset{(D.6)}{\leq} \frac{2^9\Upsilon_H^2}{\gamma_H^4}(F_t - F^\star)^2 + \frac{2^9 C_{g,1}}{\gamma_H^6}(F_t - F^\star)^2 + \frac{2^7 C_{g,2}}{\gamma_H^4} \\
&= \frac{2^9(\Upsilon_H^2 + C_{g,1}/\gamma_H^2)}{\gamma_H^4}(F_t - F^\star)^2 + \frac{2^7 C_{g,2}}{\gamma_H^4}.
\end{aligned}$$

Combining the above two displays and taking full expectation, we obtain the recursion:

$$\begin{aligned}
&\mathbb{E}[(F_{t+1} - F^\star)^2] \\
&\leq \left(1 - \frac{2\gamma_H}{\Upsilon_H}\beta_t + \frac{\gamma_H}{2\Upsilon_H}\chi_t + \frac{4\Upsilon_H^{1/2}}{\gamma_H}\left(\Upsilon_H^{1/2} + \frac{C_{g,1}^{1/4}}{\gamma_H^{1/2}}\right)\chi_t + \frac{2^{12}\Upsilon_H^3(\Upsilon_H^2 + C_{g,1}/\gamma_H^2)}{\gamma_H^5}\eta_t^3\right)\mathbb{E}[(F_t - F^\star)^2] \\
&\quad + \frac{3^4\Upsilon_H^5 C_{g,2}}{\gamma_H^7}\frac{\chi_t^4}{\beta_t^3} + \frac{2^{10}\Upsilon_H^3 C_{g,2}}{\gamma_H^5}\eta_t^3.
\end{aligned}$$

We apply the above inequality recursively until $(F_0 - F^\star)^2$ and then apply Lemma B.4 to compute the rate of $\mathbb{E}[(F_t - F^\star)^2]$. We first verify the assumptions. Since $\chi_t = o(\beta_t)$ by $\chi > \beta$, we know

$$\lim_{i\to\infty} \frac{\beta_i - \frac{\Upsilon_H}{2\gamma_H}\left(\frac{\gamma_H}{2\Upsilon_H}\chi_i + \frac{4\Upsilon_H^{1/2}}{\gamma_H}\left(\Upsilon_H^{1/2} + \frac{C_{g,1}^{1/4}}{\gamma_H^{1/2}}\right)\chi_i + \frac{2^{12}\Upsilon_H^3(\Upsilon_H^2 + C_{g,1}/\gamma_H^2)}{\gamma_H^5}\eta_i^3\right)}{\beta_i} = 1.$$

Since $\beta \in (0,1)$, we have $\lim_{i\to\infty} i\beta_i = \infty$ and (B.2) holds naturally. Furthermore, since $\lim_{i\to\infty} i(1-\beta_{i-1}/\beta_i) = -\beta$ and $\lim_{i\to\infty} i(1 - \chi_{i-1}/\chi_i) = -\chi$, we obtain from Lemma B.1 that $\lim_{i\to\infty} i(1 - \beta_{i-1}^4/\beta_i^4) = -4\beta$ and $\lim_{i\to\infty} i(1 - \chi_{i-1}^4/\chi_i^4) = -4\chi$. Thus, we have

$$\begin{aligned}
\lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}^4/\beta_{i-1}^4}{\chi_i^4/\beta_i^4}\right) &= \lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}^4}{\chi_i^4} + \frac{\chi_{i-1}^4}{\chi_i^4}\left\{1 - \frac{1/\beta_{i-1}^4}{1/\beta_i^4}\right\}\right) = 4(\beta - \chi) < 0, \\
\lim_{i\to\infty} i\left(1 - \frac{\eta_{i-1}^3/\beta_{i-1}}{\eta_i^3/\beta_i}\right) &= \lim_{i\to\infty} i\left(1 - \frac{\eta_{i-1}^3}{\eta_i^3} + \frac{\eta_{i-1}^3}{\eta_i^3}\left\{1 - \frac{1/\beta_{i-1}}{1/\beta_i}\right\}\right) = -2\beta < 0.
\end{aligned}$$

This suggests that (B.1) also holds. Now, we apply Lemma B.4 and obtain

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4\big] &\overset{(D.6)}{\lesssim} \frac{1}{\gamma_H^2}\mathbb{E}\big[(F_t - F^\star)^2\big] \lesssim \frac{1}{\gamma_H^2}\cdot\frac{\Upsilon_H}{\gamma_H}\left(\frac{\Upsilon_H^5 C_{g,2}}{\gamma_H^7}\cdot\frac{\chi_t^4}{\beta_t^4} + \frac{\Upsilon_H^3 C_{g,2}}{\gamma_H^5}\cdot\frac{\eta_t^3}{\beta_t}\right) \\
&\lesssim \frac{\Upsilon_H^4 C_{g,2}}{\gamma_H^8}\beta_t^2 + \frac{\Upsilon_H^6 C_{g,2}}{\gamma_H^{10}}\cdot\frac{\chi_t^4}{\beta_t^4} = O\left(\beta_t^2 + \frac{\chi_t^4}{\beta_t^4}\right).
\end{aligned} \tag{E.7}$$

Part 2: Bound of $\mathbb{E}[\|B_t - B^\star\|^4]$. By the construction of B_t in (7), we have

$$\mathbb{E}\big[\|B_t - B^\star\|^4\big] \lesssim \mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 F_i)\right\|^4\right] + \mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 F_i - \nabla^2 F^\star)\right\|^4\right]. \tag{E.8}$$

For the first term, we note that $\bar{H}_i - \nabla^2 F_i$ is a martingale difference sequence and (14) implies

$$\mathbb{E}\big[\,\big\|\bar{H}_i - \nabla^2 F_i\big\|_F^4\,\big] \lesssim \mathbb{E}\big[\,\big\|\bar{H}_i - \nabla^2 F_i\big\|^4\,\big] \overset{(14)}{\lesssim} C_{H,1}\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4] + C_{H,2}.$$

Therefore, same as in (Chen et al., 2020b, (63)), we apply (Rio, 2008, Theorem 2.1) and obtain

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 F_i)\Big\|^4\Big] &\le \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 F_i)\Big\|_F^4\Big] \lesssim \frac{1}{t^4}\Big[\sum_{i=0}^{t-1}\Big(\mathbb{E}[\|\bar{H}_i - \nabla^2 F_i\|_F^4]\Big)^{1/2}\Big]^2 \\
&\lesssim \frac{1}{t^4}\Big(\sum_{i=0}^{t-1} C_{H,1}^{1/2}(\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4])^{1/2}\Big)^2 + \frac{1}{t^4}\Big(\sum_{i=0}^{t-1} C_{H,2}^{1/2}\Big)^2 \\
&\overset{(E.7)}{\lesssim} \frac{1}{t^2}\Big(\frac{\Upsilon_H^4 C_{g,2}C_{H,1}}{\gamma_H^8}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\beta_i\Big)^2 + \frac{\Upsilon_H^6 C_{g,2}C_{H,1}}{\gamma_H^{10}}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\frac{\chi_i^2}{\beta_i^2}\Big)^2\Big) + \frac{C_{H,2}}{t^2}.
\end{aligned}$$

We only consider the case where $\chi \le 1.5\beta$, otherwise $\chi_t^2/\beta_t^2 = o(\beta_t)$ and all χ_t^2/β_t^2 terms in the following can be absorbed into β_t. We note that

$$\frac{1}{t}\sum_{i=0}^{t-1}\beta_i = \frac{1}{t}\beta_0 + \sum_{i=1}^{t-1}\prod_{j=i+1}^{t-1}\Big(1-\frac{1}{j}\Big)\frac{1}{i}\beta_i \overset{(B.3)}{\lesssim} \frac{1}{1-\beta}\beta_t \quad \text{and} \quad \frac{1}{t}\sum_{i=0}^{t-1}\frac{\chi_i^2}{\beta_i^2} \overset{(B.3)}{\lesssim} \frac{1}{1-2(\chi-\beta)}\cdot\frac{\chi_t^2}{\beta_t^2}, \tag{E.9}$$

where we are able to apply Lemma B.2 since the condition (B.2) is satisfied by $\beta < 1$ and $\chi \le 1.5\beta \Rightarrow 2\chi - 2\beta < 1$. Thus, we combine the above two displays and have

$$\frac{1}{t^2}\Big(\frac{\Upsilon_H^4 C_{g,2}C_{H,1}}{\gamma_H^8}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\beta_i\Big)^2 + \frac{\Upsilon_H^6 C_{g,2}C_{H,1}}{\gamma_H^{10}}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\frac{\chi_i^2}{\beta_i^2}\Big)^2\Big) = o\left(\frac{1}{t^2}\right) \quad \text{and} \quad \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 F_i)\Big\|^4\Big] \lesssim \frac{C_{H,2}}{t^2}. \tag{E.10}$$

For the second term on the right hand side in (E.8), the Υ_L-Lipschitz continuity of $\nabla^2 F(\boldsymbol{x})$ leads to

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 F_i - \nabla^2 F^\star)\Big\|^4\Big] &\le \mathbb{E}\Big[\Big(\frac{1}{t}\sum_{i=0}^{t-1}\|\nabla^2 F_i - \nabla^2 F^\star\|\Big)^4\Big] \le \frac{\Upsilon_L^4}{t^4}\mathbb{E}\Big[\Big(\sum_{i=0}^{t-1}\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|\Big)^4\Big] \\
&\le \frac{\Upsilon_L^4}{t^4}\Big(\sum_{i=0}^{t-1}(\mathbb{E}[\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^4])^{1/4}\Big)^4 \quad \text{(by Hölder's inequality)} \\
&\overset{(E.7)}{\lesssim} \frac{\Upsilon_L^4\Upsilon_H^4 C_{g,2}}{\gamma_H^8}\left(\frac{1}{t}\sum_{i=0}^{t-1}\beta_i^{1/2}\right)^4 + \frac{\Upsilon_L^4\Upsilon_H^6 C_{g,2}}{\gamma_H^{10}}\left(\frac{1}{t}\sum_{i=0}^{t-1}\frac{\chi_i}{\beta_t}\right)^4 \\
&\overset{(E.9)}{\lesssim} \frac{\Upsilon_L^4\Upsilon_H^4 C_{g,2}}{\gamma_H^8}\beta_t^2 + \frac{\Upsilon_L^4\Upsilon_H^6 C_{g,2}}{\gamma_H^{10}}\cdot\frac{\chi_t^4}{\beta_t^4} \quad \text{(by } 1-\beta/2 > 1/2 \text{ and } 1-(\chi-\beta) > 1/2\text{)}.
\end{aligned} \tag{E.11}$$

Plugging (E.10) and (E.11) into (E.8), we have

$$\mathbb{E}[\|B_t - B^\star\|^4] \lesssim \frac{\Upsilon_L^4\Upsilon_H^4 C_{g,2}}{\gamma_H^8}\beta_t^2 + \frac{\Upsilon_L^4\Upsilon_H^6 C_{g,2}}{\gamma_H^{10}}\cdot\frac{\chi_t^4}{\beta_t^4} = O\left(\beta_t^2 + \frac{\chi_t^4}{\beta_t^4}\right). \tag{E.12}$$

This completes the proof.

E.2 Analysis of a dominant term in $\widehat{\Xi}_t$

In this section, we focus on a dominant term of $\widehat{\Xi}_t$ and show that the dominate term converges to the limiting covariance matrix $\Xi^\star$. We first introduce a decomposition of the iterate $\boldsymbol{x}_t$.

Lemma E.1. (Na and Mahoney, 2022, Lemma 5.1) The iterate sequence (6) can be decomposed as

$$\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star = \mathcal{I}_{1,t} + \mathcal{I}_{2,t} + \mathcal{I}_{3,t}, \tag{E.13}$$

where

$$\mathcal{I}_{1,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I - C^\star)\}\, \varphi_i \boldsymbol{\theta}^i, \tag{E.14}$$

$$\mathcal{I}_{2,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I - C^\star)\}\, (\bar{\alpha}_i - \varphi_i)\, \bar{\Delta}\boldsymbol{x}_i, \tag{E.15}$$

$$\mathcal{I}_{3,t} = \prod_{i=0}^{t} \{I - \varphi_i(I - C^\star)\}\, (\boldsymbol{x}_0 - \boldsymbol{x}^\star) + \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I - C^\star)\}\, \varphi_i \boldsymbol{\delta}^i, \tag{E.16}$$

and

$$C^\star = (I - \mathbb{E}[B^\star S(S^T(B^\star)^2 S)^\dagger S^T B^\star])^\tau, \tag{E.17}$$

$$\boldsymbol{\theta}^i = \bar{\Delta}\boldsymbol{x}_i - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_i \mid \mathcal{F}_{i-1}] = -(I - \widetilde{C}_i)B_i^{-1}\bar{g}_i + (I - C_i)B_i^{-1}\nabla F_i, \tag{E.18}$$

$$\boldsymbol{\delta}^i = -(I - C_i)\big\{(B^\star)^{-1}\boldsymbol{\psi}^i + \{B_i^{-1} - (B^\star)^{-1}\}\nabla F_i\big\} + \big(C_i - C^\star\big)(\boldsymbol{x}_i - \boldsymbol{x}^\star), \tag{E.19}$$

$$\boldsymbol{\psi}^i = \nabla F_i - B^\star(\boldsymbol{x}_i - \boldsymbol{x}^\star). \tag{E.20}$$

Here, $\mathcal{I}_{1,t}$ includes the summation of martingale difference sequence; $\mathcal{I}_{2,t}$ characterizes the influence of the adaptive stepsize; and $\mathcal{I}_{3,t}$ encompasses all remaining errors. Based on (E.13), we decompose the following matrix as

$$\frac{1}{t}\sum_{i=1}^{t} \frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i - \boldsymbol{x}^\star)(\boldsymbol{x}_i - \boldsymbol{x}^\star)^T = \sum_{k=1}^{3}\sum_{l=1}^{3} \frac{1}{t}\sum_{i=0}^{t-1} \frac{1}{\varphi_i}\mathcal{I}_{k,i}\mathcal{I}_{l,i}. \tag{E.21}$$

We study the dominant term $\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{1,i}\mathcal{I}_{1,i}^T$ in this section and defer the analysis on the remaining terms to Appendix E.3. The next lemma shows consistency of the dominant term and establishes the convergence rate, with proof provided in Appendix E.2.1.

Lemma E.2. Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \ge \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1 - \gamma_S$, and the stepsize parameters satisfy $\beta \in (0,1)$, $\chi > \beta$, and $c_\beta, c_\chi > 0$. Then, we have

$$\mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{1,i}\mathcal{I}_{1,i}^T - \Xi^\star\right\|\right] \lesssim \begin{cases} \sqrt{\beta_t} + \chi_t/\beta_t, & \beta \in (0, 0.5], \\ 1/\sqrt{t\beta_t} + \chi_t/\beta_t, & \beta \in (0.5, 1). \end{cases}$$

E.2.1 Proof of Lemma E.2

We define

$$\widetilde{C}_{k,j}^{\star} = I - \left(B^{\star}S_{k,j}(S_{k,j}^{T}(B^{\star})^{2}S_{k,j})^{\dagger}S_{k,j}^{T}B^{\star}\right) \quad \text{and} \quad \widetilde{C}_{k}^{\star} = \prod_{j=1}^{\tau}\widetilde{C}_{k,j}^{\star}, \tag{E.22}$$

where $S_{k,j}$ is the same sketching matrix in $\widetilde{C}_{k,j}$ at (B.5). It is easy to verify $\mathbb{E}\widetilde{C}_{k}^{\star} = C^{\star}$ with $C^{\star}$ defined in (E.17). We also define

$$\widetilde{\boldsymbol{\theta}}^{k} = -(I - \widetilde{C}_{k}^{\star})(B^{\star})^{-1}\nabla f(\boldsymbol{x}^{\star};\xi_{k}) \quad \text{and} \quad \widehat{\boldsymbol{\theta}}^{k} = \boldsymbol{\theta}^{k} - \widetilde{\boldsymbol{\theta}}^{k}. \tag{E.23}$$

Basically, $\widetilde{\boldsymbol{\theta}}^{k}$ and $\boldsymbol{\theta}^{k}$ share the same randomness but $\widetilde{\boldsymbol{\theta}}^{k}$ is constructed at $\boldsymbol{x}^{\star}$ instead of $\boldsymbol{x}_{k}$, which means we use the *iid* copies $\{\widetilde{\boldsymbol{\theta}}^{k}\}_{k}$ to approximate the martingale difference sequence $\{\boldsymbol{\theta}^{k}\}_{k}$. We decompose $\mathcal{I}_{1,i}$ as

$$\mathcal{I}_{1,i} = \sum_{k=0}^{i}\prod_{l=k+1}^{i}\{I - \varphi_{l}(I - C^{\star})\}\varphi_{k}\widetilde{\boldsymbol{\theta}}^{k} + \sum_{k=0}^{i}\prod_{l=k+1}^{i}\{I - \varphi_{l}(I - C^{\star})\}\varphi_{k}\widehat{\boldsymbol{\theta}}^{k} =: \widetilde{\mathcal{I}}_{1,i} + \widehat{\mathcal{I}}_{1,i}. \tag{E.24}$$

Intuitively, as $\boldsymbol{x}_{i}$ converges to $\boldsymbol{x}^{\star}$, $\widetilde{\mathcal{I}}_{1,i}$ should be a good approximation to $\mathcal{I}_{1,i}$ and $\widehat{\mathcal{I}}_{1,i}$ should be negligible. The next two lemma provide bounds for $\widetilde{\mathcal{I}}_{1,i}$ and $\widehat{\mathcal{I}}_{1,i}$, respectively. The proofs are provided in Appendices E.2.2 and E.2.3.

Lemma E.3. Under the assumptions of Lemma E.2, we have

$$\mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\widetilde{\mathcal{I}}_{1,i}\widetilde{\mathcal{I}}_{1,i}^{T} - \Xi^{\star}\right\|\right] \lesssim \begin{cases}\beta_{t}, & \beta\in(0,1/3],\\ 1/\sqrt{t\beta_{t}}, & \beta\in(1/3,1).\end{cases}$$

Lemma E.4. Under the assumptions of Lemma E.2, we have

$$\mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\widehat{\mathcal{I}}_{1,i}\widehat{\mathcal{I}}_{1,i}^{T}\right\|\right] \leq \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\mathbb{E}[\|\widehat{\mathcal{I}}_{1,i}\|^{2}] \lesssim \beta_{t} + \frac{\chi_{t}^{2}}{\beta_{t}^{2}}.$$

By the decomposition (E.24), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\mathcal{I}_{1,i}\mathcal{I}_{1,i}^{T} - \Xi^{\star}\right\|\right] &\leq \mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\widetilde{\mathcal{I}}_{1,i}\widetilde{\mathcal{I}}_{1,i}^{T} - \Xi^{\star}\right\|\right]\\ &\quad + \mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\widehat{\mathcal{I}}_{1,i}\widehat{\mathcal{I}}_{1,i}^{T}\right\|\right] + 2\mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\widetilde{\mathcal{I}}_{1,i}\widehat{\mathcal{I}}_{1,i}^{T}\right\|\right].\end{aligned} \tag{E.25}$$

We apply Hölder's inequality twice to the last term and obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\widetilde{\mathcal{I}}_{1,i}\widehat{\mathcal{I}}_{1,i}^{T}\right\|\right] &\leq \mathbb{E}\left[\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\|\widetilde{\mathcal{I}}_{1,i}\|\|\widehat{\mathcal{I}}_{1,i}\|\right]\\ &\leq \mathbb{E}\left[\sqrt{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\|\widetilde{\mathcal{I}}_{1,i}\|^{2}}\sqrt{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\|\widehat{\mathcal{I}}_{1,i}\|^{2}}\right] \leq \sqrt{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\mathbb{E}\|\widetilde{\mathcal{I}}_{1,i}\|^{2}}\sqrt{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i}}\mathbb{E}\|\widehat{\mathcal{I}}_{1,i}\|^{2}}.\end{aligned} \tag{E.26}$$

Given Lemmas E.3 and E.4, it suffices to bound $\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\|\widetilde{\mathcal{I}}_{1,i}\|^2$. We have

$$\begin{aligned}
\mathbb{E}\big[\|\widetilde{\mathcal{I}}_{1,i}\|^2\big] &\overset{(E.24)}{=} \sum_{k_1,k_2=0}^{i}\varphi_{k_1}\varphi_{k_2}\mathbb{E}\Big[(\widetilde{\boldsymbol{\theta}}^{k_1})^T\Big(\prod_{l_1=k_1+1}^{i}\{I-\varphi_{l_1}(I-C^\star)\}\Big)^T\Big(\prod_{l_2=k_2+1}^{i}\{I-\varphi_{l_2}(I-C^\star)\}\Big)\widetilde{\boldsymbol{\theta}}^{k_2}\Big] \\
&= \sum_{k=0}^{i}\varphi_k^2\mathbb{E}\Big[\Big\|\prod_{l=k+1}^{i}\{I-\varphi_l(I-C^\star)\}\widetilde{\boldsymbol{\theta}}^k\Big\|^2\Big] \leq \sum_{k=0}^{i}\varphi_k^2\prod_{l=k+1}^{i}\big\|I-\varphi_l(I-C^\star)\big\|^2\mathbb{E}\big[\|\widetilde{\boldsymbol{\theta}}^k\|^2\big] \\
&\leq \sum_{k=0}^{i}\prod_{l=k+1}^{i}\big(1-(1-\rho)^\tau\varphi_l\big)^2\varphi_k^2\mathbb{E}\big[\|\widetilde{\boldsymbol{\theta}}^k\|^2\big], \qquad \text{(E.27)}
\end{aligned}$$

where the second equality uses the fact that $\{\widetilde{\boldsymbol{\theta}}^k\}_k$ are mean zero and independent, and the last inequality uses the fact $\varphi_t \leq \eta_t \leq 1$ (cf. Appendix 4.2) and Lemma B.5(d). Note that $\varphi_t \leq 1$ is not essential; given $\varphi_t \to 0$, we can apply Lemma B.4 to derive the same results without this condition. Next, we bound the moment of $\|\widetilde{\boldsymbol{\theta}}^k\|$. We note for $m = 2, 4$ and any $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^m\big] \lesssim \mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_k)-\nabla f(\boldsymbol{x}_k;\xi_k)\|^m\big] + \mathbb{E}\big[\|\nabla f(\boldsymbol{x}_k;\xi_k)-\nabla F_k\|^m\big] + \mathbb{E}\big[\|\nabla F_k-\nabla F^\star\|^m\big] \\
&\lesssim \Upsilon_H^m\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^m\big] + C_{g,1}^{m/4}\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^m\big] + C_{g,2}^{m/4} + \Upsilon_H^m\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^m\big] \quad \text{(Assumptions 3.2, 3.3)} \\
&\lesssim C_{g,2}^{m/4} \qquad \big(\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^m\big] = o(1) \text{ by Lemma 4.2}\big). \qquad \text{(E.28)}
\end{aligned}$$

Then, by (15) and Lemma B.5(c), we have for $m = 2, 4$

$$\mathbb{E}\big[\|\widetilde{\boldsymbol{\theta}}^k\|^m\big] \overset{(E.23)}{\leq} \mathbb{E}\big[\|I-\widetilde{C}^\star\|^m\|(B^\star)^{-1}\|^m\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^m\big] \leq \frac{2^m}{\gamma_H^m}\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^m\big] \overset{(E.28)}{\lesssim} \frac{C_{g,2}^{m/4}}{\gamma_H^m}. \qquad \text{(E.29)}$$

Plugging (E.29) into (E.27), we get

$$\begin{aligned}
\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\widetilde{\mathcal{I}}_{1,i}\|^2\big] &\leq \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\varphi_k^2\mathbb{E}\big[\|\widetilde{\boldsymbol{\theta}}^k\|^2\big] \\
&\lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\frac{1}{\varphi_i}\sum_{k=0}^{t}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\varphi_k^2}_{\longrightarrow 0.5/(1-\rho^\tau) \quad \text{by Lemma B.2}} \lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)}, \qquad \text{(E.30)}
\end{aligned}$$

where the last inequality uses the fact that $\lim_{t\to\infty}\sum_{i=0}^{t-1}a_i/t = a$ if $\lim_{t\to\infty}a_t = a$. Combining (E.26), (E.30), and Lemma E.4 (particularly (E.40) in the proof), we derive

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{I}}_{1,i}\widehat{\mathcal{I}}_{1,i}^T\Big\|\Big] \lesssim \frac{C_{g,2}^{1/4}C_{\widehat{\boldsymbol{\theta}}}^{1/2}}{\gamma_H(1-\rho^\tau)}\left(\frac{1}{(1-\beta)^{1/2}}\sqrt{\beta_t} + \frac{\Upsilon_H^{1/2}\mathbf{1}_{\{\chi\leq 1.5\beta\}}}{\gamma_H^{1/2}(1-2(\chi-\beta))^{1/2}}\cdot\frac{\chi_t}{\beta_t}\right)$$

with a constant $C_{\widehat{\boldsymbol{\theta}}} > 0$ defined later in (E.39). Finally, combining Lemma E.3 ((E.38) in the proof), Lemma E.4 ((E.40) in the proof), and (E.25), we have

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{1,i}\mathcal{I}_{1,i}^T-\Xi^\star\Big\|\Big]$$
$$\lesssim \begin{cases} \frac{C_{g,2}^{1/4}C_{\widehat{\boldsymbol{\theta}}}^{1/2}}{\gamma_H(1-\rho^\tau)}\sqrt{\beta_t}+\frac{\Upsilon_H^{1/2}C_{g,2}^{1/4}C_{\widehat{\boldsymbol{\theta}}}^{1/2}\mathbf{1}_{\{\chi\le 1.5\beta\}}}{\gamma_H^{3/2}(1-\rho^\tau)}\cdot\frac{\chi_t}{\beta_t}=O\left(\sqrt{\beta_t}+\frac{\chi_t}{\beta_t}\right), & \beta\in(0,0.5),\\ \max\left(\frac{C_{g,2}^{1/4}C_{\widehat{\boldsymbol{\theta}}}^{1/2}}{\gamma_H(1-\rho^\tau)},\frac{\max(\|\Lambda\|_F,C_{g,2}^{1/2}/\gamma_H^2)}{c_\beta(1-\rho^\tau)^{3/2}}\right)\sqrt{\beta_t}+\frac{\Upsilon_H^{1/2}C_{g,2}^{1/4}C_{\widehat{\boldsymbol{\theta}}}^{1/2}\mathbf{1}_{\{\chi\le 1.5\beta\}}}{\gamma_H^{3/2}(1-\rho^\tau)}\cdot\frac{\chi_t}{\beta_t}=O\left(\sqrt{\beta_t}+\frac{\chi_t}{\beta_t}\right), & \beta=0.5,\\ \frac{\max(\|\Lambda\|_F,C_{g,2}^{1/2}/\gamma_H^2)}{(1-\rho^\tau)^{3/2}}\cdot\frac{1}{\sqrt{t\beta_t}}+\frac{\Upsilon_H^{1/2}C_{g,2}^{1/4}C_{\widehat{\boldsymbol{\theta}}}^{1/2}\mathbf{1}_{\{\chi\le 1.5\beta\}}}{\gamma_H^{3/2}(1-\rho^\tau)(1-2(\chi-\beta))^{1/2}}\cdot\frac{\chi_t}{\beta_t}=O\left(\frac{1}{\sqrt{t\beta_t}}+\frac{\chi_t}{\beta_t}\right), & \beta\in(0.5,1), \end{cases} \tag{E.31}$$

where $\Lambda=\mathbb{E}[(I-\widetilde{C}^*)\Omega^\star(I-\widetilde{C}^*)^T]$. This completes the proof.

E.2.2 Proof of Lemma E.3

By the eigenvalue decomposition $I-C^\star=U\Sigma U^T$ with $\Sigma=\mathrm{diag}(\sigma_1,\ldots,\sigma_d)$ in (17), we have

$$\widetilde{\mathcal{I}}_{1,i}=\sum_{k=0}^{i}\prod_{l=k+1}^{i}\{I-\varphi_l(I-C^\star)\}\varphi_k\widetilde{\boldsymbol{\theta}}^k=U\sum_{k=0}^{i}\prod_{l=k+1}^{i}\{I-\varphi_l\Sigma\}\varphi_kU^T\widetilde{\boldsymbol{\theta}}^k. \tag{E.32}$$

Let $\widetilde{\mathcal{Q}}_i=U^T\widetilde{\mathcal{I}}_{1,i}$ and $\Gamma=U^T\Lambda U$ with $\Lambda=\mathbb{E}[(I-\widetilde{C}^*)\Omega^\star(I-\widetilde{C}^*)^T]$. Recalling the expression of $\Xi^\star$ in (20), we get

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{I}}_{1,i}\widetilde{\mathcal{I}}_{1,i}^T-\Xi^\star\Big\|\Big]=\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_i\widetilde{\mathcal{Q}}_i^T-\Theta\circ\Gamma\Big\|\Big]\le\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_i\widetilde{\mathcal{Q}}_i^T-\Theta\circ\Gamma\Big\|_F\Big]$$
$$\le\sqrt{\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_i\widetilde{\mathcal{Q}}_i^T-\Theta\circ\Gamma\Big\|_F^2\Big]}\quad\text{(by Hölder's inequality)}.$$

We perform bias-variance decomposition on this term:

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_i\widetilde{\mathcal{Q}}_i^T-\Theta\circ\Gamma\Big\|_F^2\Big]=\sum_{p,q=1}^{d}\mathbb{E}\Big[\Big(\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_{i,p}\widetilde{\mathcal{Q}}_{i,q}-\Theta_{p,q}\Gamma_{p,q}\Big)^2\Big]=:I+II, \tag{E.33}$$

with

$$I=\sum_{p,q=1}^{d}\left\{\mathbb{E}\Big[\Big(\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_{i,p}\widetilde{\mathcal{Q}}_{i,q}\Big)^2\Big]-\Big(\mathbb{E}\Big[\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_{i,p}\widetilde{\mathcal{Q}}_{i,q}\Big]\Big)^2\right\}\quad\text{(variance)},$$
$$II=\sum_{p,q=1}^{d}\Big(\mathbb{E}\Big[\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_{i,p}\widetilde{\mathcal{Q}}_{i,q}\Big]-\Theta_{p,q}\Gamma_{p,q}\Big)^2\quad\text{(bias}^2\text{)},$$

and $\widetilde{\mathcal{Q}}_{i,p}$ and $\widetilde{\mathcal{Q}}_{i,q}$ represent the p-th and q-th elements in $\widetilde{\mathcal{Q}}_i$. We first look at II. By (E.32), we get

$$\mathbb{E}\Big[\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_{i,p}\widetilde{\mathcal{Q}}_{i,q}\Big] = \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\sum_{k_1=0}^{i}\sum_{k_2=0}^{i}\prod_{l_1=k_1+1}^{i}(1-\sigma_p\varphi_{l_1})\cdot \prod_{l_2=k_2+1}^{i}(1-\sigma_q\varphi_{l_2})\varphi_{k_1}\varphi_{k_2}\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1}\widetilde{\boldsymbol{\theta}}^{k_2^T}U\big)_{p,q}\Big].$$

Given the definition of $\widetilde{\boldsymbol{\theta}}^k$ in (E.23) and the independence among $\{\widetilde{\boldsymbol{\theta}}^k\}_k$, it is observed that

$$\mathbb{E}[U^T\widetilde{\boldsymbol{\theta}}^{k_1}\widetilde{\boldsymbol{\theta}}^{k_2^T}U] = 0 \text{ for } k_1\neq k_2 \qquad\text{and}\qquad \mathbb{E}[U^T\widetilde{\boldsymbol{\theta}}^{k}\widetilde{\boldsymbol{\theta}}^{kT}U] = \Gamma.$$

Thus, combining the above two displays leads to

$$\mathbb{E}\Big[\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{Q}}_{i,p}\widetilde{\mathcal{Q}}_{i,q}\Big] = \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-\sigma_p\varphi_l)(1-\sigma_q\varphi_l)\varphi_k^2\Gamma_{p,q}.$$

We plug the above display into the term II and apply Lemma B.3 to bound it. For $\beta\in(0.5,1)$, we have

$$\begin{aligned}|II| &\leq \sum_{p,q=1}^{d}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\Big|\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-\sigma_p\varphi_l)(1-\sigma_q\varphi_l)\varphi_k^2-\Theta_{p,q}\Big|\Big)^2\Gamma_{p,q}^2\\ &\lesssim \sum_{p,q=1}^{d}\Big(\frac{\beta}{(\sigma_p+\sigma_q)^2}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{i\varphi_i}\Big)^2\Gamma_{p,q}^2 \overset{\text{(E.9)}}{\lesssim} \sum_{p,q=1}^{d}\Big(\frac{\beta}{(\sigma_p+\sigma_q)^2}\cdot\frac{1}{1-(1-\beta)}\cdot\frac{1}{t\varphi_t}\Big)^2\Gamma_{p,q}^2\\ &\lesssim \frac{\|\Gamma\|_F^2}{(1-\rho^\tau)^4}\cdot\frac{1}{t^2\varphi_t^2}\lesssim\frac{\|\Lambda\|_F^2}{(1-\rho^\tau)^4}\cdot\frac{1}{t^2\beta_t^2}\quad\text{(by Lemma B.5(d) and }\chi_t=o(\beta_t)).\end{aligned}\tag{E.34}$$

Applying Lemma B.3 for $\beta\in(0,0.5)$ and $\beta=0.5$, we similarly obtain

$$|II|\lesssim\|\Lambda\|_F^2\beta_t^2 \text{ for } \beta\in(0,0.5)\qquad\text{and}\qquad |II|\lesssim\Big(1+\frac{\beta/c_\beta^2}{2(1-\rho^\tau)^2}\Big)^2\|\Lambda\|_F^2\beta_t^2 \text{ for } \beta=0.5.\tag{E.35}$$

Now we deal with the term I. By (E.32), we expand I as

$$I=\sum_{p,q=1}^{d}\frac{1}{t^2}\sum_{i_1,i_2=0}^{t-1}\frac{1}{\varphi_{i_1}}\frac{1}{\varphi_{i_2}}\sum_{k_1,k_1'=0}^{i_1}\sum_{k_2,k_2'=0}^{i_2}\prod_{l_1=k_1+1}^{i_1}(1-\sigma_p\varphi_{l_1})\prod_{l_1'=k_1'+1}^{i_1}(1-\sigma_q\varphi_{l_1'})\prod_{l_2=k_2+1}^{i_2}(1-\sigma_p\varphi_{l_2})\prod_{l_2'=k_2'+1}^{i_2}(1-\sigma_q\varphi_{l_2'})$$
$$\varphi_{k_1}\varphi_{k_1'}\varphi_{k_2}\varphi_{k_2'}\Big\{\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1}\widetilde{\boldsymbol{\theta}}^{k_1'^T}U\big)_{p,q}\big(U^T\widetilde{\boldsymbol{\theta}}^{k_2}\widetilde{\boldsymbol{\theta}}^{k_2'^T}U\big)_{p,q}\Big]-\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1}\widetilde{\boldsymbol{\theta}}^{k_1'^T}U\big)_{p,q}\Big]\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_2}\widetilde{\boldsymbol{\theta}}^{k_2'^T}U\big)_{p,q}\Big]\Big\}.$$

It is noteworthy that the term in the curly braces is nonzero only when the indices k_1,k_1',k_2,k_2' are pairwise identical. Thus, we decompose I into four terms I_1,I_2,I_3,I_4 by classifying the indices.

- **Term 1:** $k_1=k_1'=k_2=k_2'$.

Summing over all the indices under this case, we get

$$
\begin{aligned}
|I_1| = \sum_{p,q} \frac{1}{t^2} \sum_{i_1=0}^{t-1} \sum_{i_2=0}^{t-1} \frac{1}{\varphi_{i_1}} \frac{1}{\varphi_{i_2}} \sum_{k=0}^{i_1 \wedge i_2} \prod_{l_1=k+1}^{i_1} (1-\sigma_p \varphi_{l_1})(1-\sigma_q \varphi_{l_1}) \cdot \\
\prod_{l_2=k+1}^{i_2} (1-\sigma_p\varphi_{l_2})(1-\sigma_q\varphi_{l_2})\varphi_k^4 \Big\{ \mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^k \widetilde{\boldsymbol{\theta}}^{k^T} U\big)_{p,q}^2 \Big] - \Big(\mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^k \widetilde{\boldsymbol{\theta}}^{k^T} U\big)_{p,q} \Big] \Big)^2 \Big\} \\
\leq \frac{1}{t^2} \sum_{i_1=0}^{t-1} \sum_{i_2=0}^{t-1} \frac{1}{\varphi_{i_1}} \frac{1}{\varphi_{i_2}} \sum_{k=0}^{i_1 \wedge i_2} \prod_{l_1=k+1}^{i_1} (1-(1-\rho^\tau)\varphi_{l_1})^2 \prod_{l_2=k+1}^{i_2} (1-(1-\rho^\tau)\varphi_{l_2})^2 \varphi_k^4 \mathbb{E}\Big[\sum_{p,q} \big(U^T \widetilde{\boldsymbol{\theta}}^k \widetilde{\boldsymbol{\theta}}^{k^T} U\big)_{p,q}^2 \Big].
\end{aligned}
$$

Here, the equality holds because $1 - \sigma_k \varphi_t > 0$ for any $1 \leq k \leq d$ and $t \geq 0$ following the same discussion as in (E.27). By (E.29), we know

$$
\mathbb{E}\Big[\sum_{p,q} \big(U^T \widetilde{\boldsymbol{\theta}}^k \widetilde{\boldsymbol{\theta}}^{k^T} U\big)_{p,q}^2 \Big] = \mathbb{E}[\|\widetilde{\boldsymbol{\theta}}^k\|^4] \lesssim \frac{C_{g,2}}{\gamma_H^4}.
$$

Due to the symmetry between the indices i_1 and i_2, $|I_1|$ can be further bounded by

$$
\begin{aligned}
|I_1| \leq \frac{2}{t^2} \sum_{i_1=0}^{t-1} \frac{1}{\varphi_{i_1}} \sum_{i_2=0}^{i_1} \frac{1}{\varphi_{i_2}} \prod_{l_2=i_2+1}^{i_1} (1-(1-\rho^\tau)\varphi_{l_2})^2 \underbrace{\sum_{k=0}^{i_2} \prod_{l_1=k+1}^{i_2} (1-(1-\rho^\tau)\varphi_{l_1})^4 \varphi_k^4}_{\lesssim \varphi_{i_2}^3/(1-\rho^\tau) \quad \text{by Lemma B.2}} \mathbb{E}[\|\widetilde{\boldsymbol{\theta}}^k\|^4] \\
\lesssim \frac{C_{g,2}}{\gamma_H^4 (1-\rho^\tau)} \cdot \frac{1}{t^2} \sum_{i_1=0}^{t-1} \underbrace{\frac{1}{\varphi_{i_1}} \sum_{i_2=0}^{i_1} \prod_{l_2=i_2+1}^{i_1} (1-(1-\rho^\tau))\varphi_{l_2})^2 \varphi_{i_2}^2}_{\longrightarrow 0.5/(1-\rho^\tau) \quad \text{by Lemma B.2}} \lesssim \frac{C_{g,2}}{\gamma_H^4 (1-\rho^\tau)^2} \cdot \frac{1}{t}.
\end{aligned}
$$

- **Term 2:** $k_1 = k_1', k_2 = k_2', k_1 \neq k_2$.

We note that

$$
\mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^{k_1} \widetilde{\boldsymbol{\theta}}^{k_1^T} U\big)_{p,q} \big(U^T \widetilde{\boldsymbol{\theta}}^{k_2} \widetilde{\boldsymbol{\theta}}^{k_2^T} U\big)_{p,q} \Big] = \mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^{k_1} \widetilde{\boldsymbol{\theta}}^{k_1^T} U\big)_{p,q} \Big] \mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^{k_2} \widetilde{\boldsymbol{\theta}}^{k_2^T} U\big)_{p,q} \Big] \ \text{ for } k_1 \neq k_2.
$$

This indicates that $I_2 = 0$.

- **Term 3:** $k_1 = k_2, k_1' = k_2', k_1 \neq k_1'$.

In this case, it is observed that

$$
\mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^{k_1} \widetilde{\boldsymbol{\theta}}^{k_1'^T} U\big)_{p,q} \Big] \mathbb{E}\Big[\big(U^T \widetilde{\boldsymbol{\theta}}^{k_1} \widetilde{\boldsymbol{\theta}}^{k_1'^T} U\big)_{p,q} \Big] = 0.
$$

Thus, we have

$$
\begin{aligned}
|I_3| \leq \frac{1}{t^2} \sum_{i_1=0}^{t-1} \sum_{i_2=0}^{t-1} \frac{1}{\varphi_{i_1}} \frac{1}{\varphi_{i_2}} \sum_{k_1=0}^{i_1 \wedge i_2} \prod_{l_1=k_1+1}^{i_1} (1-(1-\rho^\tau)\varphi_{l_1}) \prod_{l_2=k_1+1}^{i_2} (1-(1-\rho^\tau)\varphi_{l_2})\varphi_{k_1}^2 \cdot \\
\sum_{k_1'=0, k_1 \neq k_1'}^{i_1 \wedge i_2} \prod_{l_1'=k_1'+1}^{i_1} (1-(1-\rho^\tau)\varphi_{l_1'}) \prod_{l_2'=k_1'+1}^{i_2} (1-(1-\rho^\tau)\varphi_{l_2'})\varphi_{k_1'}^2 \mathbb{E}\Big[\sum_{p,q} \big(U^T \widetilde{\boldsymbol{\theta}}^{k_1}\big)_p^2 \big(U^T \widetilde{\boldsymbol{\theta}}^{k_1'}\big)_q^2 \Big].
\end{aligned}
$$

Since $k_1 \neq k_1'$, we have

$$\mathbb{E}\Big[\sum_{p,q}\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1}\big)_p^2\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1'}\big)_q^2\Big] = \mathbb{E}\big[\|\widetilde{\boldsymbol{\theta}}^{k_1}\|^2\big]\mathbb{E}\big[\|\widetilde{\boldsymbol{\theta}}^{k_1'}\|^2\big] \overset{(E.29)}{\lesssim} \frac{C_{g,2}}{\gamma_H^4}.$$

By the symmetry of the indices i_1 and i_2, we can further bound $|I_3|$ by

$$\begin{aligned}
|I_3| &\lesssim \frac{1}{t^2}\sum_{i_1=0}^{t-1}\frac{1}{\varphi_{i_1}}\sum_{i_2=0}^{i_1}\frac{1}{\varphi_{i_2}}\prod_{l=i_2+1}^{i_1}(1-(1-\rho^\tau)\varphi_l)^2\underbrace{\Big\{\sum_{k_1=0}^{i_2}\prod_{l_1=k_1+1}^{i_2}(1-(1-\rho^\tau)\varphi_{l_1})^2\varphi_{k_1}^2\Big\}}_{\lesssim\varphi_{i_2}/(1-\rho^\tau)\quad\text{by Lemma B.2}}{}^2\cdot\frac{C_{g,2}}{\gamma_H^4} \\
&\lesssim \frac{C_{g,2}}{\gamma_H^4(1-\rho^\tau)^2}\cdot\frac{1}{t^2}\sum_{i_1=0}^{t-1}\frac{1}{\varphi_{i_1}}\underbrace{\sum_{i_2=0}^{i_1}\prod_{l=i_2+1}^{i_1}(1-(1-\rho^\tau)\varphi_l)^2\varphi_{i_2}}_{\longrightarrow 0.5/(1-\rho^\tau)\quad\text{by Lemma B.2}} \\
&\lesssim \frac{C_{g,2}}{\gamma_H^4(1-\rho^\tau)^3}\cdot\frac{1}{t^2}\sum_{i_1=0}^{t-1}\frac{1}{\varphi_{i_1}} \lesssim \frac{C_{g,2}}{c_\beta\gamma_H^4(1-\rho^\tau)^3}\cdot\frac{1}{t^2}\sum_{i_1=0}^{t-1}(i_1+1)^\beta \lesssim \frac{C_{g,2}}{\gamma_H^4(1-\rho^\tau)^3}\cdot\frac{1}{t\beta_t}.
\end{aligned} \tag{E.36}$$

• **Term 4:** $k_1 = k_2', k_2 = k_1', k_1 \neq k_2$.
In this case, we have

$$\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1}\widetilde{\boldsymbol{\theta}}^{k_1'T}U\big)_{p,q}\Big]\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1'}\widetilde{\boldsymbol{\theta}}^{k_1^T}U\big)_{p,q}\Big] = 0.$$

The analysis of I_4 is almost identical to I_3, only with the expectation term being replaced by

$$\sum_{p,q}\mathbb{E}\Big[\big(U^T\widetilde{\boldsymbol{\theta}}^{k_1}\widetilde{\boldsymbol{\theta}}^{k_1^T}U\big)_{p,q}\big(U^T\widetilde{\boldsymbol{\theta}}^{k_2}\widetilde{\boldsymbol{\theta}}^{k_2^T}U\big)_{p,q}\Big] = \sum_{p,q}\Gamma_{p,q}^2 = \|\Gamma\|_F^2 = \|\Lambda\|_F^2 \text{ when } k_1 \neq k_2.$$

Therefore, we conclude that

$$|I_4| \lesssim \frac{\|\Lambda\|_F^2}{(1-\rho^\tau)^3}\cdot\frac{1}{t\beta_t}.$$

Combining the analyses of four terms, we obtain

$$|I| \le \sum_{i=1}^4 |I_i| \lesssim \frac{\max(\|\Lambda\|_F^2, C_{g,2}/\gamma_H^4)}{(1-\rho^\tau)^3}\cdot\frac{1}{t\beta_t}. \tag{E.37}$$

Plugging (E.34), (E.35), and (E.37) into to (E.33), we have

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\widetilde{\mathcal{I}}_{1,i}\widetilde{\mathcal{I}}_{1,i}^T - \Xi^\star\Big\|\Big] \lesssim \begin{cases} \|\Lambda\|_F\beta_t = O(\beta_t), & \beta\in(0,1/3), \\ \max\left(\|\Lambda\|_F, \dfrac{\max(\|\Lambda\|_F, C_{g,2}^{1/2}/\gamma_H^2)}{c_\beta^{3/2}(1-\rho^\tau)^{3/2}}\right)\beta_t = O(\beta_t), & \beta = 1/3, \\ \dfrac{\max(\|\Lambda\|_F, C_{g,2}^{1/2}/\gamma_H^2)}{(1-\rho^\tau)^{3/2}}\cdot\dfrac{1}{\sqrt{t\beta_t}} = O(1/\sqrt{t\beta_t}), & \beta\in(1/3,1). \end{cases} \tag{E.38}$$

We complete the proof.

E.2.3 Proof of Lemma E.4

We present the following lemma to bound $\widehat{\boldsymbol{\theta}}^k$ defined in (E.23), with proof deferred to Appendix E.2.4.

Lemma E.5. Under the assumptions of Lemma E.2, we have

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^k\|^2\big] \lesssim \frac{\Upsilon_H^2}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2\big] + \frac{\tau^2\Upsilon_S C_{g,2}^{1/2}}{\gamma_H^4}\mathbb{E}\big[\|B_k - B^\star\|^2\big].$$

This lemma indicates that the difference between the martingale difference $\boldsymbol{\theta}^k$ and its approximation $\widetilde{\boldsymbol{\theta}}^k$ vanishes. Combining Lemma E.5 with (E.7) and (E.12) in the proof of Lemma 4.2, we get

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^k\|^2\big] \lesssim C_{\widehat{\boldsymbol{\theta}}}\left(\beta_k + \frac{\Upsilon_H}{\gamma_H}\cdot\frac{\chi_k^2}{\beta_k^2}\right) \quad \text{with} \quad C_{\widehat{\boldsymbol{\theta}}} = \frac{\Upsilon_H^2 C_{g,2}^{1/2}}{\gamma_H^6}\max\left(\Upsilon_H^2, \frac{\tau^2\Upsilon_S\Upsilon_L^2 C_{g,2}^{1/2}}{\gamma_H^2}\right). \tag{E.39}$$

Recall the expression of $\widehat{\mathcal{I}}_{1,i}$ in (E.24). Since $\{\widehat{\boldsymbol{\theta}}^k\}_k$ is a martingale difference sequence, we follow the analysis in (E.27) and (E.30), and obtain

$$\begin{aligned}
&\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\widehat{\mathcal{I}}_{1,i}\|^2\big] \le \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\varphi_k^2\mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^k\|^2\big] \\
&\lesssim C_{\widehat{\boldsymbol{\theta}}}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\varphi_k^2\beta_k}_{\lesssim\beta_i/(1-\rho^\tau)\quad\text{by Lemma B.2}} + \frac{\Upsilon_H C_{\widehat{\boldsymbol{\theta}}}}{\gamma_H}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\frac{\varphi_k^2\chi_k^2}{\beta_k^2}}_{\lesssim(\chi_i^2/\beta_i^2)/(1-\rho^\tau)\quad\text{by Lemma B.2}} \\
&\overset{\text{(E.9)}}{\lesssim} \frac{C_{\widehat{\boldsymbol{\theta}}}}{(1-\rho^\tau)(1-\beta)}\beta_t + \frac{\Upsilon_H C_{\widehat{\boldsymbol{\theta}}}\mathbf{1}_{\{\chi\le1.5\beta\}}}{\gamma_H(1-\rho^\tau)(1-2(\chi-\beta))}\cdot\frac{\chi_t^2}{\beta_t^2} = O\left(\beta_t + \frac{\chi_t^2}{\beta_t^2}\right).
\end{aligned} \tag{E.40}$$

We complete the proof.

E.2.4 Proof of Lemma E.5

We expand $\widehat{\boldsymbol{\theta}}^k$ based on its definition in (E.23) as

$$\widehat{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^k - \widetilde{\boldsymbol{\theta}}^k = (I - C_k)B_k^{-1}\nabla F_k - (I - \widetilde{C}_k)B_k^{-1}\nabla f(\boldsymbol{x}_k;\xi_k) + (I - \widetilde{C}_k^\star)(B^\star)^{-1}\nabla f(\boldsymbol{x}^\star;\xi_k).$$

Then, we can bound $\|\widehat{\boldsymbol{\theta}}^k\|^2$ as

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}^k\|^2 &\lesssim \|I - C_k\|^2\|B_k^{-1}\|^2\|\nabla F_k\|^2 + \|I - \widetilde{C}_k\|^2\|B_k^{-1}\|^2\|\nabla f(\boldsymbol{x}_k;\xi_k) - \nabla f(\boldsymbol{x}^\star;\xi_k)\|^2 \\
&\quad + \|(I - \widetilde{C}_k)B_k^{-1} - (1 - \widetilde{C}_k^\star)(B^\star)^{-1}\|^2\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^2 =: I + II + III.
\end{aligned}$$

For the first two terms, by Assumption 3.3 and Lemma B.5(c), we get

$$\mathbb{E}[I] \lesssim \frac{\Upsilon_H^2}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2\big] \qquad \text{and} \qquad \mathbb{E}[II] \lesssim \frac{\Upsilon_H^2}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2\big]. \tag{E.41}$$

Regarding the term III, we have

$$\begin{aligned}
\|(I - \widetilde{C}_k)B_k^{-1} - (1 - \widetilde{C}_k^\star)(B^\star)^{-1}\|^2 &\le \|I - \widetilde{C}_k\|^2\|B_k^{-1}\|^2\|(B^\star)^{-1}\|^2\|B_k - B^\star\|^2 \\
&\quad + \|(B^\star)^{-1}\|^2\|\widetilde{C}_k - \widetilde{C}_k^\star\|^2 \lesssim \frac{1}{\gamma_H^4}\|B_k - B^\star\|^2 + \frac{1}{\gamma_H^2}\|\widetilde{C}_k - \widetilde{C}_k^\star\|^2.
\end{aligned}$$

Then, we apply the tower property of conditional expectation to bound $\mathbb{E}[III]$ by first conditioning on $\mathcal{F}_{k-1}$, and have

$$\mathbb{E}[III] \lesssim \mathbb{E}\Big[\mathbb{E}\Big[\Big(\frac{1}{\gamma_H^4}\|B_k - B^\star\|^2 + \frac{1}{\gamma_H^2}\|\widetilde{C}_k - \widetilde{C}_k^\star\|^2\Big)\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^2 \mid \mathcal{F}_{k-1}\Big]\Big]$$
$$= \frac{1}{\gamma_H^4}\mathbb{E}\Big[\|B_k - B^\star\|^2\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^2 \mid \mathcal{F}_{k-1}\big]\Big] + \frac{1}{\gamma_H^2}\mathbb{E}\Big[\mathbb{E}\big[\|\widetilde{C}_k - \widetilde{C}_k^\star\|^2 \mid \mathcal{F}_{k-1}\big]\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_k)\|^2 \mid \mathcal{F}_{k-1}\big]\Big]$$
$$\overset{\text{(E.28)}}{\lesssim} \frac{C_{g,2}^{1/2}}{\gamma_H^4}\mathbb{E}\big[\|B_k - B^\star\|^2\big] + \frac{C_{g,2}^{1/2}}{\gamma_H^2}\mathbb{E}\big[\|\widetilde{C}_k - \widetilde{C}_k^\star\|^2\big]. \tag{E.42}$$

Here, the second equality is due to $\sigma(\|B_k - B^\star\|) \in \mathcal{F}_{k-1}$ and the independence between ξ_k and the sketching matrices $\{S_{k,j}\}_{j=0}^{\tau}$. Plugging in the definition of $\widetilde{C}_k$ (B.5) and $\widetilde{C}_k^\star$ (E.22), we have

$$\|\widetilde{C}_k - \widetilde{C}_k^\star\| = \Big\|\prod_{j=0}^{\tau-1}\widetilde{C}_{k,j} - \prod_{j=0}^{\tau-1}\widetilde{C}_{k,j}^\star\Big\| \leq \Big\|\prod_{j=0}^{\tau-2}\widetilde{C}_{k,j} - \prod_{j=0}^{\tau-2}\widetilde{C}_{k,j}^\star\Big\| \cdot \|C_{k,\tau-1}^\star\| + \Big\|\prod_{j=0}^{\tau-2}\widetilde{C}_{k,j}\Big\| \cdot \|\widetilde{C}_{k,\tau-1} - \widetilde{C}_{k,\tau-1}^\star\|$$
$$\leq \cdots \leq \sum_{j=0}^{\tau-1}\Big\|\widetilde{C}_{k,j} - \widetilde{C}_{k,j}^\star\Big\| \quad (\text{by } \|C_{k,\tau-1}^\star\| \leq 1 \text{ and } \|\widetilde{C}_{k,j}\| \leq 1).$$

Applying (Na and Mahoney, 2022, Lemma 5.2) and Assumption 3.4, we obtain

$$\mathbb{E}\big[\|\widetilde{C}_k - \widetilde{C}_k^\star\|^2 \mid \mathcal{F}_{k-1}\big] \leq \frac{4\|B_k - B^\star\|^2}{\gamma_H^2}\mathbb{E}[(\sum_{j=0}^{\tau-1}\|S_{k,j}\|\|S_{k,j}^\dagger\|)^2] \lesssim \frac{\tau^2\Upsilon_S}{\gamma_H^2}\|B_k - B^\star\|^2.$$

Combining the above display to (E.42), we get

$$\mathbb{E}[III] \lesssim \frac{\tau^2\Upsilon_S C_{g,2}^{1/2}}{\gamma_H^4}\mathbb{E}\big[\|B_k - B^\star\|^2\big]. \tag{E.43}$$

Combining (E.41) and (E.43) completes the proof.

E.3 Proof of Theorem 4.3

The weighted sample covariance matrix $\widehat{\Xi}_t$ can be decomposed as

$$\widehat{\Xi}_t = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i - \boldsymbol{x}^\star)(\boldsymbol{x}_i - \boldsymbol{x}^\star)^T + \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star)^T$$
$$- \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i - \boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star)^T - \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star)(\boldsymbol{x}_i - \boldsymbol{x}^\star)^T. \tag{E.44}$$

The next two lemmas show that $\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i - \boldsymbol{x}^\star)(\boldsymbol{x}_i - \boldsymbol{x}^\star)^T$ converges to $\Xi^\star$, and the remaining terms are negligible as $\bar{\boldsymbol{x}}_t$ converges to $\boldsymbol{x}^\star$. The proofs are in Appendices E.3.1 and E.3.2.

Lemma E.6. Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \geq \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1-\gamma_S$, and the stepsize parameters satisfy $\beta\in(0,1)$, $\chi > 1.5\beta$, and $c_\beta, c_\chi > 0$. Then, we have

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\boldsymbol{x}^\star)(\boldsymbol{x}_i-\boldsymbol{x}^\star)^T-\Xi^\star\Big\|\Big] \lesssim \begin{cases}\sqrt{\beta_t}+\chi_t/\beta_t^{1.5}, & \beta\in(0,0.5],\\ 1/\sqrt{t\beta_t}+\chi_t/\beta_t^{1.5}, & \beta\in(0.5,1),\end{cases} \tag{E.45}$$

and

$$\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i-1}}\mathbb{E}\big[\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2\big] = O(1). \tag{E.46}$$

Lemma E.7. Suppose Assumptions 3.1 – 3.4 hold, the number of sketches satisfies $\tau \geq \log(\gamma_H/4\Upsilon_H)/\log\rho$ with $\rho = 1-\gamma_S$, and the stepsize parameters satisfy $\beta\in(0,1)$, $\chi > 1.5\beta$, and $c_\beta, c_\chi > 0$. Then, we have

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star)^T\Big\|\Big] \leq \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i-1}}\mathbb{E}\big[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star\|^2\big] \lesssim \begin{cases}\beta_t+\chi_t^2/\beta_t^3, & \beta\in(0,0.5],\\ 1/t\beta_t+\chi_t^2/\beta_t^3, & \beta\in(0.5,1).\end{cases}$$

By the decomposition (E.44), we follow the derivations in (E.25) and (E.26) and obtain

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\Xi}_t-\Xi^\star\|\big] \leq{}& \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\boldsymbol{x}^\star)(\boldsymbol{x}_i-\boldsymbol{x}^\star)^T-\Xi^\star\Big\|\Big]+\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star)^T\Big\|\Big]\\ &+2\sqrt{\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}\mathbb{E}\big[\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2\big]}\sqrt{\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}\mathbb{E}\big[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star\|^2\big]}.\end{aligned}$$

Plugging (E.50) and (E.51) in the proof of Lemma E.6 and (E.60) in the proof of Lemma E.7 into the above display, we obtain

$$\begin{aligned}&\mathbb{E}\big[\|\widehat{\Xi}_t-\Xi^\star\|\big]\\ &\lesssim\begin{cases}\dfrac{C_{g,2}^{1/4}}{\gamma_H(1-\rho^\tau)}\max\left(C_{\widehat{\boldsymbol{\theta}}}^{1/2},\dfrac{C_{\boldsymbol{\delta}}^{1/2}}{(1-\rho^\tau)^{1/2}}\right)\sqrt{\beta_t}+\dfrac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\leq2\beta\}}}{\gamma_H^2(1-\rho^\tau)^{3/2}}\sqrt{\dfrac{\chi_t^2}{\beta_t^3}}=O\left(\sqrt{\beta_t}+\dfrac{\chi_t}{\beta_t^{1.5}}\right), & \beta\in(0,0.5),\\ \max\left(\dfrac{C_{g,2}^{1/4}}{\gamma_H(1-\rho^\tau)}\max\left(C_{\widehat{\boldsymbol{\theta}}}^{1/2},\dfrac{C_{\boldsymbol{\delta}}^{1/2}}{(1-\rho^\tau)^{1/2}}\right),\dfrac{\max(\|\Lambda\|_F,C_{g,2}^{1/2}/\gamma_H^2)}{c_\beta(1-\rho^\tau)^{3/2}}\right)\sqrt{\beta_t}+\dfrac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\leq2\beta\}}}{\gamma_H^2(1-\rho^\tau)^{3/2}}\sqrt{\dfrac{\chi_t^2}{\beta_t^3}}=O\left(\sqrt{\beta_t}+\dfrac{\chi_t}{\beta_t^{1.5}}\right), & \beta=0.5,\\ \dfrac{\max(\|\Lambda\|_F,C_{g,2}^{1/2}/\gamma_H^2)}{(1-\rho^\tau)^{3/2}}\cdot\dfrac{1}{\sqrt{t\beta_t}}+\dfrac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\leq\beta+0.5\}}}{\gamma_H^2(1-\rho^\tau)^{3/2}}\sqrt{\dfrac{\chi_t^2}{\beta_t^3}}=O\left(\dfrac{1}{\sqrt{t\beta_t}}+\dfrac{\chi_t}{\beta_t^{1.5}}\right), & \beta\in(0.5,1),\end{cases}\end{aligned}$$

with constants $C_{\widehat{\boldsymbol{\theta}}} > 0$ defined in (E.39) and $C_{\boldsymbol{\delta}} > 0$ later defined in (E.58). This completes the proof.

E.3.1 Proof of Lemma E.6

Recalling the decomposition (E.21), we have proved the consistency of the dominant term $\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{1,i}\mathcal{I}_{1,i}^T$ in Appendix E.2. The next two lemmas suggest that the terms involving $\{\mathcal{I}_{2,i}\}_i$ and $\{\mathcal{I}_{3,i}\}_i$ are higher order errors, the proofs of which are deferred to Appendices E.3.3 and E.3.4.

Lemma E.8. Suppose the assumptions in Lemma E.6 hold, we have

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{2,i}\mathcal{I}_{2,i}^T\Big\|\Big] \leq \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{2,i}\|^2\big] = O(\chi_t^2/\beta_t^3)\cdot\mathbf{1}_{\{\chi<1.5\beta+0.5\}}+o(\beta_t)\cdot\mathbf{1}_{\{\chi\geq1.5\beta+0.5\}}.$$

Lemma E.9. Suppose the assumptions in Lemma E.6 hold, we have

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{3,i}\mathcal{I}_{3,i}^T\Big\|\Big] \leq \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{3,i}\|^2\big] \lesssim \beta_t.$$

With the above two lemmas, we separate the proof Lemma E.6 by two parts.

Part 1: Proof of (E.45). By the decomposition (E.21), we follow (E.25) and (E.26) and have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\boldsymbol{x}^\star)(\boldsymbol{x}_i-\boldsymbol{x}^\star)^T-\Xi^\star\Big\|\Big]\\
&\leq \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{1,i}\mathcal{I}_{1,i}^T-\Xi^\star\Big\|\Big]+\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{2,i}\mathcal{I}_{2,i}^T\Big\|\Big]+\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathcal{I}_{3,i}\mathcal{I}_{3,i}^T\Big\|\Big]\\
&+2\sum_{1\leq r<s\leq 3}\sqrt{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\mathcal{I}_{r,t}\|^2\big]}\sqrt{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\mathcal{I}_{s,i}\|^2\big]}.
\end{aligned}\tag{E.47}$$

Given Lemmas E.2, E.8, and E.9, it is sufficient to establish the bound for $\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{1,i}\|^2\big]$. We first bound the moment for $\|\boldsymbol{\theta}^k\|$. Based on its definition (E.18), we have

$$\boldsymbol{\theta}^k=-(I-\widetilde{C}_k)B_k^{-1}(\bar{g}_k-\nabla F_k)+(\widetilde{C}_k-C_k)B_k^{-1}\nabla F_k.$$

Furthermore, by $\|\widetilde{C}_k\|\leq 1$, $\|C_k\|\leq 1$, and (15), we get

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{\theta}^k\|^2\big] &\lesssim \frac{1}{\gamma_H^2}\mathbb{E}\big[\|\bar{g}_k-\nabla F_k\|^2\big]+\frac{1}{\gamma_H^2}\mathbb{E}\big[\|\nabla F_k\|^2\big]\\
&\leq \frac{C_{g,1}^{1/2}}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2\big]+\frac{C_{g,2}^{1/2}}{\gamma_H^2}+\frac{\Upsilon_H^2}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2\big]\quad\text{(by Assumptions 3.2 and 3.3)}\\
&\lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2}\quad\big(\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2\big]=o(1)\text{ by Lemma 4.2}\big).
\end{aligned}\tag{E.48}$$

Since $\boldsymbol{\theta}^k$ is a martingale difference sequence, we follow (E.40) and get

$$\begin{aligned}
\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{1,i}\|^2\big] &\leq \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\varphi_k^2\mathbb{E}\big[\|\boldsymbol{\theta}^k\|^2\big]\\
&\lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2}\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\frac{1}{\varphi_i}\sum_{k=0}^{t}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)^2\varphi_k^2}_{\to 0.5/(1-\rho^\tau)\quad\text{by Lemma B.2}}\lesssim\frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)}.
\end{aligned}\tag{E.49}$$

Combining the above display, Lemma E.2 ((E.31) in the proof), Lemma E.8 ((E.61) in the proof), and Lemma E.9 ((E.62) in the proof), and plugging them into (E.47), we get

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i-\boldsymbol{x}^\star)(\boldsymbol{x}_i-\boldsymbol{x}^\star)^T-\Xi^\star\Big\|\Big]\\
&\lesssim\begin{cases}
\frac{C_{g,2}^{1/4}}{\gamma_H(1-\rho^\tau)}\max\left(C_{\widehat{\boldsymbol{\theta}}}^{1/2},\frac{C_{\boldsymbol{\delta}}^{1/2}}{(1-\rho^\tau)^{1/2}}\right)\sqrt{\beta_t}+\frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\le 2\beta\}}}{\gamma_H^2(1-\rho^\tau)^{3/2}}\sqrt{\frac{\chi_t^2}{\beta_t^3}}=O\left(\sqrt{\beta_t}+\frac{\chi_t}{\beta_t^{1.5}}\right), & \beta\in(0,0.5),\\
\max\left(\frac{C_{g,2}^{1/4}}{\gamma_H(1-\rho^\tau)}\max\left(C_{\widehat{\boldsymbol{\theta}}}^{1/2},\frac{C_{\boldsymbol{\delta}}^{1/2}}{(1-\rho^\tau)^{1/2}}\right),\frac{\max(\|\Lambda\|_F,C_{g,2}^{1/2}/\gamma_H^2)}{c_\beta(1-\rho^\tau)^{3/2}}\right)\sqrt{\beta_t}+\frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\le 2\beta\}}}{\gamma_H^2(1-\rho^\tau)^{3/2}}\sqrt{\frac{\chi_t^2}{\beta_t^3}}=O\left(\sqrt{\beta_t}+\frac{\chi_t}{\beta_t^{1.5}}\right), & \beta=0.5,\\
\frac{\max(\|\Lambda\|_F,C_{g,2}^{1/2}/\gamma_H^2)}{(1-\rho^\tau)^{3/2}}\cdot\frac{1}{\sqrt{t\beta_t}}+\frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\le\beta+0.5\}}}{\gamma_H^2(1-\rho^\tau)^{3/2}}\sqrt{\frac{\chi_t^2}{\beta_t^3}}=O\left(\frac{1}{\sqrt{t\beta_t}}+\frac{\chi_t}{\beta_t^{1.5}}\right), & \beta\in(0.5,1),
\end{cases}
\end{aligned}
\tag{E.50}
$$

where constants $C_{\widehat{\boldsymbol{\theta}}}>0$ is defined in (E.39) and $C_{\boldsymbol{\delta}}>0$ will be later defined in (E.58). Here, we also use the observation that $\chi_t^2/\beta_t^3=o(\beta_t)$ when $\chi>2\beta$ and $\beta\in(0,0.5]$, and $\chi_t^2/\beta_t^3=o(1/t\beta_t)$ when $\chi>\beta+0.5$ and $\beta\in(0.5,1)$.

Part 2: Proof of (E.46). By the decomposition (E.21), we have

$$
\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_{i-1}}\mathbb{E}\big[\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2\big]\lesssim\sum_{k=1}^{3}\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{k,i}\|^2\big]\lesssim\frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)},
\tag{E.51}
$$

where the last inequality follows from (E.49), and Lemmas E.8 and E.9. We complete the proof.

E.3.2 Proof of Lemma E.7

By (E.13), we decompose $\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star$ as

$$
\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star=\frac{1}{t}\sum_{i=0}^{t-1}\mathcal{I}_{1,i}+\frac{1}{t}\sum_{i=0}^{t-1}\mathcal{I}_{2,i}+\frac{1}{t}\sum_{i=0}^{t-1}\mathcal{I}_{3,i}=:\bar{\mathcal{I}}_{1,t}+\bar{\mathcal{I}}_{2,t}+\bar{\mathcal{I}}_{3,t}.
\tag{E.52}
$$

We expand $\bar{\mathcal{I}}_{1,t}$ by plugging in (E.14) and exchange the indices. Then, we obtain

$$
\bar{\mathcal{I}}_{1,t}=\frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=0}^{i}\prod_{l=k+1}^{i}\{I-\varphi_l(I-C^\star)\}\varphi_k\boldsymbol{\theta}^k=\frac{1}{t}\sum_{k=0}^{t-1}\sum_{i=k}^{t-1}\prod_{l=k+1}^{i}\{I-\varphi_l(I-C^\star)\}\varphi_k\boldsymbol{\theta}^k.
$$

Since $\boldsymbol{\theta}^k$ is a martingale difference sequence, the interaction terms in $\mathbb{E}\big[\|\bar{\mathcal{I}}_{1,t}\|^2\big]$ are vanished. Thus, we have

$$
\begin{aligned}
\mathbb{E}\big[\|\bar{\mathcal{I}}_{1,t}\|^2\big]&=\frac{1}{t^2}\sum_{k=0}^{t-1}\varphi_k^2\mathbb{E}\Big[\Big\|\sum_{i=k}^{t-1}\prod_{l=k+1}^{i}\{I-\varphi_l(I-C^\star)\}\boldsymbol{\theta}^k\Big\|^2\Big]\\
&\le\frac{1}{t^2}\sum_{k=0}^{t-1}\Big(\sum_{i=k}^{t-1}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\Big)^2\varphi_k^2\mathbb{E}\big[\|\boldsymbol{\theta}^k\|^2\big]=:(\#)\quad\text{(by Lemma B.5(d))}.
\end{aligned}
$$

We rewrite the above display by exchanging the indices, and obtain

$$
\begin{aligned}
(\#) &= \frac{1}{t^2}\sum_{k=0}^{t-1}\sum_{i_1=k}^{t-1}\sum_{i_2=k}^{t-1}\prod_{l_1=k_1+1}^{i_1}(1-(1-\rho^\tau)\varphi_{l_1})\prod_{l_2=k_2+1}^{i_2}(1-(1-\rho^\tau)\varphi_{l_2})\varphi_k^2\mathbb{E}\big[\|\boldsymbol{\theta}^k\|^2\big] \\
&= \frac{1}{t^2}\sum_{i_1=0}^{t-1}\sum_{i_2=0}^{t-1}\sum_{k=0}^{i_1\wedge i_2}\prod_{l_1=k+1}^{i_1}(1-(1-\rho^\tau)\varphi_{l_1})\prod_{l_2=k+1}^{i_2}(1-(1-\rho^\tau)\varphi_{l_2})\varphi_k^2\mathbb{E}\big[\|\boldsymbol{\theta}^k\|^2\big] \\
&\leq \frac{2}{t^2}\sum_{i_1=0}^{t-1}\sum_{i_2=0}^{i_1}\prod_{l_1=i_2+1}^{i_1}(1-(1-\rho^\tau)\varphi_{l_1})\sum_{k=0}^{i_2}\prod_{l_2=k+1}^{i_2}(1-(1-\rho^\tau)\varphi_{l_2})^2\varphi_k^2\mathbb{E}\big[\|\boldsymbol{\theta}^k\|^2\big],
\end{aligned}
$$

where the last inequality comes from the symmetry between the indices i_1 and i_2. We plug in (E.48) and get

$$
\begin{aligned}
\mathbb{E}\big[\|\bar{\mathcal{I}}_{1,t}\|^2\big] &\lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2}\cdot\frac{1}{t^2}\sum_{i_1=0}^{t-1}\sum_{i_2=0}^{i_1}\prod_{l_1=i_2+1}^{i_1}(1-(1-\rho^\tau)\varphi_{l_1})\underbrace{\sum_{k=0}^{i_2}\prod_{l_2=k+1}^{i_2}(1-(1-\rho^\tau)\varphi_{l_2})^2\varphi_k^2}_{\lesssim\varphi_{i_2}/(1-\rho^\tau)\quad\text{by Lemma B.2}} \\
&\lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)}\cdot\frac{1}{t^2}\sum_{i_1=0}^{t-1}\underbrace{\sum_{i_2=0}^{i_1}\prod_{l_1=i_2+1}^{i_1}(1-(1-\rho^\tau)\varphi_{l_1})\varphi_{i_2}}_{\longrightarrow 1/(1-\rho^\tau)\quad\text{by Lemma B.2}} \lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{1}{t}. \qquad \text{(E.53)}
\end{aligned}
$$

For the term $\bar{\mathcal{I}}_{2,t}$, we plug in (E.15) and get

$$
\bar{\mathcal{I}}_{2,t} = \frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=0}^{i}\prod_{l=k+1}^{i}\{I-\varphi_l(I-C^\star)\}(\bar{\alpha}_k-\varphi_k)\bar{\Delta}\boldsymbol{x}_k.
$$

Furthermore, by Lemma B.5(d) and the fact that $|\bar{\alpha}_k-\varphi_k|\leq\chi_k/2$, we know

$$
\begin{aligned}
\mathbb{E}\big[\|\bar{\mathcal{I}}_{2,t}\|^2\big] &\lesssim \mathbb{E}\Bigg[\Big(\frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\chi_k\|\bar{\Delta}\boldsymbol{x}_k\|\Big)^2\Bigg] \\
&\leq \Big(\frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\chi_k\sqrt{\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_k\|^2\big]}\Big)^2 \quad \text{(by Hölder's inequality).} \qquad \text{(E.54)}
\end{aligned}
$$

Using $\|\widetilde{C}_k\|\leq 1$ and (15), we bound $\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_k\|^2\big]$ as

$$
\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_k\|^2\big] \leq \mathbb{E}\big[\|I-\widetilde{C}_k\|^2\|B_k^{-1}\|^2\|\bar{g}_k\|^2\big] \lesssim \frac{1}{\gamma_H^2}\big(\mathbb{E}\big[\|\bar{g}_k-\nabla F_k\|^2\big]+\mathbb{E}\big[\|\nabla F_k\|^2\big]\big) \overset{\text{(E.48)}}{\lesssim} \frac{C_{g,2}^{1/2}}{\gamma_H^2}. \qquad \text{(E.55)}
$$

Consequently, we obtain

$$\mathbb{E}\big[\|\bar{\mathcal{I}}_{2,t}\|^2\big]\cdot \mathbf{1}_{\{\chi<\beta+1\}} \lesssim \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi<\beta+1\}}}{\gamma_H^2}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\sum_{k=0}^{i}\prod_{l=k+1}^{i}\left(1-\left(1-\rho^\tau\right)\varphi_l\right)\varphi_k\cdot\frac{\chi_k}{\varphi_k}}_{\lesssim(\chi_i/\beta_i)/(1-\rho^\tau)\quad \text{by Lemma B.2}}\Big)^2$$

$$\lesssim \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi<\beta+1\}}}{\gamma_H^2(1-\rho^\tau)^2}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\frac{\chi_i}{\beta_i}\Big)^2 \overset{\text{(E.9)}}{\lesssim} \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi<\beta+1\}}}{\gamma_H^2(1-\rho^\tau)^2(1-(\chi-\beta))^2}\cdot\frac{\chi_t^2}{\beta_t^2} \tag{E.56}$$

$$\mathbb{E}\big[\|\bar{\mathcal{I}}_{2,t}\|^2\big]\cdot \mathbf{1}_{\{\chi\geq\beta+1\}} = \Big(\frac{1}{t}\sum_{i=0}^{t-1}o(\beta_i)\Big)^2\cdot\mathbf{1}_{\{\chi\geq\beta+1\}} \overset{\text{(E.9)}}{=} o(\beta_t^2)\cdot\mathbf{1}_{\{\chi\geq\beta+1\}}.$$

Here, we use the fact that $\chi \geq \beta+1 > 2\beta \Rightarrow \chi_t/\beta_t = o(\beta_t)$. For the term $\bar{\mathcal{I}}_{3,t}$, (E.16) gives us the following expansion

$$\bar{\mathcal{I}}_{3,t} = \frac{1}{t}\sum_{i=0}^{t-1}\prod_{k=0}^{i}\{I-\varphi_k(I-C^\star)\}(\boldsymbol{x}_0-\boldsymbol{x}^\star)+\frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=0}^{i}\sum_{l=1+1}^{i}\{I-\varphi_l(I-C^\star)\}\varphi_k\boldsymbol{\delta}^k.$$

Similar to (E.54), by Hölder's inequality, we have

$$\mathbb{E}\big[\|\bar{\mathcal{I}}_{3,t}\|^2\big] \lesssim \Big(\frac{1}{t}\sum_{i=0}^{t-1}\prod_{k=0}^{i}(1-(1-\rho^\tau)\varphi_k)\Big)^2\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2$$
$$+\Big(\frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\varphi_k\sqrt{\mathbb{E}\big[\|\boldsymbol{\delta}^k\|^2\big]}\Big)^2. \tag{E.57}$$

Next, we bound the rate of $\mathbb{E}\big[\|\boldsymbol{\delta}^k\|^2\big]$. By the definition of $\boldsymbol{\delta}^k$ in (E.19) and $\boldsymbol{\psi}^k$ in (E.20), we have

$$\|\boldsymbol{\delta}^k\|^2 \lesssim \|C_k-C^\star\|^2\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2+\|(B^\star)^{-1}\|^2\|\boldsymbol{\psi}^k\|^2+\|B_k^{-1}\|^2\|(B^\star)^{-1}\|^2\|B_t-B^\star\|^2\|\nabla F_k\|^2$$
$$\leq \frac{\tau^2\Upsilon_S}{\gamma_H^2}\|B_k-B^\star\|^2\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2+\frac{1}{\gamma_H^2}\cdot\frac{\Upsilon_L^2}{4}\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^4+\frac{\Upsilon_H^2}{\gamma_H^4}\|B_k-B^\star\|^2\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2.$$

The second inequality is due to 1) $\|C_k-C^\star\| \leq \tau\Upsilon_S^{1/2}\|B_k-B^\star\|/\gamma_H$ (Na and Mahoney, 2022, Lemma 5.2), the Υ_L-Lipschitz continuity of $\nabla F^2(\boldsymbol{x})$, and (15). Thus, we take expectation and obtain

$$\mathbb{E}\big[\|\boldsymbol{\delta}^k\|^2\big] \lesssim \Big(\frac{\tau^2\Upsilon_S}{\gamma_H^2}+\frac{\Upsilon_H^2}{\gamma_H^4}\Big)\mathbb{E}\big[\|B_k-B^\star\|^2\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^2\big]+\frac{\Upsilon_L^2}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^4\big]$$
$$\lesssim \Big(\frac{\tau^2\Upsilon_S}{\gamma_H^2}+\frac{\Upsilon_H^2}{\gamma_H^4}\Big)\sqrt{\mathbb{E}\big[\|B_k-B^\star\|^4\big]}\sqrt{\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^4\big]}+\frac{\Upsilon_L^2}{\gamma_H^2}\mathbb{E}\big[\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|^4\big]$$
$$\lesssim \Big(\frac{\tau^2\Upsilon_S}{\gamma_H^2}+\frac{\Upsilon_H^2}{\gamma_H^4}\Big)\cdot\frac{\Upsilon_L^2\Upsilon_H^4C_{g,2}}{\gamma_H^8}\beta_t^2 =: C_{\boldsymbol{\delta}}\beta_t^2, \tag{E.58}$$

where the last inequality follows from Lemma 4.2 (particularly (E.7) and (E.12) in the proof) and the observation $\chi > 1.5\beta \Rightarrow \chi_t^4/\beta_t^4 = o(\beta_t^2)$. We plug (E.58) into (E.57), apply Lemma B.2, and get

$$\mathbb{E}\big[\|\bar{\mathcal{I}}_{3,t}\|^2\big] \lesssim \Big(\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\prod_{k=0}^{i}(1-(1-\rho^\tau)\varphi_k)}_{=o(\chi_t^2/\beta_t^2+\beta_t)\quad \text{by (B.4)}}\Big)^2\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2 + C_{\boldsymbol{\delta}}\Big(\frac{1}{t}\sum_{i=0}^{t-1}\underbrace{\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\varphi_k\cdot\beta_k}_{\lesssim\beta_i/(1-\rho^\tau)\quad \text{by (B.3)}}\Big)^2$$
$$\overset{(E.9)}{\lesssim} \frac{C_{\boldsymbol{\delta}}}{(1-\rho^\tau)^2(1-\beta)^2}\beta_t^2. \tag{E.59}$$

We recall the fact that $\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\lesssim 1/\beta_t$ in (E.36), combine (E.52), (E.53), (E.56), and (E.59) together, and obtain

$$\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star\|^2\big] \lesssim \begin{cases} \frac{C_{\boldsymbol{\delta}}}{(1-\rho^\tau)^2}\beta_t + \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\le 2\beta\}}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{\chi_t^2}{\beta_t^3} = O\left(\beta_t+\frac{\chi_t^2}{\beta_t^3}\right), & \beta\in(0,0.5),\\ \max\left(\frac{C_{g,2}^{1/2}}{c_\beta^2\gamma_H^2(1-\rho^\tau)^2}, \frac{C_{\boldsymbol{\delta}}}{(1-\rho^\tau)^2}\right)\beta_t + \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\le 2\beta\}}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{\chi_t^2}{\beta_t^3} = O\left(\beta_t+\frac{\chi_t^2}{\beta_t^3}\right) & \beta=0.5,\\ \frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{1}{t\beta_t} + \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi\le\beta+0.5\}}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{\chi_t^2}{\beta_t^3} = O\left(\frac{1}{t\beta_t}+\frac{\chi_t^2}{\beta_t^3}\right), & \beta\in(0.5,1). \end{cases} \tag{E.60}$$

Here follows the same discussion as in (E.50). This completes the proof.

E.3.3 Proof of Lemma E.8

Based on the definition of $\mathcal{I}_{2,i}$ in (E.15), we apply Lemma B.5(d) and the fact that $|\bar{\alpha}_k-\varphi_k|\le\chi_k/2$, then we have

$$\begin{aligned}\mathbb{E}\|\mathcal{I}_{2,i}\|^2 &\le \mathbb{E}\left[\left(\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\frac{\chi_k}{2}\|\bar{\Delta}\boldsymbol{x}_k\|\right)^2\right]\\ &\lesssim \Big(\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\chi_k\sqrt{\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_k\|^2\big]}\Big)^2 \quad \text{(by Hölder's inequality)}\\ &\overset{(E.55)}{\lesssim} \frac{C_{g,2}^{1/2}}{\gamma_H^2}\Big(\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\varphi_k\cdot\frac{\chi_k}{\varphi_k}\Big)^2 \lesssim \frac{C_{g,2}^{1/2}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{\chi_i^2}{\varphi_i^2} \quad \text{(by Lemma B.2).}\end{aligned}$$

With the above display, we obtain

$$\begin{aligned}\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{2,i}\|^2\big]\cdot\mathbf{1}_{\{\chi<1.5\beta+0.5\}} &\lesssim \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi<1.5\beta+0.5\}}}{\gamma_H^2(1-\rho^\tau)^2}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\frac{\chi_i^2}{\varphi_i^3}\\ &\overset{(E.9)}{\lesssim} \frac{C_{g,2}^{1/2}\mathbf{1}_{\{\chi<1.5\beta+0.5\}}}{\gamma_H^2(1-\rho^\tau)^2(1-(2\chi-3\beta))}\cdot\frac{\chi_t^2}{\beta_t^3} = O\left(\frac{\chi_t^2}{\beta_t^3}\right)\cdot\mathbf{1}_{\{\chi<1.5\beta+0.5\}},\\ \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{2,i}\|^2\big]\cdot\mathbf{1}_{\{\chi\ge1.5\beta+0.5\}} &= \frac{1}{t}\sum_{i=0}^{t-1}o(\beta_t)\cdot\mathbf{1}_{\{\chi\ge1.5\beta+0.5\}} \overset{(E.9)}{=} o(\beta_t)\cdot\mathbf{1}_{\{\chi\ge1.5\beta+0.5\}}.\end{aligned} \tag{E.61}$$

This completes the proof.

E.3.4 Proof of Lemma E.9

Given the expression of $\mathcal{I}_{3,i}$ in (E.16), we apply Lemma B.5(d) and have

$$\begin{aligned}\mathbb{E}\big[\|\mathcal{I}_{3,i}\|^2\big] &\lesssim \prod_{k=0}^{i}(1-(1-\rho^\tau)\varphi_k)^2\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2+\Big(\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho)^\tau)\varphi_l)\varphi_k\|\boldsymbol{\delta}^k\|\Big)^2\\ &\leq \prod_{k=0}^{i}(1-(1-\rho^\tau)\varphi_k)^2\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2+\Big(\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\varphi_k\sqrt{\mathbb{E}\big[\|\boldsymbol{\delta}^k\|^2\big]}\Big)^2,\end{aligned}$$

where the last inequality is due to Hölder's inequality. We plugging in (E.58), apply Lemma B.2, and get

$$\begin{aligned}\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\mathbb{E}\big[\|\mathcal{I}_{3,i}\|^2\big] &\lesssim \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\Big(\underbrace{\prod_{k=0}^{i}(1-(1-\rho^\tau)\varphi_k)}_{=o(\beta_i)\quad\text{by (B.3)}}\Big)^2\cdot\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2\\ &\quad+\frac{C_{\boldsymbol{\delta}}}{t}\sum_{i=0}^{t-1}\frac{1}{\varphi_i}\Big(\underbrace{\sum_{k=0}^{i}\prod_{l=k+1}^{i}(1-(1-\rho^\tau)\varphi_l)\varphi_k\cdot\beta_k}_{\lesssim\beta_i/(1-\rho^\tau)\quad\text{by (B.4)}}\Big)^2\\ &\lesssim \frac{C_{\boldsymbol{\delta}}}{(1-\rho^\tau)^2}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\beta_i \overset{\text{(E.9)}}{\lesssim} \frac{C_{\boldsymbol{\delta}}}{(1-\rho^\tau)^2(1-\beta)}\beta_t = O\left(\beta_t\right).\end{aligned} \tag{E.62}$$

This completes the proof.

F Additional Experiment Results

F.1 Regression problems

We follow the experiment settings in Section 5.1 and 5.2 and provide a comprehensive comparison of three estimators: batch-means estimator $\bar{\Xi}_t$ based on ASGD, and plug-in estimator $\widetilde{\Xi}_t$ and weighted sample covariance matrix $\widehat{\Xi}_t$ based on (sketched) Newton method.

The empirical coverage rates of confidence intervals across varying d, r, and τ are summarized in Table 4-7. The results align with the observations from Section 5.1 and 5.2: confidence intervals based on $\widehat{\Xi}_t$ achieve approximately 95% coverage in most cases, while those based on $\bar{\Xi}_t$ often fall below the target confidence level. The bias in $\widetilde{\Xi}_t$ significantly impacts its performance in statistical inference. Comparing the tables, we note that the coverage rate for $\bar{\Xi}_t$ is lower in logistic regression than in linear regression, whereas no significant difference is observed between the two regression types for Newton methods. This suggests that Newton methods are generally more robust. Additionally, the coverage rates for $\widetilde{\Xi}_t$ are lower for the Equi-correlation Σ_a compared to Topelitz Σ_a, indicating that the Equi-correlation Σ_a problem is more challenging. Despite this, the performance of $\widehat{\Xi}_t$ remains stable across all settings.

F.2 CUTEst problems

The experimental setup is described in Section 5.3. The trajectories of the limiting covariance estimation error are presented in Figure 3. The results indicate that the estimation error of $\widehat{\Xi}_t$

Toeplitz Σ_a	Dim	ASGD	Sketched Newton							
			$\tau=\infty$		$\tau=10$		$\tau=20$		$\tau=40$	
		$\bar{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
$r=0.4$	20	88.50	95.50	96.00	88.00	94.50	86.00	92.50	92.50	94.50
	40	89.00	96.00	96.50	92.50	96.50	90.50	94.50	90.00	94.50
	60	89.50	95.00	95.00	84.50	92.50	88.50	94.00	88.50	95.50
	100	91.50	100.0	100.0	89.00	94.00	90.50	95.50	91.50	95.00
$r=0.5$	20	90.00	96.00	96.50	89.50	95.50	94.00	97.00	90.50	96.00
	40	88.50	94.50	94.50	94.00	97.00	89.00	96.00	90.00	96.50
	60	91.00	96.00	96.00	91.00	97.00	85.00	97.00	87.00	93.50
	100	90.50	100.0	100.0	92.50	97.50	86.50	94.50	88.50	93.50
$r=0.6$	20	88.50	89.50	90.00	90.50	96.50	92.50	94.50	91.50	95.00
	40	91.00	95.00	95.00	88.50	98.00	83.00	92.00	88.00	95.50
	60	87.50	95.00	94.50	86.50	95.50	85.50	95.00	87.50	94.50
	100	91.50	100.0	100.0	93.50	99.00	97.00	99.50	95.00	98.50

Table 4: *The empirical coverage rate (%) for linear regression problems with Toeplitz matrix* Σ_a *(with various* r*) at confidence level* 95%.

Equi-corr Σ_a	Dim	ASGD	Sketched Newton							
			$\tau=\infty$		$\tau=10$		$\tau=20$		$\tau=40$	
		$\bar{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
$r=0.1$	20	93.50	95.50	95.00	84.50	94.00	90.00	95.50	91.50	95.50
	40	89.00	98.00	97.00	76.50	94.50	79.50	94.00	83.50	95.50
	60	87.00	93.00	93.00	73.00	94.50	76.50	94.50	76.50	94.50
	100	91.50	100.0	100.0	67.50	95.50	71.50	97.00	68.50	91.00
$r=0.2$	20	91.50	96.00	96.50	82.50	94.50	84.00	96.00	89.50	96.00
	40	88.50	98.00	98.50	71.00	92.50	74.50	94.00	79.50	98.00
	60	86.00	97.00	96.00	70.00	96.00	75.00	94.50	77.00	96.00
	100	83.50	100.0	100.0	71.00	95.50	70.00	96.00	68.50	93.50
$r=0.3$	20	86.50	95.50	96.00	82.00	97.00	85.00	95.00	95.00	97.00
	40	90.50	98.50	98.00	75.00	95.50	74.00	94.50	75.00	94.00
	60	90.50	93.00	93.50	77.50	98.00	76.50	95.00	80.50	97.50
	100	87.50	100.0	100.0	83.00	97.00	74.00	94.50	73.50	92.50

Table 5: *The empirical coverage rate (%) for linear regression problems with Equi-correlation matrix* Σ_a *(with various* r*) at confidence level* 95%.

Toeplitz Σ_a	Dim	ASGD	Sketched Newton							
			$\tau=\infty$		$\tau=10$		$\tau=20$		$\tau=40$	
		$\bar{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
$r=0.4$	20	81.00	94.50	94.00	92.00	96.50	95.00	95.50	95.50	97.50
	40	84.50	93.50	93.00	92.00	95.00	90.50	95.50	86.50	91.50
	60	87.50	95.50	95.50	92.50	95.50	91.50	94.00	94.50	97.00
	100	84.00	96.50	96.00	92.00	97.00	91.00	93.50	92.00	95.50
$r=0.5$	20	85.50	95.00	95.00	90.50	96.00	94.50	97.00	95.00	96.50
	40	85.50	93.50	93.50	89.50	96.00	93.00	95.50	90.00	94.00
	60	88.00	94.50	94.00	93.00	95.50	87.50	93.50	91.00	96.00
	100	85.00	93.00	93.00	93.00	96.00	89.00	93.50	92.00	95.50
$r=0.6$	20	84.50	94.50	94.50	93.00	98.00	89.50	92.50	96.00	97.00
	40	87.50	95.00	95.00	90.50	94.00	91.00	95.00	91.00	95.50
	60	92.00	94.50	95.00	86.50	93.50	89.00	93.00	93.50	96.00
	100	83.50	97.50	97.50	86.00	92.00	87.00	92.00	89.00	94.00

Table 6: *The empirical coverage rate (%) for logistic regression problems with Toeplitz matrix* Σ_a *(with various* r*) at confidence level* 95%.

Equi-corr Σ_a	Dim	ASGD	Sketched Newton							
			$\tau=\infty$		$\tau=10$		$\tau=20$		$\tau=40$	
		$\bar{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$	$\widetilde{\Xi}_t$	$\widehat{\Xi}_t$
$r=0.1$	20	89.00	93.50	93.50	88.50	95.50	92.50	95.00	95.00	95.00
	40	83.50	97.50	97.50	85.00	97.50	85.50	94.50	89.00	93.50
	60	84.00	96.00	97.00	77.00	93.00	81.00	95.50	82.50	95.50
	100	85.50	92.50	91.50	82.00	95.00	81.00	94.50	81.50	93.50
$r=0.2$	20	87.50	96.50	96.00	85.50	96.00	89.00	95.00	92.50	96.50
	40	86.00	96.00	95.00	81.50	92.50	86.50	97.00	94.50	95.50
	60	85.00	93.50	93.50	79.00	94.00	80.00	97.50	79.00	95.50
	100	76.00	96.50	96.00	74.00	92.50	77.00	92.50	73.00	94.50
$r=0.3$	20	90.50	93.00	93.50	85.50	93.00	87.50	93.50	95.00	96.50
	40	86.00	95.00	94.50	81.00	94.00	82.00	93.50	86.00	93.50
	60	82.00	94.00	94.50	83.50	95.50	88.00	98.00	82.50	98.00
	100	62.00	92.00	90.50	77.00	96.00	72.50	97.00	72.00	95.50

Table 7: *The empirical coverage rate (%) for logistic regression problems with Equi-correlation matrix* Σ_a *(with various* r*) at confidence level* 95%.

decreases to small values across all settings, highlighting its effectiveness in constrained optimization. While $\widetilde{\Xi}_t$ demonstrates faster convergence, its error stabilizes at a constant value when bias is significant.

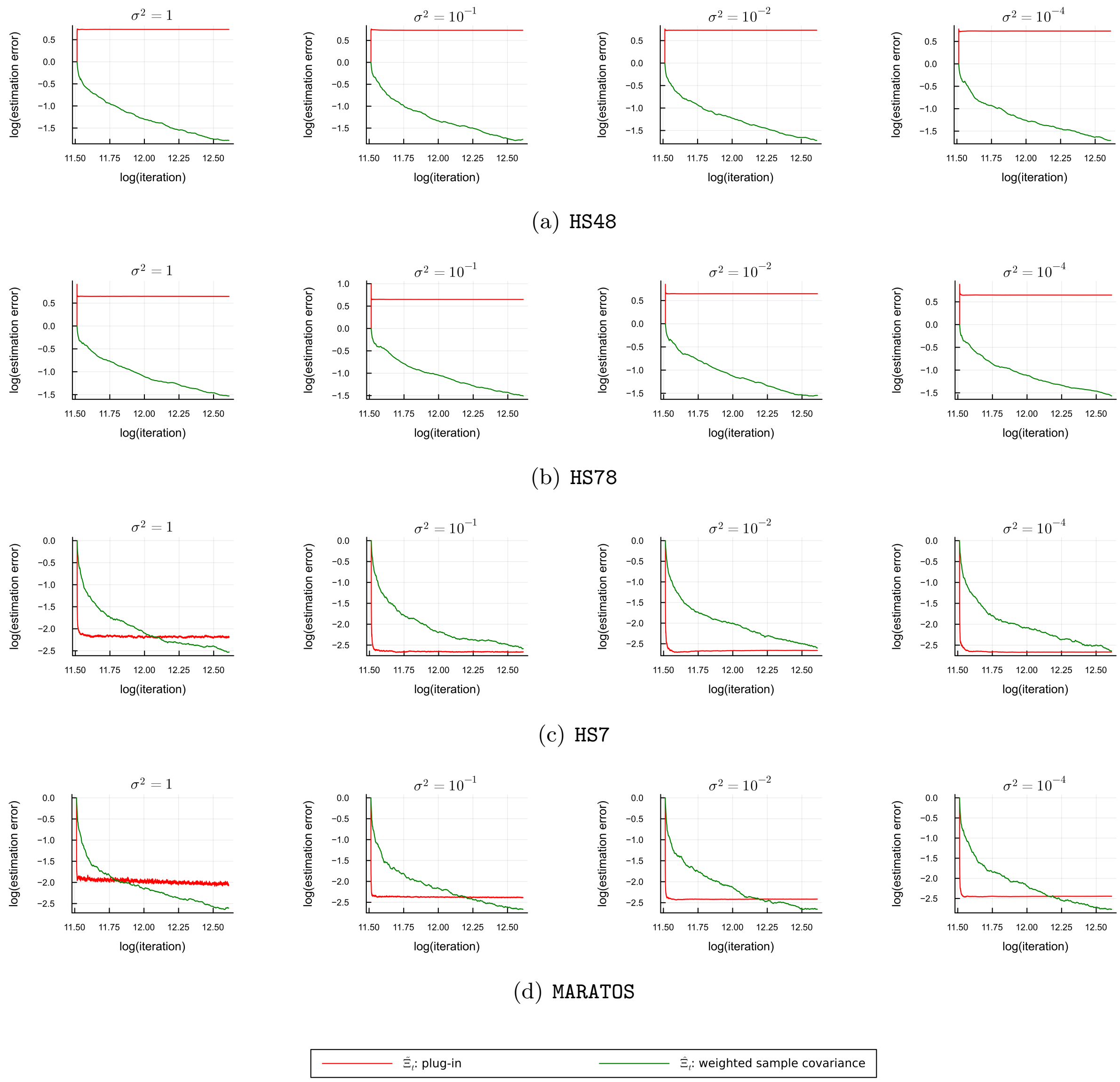

Figure 3: *Covariance estimation error for CUTEst problems. Each row represents a problem, with the four figures in each row corresponding to different noise levels. Each figure displays the estimation error trajectory for the plug-in estimator $\widetilde{\Xi}_t$ (red line) and the weighted sample covariance matrix $\widehat{\Xi}_t$ (green line).*